iPRES 2019

16th International Conference on

Digital Preservation

Proceedings

Amsterdam // 16 – 20 September 2019

www.ipres2019.org

iPRES 2019

Proceedings of the 16th International Conference on Digital Preservation. 2019, Marcel Ras, Barbara Sierman, Angela Puggioni (editors).

ISBN (PDF): 9789062590438
ISBN (epub): 9789062590452

These proceedings contain only the bundled peer reviewed submissions of the 16th International Conference on Digital Preservation. All other materials of the conference are published on the Open Science Framework iPRES 2019 Conference pages: https://osf.io/6ern4/. These pages contain all submitted papers, panels, posters, workshops, tutorials, ad-hoc proposals, hackathons, as well as presenters' slides and optional additionals and the collaborative notes taken during the conference.
The three iPRES 2019 Keynote presentations were recorded and published on the YouTube channel of the Dutch Digital Heritage Network: https://www.youtube.com/channel/UCvHDarEQeREkR-VLdzJSbKWw/videos.

Conference Photo albums are available on flickr: https://www.flickr.com/photos/ipres2019/albums/.
A searchable Twitter Archive is available: https://hawksey.info/tagsexplorer/arc.html?key=1xOqbbwS0W-xVy_qMezz3y-wKqpPGjAYbdRCYqwXl-xL0&gid=400689247, also searchable with a TAGSExplorer: https://hawksey.info/tagsexplorer/?key=1xOqbbwS0W-xVy_qMezz3ywKqpPGjAYbdRCYqwXl-xL0&gid=400689247

IPRES 2019 IS HOSTED BY THE FOLLOWING PARTNERS WORKING TOGETHER IN THE DUTCH DIGITAL HERITAGE NETWORK

National Archives of the Netherlands
Ministry of Education, Culture and Science

KB national library of the netherlands

KNAW Humanities Cluster

Cultural Heritage Agency
Ministry of Education, Culture and Science

SOUND AND VISION

Ministry of Education, Culture and Science

dutch digital heritage network

IPRES 2019 IS PROUDLY SPONSERED BY

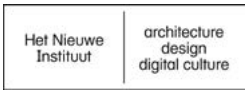


Table of contents

2 **Proceedings**
3 **Sponsors**
4 **Table of contents**
13 **Introduction**
22 **Acknowledgements**

Longpapers

25 **Recovering '90S Data Tapes - Experiences From the KB Web Archaeology project**
Johan van der Knijff

37 **Australian Law Implications on Digital Preservation**
Timothy Robert Hart, Denise de Vries, Carl Mooney

46 **Building the future of digital preservation in french archival services - Processes, functions and staffing for an effective digital preservation**
Marion Humbert, Stéphanie Roussel, Édouard Vasseur

53 **Preservation of Metadata - A case study of a strategy to ensure technology agnostic metadata preservation**
Eld Zierau

64 **Extensive Extensions - Exploring File Extensions in Library of Congress Collections**
Mark Cooper, Aly DesRochers, Jesse Johnston, Camille Salas, Trevor Owens

72 **Aligning the eARK4All Archival Information Package and Oxford Common File Layout Specifications - Complementary rather than competing approaches**
Neil Jefferies, Karin Bredenberg, Angela Dappert

81 **A Pragmatic Application Of PREMIS - Mapping the key concepts to a real-world system**
Jack O'Sullivan, Richard Smith, Alan Gairey, Kevin O'Farrelly

92 **CoreTrustSealCertified Repositories - Enabling Findable, Accessible, Interoperable, and Reusable (FAIR) Data**
Mustapha Mokrane, Jonas Recker

101 **Setting Up Open Access Repositories - Challenges and Lessons from Palestine**
Rawia Awadallah, Iyad ALAgha, Tomasz Miksa, Andreas Rauber, Joy Davidson, Adnan Yahya, Anna Maria Tammaro, Stefano Caselli, Janet Anderson, David Anderson, Nael Salman, Yousef AbuZir

114 **Supporting Virtual Reality and 3D in Academic Libraries - Defining Preservation and Curation Challenges**
Zack Lischer-Katz, Matt Cook, Nathan Hall, Juliet Hardesty, Jamie Wittenberg, Jennifer Johnson, Robert McDonald

126 **Archivematica as a Case Study for Sustained Digital Preservation**
Ashley Blewer, Sarah Romkey, Ross Spencer

134 **Preservation Planning, Beacons for a TDR - Three Cases on Archiving Emerging Media**
Marjolein Steeman

142 **The Integrated Preservation Suite - Scaled and automated preservation planning for highly diverse digital collections**
Peter May, Maureen Pennock, David A. Russo

155 **Digital preservation and enterprise architecture collaboration at the University of Melbourne - A meeting of mindsets**
Jaye Weatherburn, Lyle Winton, Sean Turner

167 **FAIR data work @ DANS - Implementing Metrics, Reviews, and More**
Peter Doorn, Marjan Grootveld, Francesca Morselli

179 **Preservation strategies for an internet-based artwork yesterday today and tomorrow**
Claudia Roeck , Rafael Gieschke, Klaus Rechert, Julia Noordegraaf

191 **Towards a Universal Virtual Interactor (UVI) for Digital Objects**
Euan Cochrane, Klaus Rechert, Seth Anderson, Jessica Meyerson, Ethan Gates

201 **Who is asking? Humans and machines experience a different scholarly web**
Martin Klein, Lyudmila Balakireva, Harihar Shankar

212 **Considerations on the Acquisition and Preservation of eBook Mobile Apps**
Maureen Pennock, Peter May, Michael Day

221 **Eye on CoreTrustSeal - Recommendations for Criterion R0 from Digital Preservation and Research Data Management Perspectives**
Michelle Lindlar, Pia Rudnik

234 **Integrating Dataverse and Archivematica for Research Data Preservation**
Meghan Goodchild, Grant Hurley

Shortpapers

245 **Sustainability through Community - ffmprovisr and the Case for Collaborative Knowledge Transfer**
Andrew Weaver, Ashley Blewer

250 **Fedora and the Oxford Common File Layout - Enhancing Support for Digital Preservation**
David Wilcox

254 **Oais version 3 draftupdates**
David Giaretta, John Garrett, Mark Conrad, Eld Zierau, Terry Longstreth, J Steven Hughes, Felix Engel

260 **Data Stewards and Digital Preservation in Everyday Research Practice**
Esther Plomp, Yan Wang, J.S. Love

265 **What is the Standard Format for Digitized Audio? - Approaches for Storing Complex Audio Objects**
Nick Krabbenhoeft

271 **The Matterhorn RDF Data Model- Formalizing Archival Metadata With SHACL**
Tobias Wildi, Alain Dubois

276 **An Overview of the Digital Preservation Storage Criteria and Usage Guide**
Eld Zierau, Sibyl Schaefer, Nancy Y McGovern, Andrea Goethals

282 **Giving Access to Multimedia Objects in Libraries with Emulation - Implementing a Solution**
Tobias Steinke

286 **A Perspective from Digital Preservation on Sustainability of E-Infrastructures**
Simon Lambert

290 **Together Forever, or How We Created a Common and Collaborative - Digital Preservation Service**
Johan Kylander, Heikki Helin, Kimmo Koivunen, Juha Lehtonen

297 **Provenance for Internet Art - Using the W3C PROV data model**
Lozana Rossenova, Dr Karin de Wild, Dragan Espenschied

305 **Saving Data Journalism - Using ReproZip-Web to Capture Dynamic Websites for Future Reuse**
Katherine Boss, Vicky Steeves, Rémi Rampin, Fernando Chirigati, Brian Hoffman

311 **Jisc Open Research Hub - Integrate, Deposit, Discover and Preserve**
Paul Stokes, John Kaye, Dom Fripp, Tamsin Burland, Tom Davey, Matt McGowen

317 **The Portico Straight-to-Ingest Project**
Sheila Morrissey, Amy Kirchhoff

323 **Getting digital preservation data outWikidata**
Katherine Thornton, Kenneth Seals-Nutt

329 **Reformat Required: Addressing “Paper Minds” in Archives through a New Collaborative Model for Digital Knowledge Transfer**
Angela Beking

335 **One repository solution doesn’t fit all - Towards a distributed and collaborative approach to digital preservation services at the**

Amsterdam City Archives
Ana van Meegen Silva

340 **Building Network Capacity Among Memory Institutions - A Multi-strand Development Approach**
Melinda Haunton, Jo Pugh, James Travers

348 **People Get Ready - Building Sustainability into Digital Preservation Workforce Development**
Sharon McMeekin

354 **Passive Digital Preservation Now & Later - Microfilm, Micr'Olonys and DNA**
Vincent Joguin

360 **Building Resilience at the National Geoscience Data Center - Enhancing Digital Data Continuity Through Research Data Management Training**
Jaana Pinnick

366 **Building a User Group for a Digital Preservation Product - Lessons from 11 years of User Group History**
Jonathan Tilbury, Remke Verdegem, Euan Cochrane

372 **Building Sustainable Digital Preservation - A Typology of Digital Collections**
Helen Hockx-Yu, Donald Brower

377 **Development and improvement of image processing scheme for archiving inscription**
Hideyuki Uesugi, Masayuki Uesugi, Teruhiro Tani

382 **Early Exit Strategies in Digital Preservation**
Ashley Adair, Maria Esteva, Benn Chang

388 **The case for a standard that's old news - RECOMMENDATION OF PDF/A FOR DIGITIZED NEWSPAPER PRESERVATION**
Anna Oates, William Schlaack

Posters

396 **Significant Properties Of Spreadsheets - An Update On The Work Of The Open Preservation Foundation's Archives Interest Group**
Remco van Veenendaal, Frederik Holmelund Kjærskov, Kati Sein, Jack O'Sullivan, Anders Bo Nielsen, Phillip Mike Tømmerholt , Jacob Takema

399 **Preferred, Obsolete Or In-Between? Developing A Criteria Catalogue For AV-Material - Preservation planning at the German National Library of Science and Technology (tib)**
Merle Friedrich

401 **Engaging Decision Makers - An Executive Guide on Digital Preservation**
Sarah Middleton, Sharon McMeekin

404 **Safeguarding the nation's digital memory - A bayesian model of digital preservation risk for the disruptive digital archive**
Alec Mulinder, Sonia Ranade, David H Underdown

406 **Enhancing Services to Preserve New Forms of Scholarship**
Kate Wittenberg, David Millman, Karen Hanson, Craig Van Dyck, Susan Doerr

409 **The australasia preserves story**
Building a digital preservation community of practice in the Australasian region
Jaye Weatherburn

412 **The Web Curator Tool Relaunch - The Next Generation of Web Crawling**
Jeffrey van der Hoeven, Ben O'Brien

414 **ARCHIVER - Archiving and Preservation for Research Environments**
João Fernandes - Jamie Shiers - Bob Jones - Sara Pitonnet Gaiarin

417 **Software Preservation Services in Cultural Heritage Organizations: Mapping the Landscape**
Wendy Hagenmaier, Christa Williford, Monique Lassere, Lauren Work, Jessica G. Benner, Seth Erickson

420 **CREATING CONTINUITY FOR DIGITAL PRESERVATION PROJECTS- Steps for archiving project outputs when Business-As-Usual has not yet been achieved**
Edith Halvarsson, Sarah Mason

423 **Long-term Preservation of pdf files in institutional repositories in Japan**
Teru Agata, Yosuke Miyata, Atsushi Ikeuchi

426 **Concept of a Preservation System for Scientific Experiments in HPC**
Kyryll Udod, Volodymyr Kushnarenko, Stefan Wesner

429 **Malware threats in digital preservation - Extending the evidence base**
Maureen Pennock, Michael Day, Evanthia Samaras

432 **Digital Preservation in a High Security Environment - Student Records, Encryption, and Preservation**
Annalise Berdini

434 **Videotex Art Restoration - Technical and Conceptual Challenges**
John Durno

437 **CD-Ark - A Tool for Cooperative Processing of Optical Discs**
Zdeněk Hruška

439 **Safe Havens For Archives At Risk - Guidelines, Principles and**

Approaches
Afelonne Doek, Tim Gollins

443 **ARCHIVING THE SCHOLARLY GIT EXPERIENCE - An Environmental Scan of Git Preservation Efforts**
Vicky Steeves, Genevieve Milliken

445 **A versatile solution for long-term preservation of research data - Data Life-Cycle Management: The Swiss Way**
Pierre-Yves Burgi, Hugues Cazeaux, Lydie Echernier

448 **Introduction on authorized preservation practice of the National Digital Preservation Program in China**
Chao Wang, Zhenxin Wu, Jiancheng Zheng

450 **Preservation Metadata Dictionary - PREMIS implementation in practice**
Marjolein Steeman, Yvette Hollander

Demonstrations

451 **The Integrated Preservation Suite - Demonstrating a scalable preservation planning toolset for diverse digital collections**
Peter May, Maureen Pennock, David Russo

453 **Review, Appraisal, And Triage of Mail: A Tutorial**
Christopher A. Lee, Kam Woods

455 **From Preservation To Access In One Step - Aviary: An Access Platform For Audiovisual Content - Demo Session**
Pamela Vízner, Bertram Lyons

458 **Jisc Open Research Hub - Demonstration—from Repository to Preservation**
Paul Stokes, John Kaye, Tamsin Burland, Dom Fripp, Tom Davey

Panels

460 **Score Model for the preservation of digital collections - A case for another digital preservation self-evaluation tool**
Bert Lemmens, Robert Gillesse

463 **Dawn of Digital Repository Certification Under ISO 16363 Exploring the Horizon and Beyond - Perspectives From Three Institutions**
Lisa LaPlant, Jamie Shiers, Irfan Zuberi, Maureen Pennock

466 **Preserving eBooks: Past, Present and Future - A Series of National Library Perspectives**
Trevor Owens, Maureen Pennock, Faye Lemay, Tobias Steinke

469 **(I Can't Get No) Satisfaction - Measuring Success in Digital Stewardship Programs**
Karl-Rainer Blumenthal, Peggy Griesinger, Julia Kim, Shira Peltzman, Vicky Steeves, Erwin Verbruggen

472 **The People And Processes of Digital Preservation- International organizations leveraging internal wisdom to build support for digital records**
Jeanne Kramer-Smyth, Montserrat Canela, Ineke Deserno

475 **Achieving Criticality of Preservation Knowledge - Sustaining Digital Preservation in the Nuclear Field**
Elizabeth Kata, William Kilbride, Jenny Mitcham, Jim Moye

478 **Stronger together; collaboration as a matter of course Publishers Round Table**
Eefke Smit

480 **PIDs & Preservation - Incorporating persistent identifiers in a preservation strategy**
Helena Cousijn, Ginny Hendricks, Alice Meadows

483 **CLOUD ATLAS - Navigating the Cloud for Digital Preservation**
Andrea Goethals, Jefferson Bailey, Roslynn Ross, Nicholas Taylor

Workshops & Tutorials

486 **Understanding and Implementing PREMIS - A tutorial**
Karin Bredenberg, Eld Zierau, Angela Di Iorio

488 **Understanding and Implementing METS**
Karin Bredenberg, Sean Mosely, Juha Lehtonen

491 **Preserving complex digital objects - Workshop**
Caylin Smith, Patricia Falcao, Sara Day Thomson

494 **A cost model for analysing and controlling the cost of long-term digital accessibility as a strategic decision-making tool - Workshop by Eye Filmmuseum and BMC**
H. Uffen, W. Swagemakers, E. van Velzen

496 **Machine Learning For Big Text - A Tutorial On Using Predictive Coding Tools To Process Large Archival Datasets**
Brent West, Joanne Kaczmarek

498 **Achieving Criticality of Preservation Knowledge - Sustaining Digital Preservation in the Nuclear Field**
Elizabeth Kata, William Kilbride, Jenny Mitcham, Jim Moye

501 **Preservation Metadata Workshop - A workshop**
Karin Bredenberg, Neil Jefferies, Sarah Romkey, Andreas Nef

504 **Preservation Action Rules Workshop - Parcore: See one, Do one, Teach one**
Justin Simpson, Matthew Addis, Jack O'Sullivan, Carl Wilson, Sarah Romkey, Jon Tilbury

508 **Introducing the memento tracer framework for scalable high-quality web archiving**
Martin Klein, Herbert Van de Sompel

510 **Managing And Preserving Digital - Content With Fedora**
David Wilcox, Andrew Woods, Daniel Bernstein

512 **Level up on preservation: Updating and Mapping the next generation of the *Levels of Preservation***
Bradley J. Daigle, Aaron Collie, Corey Davis, Helen Tibbo, Lauren Work

Introduction

As Programme Organization, we are delighted to present you the Proceedings of the 16th International Conference on Digital Preservation, iPRES 2019. Following a good tradition, the iPRES 2019 Programme consists of research and practice submissions, which ensures visibility and promotion of both academic research work and the projects and initiatives of institutions involved in digital preservation practices. Furthermore, workshops and tutorials provide opportunities for participants to share information, knowledge and best practices, and explore opportunities for collaboration on new approaches. Whereas panels reflect on the various topics, discussing practical solutions, ways to move forward and future developments.

iPRES brings together scientists, students, researchers, archivists, librarians, service providers, vendors and other experts to share recent developments and innovative projects in a wide variety of topics in digital preservation; from strategy to implementation, and from international collaboration to local initiatives. Year on year the debate and research profiled at iPRES have moved digital preservation from a technology driven niche specialism of experts to a global challenge. A continuously growing community is conducting this debate.

The Dutch Digital Heritage Network

iPRES 2019 was hosted by the Dutch Digital Heritage Network (DDHN) , a collaborative effort of an increasing number of network partners: museums, archives, libraries, research institutes and other parties from the heritage sector. Heritage collections include a wealth of information that is of interest for newly developed digital applications, for example all information about an artist such as Vincent van Gogh, or the historical development of a location such as the Amsterdamse IJ. The objective is to make this fully available and permanently accessible. The Dutch Digital Heritage Network has

established a national strategy and is working on ways to increase the 'visibility, usability and sustainability' of digital heritage collections. The ultimate goal is to develop a network of common facilities, services and knowledge base. The DDHN considered it a great honour to be the Host of the 16th iPRES conference as the conference and the Network are all about preserving our digital heritage for the future.

The general Conference Theme was **Eye on the Horizon**, aiming to broaden the voices and approaches participating in the conference, reflecting the venue and looking forward to the future. The future of digital collections but also the future of the iPRES community. This main theme was divided into five sub-themes:

- Collaboration: a Necessity, an Opportunity or a Luxury?
- Designing and Delivering Sustainable Digital Preservation
- Exploring New Horizons
- Building Capacity, Capability and Community
- The Cutting Edge: Technical Infrastructure and Implementation

Keynotes

Among emerging topics that preoccupy many are the fast proliferation of digital technologies and the ever increasing production and usage of digital content. These phenomena cause a growing concern about the management of that digital content both for present and future use. But also a growing concern about the question of what to keep, how to select and who is selecting?

Arguably, the issues of safekeeping digital content that have traditionally stayed within the realm of historical records and memory institutions, are now part of everyday life, posing challenging questions. What happens to our social content? Where will our memories be kept in 50 years? How will the public and scholars, including historians, researchers, and genealogists, know what life was like in the early 21st century? How will we ensure the reproducibility and reuse of scientific output in the future? The three keynote presenters addressed these issues each in their specific ways.

Geert Lovink is the founding director of the Institute of Network Cultures, whose goals are to explore, document and feed the potential for socio-economical change of the new media field through events, publications and open dialogue. As theorist, activist and net critic, Lovink has made an effort in helping to shape the development of the web. Lovink is a Research Professor of Interactive Media at the Hogeschool van Amsterdam (HvA) and a Professor of Media Theory at the European Graduate School. Geert Lovink reflects on the use of social media, the social consequences and the role of the World leaders in developing social media tools have. The full recordings of the presentation "Sad by Design: Politics and Psychology of the Social Media Age" have been published on the YouTube Channel of the Dutch Digital Heritage Network. URL: https://youtu.be/7LDb21An-Ag

Michelle Caswell is an Associate Professor of Archival Studies in the Department of Information Studies at the University of California Los Angeles (UCLA), where she also holds a joint appointment with Asian American studies. In her keynote she

reflected on her work in critical archival studies which engages how individuals and communities activate archives to forge identities, create robust representations, and produce feelings of belonging. Caswell directs a team of students at UCLA's Community Archives Lab, which explores the ways that independent, identity-based memory organizations document, shape, and provide access to the histories of minoritized communities, with a particular emphasis on understanding their affective, political, and artistic impact. The full recordings of the presentation "Whose Digital Preservation? Locating Our Standpoints to Reallocate Resources" have been published on the YouTube Channel of the Dutch Digital Heritage Network.
URL: https://youtu.be/atX14DDvKbw

The third keynote speaker, Eliot Higgins, is the founder and CEO of Bellingcat. Launched in 2014 Bellingcat has played a major role in the growth of online open source investigation, making key findings in the shooting down of MH17 in 2014, and conflict in Syria, and the poisoning of Sergei Skripal in the UK. Bellingcat has recently opened a new headquarter in The Hague, and is currently focused on a major new project documenting the conflict in Yemen, and working with Dutch organisations to spread the use of online open source investigation at a local level. The full recordings of the presentation "Bellingcat and Beyond - The Future for Bellingcat and Online Open Source Investigation" have been published on the YouTube Channel of the Dutch Digital Heritage Network. URL: https://youtu.be/kZAb7CVGmXM]

Peer Reviewed Programme

The conference programme included sessions of paper presentations, panels, posters, demonstrations and hackathons, preceded by workshops and tutorials on Monday. We received a total of 148 submissions this year and were able to accept 99 of them after a peer review process. The categories of accepted and rejected submissions are detailed in the graph below, with an overall acceptance percentage of 66%. (Figure 1)

The DDHN is all about collaboration, as is the iPRES community. Collaboration is reflected in the conference submissions as we see a large number of papers written by two or more Authors. The 99 accepted peer reviewed submissions are written by 279 Authors, from which 227 are unique Authors. The majority of accepted submissions have multiple Authors, in some cases even more then five. These 227 Authors are based in 22 different countries, with the main body of Authors being based in the US, UK, The Netherlands, Germany and Canada. The majority of Authors of accepted papers are employed at Universities and Research Institutes (102), with LAM Institutes as second in row (98). Also interesting is that 15% of the Authors of accepted submissions are Authors of commercial vendors (42). This shows the diversity of the community and the diversity of submissions and topics of the conference. (Figure 2)

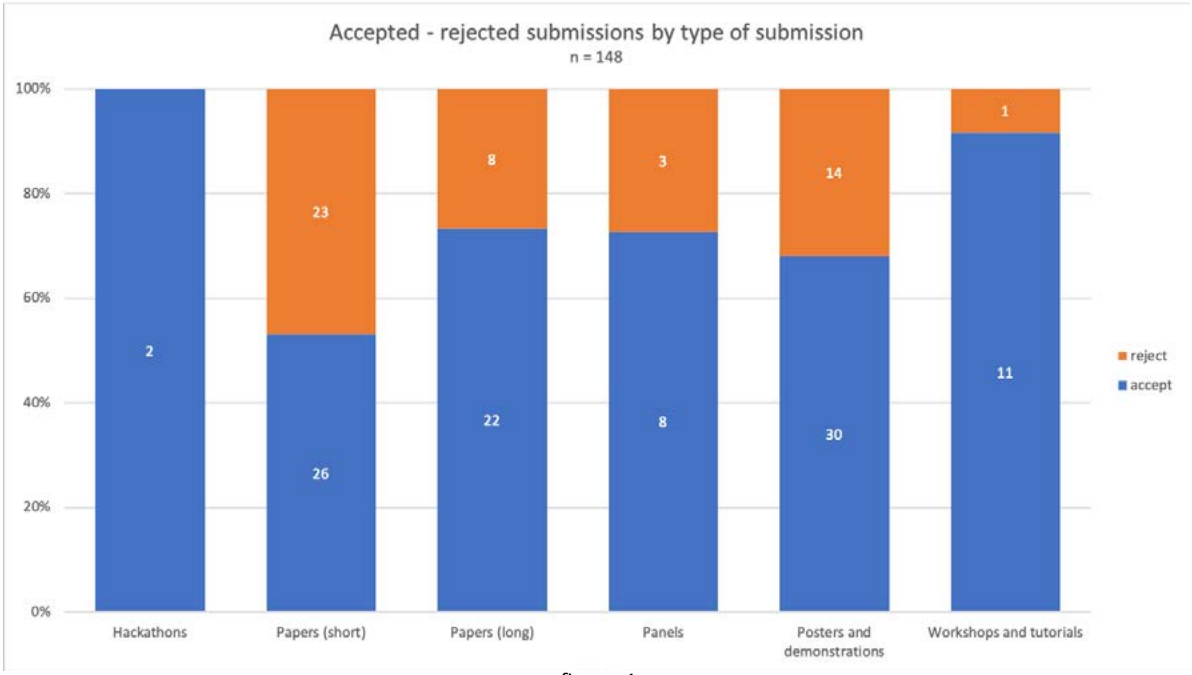


figure 1

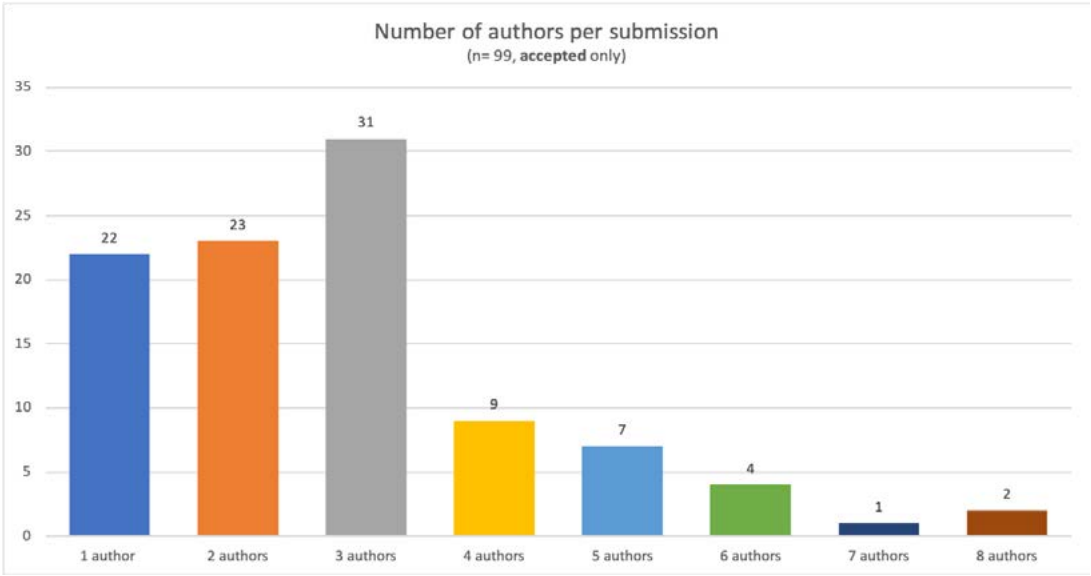


figure 2

Focussing a bit more on collaboration, we also looked at the number of collaborative submissions. Submissions written by multiple Authors based in different Institutes. The graph below shows that there are quite some submissions written by Authors from multiple Institutes. 43 of the 99 accepted submissions are written by representatives of multiple Institutes. That truly is collaboration and a representation of the Digital Preservation community. (Figure 3)

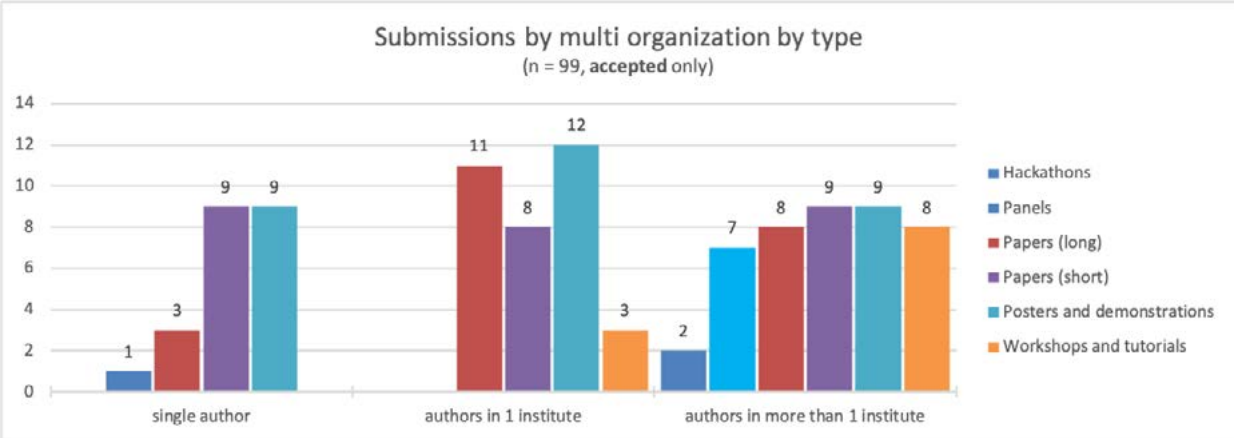


figure 3

Non-Peer Reviewed Programme

In addition to the peer reviewed programme, iPRES 2019 also had a non-peer reviewed programme in which we introduced two novelties. The Great Digital Preservation Bake-Off on Wednesday and Thursday, and the Professional Visits on Friday.

The Great Digital Preservation Bake-Off

After the successful Digital Preservation System Showcase sessions held at the iPRES conferences in Melbourne and Chapel Hill, the demonstrations on software solutions have returned to iPRES as the Great Digital Preservation Bake-Off! The Bake-Off was the chance for vendors and developers to present – and curious digital preservation minds to inquire – how tools fare in comparable situations. The Bake-Off consisted of a number of showcase sessions covering both open source and commercial tools, which were divided into two categories: full end-to-end solutions and content or workflow specific software. Eight vendors and thirteen developers demonstrated preservation tools and implementations of tools and services in front of a critical audience in a fine setting. Vendors demonstrated their solutions based on a test data-set the conference organizers created for them. By providing a test data-set, demonstrations became more comparable. The full programme of the Bake-Off sessions, held on Wednesday 18 September and Thursday 19 September, is to be found on the Open Science Framework pages on the sessions.
URL: https://osf.io/b6eav/

Professional Visits

iPRES 2019 introduced another novelty on Friday. Just after the main conference the attendees were invited to visit Dutch Cultural Heritage Institutes in

the Professional Visits Programme. This Programme allowed all iPRES 2019 attendees to take a peek into the digital preservation world of The Netherlands and to see how some of the Dutch best collections are being preserved. Twelve Dutch Cultural Heritage Institutes in Amsterdam, The Hague, Delft, Rotterdam and Hilversum participated in the programme, with the majority of conference attendees visiting one or two of them.

Ad Hoc Programme

Back on the programme in 2019 was the Ad Hoc session with 14 short presentations of five minutes each and one digital preservation game. The programme of the Ad Hoc sessions is to be found on the Open Science Framework pages of the iPRES 2019 conference.
URL: https://osf.io/v8xfz/

iPRES 2019 Hackathon

Another novelty of iPRES 2019 was the introduction of the Hackathon. A hackathon is a hands on experiment in collaborative learning. It provides a space throughout the conference to rethink, retool, or develop new methods. This is where the early-adopters and true pioneers get together. The iPRES 2019 Hackathon started on Tuesday and went on until Thursday as an in-depth and extended version of a workshop and together with other participants and guided by two teams of experienced professionals the hackathons engaged in collaborative work. The results of the work of the two Hackathon teams were presented on Thursday.

iPRES 2019 Awards

It is a good tradition for the iPRES conference to recognize outstanding contributions and to celebrate these in awards. This year at the iPRES 2019 Conference there were no less than five awards to celebrate. iPRES 2019 recognizes the following outstanding contributions:

- The Best Paper Award was for 'The Integrated Preservation Suite: Scaled and automated preservation planning for highly diverse digital collections' by Peter May, Maureen Pennock and David A. Russo (The British Library). Special thanks to nestor for sponsoring this award

- The Best Poster Award went to: 'ENGAGING DECISION MAKERS: An Executive Guide on Digital Preservation' by Sarah Middleton and Sharon McMeekin of the Digital Preservation Coalition. Special thanks to DANS for sponsoring this award

- The Best Poster Audience Award was for: 'Significant Properties Of Spreadsheets' by Remco van Veenendaal (National Archives of the Netherlands), Frederik Holmelund Kjærskov (Danish National Archives), Kati Sein (National Archives of Estonia), Jack O'Sullivan (Preservica), Anders Bo Nielsen (Danish National Archives), Philip Mike Tømmerholt (Danish National Archives) and Jacob Takema (National Archives of the Netherlands). Special thanks to DANS for sponsoring this award

- The Best First Time Contribution Award was presented to Andrew Weaver of Washington State University for his work on 'Sustainability through Community: ffmprovisr and the Case for Collaborative Knowledge Transfer' which Andrew co-wrote with Ashley Blewer of Artefactual Systems. Special thanks to the Digital Preservation Coalition for sponsoring this award.

- The Best Tulip Photo Award was won by Jon Tilbury of Preservica. Special thanks to the Dutch Digital Heritage Network for sponsoring this award.

Conference attendees

As collaboration is the main driver for the network, the focus of the 16th iPRES conference had to be collaboration and bringing together different communities as well as a diversity in expertise. And we think this did work out very well. iPRES 2019 was attended by 419 digital preservationists from 35 different countries.

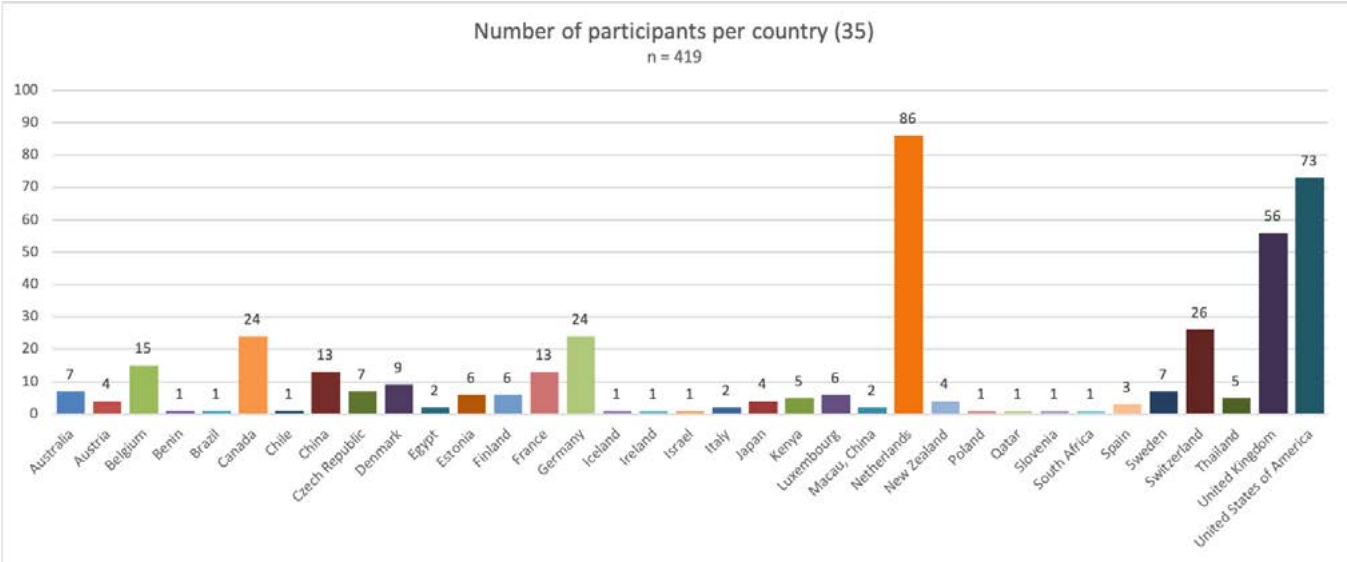


figure 4

For over 200 of these participants it was the first time they attended the conference, which is an overwhelming number of first timers and newcomers in the digital preservation community.

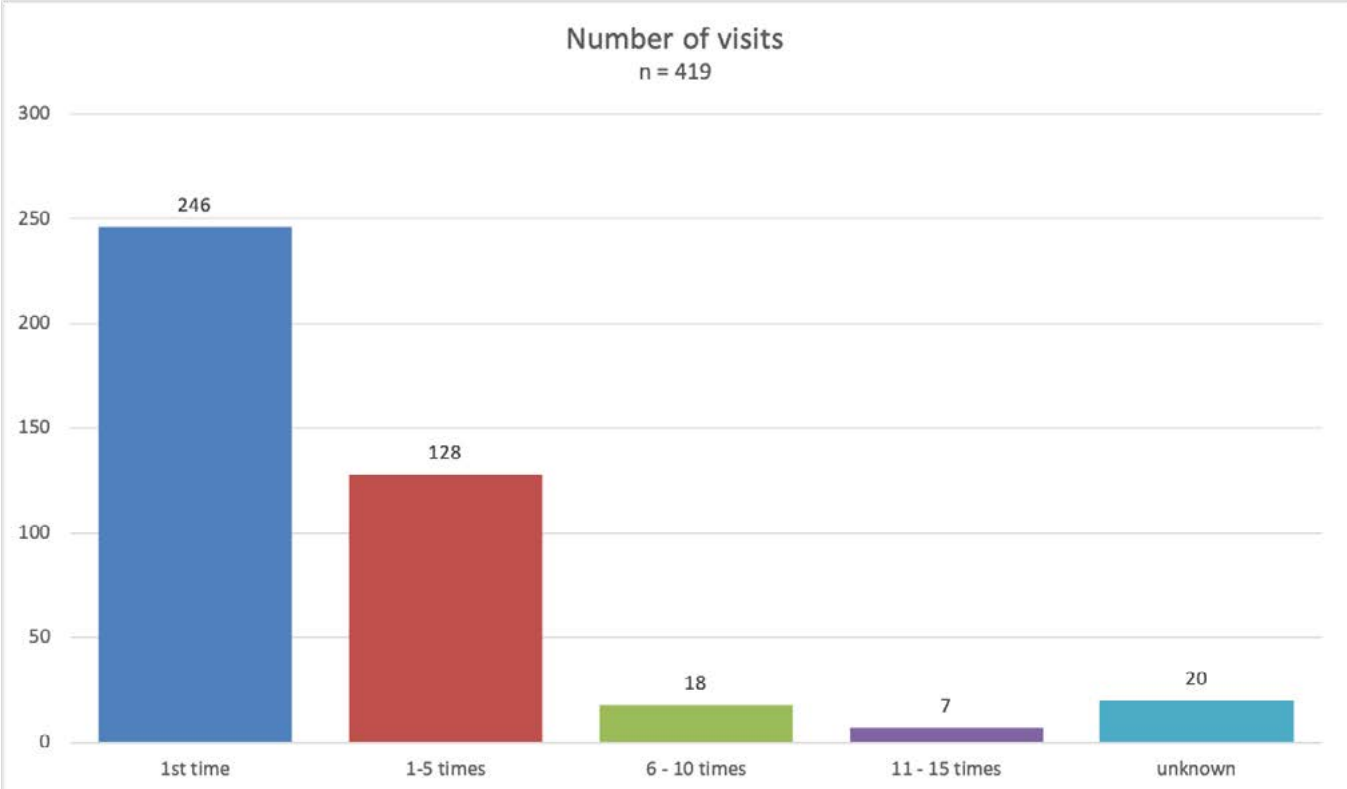


FIGURE 5

Whereas iPRES in the early day was mainly attended by Researchers and Librarians, we see the conference becoming a cross-domain event with attendees from all domains. Still mainly researchers and librarians, but a growing number of digital preservationists from Archives, Museums, Research Institutes and Vendors. That means that the debate has become more diverse with a wide variety of topics discussed from many different angles by a growing community.

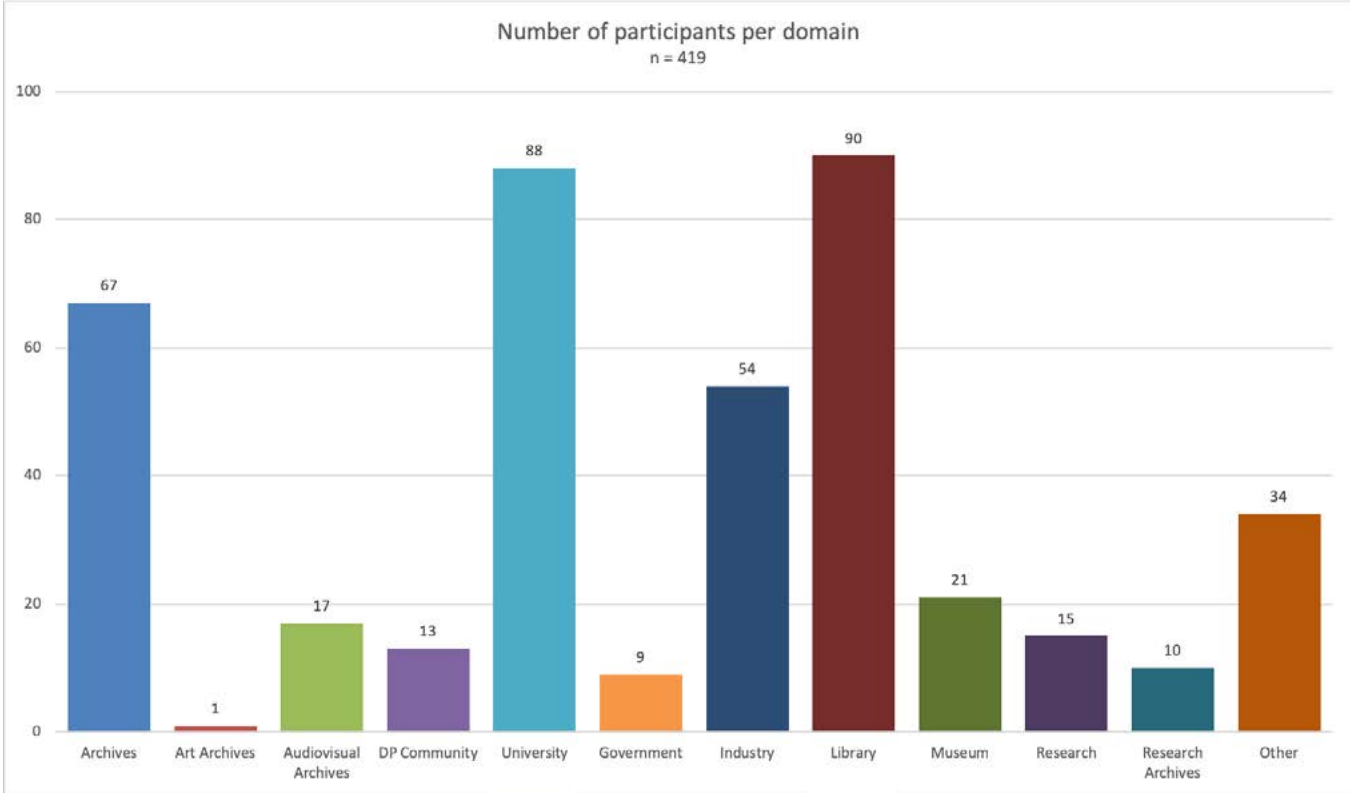


FIGURE 6

Acknowledgements

This year for the first time iPRES was organized by a group of Institutes from different domains, setting sail for the conference gathered with the Netherlands Digital Preservation Coalition (the NCDD) and from 2018 onwards in collaboration with the Dutch Digital Heritage Network. Organizing a conference together has the advantage that there is a lot of expertise available and a surplus of effort of enthusiastic colleagues at hand. We are grateful to the conference hosts who provided the experts helping in organizing the conference.

We also extend our warm thanks to all the sponsors of this years iPRES conference. With their generous support we were able to set up a truly inspiring conference.

The Organising Committee, led by the General Chair Marcel Ras, Digital Preservation Programme Manager at the Dutch Digital Heritage Network, vice-chaired by Barbara Sierman, Digital Preservation Manager at the National Library of The Netherlands and Angela Puggioni, Communications Officer at the Dutch Digital Heritage Network, is pleased with the community response to the calls for contributions and participation. We wish to express our gratitude to all the Programme Committee members who helped us ensure that iPRES 2019 was a high quality event and would also like to acknowledge the tremendous effort of the Local Organising Committee members, The Bake-Off team, the Support team members and the local organisers to ensure a smooth running of the conference and a warm welcome to all the attending delegates.

Marcel Ras	**iPRES 2019 General Chair**
Barbara Sierman	**iPRES 2019 Vice Chair**
Angela Puggioni	**Communications Officer**

CONFERENCE ORGANISATION

PROGRAMME COMMITTEE

General Chair
Marcel Ras (Dutch Digital Heritage Network)

Vice Chair
Barbara Sierman (KB National Library of The Netherlands)

Communications
Angela Puggioni (AP Tekst & Communicatie)

Paper & Panel Chairs
Andrea Goethals (National Library of New Zealand)
Barbara Signori (Swiss National Library)
Pamela Vizner (AVP)
Zhenxin Wu (Library of Chinese Academy of Science)

Workshop & Tutorial Chairs
Jesse de Vos (Netherlands Institute for Sound and Vision)
Katherine Thornton (Yale University)

Poster & Demonstration Chairs
Natalie Harrower (Digital Repository of Ireland)
Tim Walsh (Concordia University)

Hackathon Chairs
Klaus Rechert (University of Freiburg)
Annet Dekker (Amsterdam University)

Communications Team
Jessica Meyerson (Educopia Institute)
William Kilbride (Digital Preservation Coalition)

Great Digital Preservation Bake-Off Team
Robert Gillesse (International Institute for Social History)
Tim Walsh (Concordia University)
Michelle Lindlar (TIB - Leibniz Information Centre for Science and Technology)
Pepijn Lucker (National Archives Netherlands)
Jenny Mitcham (Digital Preservation Coalition)
Marcel Ras (Dutch Digital Heritage Network)
Barbara Sierman (KB National Library of The Netherlands)

LOCAL ORGANIZING COMMITTEE

Annet Dekker (Amsterdam University)
Pepijn Lucker (National Archives Netherlands)
Walter Swagemakers (Eye Film Museum)
Remco van Veenendaal (National Archives Netherlands)
Erwin Verbruggen (Netherlands Institute for Sound and Vision)
Ingeborg Verheul (LCRDM/SurfSara)
Marcel Ras (Dutch Digital Heritage Network)
Barbara Sierman (KB National Library of The Netherlands)
Angela Puggioni (AP Tekst & Communicatie)



Reviewers

Aija Vahtola, National Library of Finland
Angela Dappert, Digital Life-Cycle Management
Ben O'Brien, National Library of New Zealand
Ben Fino-Radin, small data industries
Bertram Lyons, AVP
Cal Lee, University of North Carolina at Chapel Hill
Courtney Mumma, Univ. of Texas Libraries
David Pearson, Department of Parliamentary Services (Australia)
Dorothy Waugh, Stuart A. Rose Manuscript, Archives, and Rare Book Library
Eld Zierau, Royal Danish Library
Erwin Verbruggen, Netherlands Institute for Sound and Vision
Evelyn McLellan, Artefactual
Gang Chen, Institute of High Energy Physics, CAS
Heikki Helin, CSC – IT Center for Science
Ingeborg Verheul, LCRDM – SurfSara
Jan Hutar, Archives New Zealand
Jane Kelly, Harvard University
Jaye Weatherburn, The University of Melbourne
Jenny Mitcham, Digital Preservation Coalition
Jessica Moran, National Library of New Zealand
Jiao Li, Institute of Medical Information and Library
John Richan, Concordia University
Joost van der Nat, Dutch Digital Heritage Network
José Borbinha, Lisbon University
Kate Murray, Library of Congress
Laura Molloy, Digital Curation Centre
Laurents Sesink, Leiden University Library
Libor Coufal, National Library of Australia
Mariella Guercio, University of Roma Sapienza
Martin Klein, Los Alamos National Laboratory
Matthias Priem, VIAA
Matthew Addis, Arkivum
Maureen Pennock, British Library
Michal Indrak, Moravian Library
Michelle Lindlar, Technische Informationsbibliothek
Nancy McGovern, MIT Libraries
Natalie Pang, National University of Singapore
Neil Jefferies, Bodleian Library
Perla Olivia Rodríguez Reséndiz, Universidad Nacional Autónoma de México
Peter May, British Library
Raivo Ruusalepp, National Library of Estonia
Remco van Veenendaal, National Archives Netherlands
Robert Buckley, UAE National Archives
Robert Gillesse, IISH
Sarah Cook, Glasgow University
Sarah Mason, University of Oxford
Sharon McMeekin, Digital Preservation Coalition
Sheila Morrissey, Portico
Shigeo Sugimoto, University of Tsukuba
Shira Peltzman, UCLA
Siang Hock Kia, National Library Board Singapore
Sibyl Schaefer, UCSD
Stefan Strathmann, Goettingen State and University Library
Stefana Breitwieser, Canadian Centre for Architecture
Steve Knight,National Library of New Zealand
Susan Reilly, Quatar National library
Tobias Steinke, DNB
Trevor Owens, Library of Congress
Ulla Bøgvad Kejser, Royal Library of Denmark
Xiaodong Qiao, Wanfang data Co.
Xiaolin Zhang, Shanghai Technical University
Yunpeng Qu, Library of Chinese Academy of Science
Zhixiong Zhang, Library of Chinese Academy of Science

Recovering '90s Data Tapes

Experiences From the KB Web Archaeology project

Johan van der Knijff
KB | National Library of the Netherlands
The Netherlands
johan.vanderknijff@kb.nl
https://orcid.org/0000-0003-3686337X

Abstract – the recovery of digital data from tape formats from the mid to late 'gos is not well covered by existing digital preservation and forensics literature. This paper addresses this knowledge gap with a discussion of the hardware and software that can be used to read such tapes. It introduces *tapeimgr*, a user-friendly software application that allows one to read tapes in a formatagnostic manner. It also presents workflows that integrate the discussed hardware and software components. It then shows how these workflows were used to recover the contents of a set of DDS-1, DDS-3 and DLT-IV tapes from the mid to late '90s. These tapes contain the source data of a number of "lost" web sites that the National Library of the Netherlands (KB) is planning to reconstruct at a later stage as part of its ongoing Web Archaeology project. The paper also presents some first results of sites from 1995 that have already been reconstructed from these tapes.

Keywords – tapes, digital forensics, web archaeology Conference Topics – The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

When the National Library of the Netherlands (hereafter: KB) launched its web archive in 2007, many sites from the "early" Dutch web had already gone offline. As a result, the time period between (roughly) 1992 and 2000 is under-represented in the web archive. To improve the coverage of web sites from this historically important era, the KB has started to investigate the use of tools and methods from the emerging field of "web archaeology" [1]. Analogous to how archaeologists study past cultures from excavated physical artefacts, web archaeology is about reconstructing "lost" web sites using data that are recovered from old (and often obsolete) physical carriers. It is worth noting that Ross and Gow introduced the concept of "digital archaeology" (of which web archaeology is a special case) as early as 1999 [2].

Over the last year, the KB web archiving team has reached out to a number of creators of "early" Dutch websites that are no longer online. Many of these creators still possess offline information carriers with the original source data of their sites. This would potentially allow us to reconstruct those sites, and then ingest them into the web archive. The overall approach would be similar to how we already reconstructed the first Dutch web index *NL-Menu* in 2018 [3],[4].

A few of these creators have already provided us with sample sets of carriers which, though limited in size, comprise a range of physical formats, such as CD- ROMs, floppy disks, ZIP disks, USB thumb drives, and (internal) hard disks. One sample set was provided to us by the former owners of *xxLINK*, a web development and hosting company that was founded in 1994. It was the first Dutch company that provided these services, and throughout the '90s *xxLINK* created web sites for a large number of well-known Dutch companies and institutions[1]. A particularly interesting feature of the *xxLINK* sample set is that it includes 33 data tapes.

There is a relative wealth of digital preservation and digital forensics literature on the recovery of data from physical carriers. Examples include Ross and Gow [2], Elford et al. [5], Woods and Brown [6],

[1] Elizabeth Mattijsen, old xxLINK-homepage: 16th International Conference on Digital Preservation iPRES 2019, Amsterdam, The Netherlands.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Woods et al. [7], Lee et al. [8], John [9] and Pennock et al. [10]. For many carrier types published workflow descriptions are readily available (see e.g. Prael and Wickner [11], Salo [12] and the workflows published by the *BitCurator Consortium* [13], to name but a few). Even though these cover a wide range of physical carrier types, the existing literature provides surprisingly little information on how to recover data from legacy tape formats. One of the few exceptions are De Haan [14] and De Haan et al. [15], who describe how they rescued 11 GB worth of data from three DLT tapes. However, they do not provide much detail about the hardware and software setup they used for this.

Reading these legacy tape formats presents a number of challenges. First, it requires specific hardware that is now largely obsolete. This includes not only the actual tape readers, but also host adapters that are needed to connect a tape reader to a modern forensic workstation, cables and adapter plugs. Because of this, finding the "right" hardware setup is often not straightforward. Furthermore, since the original software that was used to write (and read) legacy data tapes may not be available anymore (if it is known at all), the tapes should be read in a format-agnostic way at the block device level. This can be done with existing software tools, but these tools are not very user-friendly, and the resulting workflows can be quite unwieldy. Also, the logical interpretation of data files that have been recovered from tape requires some additional work. Finally, even though there are still various online resources that cover reading these tapes[2], the information they provide is often fragmentary, or geared towards specific backup software or hardware. This is especially true for older resources that date back to the time when these tape formats were in heavy use.

Hence, there appears to be a knowledge gap. The overall aim of this paper is to fill this gap by discussing the hardware and software that can be used to read such tapes, and by presenting practical workflows that allowed us to recover the information from the *xxLINK* tapes. These workflows are largely based on current hard- and software. They are also fully open source, and can be easily integrated into Linux-based platforms, including the popular *BitCurator*[3] environment.

II. Outline

This paper starts with a brief overview of the tape formats in the *xxLINK* sample set. This is followed by a discussion of the hardware that is needed for accessing tapes like these. This section also provides some suggestions that will hopefully be useful to others who are starting similar tape-related work. It then suggests a format-agnostic procedure for reading the data on the tapes, and presents a new software application that was developed specifically for reading tapes in a simple and user-friendly manner. Next follows a discussion of how this hardware and software setup were integrated into workflows, and how these workflows were used to recover the data on the *xxLINK* tapes. This is followed by two sections that explain the further processing of the recovered data: the extraction of the resulting container files, and the subsequent reconstruction of any "lost" web sites whose underlying data are enclosed in them. This section also shows some first results of sites that were recovered from a 1995 tape. The closing section summarizes the main conclusions.

III. Tape Formats

The majority (19) of the tapes in the xxLINK sample set are DDS tapes, most of which were written in 1995. DDS (Digital Data Storage) is a family of tape formats that are based on Digital Audio Tape (DAT). Using the product codes I was able to identify the majority of these DDS tapes as DDS-1, which was the first generation of DDS. DDS-1 was introduced in 1989, and has a maximum capacity of 2 GB (uncompressed), or 4 GB (compressed). Two tapes could be identified as DDS-3, a format which was introduced in 1996 with a maximum capacity of 12 GB (uncompressed), or 24 GB (compressed). A total of 7 DDS generations have been released, the final one being DAT320 in 2009[4]. Backward read compatibility of

[2] See e.g. the links in the "Tapes" section at: https://github.com/KBNLresearch/forensicImagingResources/blob/master/doc/df-resources.md

[3] BitCurator: https://bitcurator.net/

[4] "Digital Data Storage", Wikipedia: https://en.wikipedia.org/wiki/Digital_Data_Storage

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DDS drives is typically limited to 2 or 3 generations[5]. The *xxLINK* set also contains 14 DLT-IV tapes which were mostly written in 1999. DLT-IV is a member of the Digital Linear Tape (DLT) family of tape formats, which dates back to 1984. DLT-IV was first introduced in 1994[6], and has a capacity of up to 40 GB (uncompressed), or 80 GB (compressed)[7]. Figure 1 shows what these tapes look like.

Figure 1: DDS-1 (left) and DLT-IV (right) tape

IV. Hardware

For all data recovery workflows that are part of the web archaeology project I set up a dedicated forensic workstation that is running the *BitCurator* environment. Reading the vintage tape formats in the *xxLINK* sample set requires some specific additional hardware, most of which can be bought used online at a low to moderate cost. Luckily, it turned out our IT department was still in the possession of an old (DDS-1-compatible) DDS-2 drive, as well as a DLT-IV drive. Both drives are shown in Figure 2.

In order to read the DDS-3 tapes, I purchased an additional used DAT-72 drive that has backward-compatibility with DDS-3.

A. *SCSI host adapter*

As all three tape drives have parallel SCSI[8] connectors, I needed a SCSI host adapter ("SCSI card") to connect them to the forensic workstation. Used SCSI cards can be easily found online, and they are usually sold cheaply. Nevertheless, finding a model that was compatible with both our workstation and the tape drives turned out to be somewhat complicated. This is due to a number of reasons.

First, SCSI cards often have interfaces that are not compatible with current hardware. Many older models have a conventional PCI interface[9], but PCI has been largely replaced by PCI Express[10] on modern motherboards and desktop machines. Some cards have a 64-bit PCI interface, which is only compatible with enterprise servers.

Even if the interface is compatible, the physical dimensions of the card can cause further complications. Older "full-height" PCI Express cards will not fit into a "low-profile" (also known as "half-height") slot, and vice versa (most modern machines only support "low-profile" cards). Many cards were originally sold with both a "fullheight" and a "low-profile" bracket, which allows one to easily change the bracket to fit the target machine. Buying second-hand, it is not uncommon to find that either of the original brackets are missing.

Online sellers do not always explicitly mention characteristics like these, and even if they do this information is not necessarily accurate. A useful resource in this regard is the web site of the Microsemi company, which has the technical specifications of the full range of Adaptec SCSI adapters[11]. Figure 3 shows the PCI Express host adapter that we are using in our workstation.

[5] "HP StorageWorks *Dd*S/DAT Media *Dd*S/DAT Media Compatibility Matrix", Hewlett Packard: https://support.hpe.com/hpsc/doc/public/display?docId=emr_na-lpg50457

[6] "Digital Linear Tape", Wikipedia: https://en.wikipedia.org/wiki/Digital_Linear_Tape

[7] "DLT Drive Media and Cleaning Tape Compatibility Guide", TapeandMedia.com: https://www.tapeandmedia.com/dlt_capacity_info.asp

[8] "Parallel SCSI", Wikipedia: https://en.wikipedia.org/wiki/Parallel_SCSI

[9] "Conventional PCI", Wikipedia: https://en.wikipedia.org/wiki/Conventional_PCI

[10] "PCI Express", Wikipedia: https://en.wikipedia.org/wiki/PCI_Express

[11] Adaptec Support", Microsemi: https://storage.microsemi.com/en-us/support/scsi/

Figure 2: DLT-IV (bottom) and DDS-2 (top) tape drives

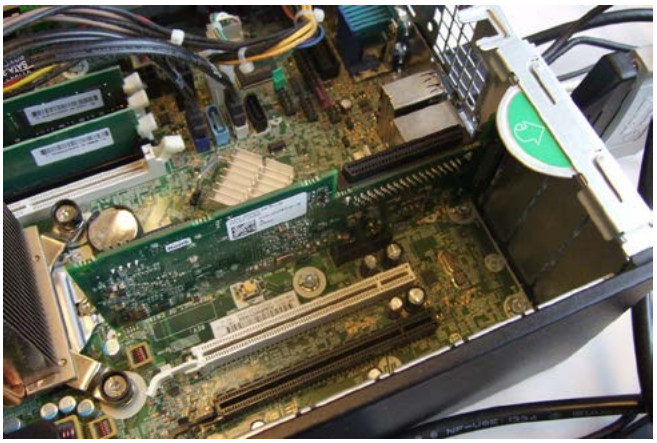


Figure 3: PCI Express SCSI host adapter

B. *SCSI connectors and terminators*

Rather than being one well-defined standard, parallel SCSI is actually a family of related standards that comprise a host of different interfaces, not all of which are mutually compatible[12], [13]. None of these standards specify what connectors should be used to interconnect SCSI devices. Over time, this has resulted in a myriad of connector types that have been developed by different manufacturers[14]. These are typically identified by multiple names. As an example, the commonly used 68-pin "DB68" connector is also referred to as "MD68", "HighDensity", "HD 68", "Half-Pitch" and "HP68", whereas the "50-contact, Centronics-type" connector is alternatively known as a "SCSI-1" or "Alternative 2, A-cable connector". This complicates both identifying the connector type on a particular device, as well as finding suitable adapter plugs and cables. For identifying a connector, the web site of Paralan provides a useful illustrated overview of the most common types[15].

If the tape reader is the last device at either end of the SCSI chain, it must be fitted with a "terminator"[16], which is a resistor circuit that prevents the electrical signal from reflecting back from the ends of the bus. Without a terminator, the tape drive will not work properly, or, more likely, it will not work at all. External SCSI devices like our tape drives use terminator plugs, as shown in Figure 4. For internal devices, termination is often achieved through jumper settings, or by physically removing the terminating resistors from their sockets[17].

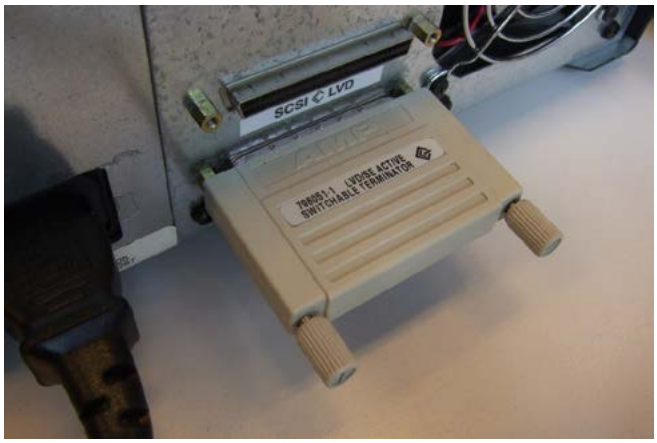


Figure 4: SCSI terminator attached to DLT-IV drive

C. *Cleaning cartridges*

Over time, the heads of a tape drive will get dirty due to a gradual accumulation of dust, and sometimes also residue from the tapes that are used. As this ultimately results in read errors, it is important to periodically clean the drive with a dedicated cleaning

[12] "Parallel SCSI", Wikipedia: https://en.wikipedia.org/wiki/Parallel_SCSI

[13] "LVD, SE, HVD, SCSI Compatibility Or Not", Paralan: http://www.paralan.com/scsiexpert.html

[14] "SCSI connector", Wikipedia: https://en.wikipedia.org/wiki/SCSI_connector

[15] "What kind of SCSI do I have?", Paralan: http://www.paralan.com/sediff.html

[16] "Parallel SCSI", Wikipedia: https://en.wikipedia.org/wiki/Parallel_SCSI

[17] "SCSI termination Q&A", Adaptec: https://storage.microsemi. com/en-us/support/scsi/3940/aha-3940uwd/hw_install/scsi_termination.htm

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

cartridge. Most drives have an indicator that lights up when cleaning is needed. The cleaning procedure is usually very simple, and involves nothing more than inserting the cleaning cartridge into the machine, which then automatically starts the cleaning cycle. A single cleaning cartridge can be used multiple (typically about 50) times. Although I was able to purchase cleaning cartridges for both the DDS and the DLT-IV drives online, it is unclear whether new cartridges are still manufactured today. Since it is not easy or even recommended to clean these drives manually (in fact this is likely to result in damage), the availability of cleaning cartridges could be a concern for keeping older tape formats like these accessible in the long run.

V. Software

Once the hardware is set up, a number of options are available for reading the data from the tapes. Often, tapes contain backup archives that were written by backup utilities such as *tar*, *cpio*, *dump* or *NTBackup*, to name but a few. One approach would be to restore the contents of each tape using the original software that was used to write it. Even though many of these utilities are still available today (especially the Unix-based ones), this approach is not a practical one. First of all, it would require prior knowledge of the tape's archive format. Although we may sometimes have this knowledge (e.g., the writing on a tape's label may indicate it was created with the *tar* utility), in practice we often simply don't know how a tape was written at all. Also, this approach would complicate things, because each format would require its own custom workflow. Finally, it would not work with formats for which the original software is not readily available on the forensic workstation (e.g. the Microsoft Tape Format that was used by Windows *NTBackup*).

A better approach is to use tools like *dd*[18] which are able to read data directly at the block device level. This way, tapes can be read in a format-agnostic manner. The general idea here is that we use *dd* to read all archive files on a tape, irrespective of their format. We then use format identification tools such as *file(1)*[19], *Apache Tika*[20], *Fido*[21] or *Siegfried*[22] to establish the format of each archive file, and subsequently use dedicated, format- specific utilities to extract their contents. This is similar to existing forensic workflows that are used for other carrier types in e.g. *BitCurator*.

A. *Reading a tape with dd and mt*

In the simplest case, reading data from a tape involves nothing more than a *dd* command line such as this one:

```
dd if=/dev/nst0 of=file0001.dd bs=16384
```

Here, the "`if`" argument tells *dd* to read input from the non-rewind block device `/dev/nst0`, and the value of "`of`" defines the file where output is written. The "`bs`" argument defines a block size (here in bytes), and this is where the first complication arises. The actual value that must be used here depends on the software that was used to write the tape, and its settings. If *dd* is invoked with a value that is too small, it will fail with a "cannot allocate memory" error. After some experimentation I was able to establish the block size using the following iterative procedure:

1. Starting with a block size of 512 bytes, try to read one single record (and direct the output to the null device, as we don't need it):

```
dd if=/dev/nst0 of=/dev/null bs=512 count=1
```

2. Position the tape 1 record backward using the *mt*[23] tool (this resets the read position to the start of the current session):

```
mt -f /dev/nst0 bsr1
```

3. If step 1 raised an error in *dd*, increase the block size value by 512 bytes, and repeat from step 1.

Repeating these steps until *dd* exits without

[18] "*dd* (Unix)", Wikipedia: https://en.wikipedia.org/wiki/Dd_%28Unix%29

[19] "file (command)", Wikipedia: https://en.wikipedia.org/wiki/File_(command)

[20] Apache Tika: https://tika.apache.org/

[21] Fido: http://fido.openpreservation.org/

[22] Siegfried: https://www.itforarchivists.com/siegfried/

[23] "mt(1) Linux man page", die.net: https://linux.die.net/man/1/mt

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

errors will yield the correct block size. Re-running the *dd* command at the start of this section with this value will recover the first session on the tape to a single output file.

This leads to a second complication: a tape may contain additional sessions. We can test for this by positioning the tape 1 record forward with the *mt* tool:

```
mt -f /dev/nst0 fsr1
```

If the *mt* call doesn't result in an error (i.e. *mt*' s exit code equals zero), at least one additional session exists. In that case we use *mt* again to position the tape 1 record backward (the start of the second session). We then repeat the block-estimation procedure for the second session, and read the data with *dd*.

All of the above steps are repeated until *mt*'s `fsr` command results in a non-zero exit code, which means no additional sessions exist. The end result is that for each session on the tape we now have a corresponding output file.

B. *Tapeimgr*

Even though the above procedure is not particularly complicated, having to go through all these steps by hand would be very cumbersome. Moreover, *dd*'s ability to overwrite entire block devices with one single command introduces a high risk of accidental data loss (hence its "destroy disk" nickname). Also, it would be useful to have a more user-friendly method for reading data tapes. Because of these reasons, I developed *tapeimgr*[24], which is a software application that allows one to read data tapes using a simple, easy-to-use graphical user interface. Written in Python, it was loosely inspired by the popular *Guymager* forensic imaging software[25]. Internally, *tapeimgr* just wraps around *dd* and *mt*, but it completely hides the complexities of these tools from the user. The software runs on any Linux distribution, and can be installed with *pip*, Python's default package manager. Its only dependencies are a recent version of *Python* (3.2 or more recent), the *TkInter* package, and *dd* and *mt*. All of these are present by default on most modern Linux distros.

Figure 5 shows *tapeimgr*'s interface. At the very minimum, a user must select a directory to which all output for a given tape is written. If necessary, the read process can be further customized using a number of options that are described in detail in the *tapeimgr* documentation. There are also entry fields for descriptive metadata, and the values that are entered here are written to a metadata file in JSON format. This file also contains some basic event and technical metadata, including SHA- 512 checksums of each session (which is represented as a file) that is read from the tape. After the user presses the *Start* button, the progress of the tape reading procedure can be monitored from a widget at the bottom of the interface; the information displayed here is also written to a log file. When *tapeimgr* has finished reading a tape, it displays a prompt that tells the user whether the read process completed successfully. In case of any problems, the log file contains detailed information about all steps in the tape reading process. In addition to the graphical user interface, *tapeimgr* also has a command-line interface, which makes it possible to integrate it into other applications.

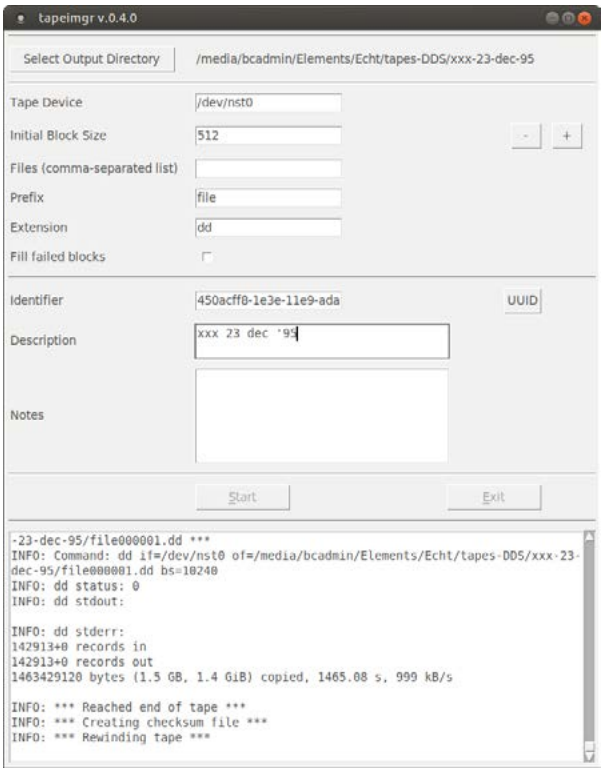


Figure 5: The tapeimgr interface

[24] "*Tapeimgr*": https://github.com/KBNLresearch/tapeimgr

[25] "Guymager homepage": https://guymager.sourceforge.io/

C. *Limitations of tapeimgr*

At this stage, *tapeimgr* has two important limitations. First, it only supports tapes for which the block size is constant within each session. More recent tape drives are often capable of writing tapes in "variable block" mode, where the block size within a session varies. This is not supported, although a possible (but so far untested) workaround may be to set the Initial Block Size to some arbitrary large value that is equal to or larger than the overall largest block size on the tape[26]. Variations in block size *between* sessions are no problem, and are fully supported. Also, *tapeimgr* is not able to recover data from tapes that were partially overwritten. As an example, suppose that a tape originally contained 3 sessions with a size of 200 MB each. If someone later overwrote part of the first session with a 10 MB session at the start of that tape, running the tape through *tapeimgr* will only recover the first 10 MB session, and ignore the remaining sessions. The reason for this is that each write action adds an "End Of Media" (EOM) marker just beyond the end of the written data, and once an EOM is written, any previously recorded data beyond that point are no longer accessible (reportedly workarounds exist, but these are specific to kernel drivers)[27].

Table 1: Summary of *xxLINK* tapes read results

	*Dd*S-1	*Dd*S-3	DLT-IV
# tapes	17	2	14
# read successfully	16	2	7

Table 1: Summary of *xxLINK* tapes read results

IV. Reading The Xxlink Tapes

With all the hardware and software in place, I first experimented with reading some unimportant, disposable DDS-1 and DLT-IV test tapes. Based on these tests I designed processing workflows, which I then documented by creating detailed descriptions that cover all steps that have to be followed to read a tape[28]. Once I was confident that the workflows were suZciently robust, I applied them to the *xxLINK* tapes. All but one of the DDS-1 tapes could be read without problems. For one tape, the recovery resulted in a 10-kB file with only null bytes, which most likely means the tape is faulty. The two DDS-3 tapes could be read successfully as well. Most of these tapes contained multiple (up to 4) sessions. Of the 14 DLT-IV tapes, only 7 could be read without problems. For the remaining 7, the reading procedure resulted in a zerolength file, which means the tape drive interprets them as empty. A common characteristic of all "failed" DLT-IV tapes is that they were written at 40 GB capacity, whereas the other tapes were written at 35 GB capacity. This is odd, as our DLT-IV drive does in fact support 40 GB capacity tapes (this was confirmed by writing some data to a blank test tape at 40 GB capacity, which could subsequently be read without problems). Although the exact cause is unknown at this stage, it is possible that these tapes are simply faulty, or perhaps they were erased or overwritten at some point. Interestingly, the label on at least one of the problematic tapes contains some writing that suggests it was already faulty around the time it was written. Table 1 gives a brief summary of the above results.

V. Extraction Of Recovered Container Files

It is important to stress that the above *tapeimgr*-based workflow only recovers the contents of the tapes at the bit level: for each session on the tape it results in one bitstream (file). Additional steps are needed to interpret the recovered bitstreams in a meaningful way.ran the Unix *file(1)* command (v. 5.32)[29] on all recovered files. The results are summarized in Table 2.

Table 2: Formats of recovered files according to file(1)

Format (file(1))	Number of files
new-fs dump file (big endian)	28
new-fs dump file (little endian)	8

[26] "Cannot allocate memory" when reading from SCSI tape", Unix Stack Exchange: https://unix.stackexchange.com/a/366217

[27] "Tape Driver Semantics", Amanda Wiki: https://wiki.zmanda.com/index.php/Tape_Driver_Semantics

[28] KB Forensic Imaging Resources": https://github.com/KBNLresearch/forensicImagingResources/tree/master/doc

[29] "file (command)", Wikipedia: https://en.wikipedia.org/wiki/File_(command)

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

tar archive	4
POSIX tar archive	2
POSIX tar archive (GNU)	5

Most files in the *xxLINK* data set are Unix *dump* archives[30]. *Dump* is an old backup utility, and its archive files can be extracted using the *restore* tool[31]. Even though *dump* and *restore* are largely obsolete today, the software is still available in the Debian repositories, and as a result these tools can be easily installed on most Linux-based platforms. A few words of caution: first, by default *restore* extracts the contents of a *dump* file to the system's root directory, i.e. it tries to recover a full backup. For our purposes this behaviour is clearly unwanted, and could even wreak havoc on the forensic workstation's file system. However, extraction to a user-defined directory is possible by running *restore* in "interactive" mode[32]. A disadvantage of having to use the interactive mode is that it makes bulk processing of *dump* files virtually impossible. This could be a serious problem when one has to deal with very large numbers of these files. Second, it is important to check the file system of the disk to which the container file is extracted. I initially tried to extract the *dump* files to an *NTFS*-formatted[33] external hard disk. However, it turned out that the names of some files and directories inside the archive were not compatible with *NTFS*, and as a result these files were not extracted. Since the *dump* archives were originally created from a Unix-based file system, this is not surprising. Also, any file attributes that are not supported by *NTFS* (e.g. access permissions and ownership) are lost when extracting to *NTFS*. Extraction to another disk that was formatted as *Ext4*[34] (which is the default file system for most Linux distributions) resolved this issue.

The remaining files are all *tar* archives, a format that is still widely used today. These files can be extracted by simply running the *tar*[35] command like this:

```
tar -xvf /path/to/file0001.dd > /dev/null
```

The earlier observations on the file system of the disk to which the container is extracted also apply to *tar* files.

Using *restore* and *tar* I was able to successfully extract the contents of all container files. For a small number of *dump* files, *restore* reported errors about missing files that could not be found. Although the exact cause is unknown at this stage, a possible explanation could be that in these cases a single *dump* was stored as two volumes on separate physical tapes. This will need further investigation. Nevertheless, overall the interpretation of the *xxLINK* tapes at the container level is quite straight- forward.

It is worth noting that the extraction may be more complicated for other container formats. For example, a number of Microsoft backup tools for the Windows platform (e.g. *NTBackup*) used to write data to tape using the *Microsoft Tape Format*[36]. This is a proprietary format that is only oZcially supported by the original creator software, which is not freely available, and only runs under Windows (however, a few open-source tools exist that claim to support the format). As the *xxLINK* data set does not include this format, it was not investigated as part of this work.

VI. Reconstruction of sites

Once the contents of the container files are extracted, we can start reconstructing the web sites. At this stage of the Web Archaeology project we have only just made a start with this; however, it is possible to present some first results. As a first step we need to inspect the contents of the data that were extracted from the container files, and identify any files and directories that contain web site data.

[30] "Unix dump", ArchiveTeam File Formats Wiki: http://fileformats.archiveteam.org/wiki/Unix_dump

[31] "restore(8) Linux man page", die.net: https://linux.die.net/man/8/restore

[32] A step-by-step description can be found here: https://github.com/KBNLresearch/forensicImagingResources/blob/master/doc/extract-dumpfile.md

[33] "NTFS", Wikipedia: https://en.wikipedia.org/wiki/NTFS

[34] "Ext4", Wikipedia:https://en.wikipedia.org/wiki/Ext4

[35] tar(1) Linux man page", die.net: https://linux.die.net/man/1/tar

[36] Microsoft Tape Format Speci1cation Version 100a: http://laytongraphics.com/mtf/MTF_100a.PDF

This includes not only any directories with the sites' source data, but also server configuration files, which contain valuable information about the sites. For instance, from the configuration files it is possible to see at which domains and URLs they were originally hosted, and how internal forwards were handled. With this information it is possible to host any of the old sites on a locally running web server at their original domains. A detailed discussion of the technical details is beyond the scope of this paper, but the general approach is similar to the one we used earlier to reconstruct the *NL-Menu* web index in 2018 [3], [4]. It comprises the following steps[37]:

1. Set up a web server (typically Apache)[38], and restrict access to the server to localhost (this ensures that any hosted sites are only accessible locally on the machine on which the server is running).
2. Copy the contents of the site (i.e. its directory tree) to the default root folder used by the web server (typically /var/www), and update the file permissions.
3. Configure the site by creating a configuration file (or by adding an entry to an existing configuration file).
4. Activate the configuration file.
5. Add the site's domain to the hosts file (/etc/hosts).
6. Restart the web server.

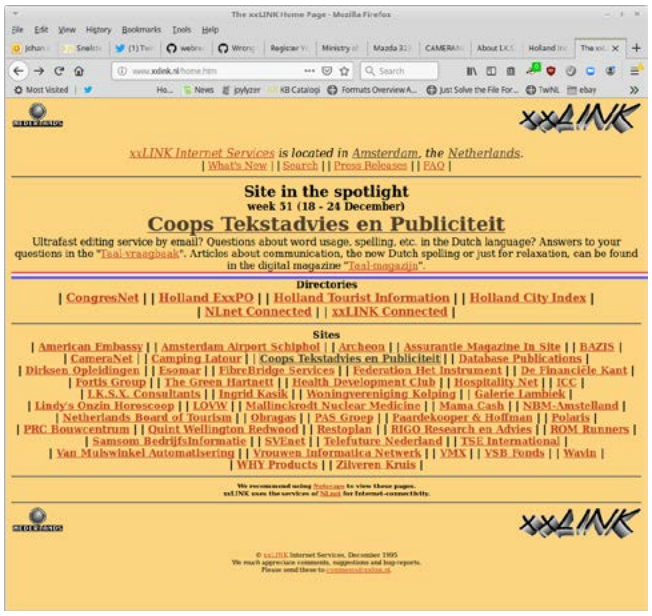


Figure 6: *xxLINK* home page

After following these steps, the site is now locally accessible at its original URL, and it can be viewed in a browser, or archived with web crawler software such as *wget*. Since the directory structure of the web site data on the *xxLINK* tapes is quite uniform, it was possible to automate these steps to a large extent. Using this approach, I have so far reconstructed about 20 sites from one of the 1995 DDS-1 tapes by hosting them on an *Apache* web server instance. A few examples will illustrate the diversity of the sites in the *xxLINK* data set. Figure 6 shows the home page of *xxLINK* 's web site. Figure 7 shows a snapshot of the home page of the web site of Schiphol Airport, which pre-dates the earliest snapshot of this site in Internet Archive[39] by several months. Figure 8 shows a report on drugs policy in the Netherlands, which was published as part of the site of the Dutch Ministry of Health, Welfare and Sport. Finally, Figure 9 shows a contest published on the site of Dutch publisher Database Publications. The objective of the contest was to correctly identify the web addresses of the home pages shown in the image; free copies of *CorelDraw 5.0* were available to five lucky winners.

The site reconstruction procedure will most likely need further refinement. For instance, most of the sites

[37] These steps are described in more detail here: https://github.com/KBNLresearch/nl-menu-resources/blob/master/doc/serving-static-website-with-Apache.md

[38] The Apache HTTP Server Project: https://httpd.apache.org/

[39] Link: https://web.archive.org/web/19961018155616/http://www.schiphol.nl/

on the 1995 tape are relatively simple static HTML sites, but a few include forms that use CGI scripts, which currently do not work. Also, it is possible that the sites on the more recent (1999) tapes are more complex, but this needs further investigation. Once the analysis and processing of the data from the remaining tapes is complete, a more in-depth report on the reconstruction procedure will be published separately.

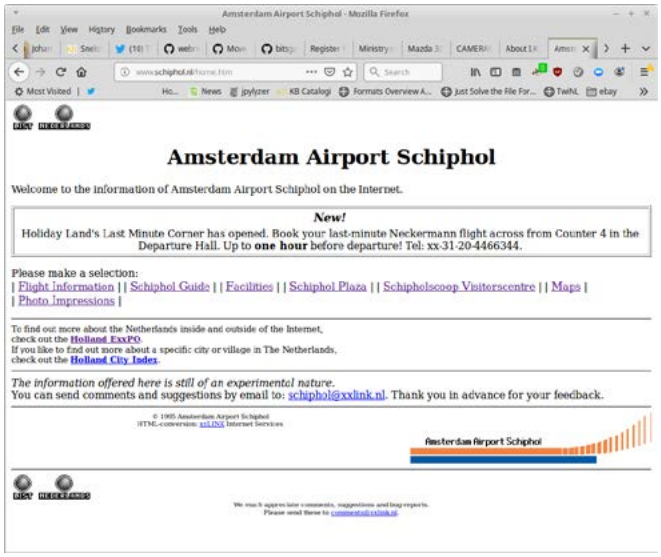


Figure 7: Home page of Schiphol Airport

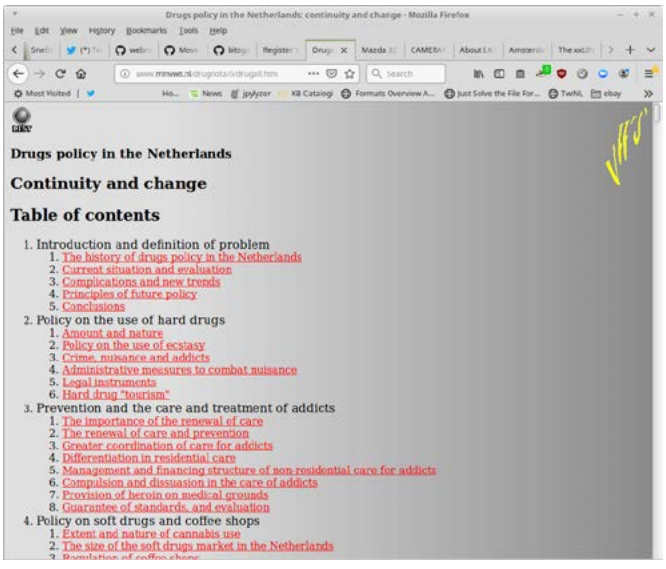


Figure 8: Report on Drugs policy in the Netherlands on web site of Dutch Ministry of Health, Welfare and Sport

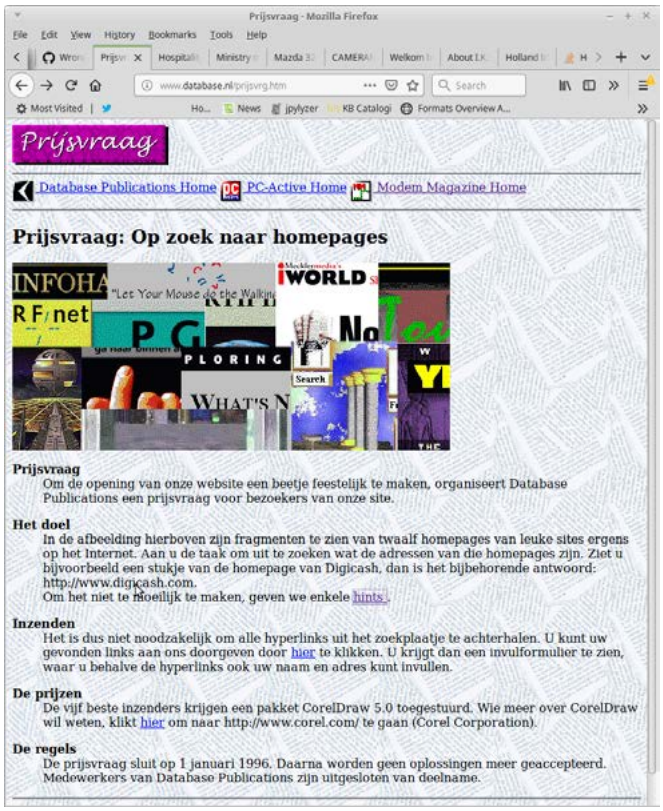


Figure 9: Home page identification contest on web site of Database Publications

VII. Conclusions

In this paper I showed how old DDS and DLT-IV tapes from the '90s can be read on a modern desktop workstation running Linux (in this case the Ubuntu-based *BitCurator* environment). I also explained how I created workflows that allow one to recover data from these tapes in a format-agnostic way, using the user-friendly *tapeimgr* software. Finally I discussed how I then extracted the contents of the resulting files, and how I used

this to reconstruct a number of "lost" web sites from 1995. The workflow descriptions are available on Github[40], and they will most likely evolve further over time. They are published under a permissive license that allows anyone to adapt them and create derivatives. They describe all aspects of the tape reading process in detail, including the hardware components used and their characteristics, links to

[40] "KB Forensic Imaging Resources": https://github.com/ KBNLresearch/forensicImagingResources/tree/master/doc

relevant documentation, instructions on how to handle the tapes (e.g. how to write-protect them), how to operate the tape readers, and how to use them in conjunction with *tapeimgr*. The level of detail provided should be suZcient to allow others to reproduce these workflows, and adapt them to their needs if necessary. Since the process of reading data tapes on Linux-based systems is quite standardized, other tape formats that are not covered by this paper can probably be processed in a similar way.

Acknowledgments

First of all thanks are due to Wendy van Dijk and Elizabeth Mattijsen for making the *xxLINK* tapes available to us, and to Kees Teszelszky, Peter de Bode and Jasper Faase for initiating the Web Archaeology project, and establishing the contact with *xxLINK*. Peter Boel and René van Egdom are thanked for their help digging out the tape drives and other obscure hardware peripherals, and Willem Jan Faber for various helpful hardware-related suggestions. Finally, thanks are due to the anonymous reviewers who provided valuable feedback to an earlier draft of this paper.

References

[1] [1] B. Sierman and K. Teszelszky, "How can we im- prove our web collection? an evaluation of we- barchiving at the KB National Library of the Nether- lands (2007-2017)," Alexandria: The Journal of Na- tional and International Library and Information Is- sues, Aug. 2017. DOI: 10.1177/0955749017725930. [Online]. Available: https://doi.org/10.1177/0955749017725930.

[2] S. Ross and A. Gow, "Digital archaeology: Rescuing neglected and damaged data resources. a jisc/npo study within the electronic libraries (elib) programme on the preservation of electronic materials.," University of Glasgow, 1999. [Online]. Available: http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2.pdf.

[3] J. van der Knijff, Resurrecting the irst Dutch web index: NL-menu revisited, 2018. [Online]. Available: https : / / www . bitsgalore . org / 2018 / 04 / 24 / resurrecting - the - first - dutch - web - index - nl-menu-revisited.

[4] ——, Crawling oline web content: The NL-menu case, 2018. [Online]. Available: https://www.bitsgalore.org/2018/07/11/crawling-offline-web-content-the-nl-menu-case.

[5] D. Elford, N. Del Pozo, S. Mihajlovic, D. Pearson, G. Clifton, and C. Webb, "Media matters: Developing processes for preserving digital objects on physical carriers at the national library of Australia," in Pro- ceedings, 74th IFLA General Conference and Council, 2008. [Online]. Available: http://archive.ifla.org/IV/ifla74/papers/084-Webb-en.pdf.

[6] K. Woods and G. Brown, "From imaging to access - effective preservation of legacy removable media," in Proceedings, Archiving 2009, 2009. [Online]. Available: https://kamwoods.net/publications/woodsbrownarch09.pdf.

[7] K. Woods, C. Lee, and S. Garfinkel, "Extending dig- ital repository architectures to support disk im- age preservation and access," in Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, ser. JCDL '11, Ottawa, Ontario, Canada: ACM, 2011, pp. 57–66, ISBN: 978-1-4503- 0744-4. DOI: 10 . 1145 / 1998076 . 1998088. [Online]. Available: http://doi.acm.org/10.1145/1998076.1998088.

[8] C. Lee, M. Kirschenbaum, A. Chassanoff, P. Olsen, and K. Woods, "BitCurator: Tools and techniques for digital forensics in collecting institutions," D-Lib Magazine, vol. 18, 5/6 2012. [Online]. Available: http://www.dlib.org/dlib/may12/lee/05lee.html.

[9] J. John, "Digital forensics and preservation," Digi- tal Preservation Coalition, 2012. [Online]. Available: http://dx.doi.org/10.7207/twr12-03 .

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

[10] M. Pennock, P. May, M. Day, K. Davies, S. Whibley, A. Kimura, and E. Halvarsson, “The flashback project: Rescuing disk-based content from the 1980’s to the current day,” in Proceedings, 11th Digital Curation Conference, 2016. [Online]. Available: https://doi.org/10.5281/zenodo.1321629.

[11] [11] A. Prael and A. Wickner, “Getting to know FRED: Introducing workflows for born-digital content,” Practical Technology for Archives, vol. 4, 2015. [Online]. Available: https://practicaltechnologyforarchives.org/issue4_prael_wickner/.

[12] [12] D. Salo, Building audio, video and data-rescue kits, 2017. [Online]. Available: https://radd.dsalo.info/wp-content/uploads/2017/10/BuildDocumentation.pdf.

[13] [13] BitCurator Consortium, Workflows. [Online]. Available: https://bitcuratorconsortium.org/workflows.

[14] [14] T. De Haan, “Project the digital city revives, a case study of web archaeology,” Proceedings of the 13th International Conference on Digital Preservation, 2016. [Online]. Available: https://doi.org/10.5281/zenodo.1255965.

[15] [15] T. De Haan, R. Jansma, and P. Vogel, Do it yourself handboek voor webarcheologie, 2017. [Online]. Available: https://hart.amsterdam/image/2017/11/17/20171116_freeze_diy_handboek.pdf.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Australian Law Implications on Digital Preservation

Timothy Robert Hart
Flinders University, Australia
tim.hart@linders.edu.au

Denise de Vries
Flinders University, Australia
Denise.deVries@Flinders.edu.au
ORCID: oooo-ooo1-9o61-6471

Carl Mooney
Flinders University, Australia
carl.mooney@linders.edu.au

Abstract – Collection institutions (Libraries, Archives, Galleries, and Museums) are responsible for storing and preserving large amounts of digital data, which can range from historical/public figure records, to state or country- wide events. The ingest process often requires sifting through large amounts of data which may not always be sorted or categorized from the source/donor. It is possible to discover information that was not intended to be disclosed should the donor not be privy to the existence of said material. This issue is typically handled by communicating with the donor, however, if they have no relation to what has been uncovered in the data, further steps may need to be taken. If the data belong to or are about someone living, that person may need to be contacted, depending on the nature of the data discovered. If the person of interest is no longer living, legally there would no issue disclosing all information uncovered, how- ever, implications for living relatives must be considered should the disclosed information be potentially revealing or harmful to them. This can include hereditary health issues, political or religious views, and other sensitive in- formation. There are significantly more variables to con- sider, such as public interest and defamation which can heavily impact the decision process following the discovery of sensitive data, all whilst guided, but not necessarily enforced by law. This remains somewhat of a gray area as the entities handling such data are often exempt from these laws and principles, making these decisions ethically and morally based more so than legally. In this article, the Australian laws and policies that surround privacy issues, defamation, and data relating to Aboriginal and Torres Strait Islander people and culture are explored. The aim is to raise awareness on potential issues that may arise in collection institutions as well as potential threats already sitting in storage and the laws and policies that may serve as guidelines to help overcome/mitigate such issues.

Keywords – access to information, defamation, privacy, sensitive information, Australian Law

Conference topics – Designing and Delivering Sustain- able Digital Preservation; Exploring New Horizons

I. INTRODUCTION

Procedures for born-digital preservation have not yet been standardized among the many institutions per- forming such actions. Some institutions are progressive and are actively making advancements in born-digital preservation, whereas others are still in their infancy when it comes to preserving born-digital content. While digitization of hard-copy material is certainly part of digi- tal preservation, as researched by LeClere [1], there are far more potential issues surrounding born-digital content. Although these issues are global, the laws which cover them are country specific. This paper is focused on the Australian jurisdiction. The main issues where Australian law may hinder the process are present dur- ing the ingest phase and the storage phase, specifically where access is made available. Examples of such issues will be discussed under the Ingest Scenarios section.

Although not always obligatory for all entities, laws and policies exist for good reason. Currently, the main entities performing digital preservation within Australia fall into this area, namely Libraries, Archives, Museums, and Universities. The material these entities store and make publicly available are exempt from the Privacy Act 1988 and the Australian Privacy Principles (APPs) within. As stated in the National Library of Australia privacy policy:

This policy sets out how the National Library of Australia (the Library) approaches and manages

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

the Australian Privacy Principles (APPs) contained in Schedule 1 to the Privacy Act 1988 (Privacy Act).

The Privacy Act regulates how Commonwealth agencies such as the Library collect, store, use and disclose personal information, and how individuals can access or correct personal information the Library holds. It requires the Library to com- ply with the APPs and take reasonable steps to implement practices, procedures and systems to protect personal information.

The Privacy Act does not apply to library material held, man- aged and made accessible by the Library, whether published (such as books, journals, newspapers and websites) or un- published (oral history interviews, photographs and archival collections). [3].

However, the Privacy Act still applies to any personal user information collected from Library services as well as from all Library employees, temporary staff, volunteers and contractors.

The records held regarding Australian Aboriginal and Torres Strait Islander people have their own surrounding issues along with protocols to guide collection institutions through them. Extra care must be taken in order to maintain the customs of Indigenous peoples and to ensure the handling of their material is done accord- ing to their cultural needs. One must again emphasize, there may not be a definitive law regarding such actions, but collection institutions should feel ethically obliged to follow relevant protocols to comply with best practice. Being aware of existing laws and the issues which they aim to prevent, is a necessity for not only adopting best practice, but preparing for any future changes to privacy law.

II. Sensitive and Identifying Information

Before understanding the laws that may affect digital preservation, it is important to understand the source of the issues and where they may arise. This understanding is crucial, as often the solution must come down to a judgment call, basing decisions on variables guided, but not often enforced, by Australian Law.

Sensitive and identifying information can be found on digital media by using forensic software tools which are freely available and easy to use, such as the BitCurator environment [4], bulk_extractor [5] and The Sleuth Kit (Autopsy) [6]. Material is often donated to collection institutions and this can lead to a range of issues. Libraries offer donor agreements which form a contractual agreement between library and donor, stipulating all conditions from both parties and how to handle the data once collected. These agreements may also pass ownership from the donor, removing them from any further say in the matter.

One issue is the discovery of sensitive data. In most scenarios the donor agreement will typically have instructions in place on how to handle this. However, there are scenarios where the solution is not so easily solved. Firstly, what data are classified as sensitive must be established along with what information can be used to identify an individual. In Part II-Division 1 of the Privacy Act [7], identifying information and sensitive information are defined as followed:

Identifying information

- Full name
- Alias/Previous name
- Date of birth
- Sex
- Last known address (including 2 previous)
- Name of current/past employer
- Driver's license

Sensitive Information

- Racial or ethnic origin
- Political opinions/membership association
- Religious beliefs/affiliations
- Philosophical beliefs
- Membership of professional/trade association or union
- Sexual orientation or practices
- Criminal record
- Health/genetic information
- Biometric information/templates

Regarding the list of sensitive information, these data can be derived by online activity and how the user in question went about their daily activities on the device on which the donated material was created. Whilst there may not be an individual element that clearly species any of these elements, there may be definitive clues. Much of this information lies deep in a system, obscure, and difficult, if

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

not impossible to find by manual means (navigating directories without the assistance of a foren- sic tool). One tool that suits this need is bulk_extractor [5] which can be used to discover anywhere between thousands to millions lines of data deemed sensitive or personal. With this tool, online activity such as websites visited, which elements within that website were viewed, and any sub-pages visited are revealed. Emails, Face- book, Web browser searches, and much more can be derived and analyzed to establish information about the user.

For example, health or genetic information could potentially be established if the user frequently researched and visited websites on a health issue. Personal information could be revealed in emails. Religious beliefs and affiliations could also be revealed by online activity, contacts, and communi- cations. Sexual orientation and practices are readily revealed should the user frequent pornographic websites. There is an abundant amount of data that are collected overtime, a digital footprint, something the average user typically will not put much effort into hiding. These data have much potential, both good and bad.

The following is a real-world example. The data have been taken from a real hard drive and processed through bulk_extractor. Any personal information has been redacted and the example has been carefully selected.

Bulk_extractor detected a high number of URL searches relating to job seeking:

"employsa.asn.au", "Job+Search", "Retail+Jobs", "resumes"

Another discovery was visits to the McDonalds' login page. It was also discovered that there was an official McDonalds' email address assigned to the user. All this information together strongly suggests the user was employed by McDonalds. Correlated against other information and further investigation, it would not be farfetched to say one could establish which workplace the user was assigned to and how much further the investigation could go.

This example shows how individual elements, typically undetectable without the aid of forensic tools, combined with other data can reveal a lot about an individual, often sensitive and personal in nature.

Note that whilst numerous records handled by collection institutions are historic and often relating to a deceased person, their sensitive information may still affect any living family members.

This relates to health and genetic information. If the information collected indicates the deceased person had a medical condition that is inheritable, this reveals possible health information for their descendants [8].

While collection institutions must abide by the laws surrounding privacy with the consumer data they hold, e.g. account information for library users and staff, the collection material itself is exempt from such law. How- ever, this does not mean the laws should not be at least considered as guidelines, influencing policies and procedures for handling sensitive data within collection institutions. The State Library of NSW provided a 'Sensitive Collections Material Policy' in 2017 that addresses this with the opening statement as follows:

As part of the Library's collections there is a signif- icant number of records containing people's personal information or, content that is considered culturally sensitive to Indigenous Australian peoples. Examples of these records include medical records, records of children in care, legal records and Indigenous cultural material. Library collection material is exempt from both the Privacy and Personal Information Project Act (1998) and Health Records and Information Privacy Act (2002), however in the spirit of this legislation and based on best practice considerations, the Library sees an ethical obligation to protect people's personal and cultural information. Of equal importance to the Library is enabling individuals to seamlessly access information about them- selves and their cultural heritage, espe- cially those who have experienced institutional or other out-of-home care. In light of both of these consider- ations, this Policy outlines access guidelines to sensitive and private records held in the Library's collections [9].

The policy goes on to address all instances of sensitive information and lists time restraints for each type of record. Using the privacy laws as guidelines for ethical obligations is something more

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

collection institutions should aim for as it provides a more trustworthy repository for people to commit to and prepares that institution for any future legal changes.

A. Ingest Scenarios

One of the key elements that must be identified is how the donated material relates to the donor and how they came into possession of it. There are many possibilities which change the severity of risk associated with handling such material.

Example 1 - The donated material belongs to and is data about the donor.
Example 2 - The donated material is of ancestry significance to the donor.
Example 3 - The donor has no relation and has discovered or purchased media in which the donated material was found (known material of significance to collection institution).

These examples relate to events prior to ingest as they would dictate how the donor agreement is written up. However, once the data have been collected and processed, further issues may arise as information can be discovered that was not intended nor covered specifically in the donor agreement. Even if the donor had searched through the material before handing it over, there is a chance they missed something. With training and the right tools, significant amounts of information can be uncovered on a system in obscure places, as well as rich amounts of metadata.

Following Example 1, once the donated material has been analyzed, should sensitive information be discovered, further decisions must be made based on what the sensitive information is. If this is covered in the donor agreement, then action should proceed as stated within the agreement. If the agreement does not cover the dis- covered data and the donor is available, the donor needs be involved with any decisions on how to proceed with the uncovered material. There are a few more variables that complicate this procedure. The information may in- criminate the donor and depending on the severity and nature of the discovery, law enforcement may need to be involved.

If this scenario were based on Example 2, this may lead to difficulties for living descendants, however, if no direct harm is caused by disclosing the information, legally there is nothing preventing it. The descendants may fight it and they may try to sue for defamation on be- half of their ancestor, or themselves. However, it should be noted that this is a gray area with an inconsistent history. This is discussed further in section III subsection B.

Another outcome, more likely to occur with Example 3, is the information discovered on donated material may be withheld from the public in their best interest. This may be relating to a public figure, loved and idolized by the country where the discovered material, whilst harmless, may alter how the public sees that figure. Alternatively, the information may need to be disclosed in the best interest of the public, commonly known as "Pub- lic Interest Disclosure" [10]. The donor would have likely signed all ownership of the material over to the collection institution as it has no relevance to them, meaning no further involvement from the donor is necessary in any decision making. There may be policies in place that help in handling such situations, but for many smaller institutions, this may be unprecedented which ultimately makes this an ethical and moral decision. It is situations like these that make this field difficult to develop definitive solutions for because no two cases will be the same, there are always gray areas and variables that complicate decision making.

III. Laws

As mentioned in the introduction, collection institutions such as national and state libraries and archives are exempt from privacy law regarding their collection mate- rial. However, it is important to familiarize oneself with the Privacy Act and the Australian Privacy Principles (APP) as well as determining if you are in fact an APP entity. The APP guidelines define an 'APP Entity' to be an organization or agency. The APP [8] define an organization to be:

An individual, a body corporate, a partnership, an unincorporated association, or a trust. This excludes organizations such as a small business operator, registered political party, state or territory authority, or a prescribed instrumentality of a state.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The APP defines Agencies as (but does not include State or Territory agencies):

A minister, a department, a federal court, Australian Federal Police, a Norfolk Island agency, the nominated Australian Government Health Service (AGHS) company, an eligible hearing service provider, or a service operator under the Healthcare Identifiers Act 2010. Individuals may also fall under the agency category if they hold or perform duties of an office established by or under a Commonwealth enact- ment, or duties for the Governor-General, a Minister, as well as bodies established or appointed by them.

The APPs outline how personal information is handled, used, and managed by APP entities. This applies to most Australian and Norfolk Island Government agencies, private sector and not-for-profit organizations (with an annual turnover greater than A$3 million), private health service providers, and some small businesses. Small businesses (A$3 million or under) have responsibilities under the act if any of the following are true:

Private sector health service providers, businesses that sell or purchase personal information, credit report- ing bodies, contracted service providers for a Common- wealth contract, employee associations registered or recognized under the Fair Work Act 2009, businesses that have opted-in to the Privacy Act, businesses that are related to another business covered by the Act, or businesses prescribed by the Privacy Regulation 2013 [11].

Both the Privacy Act and the APPs are quite extensive, so each principle will not be discussed in detail, but the 13 APPs from the Privacy Act 1988, Schedule 1 are:

- **APP 1:** open and transparent management of personal information
- **APP 2:** anonymity and pseudonymity
- **APP 3:** collection of solicited personal information
- **APP 4:** dealing with unsolicited personal information
- **APP 5:** notification of the collection of personal in- formation
- **APP 6:** use or disclosure of personal information
- **APP 7:** direct marketing
- **APP 8:** cross-border disclosure of personal information
- **APP 9:** adoption, use or disclosure of government related identifiers
- **APP 10:** quality of personal information
- **APP 11:** security of personal information
- **APP 12:** access to personal information
- **APP 13:** correction of personal information

Data security and privacy is always a current issue, ever changing, and highly desired. New Government Legislation Acts and policies are often being created, as are current ones being reviewed and amended as needed. Therefore, it is beneficial to be aware of such changes, for they may not be obligatory for your institution at the present time, but things can change. The European General Data Protection Regulation (GDPR) is a prime example as many would be aware by the policy updates from each service subscribed to. All Australian businesses need to comply if they have dealings in or with the European Union (EU). This includes having a branch in the EU, offering goods and services in the EU, and even if the business is monitoring individuals within the EU. The GDPR shares many requirements with the Privacy Act 1988, but there are additions that are not covered in the Act, one of which is the right to be forgotten [12]. Whilst compliance may not be mandatory, careful review of updated polices and requirements can lead to adopting best practices and better policies.

A. Collection Institutions

There are a few circumstances in which collection institutions need to consider law. These include holding information, making it public, and how the information is being used. The main area of focus is the publicizing of information, as this is where the biggest potential threat lies. There are also risks surrounding the content held within collection institutions, however, there are restricted sections where this information is kept from the public. These sections require special access or per- missions by the author or representatives. The National Library of Australia's restricted area, known as the "Se- cure Room – Restricted" (SRR) is said to be almost as hard to access as *a bank vault with its door shut* [13]. Content is held within the SRR for various reasons, some of the main ones according to Gidney include:

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

- Secret/Sacred Indigenous material.
- Litigation – Ongoing court cases/upheld claims (defamation).
- Commercial in confidence .
- Pornography.
- Refused classification (RC).
- Publication with significant/dangerous errors.

This list alone illustrates the need to carefully con- sider what information is made public, as you can imagine the potential risks involved, should this listed content not be made secure. Secure areas also serve as a holding place for original documents that may have had information omitted for publicly accessible versions. Gidney listed one such case where in 1997 Goodbye Jerusalem by Bob Ellis[1] had a sentence omitted that made some offensive and damaging claims. Furthermore, on the topic of making information public, the disclosure of information marked "commercial in confidence" is forbidden without permission from the supplier. This includes any information that may result in damages to a party's commercial interests, intellectual property, or trade secrets [14].

B. Defamation

Defamation is defined similarly from country to country, but one of the better definitions posted in an article in 'The News Manual', sourced from the British Defamation Act of 19s2 is defined as:

The publication of any false imputation concerning a person, or a member of his family, whether living or dead, by which (a) the reputation of that person is likely to be injured or (b) he is likely to be injured in his profession or trade or (c) other persons are likely to be induced to shun, avoid, ridicule or despise him. Publication of defamatory matter can be by (a) spoken words or audible sound or (b) words intended to be read by sight or touch or (c) signs, signals, gestures or visible representations, and must be done to a person other than the person defamed. [15]

Prior to January 2006, defamation law varied across each state in Australia, but is now covered under the Uniform Defamation Law [16]. Furthermore, there was a distinction between libel and slander prior to the uniform law, however, the distinction was already disregarded in five jurisdictions and the rest of Australia followed with the introduction of the new law [17]. Regarding organizations and companies having the right to sue for defamation, this was possible under the old act, how- ever, under the uniform law, if the corporation exceeds 10 employees, they cannot sue. This does not include not-for-profit organizations, and it does not include individuals within corporations of 10 or more employees if they are identified in the defamatory publication [18].

With all that in mind, it may seem unwise to publicize information, however, there are defenses against defamation claims and they are quite solid. First and foremost, 'truth' is the strongest defense, more so now under the uniform law as public interest is no longer a requirement needed to supplement the truth claim [18], [19]. As long as there is substantial evidence proving the information to be true, the defamation claim will not succeed. Should the claim be won, it may result in actions taken such as in the *Goodbye Jerusalem* case where the defamatory statement was omitted in the pub- lic version. The truth remains the strongest defense for collection institutions, however, it is void should 'malice' be proven, that is, if the information was published with ill-will or with harmful motives. It should also be noted, that should the published material be based on a deceased person, they cannot legally be represented in a defamatory case, even by family members. This of course can change should the published material cause harm for living family members, but they can only claim defamation on their own behalf, they cannot clear the name of their deceased family member [18].

The other defenses include: absolute privilege, qualified privilege, honest opinion, innocent dissemination (unintentional defamation), and triviality. For collection institutions, innocent dissemination is possible, but un- likely as items should be carefully reviewed before be- ing published. Triviality may also prove to be a worthy defense, but the other defenses are not as relevant. Ab- solute privilege covers speech in parliament and court proceedings, meaning whatever is said and whatever motive

[1] Australian Federal politicians Peter Costello and Tony Abbott sued publisher Random House over Bob Ellis's memoir Goodbye Jerusalem, which featured gossip falsely claiming that they had been 'lured to the Liberal Party' by a sexual liaison.

behind it cannot be used to sue for defamation. The reports of these proceedings are then protected by qualified privilege, however, only applicable if the report is honest, for the public, or the advancement of education [18].

IV. Aboriginal and Torres Strait Islander Material

Within Australian collection institutions, historical records are held containing information on Aboriginal and Torres Strait Islander affairs. There are unique policies and procedures for dealing with such records, one of which is commonly used in libraries called the Aboriginal and Torres Strait Islander Library, Information and Resource Network (ATSILIRN). The ATSILIRN protocols act as guidelines for librarians, archives, and all information services that interact with Aboriginal and Torres Strait Islander people or handle materials with such content [20].

The protocols were published in 199s by the Australian Library and Information Association (ALIA) and were then endorsed by ATSILIRN. Updates to the protocols took place in 200s and again in 2010, with 2012 being the latest revision. Once again, these serve only as guidelines, they are not definitive and must be interpreted and applied in context for each issue or situation the protocols may be needed. The protocols cover the following categories:

- Governance and management
- Content and perspectives
- Intellectual property
- Accessibility and use
- Description and classification
- Secret and sacred materials
- Offensive
- Staffing
- Developing professional practice
- Awareness of peoples and issues
- Copying and repatriation records
- The digital environment

Due to Indigenous protocol and sensitivities, some Aboriginal and Torres Strait Islander material must be locked in secure sections of collection institutions, an example of which can be found in the SRR of the NLA. Some of this material may also impose access restrictions and can only be accessed via special permissions such as content classified as 'secret men's' or 'secret women's' business, adding further conditional access [13].

In 2007, the National and State Libraries of Australasia (NLSA) developed a framework to guide National, State, and Territory libraries on how to approach Aboriginal and Torres Strait Islander library services and collections. However, this was superseded in 2014 with the 'National position statement for Aboriginal and Torres Strait Islander library services and collections' [21]. Within the position statement, it is made clear that the following policies/protocols are endorsed: The ATSILIRN, The United Nations Declaration on the Rights of Indigenous Peoples [22], and The National and State Libraries of Australasia Guidelines for Working with Community [23]. The standards that are promoted within the position state- ment include: Rights to be informed about collections relating to the people (culture, language, heritage).

The right to determine access and use of such material. Inclusion of Aboriginal and Torres Strait Islander peoples in all decision-making processes at all levels.Strategies to increase employment and retention of Aboriginal and Torres Strait Islander staff within the library and information sector.Strategies to strengthen cultural competency across the workforce, raising awareness and knowledge on issues for Aboriginal and Torres Strait Islander library users. Strategies to make usable copies of collection material to be returned to the rightful people to support cultural and language maintenance or revitalization.

In summary, the promoted standards aim to ensure rights are given to the people relating to the content, ensuring they have the rights to decide how content is handled and managed, to give the people a chance to be part of the process and to give back to the communities where possible.

Another important position statement from the NLSA is on Intellectual Property and how it differentiates Indigenous content and non-Indigenous content [24]. The World Intellectual Property Organization describes how intellectual property is expressed by Indigenous peoples with the following principles:

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

- Intellectual property is handed down, generationally (orally or by imitation).
- It reflects community cultural and social identify. It consists of characteristic elements of a community's heritage.
- It can be produced by unknown authors or by communally recognized communities/individuals that have been granted the right, responsibility, or the permissions.
- It can often be created for spiritual/religious purposes and is something that constantly evolves within the community.

How Australian collection institutions handle Indigenous material and peoples is a good example of the importance of guidelines and protocols. While not bound by definitive law, we still must consider the affect our collected material can have on others, making this about ethically based, best practice decisions. This should be standard for all material, not just that of Australian Aboriginal and Torres Strait Islander content.

V. Conclusion

Whilst many institutions are yet to encounter issues such as those mentioned in this paper, it does not mean the potential for such issues to occur is not already present. Institutions are storing data, making selected content accessible, and giving it no further thought once processed regarding sensitive material. While some processing may be involved before and during ingest to discover such data, as well as having negotiated agreements with donors in the event such material is found, it may not be enough. Manually searching material or even using built in operating system search functions is not enough for the discovery of sensitive data. Tools exist, freely available, easy to use, and extremely thorough. Tools such as bulk_extractor [5] and The Sleuth Kit (Autopsy) [6] can be introduced into workflows to significantly increase the discovery of sensitive information.

Without a thorough investigation, sensitive information may be sitting in storage that could potentially be problematic. It may be useful information, important and vital to a collection, revealing information that was previously unknown. Hypothetically, should a disk im- age be created from computing system belonging to a historical figure and the collection institution wants to discover as much about that figure as they can, forensically analyzing the system will reveal what could not be seen prior. Hobbies, interests, past-time activities, social groups, and much more can be discovered. While these forensic methods are typically used to discover questionable and illegal content, they can also be used to find the opposite. Both outcomes should be the objective of every collection institution as they may be holding information crucial to an on-going or previously dismissed criminal investigation, or it may simply reveal fascinating new information about an entity within their collection.

The way Indigenous content and people are treated should be the exemplar of how all content and people should be treated. Whilst the protocols differ from culture to culture, the example is that we should be considering all aspects, all scenarios, and all potential issues. By doing so and by following guidelines, preventive practices can be adopted, rather than dealing with issues as they unfold. Admittedly, issues such as those discussed may never surface, depending on what type of digital material an institution is dealing with. However, it is wise to be prepared, especially given that the future will be primarily digital and we do not know how it is going to change, in turn, changing digital preservation.

If we only concern ourselves with the laws to which we are bound and not those from which we are exempt, then it limits our potential to see future issues, hidden threats, best practices, and to generally consider what is best for people. There is never a one-size-fits-all solution, every issue is unique and every guideline must be applied in context. Being aware is the first step to being prepared for any issues or changes in law that may affect collection institutions. We have discussed the laws that are applicable, emphasizing how they may serve as guidelines, we also gave insight into the issues that can arise in collection institutions, providing further aware- ness of current and future threat potential. One cannot prepare for something of which you are unaware of and it is much better to prevent, than fix, making awareness something to strive for.

References

[1] E. LeClere, "Breaking rules for good? how archivists manage privacy in large-scale digitization projects," Archives and Manuscripts, vol. 46, no. 3, pp. 289– 308, 2018. [Online]. Available: https://doi.org/10.1080/01576895.2018.1547653.

[2] Australian Government, Privacy Act 1988, 2018. [Online]. Available: https://www.legislation.gov.au/Details/C2018C00292 (visited on 03/13/2018).

[3] National Library of Australia, Privacy Policy | National Library of Australia, 2018. [Online]. Avail- able: https://www.nla.gov.au/policy-and-planning/privacy-policy (visited on 03/13/2018).

[4] Bitcurator, 2018. [Online]. Available: https://bitcurator.net/bitcurator/.

[5] Bulk extractor - ForensicsWiki, 201s. [Online]. Available: https://www.forensicswiki.org/wiki/Bulk_extractor (visited on 07/11/2018).

[6] Basis Technology, The Sleuth Kit (TSK) & Autopsy: Open Source Digital Forensics Tools, 2018. [Online]. Available: https://www.sleuthkit.org/ (visited on 07/11/2018).

[7] Office of Parliamentary Counsel, "Privacy Act 1988 Compilation No. 76," pp. 19, 27, 2017.

[8] Office of the Australian Information Commissioner, Australian Privacy Principles guidelines, en- AU, 2018. [Online]. Available: https://www.oaic.gov.au/agencies-and-organizations/app-guidelines/ (visited on 06/04/2018).

[9] State Library of NSW, New Sensitive Collections Pol- icy, 2017. [Online]. Available: http://www.sl.nsw.gov.au/blogs/new-sensitive-collections-policy (visited on 06/04/2018).

[10] Queensland Ombudsman, What is a public inter- est disclosure? 2017. [Online]. Available: https://www.ombudsman.qld.gov.au/improve-public-administration/public-interest-disclosures/what-is-a-public-interest-disclosure (visited on 03/14/2018).

[11] Office of the Australian Information Commissioner, Rights and responsibilities, 2016. [Online]. Available: https://www.oaic.gov.au/privacy-law/rights-and-responsibilities (visited on 06/04/2018).

[12] ——, Privacy business resource 21: Australian businesses and the EU General Data Protection Regulation, 2018. [Online]. Available:/agencies-and-organisations/business-resources/privacy-business-resource-21-australian-businesses-and-the-eu-general-data-protection-regulation (visited on 06/0s/2018).

[13] K. Gidney, Restricted! 2016. [Online]. Available: https://www.nla.gov.au/blogs/behind-the-scenes/2016/08/24/restricted (visited on 03/19/2018).

[14] Global Negotiator, What is Commercial in confidence? Definition and meaning. [Online]. Avail- able: https://www.globalnegotiator.com/international-trade/dictionary/commercial-confidence/ (visited on 03/19/2018).

[15] D. Ingram and P. Henshall, The News Manual Chapter 69: Defamation - what you cannot do, 2016. [On- line]. Available: http://www.thenewsmanual.net/Manuals%20Volume%203/volume3_69.htm (visited on 03/20/2018).

[16] D. Doctor, The new uniform defamation laws, en, 2007. [Online]. Available: https://www.artslaw.com.au/articles/entry/the-new-uniform-defamation-laws/ (visited on 03/20/2018).

[17] D. Rolph, "A critique of the national, uniform defamation laws," Torts Law Journal, vol. 16, no. 3, pp. 207–248, 2009.

[18] D. Ingram and P. Henshall, The News Manual: Defamation in Australia, 2016. [Online]. Available: http://www.thenewsmanual.net/Resources/medialaw_in_australia_02.html (visited on 03/20/2018).

[19] L. Huan, Uniform Defamation Laws 2006, en-US, 2006. [Online]. Available: http://www.stephens.com.au/the-uniform-defamation-laws-2006/ (visited on 03/22/2018).

[20] ATSILIRN, Aboriginal and Torres Strait Islander Library and Information Resource Network, 2012. [On- line]. Available: http://atsilirn.aiatsis.gov.au/index.php (visited on 06/06/2018).

[21] National and State Libraries of Australasia, National position statement for Aboriginal and Torres Strait Islander library services and collections, 2014. [Online]. Available: https://www.nsla.org.au/publication/national-position-statement-aboriginal-and-torres-strait-islander-library-services-and (visited on 06/06/2018).

[22] UN General Assembly, Declaration on the Rights of Indigenous Peoples, Oct. 2007.

[23] National and State Libraries of Australasia, Work- ing with community: Guidelines for collaborative practice between libraries and aboriginal and tor- res strait islander communities, 2013. [Online]. Avail- able: https://www.nsla.org.au/resources/working-community (visited on 06/06/2018).

[24] ——, Position statement on Indigenous intellectual property and ownership, 2010. [Online]. Available:

[25] https://www.nsla.org.au/publication/position-statement-indigenous-intellectual-property-and-ownership (visited on 06/06/2018).

Building the future of digital preservation in French archival services

Processes, functions and staffing for an effective digital preservation

Marion Humbert
Ministry of Culture, Paris, France
marion.humbert@culture.gouv.fr

Stéphanie Roussel
Mintika, Paris, France
stephanie.roussel@mintika.fr

Édouard Vasseur
Ministry of Armed Forces – Vitam Program
Paris, France
edouard.vasseur@culture.gouv.fr

Abstract – Studied since the 1980s in French Archival services from a technical point of view, long term digital preservation strategy became a priority a few years ago in the context of Vitam program. The French Archive Interdepartmental Service, in partnership with the Vitam program team and the Ministries of Culture (National Archives), for Europe and Foreign Affairs and of Armed Forces, commissioned a study in 2018 about " The long-term sustainability of digital information ". This collective work offers a modeling of digital preservation processes and an analysis of the skills and profiles needed to implement these processes, not only for the three ministries but for any archival service wishing to diagnose its maturity.

Keywords – Capacity and skills; Preservation planning; Preservation Watch; Reference model; Workflow.

Conference Topics – Designing and Delivering Sustainable Digital Preservation; Building Capacity, Capability and Community.

I. Introduction

A. Context and objectives

Although many heritage services are now raising the issue of collecting digital records and archives, the subject of the long-term preservation of those data is still poorly understood in its concrete aspects. What perimeter should be given to the word "digital preservation"? Which processes come into play? What skills to involve?

In an attempt to answer these questions, the French Archive Interdepartmental Service (SIAF), in partnership with the Vitam program team and the Ministries of Culture (National Archives), for Europe and Foreign Affairs and of Armed Forces, commissioned a study in 2018 about " The long-term sustainability of digital information ". This study was a complement to the work done in the Digital preservation working group led by the Vitam program [1]. Indeed, thanks to the working group, perimeter of the function "preservation", tools' limits and complex nature of processes are now better understood and shared by the actors, who defined some software development priorities in order to implement preservation functionalities in the Vitam Software Solution. These monthly meetings facilitated internal implementation and change management to "make digital curation a systematic institutional function" [2].

The study's aim was to assess the workload induced by the implementation of digital preservation using the digital archiving Vitam software in sponsor organizations [3].

B. Methodology

The study was entrusted to two consulting firms specialized in supporting archives and information management services: Serda and Mintika.

It was conducted in 3 phases:

- The definition phase aimed at providing a clear, shared understanding of "digital preservation" in the broader context of information lifecycle management and to provide resources to final users (directory of experts, summary of applicable standards, bibliography)

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

- The core phase detailed the digital preservation processes required for an effective implementation of the function. It also identified the skills and expertise necessary.
- a third (non-public) chapter provided sponsor organizations with options for change management and long-term support of digital appropriation, human resource estimation and mutualization and outsourcing priorities.
- Deliverables of Phases 1 and 2 are generic enough to be used by any archival service, and were not designed solely for the three ministries. Indeed, they are published and publicly available (in French) on "FranceArchives" resource website [4].

C. Perimeter

The study started with an analysis of the perimeter of the "digital preservation" domain and a mapping of concepts related to digital preservation. This immediately reveals that formats, which are usually the first line of inquiry associated with digital preservation, represent but a small part of the information available on the subject.

After discussions between the partners, it was decided, as shown in the figure below, to include in the studied perimeter not only technical and strategical aspects of digital preservation but the starting point of the document life-cycle: acquisition and the aim of preservation: access.

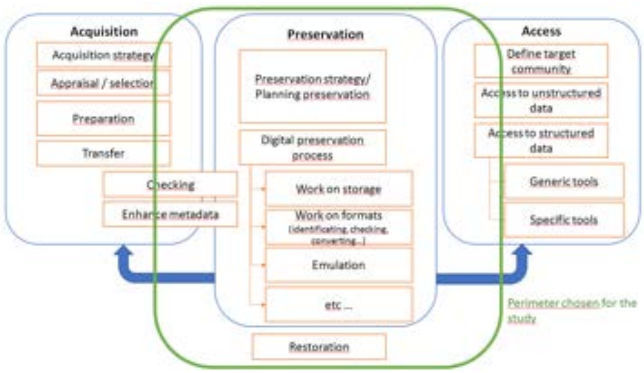


Figure 1: perimeter chosen for the study

II. Modeling processes

A. Why?

Several modelings about digital preservation already exist, especially the ISO OAIS Reference model [5], the proceedings of the international collaborative research project Interpares [6], partially translated in French [7], or the PLANET Model, designed from OAIS [8] and its projects of implementation in several institutions [9]. These approaches are interesting, but take place in a larger documentation, sometimes not translated in french. Furthermore, the Interpares UML modelling is not the most adapted for large communication when the diffusion of an all-in-one document to professionals and to the hierarchy is one of the main aims of the study.

The primary objective of this modeling was to raise awareness of the magnitude of the task throughout the network of archival public services. Indeed, while more and more electronic archiving systems are being deployed in France, the issue of digital preservation is still poorly understood by most professionals. This work is therefore intended to be educational. It is also possible to imagine future documents that will accompany this modeling to facilitate its appropriation and diffusion. Network heads thus have an important role to play in developing future digital preservation doctrines for all archiving services. It is indeed obvious that not everyone will have the means to deploy such varied skills and bear such high costs as have been highlighted in the modeling.

B. How?

The modelling was developed on the basis of this available professional literature (in particular the OAIS model and works of Interpares) as well as information gathered during interviews conducted with the Ministries of Culture (National Archives), for Europe and Foreign Affairs and of Armed Forces and a number of partners operating major electronic archiving systems (the National French Library, the National Computer Center of Higher Education and the National Audiovisual Institute). This double perspective – a conceptual approach confronted with a very operational approach – is the basis of the work.

The in-depth review of OAIS and Interpares enabled us to identify the key points of a preservation system by specifying for every process the entry and exit points. We identified these key points in the form of post-its that we were then able to arrange at will to correspond to the concrete reality of the projects and partners in the study. This synthesis work was the first step for the modelling.

Finally, OAIS and Interpares were also sources of thinking on how to model and approach the subject of preservation. It was out of the question to redo what had already been done, but we would rather propose, on the basis of this conceptual work, a very operational approach to the implementation of preservation, in the particular context of French archival practice.

Consistency with these different works, in particular with OAIS, is a guarantee of the interoperability of the model. We can thus imagine that this modeling of the digital preservation function could be appropriate for other contexts than the French context.

C. The use of the BPMN modelling language

The language used for modelling is BPMN (Business Process Model and Notation), which provides "a business process model and notation to describe an organization's value chains and business activities in the form of a standardized graphical representation" [10]. This model has the advantage of being standardized (ISO/IEC 19510). It guarantees a certain stability and interoperability, which seemed important in our context.

BPMN is also widely documented, making it easy for all users to understand and master it. Finally, BPMN is a common language in the developer community, which eases the accustomization to a community unfamiliar with purely archival issues.

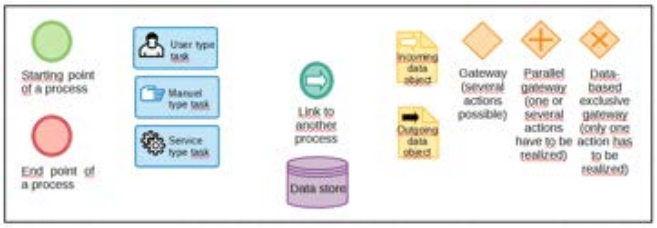


Figure 2 BPMN caption of the modelling

D. The final result

The final modelling details 19 processes, split in the following categories:

- ingest: 2 processes
- access: 1 process
- storage: 1 process
- administration: 5 processes
- preservation planning: 6 processes
- management & governance: 4 processes

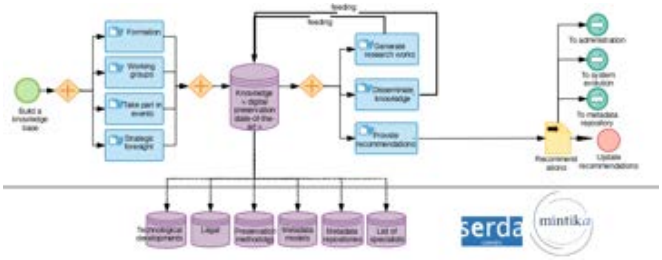


Figure 3 Extract from the modelling of the processes included in "Preservation planning"

By detailing the processes we are able to identify which ones require the more resources: appraisal, ingest, format tests, strategic foresight, international events, participation in development of collaborative tools addressing specific elements in the French environment, i.e. the Lambert conformal conic projection, initially missing from the GeoTIFF specifications.

This process-based approach was planned to be complemented with a skills and profiles thinking. In a different context, the subject has already been studied in France by the National French Library [11].

III. Designing digital preservation functions and evaluating ideal staffing

A. Why?

In France, the training in heritage preservation doesn't include a major in long-term digital preservation, neither in archives, nor in libraries or museums. Consequently, the skills needed for digital preservation are not precisely defined, or not defined at all, in the tools and repositories of professional associations [12] or administrations' human resources offices [13].

These elements of expertise had to be detailed, particularly those related to information watch operations, critical for the decision process. The technical and archival skills necessary to operate an electronic document and record management system were an important part of the reflection. Part of these skills can be found among computer scientists while others are more specific and imply thorough archival knowledge [14].

The 2018 french modelling will be completed by exploiting and translating international resources about formation in digital preservation, like the SCAPE project [15] and the DigCurV Framework [16], published in 2013 and used by the Oxford University to develop continuous training in patrimonial institutions [17]. In addition to providing surveys and methods of evaluation for active professionals,

these tools will enhance our reflection about the initial formation.

The aims with regard to formation are both increasing the number of digital preservation professionals and having a better understanding of our needs, especially in order to share skills and knowledge with smaller institutions.

B. How?

As mentioned beforehand several times, the objective of our model was to allow partners to project themselves into a future organization. To define profiles and facilitate the appropriation of the model by the study sponsors, it seemed important to use existing human resources categories, even though these frameworks will have to evolve over time to take into account new needs. These profiles were then compared with the references in the professional literature – in particular the NDSA report [18] – and mapping done whenever possible.

C. Identified functions

Profiles for each process (archivist, developer, system administrator, etc.) have been detailed in the form of a diagram and table, in order to facilitate in a given context the identification of the skills required to implement a strategy for the sustainability of digital data.

The published report focused on the profiles specifically related to digital preservation: format management before and after payment into the electronic archiving system, evaluation, conversion, standby. Resources required for the treatment and the analysis of the data themselves or the constitution of the chains of processing when it comes to releasing data from applications were considered out of scope.

Profiles are based on a list of skills needed to implement digital preservation, including archivists (taking in account the balance between archival and technical skills [19]), formats specialists, data restorers, developers and quality specialists.

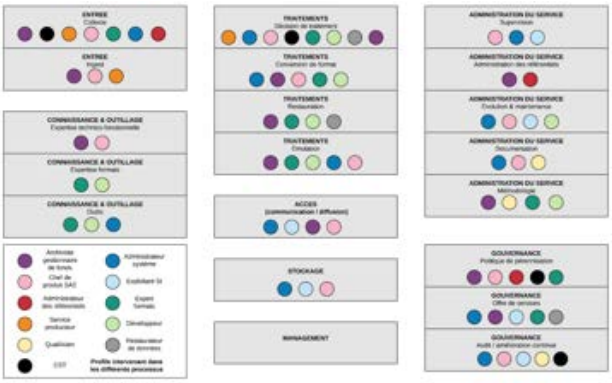


Figure 4 Profiles required for each process

Function	Description	Function in NDSA
Administrateur de référentiel	In charge of management of metadata repositories	Metadata Analyst
Administrateur système	In charge of the upkeeping, configuration, and reliable operation of computer systems	System administrator
Archiviste gestionnaire de fonds	In charge of the respect of archival science principles in the digital environment	Cataloguer / Archives & special collections curator / Digital Archivist / Collection needs analyst / Electronic record archivist
Chef de produit SAE	In charge of administration and supervision of the electronic document and record management system	Administrator / Digital Archivist
Contrôle scientifique et technique	Audit by the archival Authorities	
Développeur	In charge of software development process	Software Developer
Expert format	In charge of developing and maintaining knowledge of archived formats	Format specialist
Exploitant du système d'information	In charge of software maintenance, infra-structures and cybersecurity	
Qualiticien	In charge of quality management system	Workflow specialist
Restaurateur de données	In charge of developing technical solutions to fix damaged digital documents	Data restorer (not in NDSA report)
Service producteur	Organization who created the documents and may have expertise on some formats	Archive-producing department (not in NDSA report)

Table 1 Correlation between functions identified by NDSA report and our modeling.

The legal-related functions have been considered out of the strict perimeter of digital preservation in the study, and therefore excluded.

D. Ideal staffing

The three scenarios identified for each ministry are based on the needs and only take into account the work load directly related to the digital preservation of data.

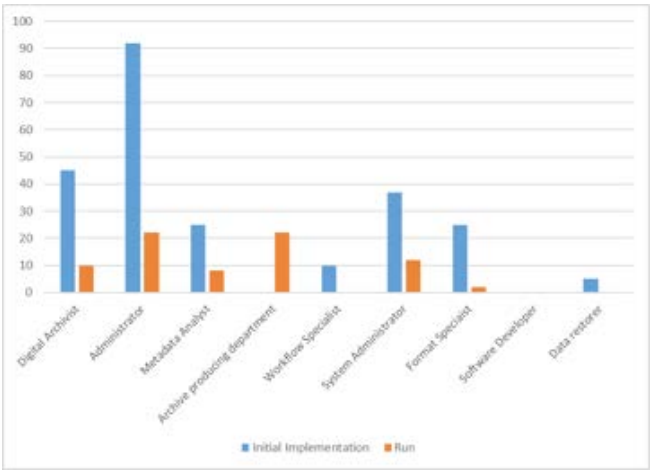


Figure 5 Estimation of the human resources (day/year)necessary to the initial implementation and run of digital preservation in the National Archives.

This projection will evolve as the sponsors will develop their maturity and set up the processes described in the modelling.

IV. Work in progress

A. Advocacy

In order to raise awareness about the subject of long-term digital preservation within sponsor organizations, a presentation of the results of the study will be submitted to the Interdepartmental Committee of French Archives in April 2019. The presentation will point out the possibilities to share resources and expertise between the three ministries, particularly for the tasks related to technology intelligence. The Ministry for Europe and Foreign Affairs presented internally the results of the study, in order to prepare change management as soon as the processes will be ready to be implemented.

These political and strategical aspects will be completed by a more technical and operational working group, created within the national PIN Group [20] (Digital Information Preservation) in order to pursue the tests and exchanges started in the Vitam program

B. Implementation

Each sponsor should now assess its situation and resources with respect to the modelling and skills identified in the study. It offers 25-year estimation in order to help anticipating the resources needed for an efficient strategy of digital preservation. Several others archives in France plan to use the study to perform diagnoses.

For example, the Ministry of Armed Forces will use the results of the study in the following ways:

- Elaboration of a digital preservation governance within the institution.
- Evaluation of the human and budgetary resources necessary for an efficient strategy of digital preservation. A part of this action consist in the identification of the current resources.
- Identification of ten actors in the preservation process and elaboration of an expert profile, who will coordinate all the actors and operations linked to digital preservation.
- Bridging the gap between the "Digital preservation" processes and the processes "Preventive and curative preservation" which already exist, and integration of these processes in the data life cycle (ingest, management, access).
- Update of existing models and internal processes, inclusion of results in ongoing projects.
- Bridging the gap between the traditional archival concepts (appraisal, description, preservation...) and the "Digital Preservation" processes.

Similarly, the National Archives are planning in 2019 to complete their digital preservation strategy thanks to the scenarios of the study. Besides, institutions with particular competences in formats and long-term digital preservation, like the National Computer Center of Higher Education are working in order to enhance the Vitam software solution and will share their developments with all the community of users.

For many institutions, especially the smallest, contributing to research and development works, creating new posts will be difficult, and the preservation processes will have to be implemented with constant resources. That is why the implementation of the modeling has two important issues: firstly, national and international coordination in order to reduce costs and have the widest possible network of expertise and then networking of existing competences.

C. Communication

Even though the study addressed internal needs of the sponsor organizations, a larger diffusion among French archival services has been made through a publication in FranceArchives, the national portal of French Archives. The model and mind-map of the scope, the processes, a bibliography and a summary of the most important standards and norms in the digital preservation area are downloadable.

Furthermore, the results of the study will be presented in April 2019 during the French Archivists Association forum. A panel associating representatives from SIAF, Vitam team, the Ministry of Armed Forces and the National Computer Center of Higher Education will present the deliverables as the foundation of many concrete applications.

D. Conclusion

Carried out at an interdepartmental level, this study took place at an ideal time, just while the Vitam project and its autonomous implementations in each ministries become operational. Without disrupting ongoing projects within each entity, it has enabled stakeholders to consolidate their strategy. They are expected in turn to adjust the practical implementation of the Vitam software solution, to identify new skills to be internalized or outsourced, to identify new profiles to develop and how to include these processes in the pre existing standards and frameworks [17].

By its concrete and quantified nature, it provides archival services with essential tools for dialogue with decision makers and enables them to consider future joint actions by pooling certain necessary functions, especially in preservation watch.

Acknowledgments

The SIAF and the Vitam program team would like to thank Pierre FUZEAU and Olivier MOLLET from the firm Serda; Baptiste NICHELE from the firm Mintika; Wilfried PRIEUR; Blandine WAGNER; Sophie RESSEGUIER and Olivier CHEVENIER from the Ministry of Armed Forces; Martine SIN BLIMA-BARRU from the National Archives; Erwann RAMONDENC and Juliet GARNIER from the Ministry for Europe and Foreign Affairs; Eleonore ALQUIER and Valérie CHAUMELLE-SERRUS from the National Audiovisual Institute; Stéphane REECHT, Thomas LEDOUX and Laurent DUPLOUY from the National French Library; Lorène BÉCHARD and Olivier ROUCHON from National Computer Center of Higher Education and Aurélien CONRAUX from the Digital Archives, Standards and Authorities Office (SIAF), for their contributions.

References

[1] M. REBOURS, “The Digital Preservation Working Group led by the interdepartmental Digital Archiving Program - Vitam”, Ipres 2018.

[2] C. PROM, “Making Digital Curation a Systematic Institutional Function”, The International Journal of Digital Curation, p. 139-152, 2011.

[3] E. LABORDE, M. REBOURS, E. VASSEUR, “Vitam Software Solution’s preservation”, Ipres 2018.

[4] ISO 14721:2012, Open Archival Information System. Traduced in French in 2017.

[5] “Préserver les données numériques : de quoi parle t-on ?”, Portail FranceArchives, 2018. URL: https://francearchives.fr/fr/article/162808225 (accessed 13 March 2019).

[6] Interpares 2 Project, Chain of Preservation Model, 2002-2007. Translated in French in 2010.

[7] C. NOUGARET, F. BANAT-BERGER, “La traduction française d’Interpares 2 : un nouvel outil conceptuel pour l’archivistique”, La Gazette des Archives, n°245, p. 67-77, 2017.

[8] B. SIERMAN, P. WHEATLEY, Evaluation of Preservation Planning within OAIS, based on the Planets Functional Model, 2010.

[9] P. LUCKER, M. RAS, B. SIERMAN, R. van VEENENDAAL, “Preservation Watch at the National Archives of Netherlands”, IPres 2018.

[10] ISO/IEC 19510:2013, Object Management Group Business Process Model and Notation, 2013.

[11] E. BERMÈS, L. FAUDUET, “The Human Face of Digital

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Preservation: Organizational and Staff Challenges, and Initiatives at the Bibliothèque nationale de France", The International Journal of Digital Curation, p. 226-237, 2011.
[12] Association des archivistes français, Référentiel métiers, la profession d'archiviste ou les métiers des archives, 2e édition, Paris, 2016.
[13] Ministère de l'action et des comptes publics, Répertoire interministériel des métiers de l'État, 2017.
[14] J. KIM, E. WARGA, W. MOEN, "Competencies Required for Digital Curation: An Analysis of Job Advertisements", International Journal of Digital Curation, p. 66-83, 2013.
[15] SCAPE, 2014, Scalable Preservation Environments (scape-project.eu)
[16] DigCurV, 2013, DigCurV Curriculum Framework (https://www.digcurv.gla.ac.uk)
[17] S. MASON, E. HALVARSSON, "Designing and implementing a digital preservation training needs assessment", Ipres 2017.
[18] A. WINSTON, et al., Staffing for digital preservation 2017, National Digital Stewardship Alliance, p.20-36, 2017.
[19] C. A. LEE, H. TIBBO, "Where's the Archivist in Digital Curation? Exploring the possibilities through a Matrix of Knowledge and Skills", Archivaria, p. 123-168, 2011.
[20] Groupe de travail "Pérennisation des informations numériques" (http://pin.association-aristote.fr/doku.php/accueil)
[21] C. BECKER, G. ANTUNES, J. BARATEIRO, R. VIEIRA, "A Capability Model for Digital Preservation. Analysing Concerns, Drivers, Constraints, Capabilities and Maturities", Ipres 2011.

Preservation of Metadata

A case study of a strategy to ensure technology agnostic metadata preservation

Eld Zierau
Royal Danish Library, Denmark
elzi@kb.dk
ORCID: 0000-0003-3406-3555

Abstract – This paper focuses on strategies for preservation of metadata; one of the major topics in the creation of a new digital preservation strategy for the merged Royal Danish Library. Preservation of metadata is important to ensure preservation of all relevant information in order to be able to access digital objects in the future. This includes all kinds of metadata, which contributes to the understanding of an object, e.g. preservation metadata as well as descriptive metadata.

The need for metadata is commonly accepted, but it is not as commonly accepted that metadata need to be preserved to the same extent as files. There are many challenges due to the fact that metadata are often updated. This is probably one of the reasons why there exist numerous examples of metadata being under backup only and not under bit preservation.

Preservation of metadata is not just needed for future access of the objects, but also for re-establishment of repository systems after major breakdowns – or for establishment of a new repository as part of an exit strategy from a previous system. The latter case may also mean that the metadata have to be structured in a way, which can be used by different systems supporting digital preservation.

This paper describes how the newly merged Royal Danish Library has created a digital preservation strategy to cope with these challenges, and discusses the background for choices made in this process.

Keywords: metadata; preservation; data model; strategy; policy; exit strategy

Conference Topics: Designing and Delivering Sustainable Digital Preservation; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

This paper describes how Royal Danish Library will ensure proper preservation of metadata. This is partly formulated in a new digital preservation policy (overall level) and strategy (more operational level) [1,2]. The policy and strategy reflect a merge of different preservation policies and strategies from two former libraries. Before 2018, Denmark had two national libraries which were each responsible for different types of materials, e.g. one library was responsible for collecting and preserving Danish Radio and TV production, while the other library was responsible for collecting and preserving Danish texts and computer games as well as donations from deceased authors. Furthermore, the two libraries were jointly responsible for the Danish web archive. In 2018, the two libraries were merged and named Royal Danish Library, and consequently three different digital preservation policies and strategies had to be merged and aligned in one set of digital preservation policy and strategy.

The purpose of this paper is to give an example of how this new strategy can align different digital preservation strategies and enable preservation of metadata along with exit strategies for different current and future systems. Additionally, the new strategy enables access to all metadata for all materials independent of the system from which they originate. Furthermore, it is our hope that this paper can form a basis for feedback and further discussion of metadata preservation strategies.

Most organizations use some sort of system to support their digital preservation. However, many systems do not support preservation of metadata,

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

but only preservation of files, when it comes to the low-level bit preservation. Examples are Preservica[1], Cumulus[2] and most (if not all) Fedora systems[3].

Even if systems do support some sort of bit preservation of metadata, the format of metadata usually depends on the system. Actually most systems have their own way of structuring and exporting metadata, therefore it is important to stress that the points made in the following description is ***not*** a critique of the above-mentioned systems, but examples of a general challenge. This is also why the systems are not addressed as preservation systems, since no current system covers all the needed digital preservation support. Examples of non-standardized formats for metadata in commercial preservation supporting systems are e.g. XIP in Preservica and DNX in Rosetta[4]. The same can be said for many locally developed systems, and systems built on open source software like Fedora. Examples are a Fedora 3 based system at Royal Danish Library in which the Fedora 3 FOXML was used as container for metadata; and another where at some point metadata were preserved from a Cumulus system with locally key/value defined metadata.

Different efforts have been made in order to deal with a standardized way to structure and pack metadata. For example, METS[5] is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library. Today, METS is used as a container format for wrapping different metadata, and several initiatives like e.g. E-ARK[6] have METS as part of their framework with an additional recommendation of how to get a more standardized way of representing metadata.

Even within the standardized ways to structure metadata, it is a well-known fact that there is no one standard to fit all, and there are many ways to combine the use of different standards, e.g. combination of MIX[7], METS and PREMIS[8] [3].

This paper describes the choices made by Royal Danish Library to fulfill the purpose of having independent metadata preservation. The descriptions are accompanied by the reasons for choices to enable a basis for discussion as well as reuse of argumentation for organizations with similar challenges.

The paper will start by providing a general description of the importance of metadata preservation along with the implied requirements for and challenges in metadata preservation. Next, the relevant strategies and policies for metadata preservation is described. To illustrate the strategy, the paper includes a detailed preservation example on several systems.

II. Metadata Preservation

Preservation of metadata is important to ensure preservation of all relevant information in order to be able to access digital objects in the future. This includes all kinds of metadata, which contribute to the understanding of an object, e.g. preservation metadata as well as descriptive metadata.

In the Open Archival Information Systems Reference Manual (OAIS), this is a question of preserving the full Archival Information Package (AIP) [4] p. 1-9:

"An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS",

where Content Information is [4] p. 1-10:

"A set of information that is the original target of preservation or that includes part or all of that information.

..."and where Preservation Description Information is [4] p. 1-14:

"The information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, Context, and Access Rights Information."

In other words, the AIP needs to include all

[1] https://preservica.com/
[2] https://www.canto.com/cumulus/
[3] https://duraspace.org/fedora/
[4] https://knowledge.exlibrisgroup.com/Rosetta
[5] http://www.loc.gov/standards/mets/
[6] http://www.eark-project.com/
[7] http://www.loc.gov/standards/mix/
[8] http://www.loc.gov/standards/premis/

relevant metadata for the package, not only for future access, but also in order to understand its context (not necessarily implemented as a package, but as information that is findable). Thus, metadata need to be preserved and bit preserved to the same extent as the data they describe.

Even at the bit preservation level, preserving metadata is not an easy task, since they are usually dynamic in the sense that they can be frequently updated. Following preservation guidelines strictly, all versions of metadata must be preserved with an audit trail containing information about what was changed and when. In practice, this may be a resource consuming task in cases where there are frequent minor corrections. A different strategy could therefore be to accept risk of loss of some information by pooling changes, and ingest them into preservation at longer time intervals and reduce the requirements of audit trail information. In any case, this would require some sort of version control of metadata.

To conduct preservation activities, all the preserved data and metadata must be available for (at least) digital preservation activities via some sort of repository system. Taking into account the rapid changes in technology during the last decades, long-term preservation will eventually imply the following requirements for a repository system: 1) repository software must be exchangeable 2) repository must be re-establishable from preserved data.

In other words, metadata must be preserved, and exit strategies for running repository systems must exist. These were also the main requirements taken into account when formulating the sections regarding metadata in the newly merged Royal Danish Library's Digital Preservation policy and strategy.

No matter which preservation supporting system a repository is currently using (commercial or non-commercial), there is a very high probability that the system will have to be replaced later on since this is about long-term preservation. This leaves some hard choices in case the preserved metadata are structured in a way that is dependent on the system, which must be replaced. One option is to let the metadata remain in the same system dependent format, which over time will result in different metadata structures from different systems. Another option is to convert the metadata. For large amounts of metadata, this can be a huge task, with the added risk of losing information during the conversion. The Cumulus based metadata in Denmark were converted, which resulted in a project lasting more than a year, even though the amount of data was relatively small[1].

III. Independence of Preservation Supporting Systems

One of the major steps for the newly merged Library was to formulate a common policy and strategy for digital preservation covering all variations of digital materials.

The goal for both former libraries has always been to ***preserve data and metadata*** in a form that can be interpreted and understood in the future. Therefore, both former strategies focused on the use of appropriate ***standards*** for metadata, implementing a ***technology watch***, and basing all digital preservation decisions on proper ***risk management***. To reach this goal, both libraries aimed at becoming a ***Trustworthy Digital Repository***. This implies an aim to create a robust ***organizational*** anchoring of the work with digital preservation, in a way that maintains consciousness and responsibility of digital preservation as one of the key tasks for the libraries. Consequently, these policies could easily be mapped into a common policy for the merged Royal Danish Library.

However, the detailed strategies to achieve these goals differed a lot at the two former libraries. While the former State and University Library had to handle daily deliveries of huge amounts of homogenous data for TV transmissions, the former Royal Library had to handle heterogeneous materials with varying levels of confidentiality and complexity. Furthermore, the libraries had chosen different systems to support their preservation; just before the merger one of the libraries signed a contract with Preservica to replace the existing Fedora 3 based internally developed system, - and the other library relied on a Cumulus repository system with preservation services built around it.

[1] About 185000 records were converted.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The merged Royal Danish Library of course aims at creating a uniform ***Technical infrastructure*** with digital preservation processes to optimize as much as possible when taking into account the requirements for bit safety, confidentiality, accessibility and the complexity of the digital materials. This means that the new strategy had to take into account an aim at as few systems as possible on the long run, and to at least get an as uniform way to handle preservation as possible. Since the market of supporting preservation systems is constantly moving, this has led to a strategy of keeping the Library as independent as possible of the systems used to support its digital preservation activities. This has resulted in a strategy of an intermediate infrastructure (as a goal for the next three to five years) illustrated in Fig. 1.

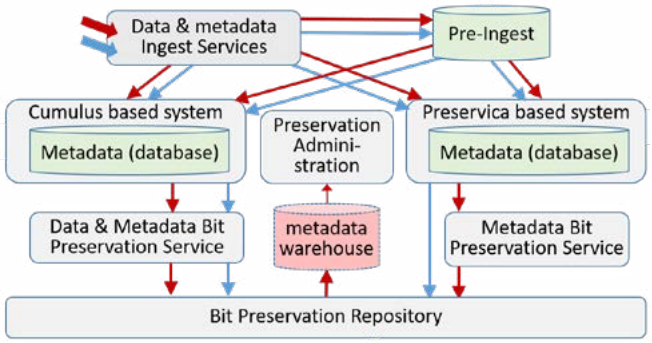


Figure 1 Strategic Royal Danish Library Technical Infrastructure.

The blue arrows specify flows of metadata, while the red arrows specify data flow.

In this infrastructure input data and metadata will be made as uniform as possible by common ingest services. Realistically, ingest of data and metadata cannot be processed by preservation supporting systems right away. The reasons for this are many, for example, the new types of material may need extra processing, implementations to the existing systems may need adjustment, and there are bulks of materials to be ingested, which must await capacity expansion. This is also why the architecture includes a pre-ingest area.

The number of supporting preservation systems will be reduced, since it is expected that all data preserved in the old Fedora 3 based system will be migrated to Preservica.

Metadata from Preservica will be bit preserved by a special metadata bit preservation application, which will extract metadata from Preservica and bit preserve them in the agreed standardized form, in the same way as the already existing metadata preservation application for Cumulus does.

Finally, a metadata warehouse is planned with an extract of the bit preserved metadata in a standardized format, which can provide an overview of all preserved metadata in a uniform way independent of the systems handling preservation. The metadata will be technology independent and can therefore cover metadata from all current systems, thus the warehouse can also be basis for support systems, e.g. preservation administration in the form of preservation planning.

IV. Metadata Strategy

Royal Danish Library has formulated a vision for data and metadata in the new digital preservation strategy: "Royal Danish Library bit preserves data and metadata in a form, which enables the material to be understood and interpreted in the future".

Bit preservation of metadata is carried out to ensure preservation of materials that are to exist on the long term, but also to protect considerable investments in the form of time and resources spent on the creation of metadata for digitized materials with existing analogue copies.

The Library's strategy is to preserve in formats suited for digital preservation (both regarding data formats and metadata formats). To the furthest extent possible, the Library will use open formats, which are standardized or internationally acknowledged. Furthermore, the strategy is to use the metadata formats in a way as close to international standards and best practice as possible. The purpose of these choices is to increase the probability of understanding the metadata in the future and the example with Cumulus data showed that non-standardized metadata can become hard to interpret very quickly.

Royal Danish Library aims to preserve its digital collections in as few metadata formats as possible. However, a full normalization of metadata formats in preservation will never be possible. The reasons are

that there are many specialized metadata formats for different types of data (e.g. MIX for still-images only) and many tools to support export of metadata in non-standardized formats exist. Furthermore, the two former libraries used different standards for descriptive metadata. The Royal Danish Library does not want to change existing metadata, since conversion of metadata from one format to another is a non-trivial task with risk of data loss.

In order to be able to interpret and understand data and metadata in the future, it is necessary to carry out a number of tasks in relation to how data and metadata are interrelated. This includes employment of a data model for metadata, to enable an independence of technology in general.

A data model includes relations and identifiers to identify related items, therefore the strategy explicitly states that relations to data must be preserved and that the Library uses universal, unique and persistent identifiers for identification of the preserved digital materials.

When digital objects and/or metadata are updated or changed, the new versions must be bit preserved along with a log of the changes, i.e. an audit trail for both material and metadata. These audit trails are regarded as metadata as well, and should therefore be bit preserved.

It is an aim for the Library to use combinations of the same metadata standards in as similar a way as possible. This will streamline the general preservation and make it easier to develop general access platforms for the preserved materials.

Generally, Royal Danish Library wants to preserve the following types of metadata:

- *Descriptive metadata*
  with information describing the content of the digital object
- *Administrative metadata*
  with necessary information for curation of the digital object, including:
  - *Technical metadata*
    e.g. file format, checksum and digitization information. These metadata are obtained by characterization of the material or delivered from the digitization process. The characterization is performed as early as possible in the lifecycle of the material. The output from the characterization is preserved along with information about the tools performing the characterization
  - *Preservation metadata*
    which includes the necessary metadata in order to perform digital preservation actions, e.g. level of bit preservation and logical preservation strategy
  - *Digital provenance*
    which includes audit trails for actions performed on the digital object. This includes metadata from the creation of an object, e.g. the scanner's serial number for digitized material, speed of a record player used for digitization, or software which creates an iso-image from a hard drive
  - *Rights metadata*
    which can be used to deduce who can be granted access to the material
- *Structural metadata*
  with information about structures, which may be inherited in the digital object (e.g. references to pages in a book object)

Royal Danish Library structures its metadata as illustrated in Fig. 2 (inspired by what The National Library of Australia has described [5]).

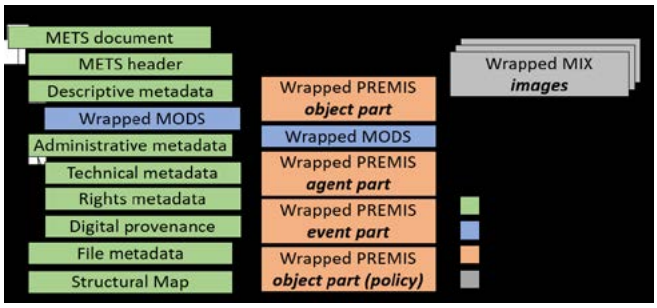


Figure 2 Metadata Model used at Royal Danish Library.

Some of the categories of metadata overlap, e.g. technical metadata are usually also regarded as preservation metadata etc. In practice this also means that metadata schemes overlap, and therefore can be used together in different ways, e.g. as described for METS and PREMIS in Ref. [3].

Royal Danish Library has chosen to use METS as container format for metadata, PREMIS for specific preservation metadata, and various XML

based specialized standards like MIX for technical metadata.

In order to ensure long-term interpretation of metadata, the metadata profiles and schemes are publicly available at the website id.kb.dk, which is harvested and preserved by the Danish web archive. The name id.kb.dk is inspired by id.loc.gov for registries hosted at the Library of Congress.

V. Data Model

The data model for bit preserved materials is designed for long-term use. It can contain data in a way, which supports delivery of whichever part of the data is needed for any application treating or publishing the material. In other words, any relation can be re-established by processing the bit preserved data. In order to ensure long-term interpretation of the model, it is designed to be as simple as possible, without the optimization that front-end use scenarios may need. Such optimizations will be placed at other levels with additional information, which can support the optimization.

The data model is a simplification of the data model used in PREMIS and the data model used in the Planets project[1] (and partly used in Preservica).

This section will describe the data model on the theoretical and conceptual level, while the next section will describe a detailed simple example of how the data model is used in practice to preserve metadata at Royal Danish Library.

The data model has three basic data model entities (illustrated in Fig. 3 below):

Digital Intellectual Entity, which expresses the top level of a digital object and unambiguously identifies a digital material. The object must be identifiable, regardless of which preservation actions, corrections or transformations have been carried out on the object through time. A Digital Intellectual Entity differs from a FRBR Intellectual Entity, since different manifestations of a FRBR Intellectual Entity would be interpreted as different Digital Intellectual Entities.

A Digital Intellectual Entity will always consist of one or more Representations, which represents versions of the Digital Intellectual Entity.

Representation which expresses a Representation of a Digital Intellectual Entity, i.e. it represents a specific version of a particular Digital Intellectual Entity. A Representation can only represent one Digital Intellectual Entity.

The contents of a Representation can differ depending on whether it describes metadata in relation to a file or whether it describes metadata in relation to other Digital Intellectual Entities. Representations for a file usually contains metadata about the file and a reference to the preserved bit sequence representing the file. Representations for relations to one or more Digital Intellectual Entities contain metadata about the referred entities and their mutual relationship (if more than one). An example is the order of pages in a book, along with descriptive metadata about the book.

File / Bit sequence, which expresses a single file / bit sequence that is bit preserved.

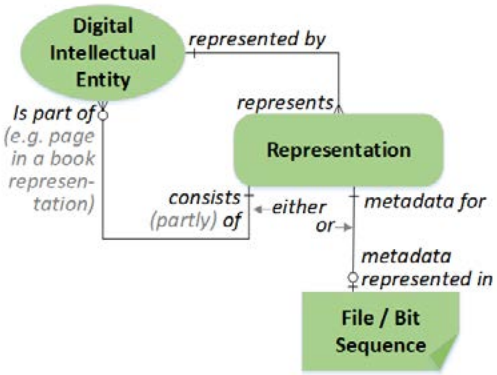


Figure 3 Data Model Entities and Relations.

There is no formula for how to model an object. Such decisions are made as part of the initial curation of the object, i.e. as preparation for or as part of ingest in a digital preservation supporting system.

VI. A Detailed Postcard Example

The example illustrates how a digitization of a postcard is represented in terms of the data model and metadata scheme described above.

[1] https://planets-project.eu/

Figure 4 Digitized Postcard from "Majus smykker" of the H. C. Ørsted award 2013, photo by Grethe Aasted Therkelsen.

The postcard example (Fig. 4) is constructed to illustrate all possible changes. The postcard is digitized and updated in various ways over time by events in the following order:

a) Digitization of the front page of the postcard
b) Edits of metadata to the front page image (correcting Danish character encodings)
c) Re-digitization of the front page (because of errors in the first scanning)
d) Adding digitization of the back page (containing additional information)

A. Modelled Postcard Example

The digitization steps are illustrated in Fig. 5 and explained in the following.

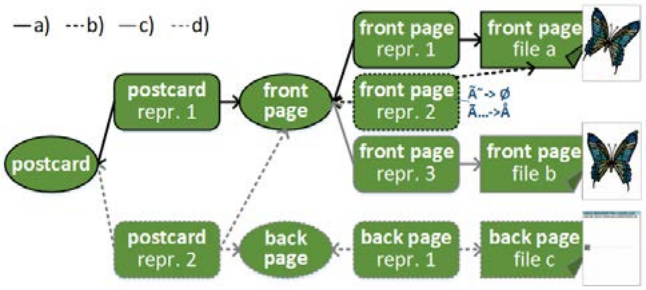


Figure 5 Data Model for steps in digitization of a postcard.

The digitization step a) produced a file and some metadata. In the data model, it produced the Digital Intellectual Entities and Representations along with the file. The Digital Intellectual Entity of the postcard only consists of an identifier, which is the valid reference for all versions (or rather Representations) of the postcard through time.

The Representation of the postcard "postcard (repr.1)" contains the relevant metadata for the postcard. This postcard Representation has no technical metadata, since no files are involved. However, it does contain "structural metadata" of the postcard pointing to the Digital Intellectual Entities of the front page. It cannot refer to the front page Representation, since this would mean that a minor page change would result in having to update the postcard Representation as well. This could easily start a chain reaction, since objects pointing to the postcard Representation would have to change as well. Consequently, it could become so large that it would require bit preservation of an additional large amount of data.

The front page Digital Intellectual Entity only consists of an identifier, which is the valid reference for all versions of the front page through time.

The Representation of the front page "front page repr. 1" contains the relevant metadata of the file. The result of the actual digitization is placed in the file "front page file a".

Step b) only consist of a metadata change, which does not affect the file itself, and thus the new Representation "front page repr. 2" is added with reference to the existing file as well as reference to the existing Digital Intellectual Entity it represents.

In step c) the "front page file a" is exchanged with "front page file b" containing a new digitization. This new file has new technical metadata, and therefore needs its own Representation "front page repr. 3", which refers to the new file and the existing Digital Intellectual Entity that it represents.

Step d) adds a back page to the postcard. A new Digital Intellectual Entity and Representation is created for the new back page in the same way as the first digitization of the front page. Since this is a change for the actual postcard, the change will also result in a new Representation for the postcard itself "postcard repr. 2", which includes the structural metadata of the postcard pointing to both the Digital Intellectual Entity of existing "front page" and the new "back page". Furthermore, it points to the Digital Intellectual Entity "postcard" which it represents.

There are many identifiers involved in this data model. The most important ones are the identifiers for the Digital Intellectual Entities, since these identifiers will be the reference point for digital material over time, in order to reference different versions (Representations) of the digital material.

B. *Files Produced for Bit Preservation*

The metadata for the digitized files are represented in XML files with metadata that obeys the XML metadata schemes. Relations between the data model entities are specified by defining identifiers and specifying relations in the XML files. For example, Representation "front page repr. 1" refers to the "front page file a" through the structural metadata of the METS metadata (in METS files and METS structmap).

The reference to the Digital Intellectual Entity from the Representation is specified as part of the PREMIS metadata as a structural relationship (relationshipType: structural, relationshipSubType: represents and UUID for the Digital Intellectual Entity in metadata for Representation).

In practice, producing the XML files results in the creation of many small files. Even though technology has come far in relation to handling many small files, there are still issues. Therefore, the metadata are packed in chunks before being bit preserved. Royal Danish Library has chosen WARC for this purpose [6].

For optimization purposes, an extra WARC record is produced containing information about the relationship between Digital Intellectual Entities, each of their Representations and files (if a file exists for the Representation). This information is preserved along with the timestamp of archiving the Representation of the Digital Intellectual Entity. This extra information is redundant information, since it can be re-produced by reading all bit preserved metadata and finding the respective identifiers. However, doing this would be a very time consuming process. An additional benefit is that WARC allows "browsing" of the metadata, where the different versions of the Digital Intellectual Entity can be distinguished by the recorded timestamp, in the same way as browsing revisited web archived materials.

WARC packages with WARC records for each step of the postcard example are publicly available at id.kb.dk[1]. The actual bit preservation of the files is obtained by use of the bitrepository.org software [7].

[1] The examples can be found at htttp://id.kb.dk/examples/

VII. Metadata Preservation based on Exports

Royal Danish Library wants to ensure bit preservation of metadata through routinely export of metadata from the preservation supporting systems (i.e. Preservica and Cumulus systems), including audit trails for the individual digital objects. The exported metadata are structured as described in Section IV about metadata (illustrated in Fig. 2).

Based on the postcard example the following section describes how extracts of metadata from the existing systems are mapped into the technology independent general data model.

A. *Postcard Exported from Cumulus*

This section describes how Royal Danish Library extracts metadata from the materials placed in a key/value based Cumulus system.

In Cumulus, the Library has defined keys for values used for transformation of the data into the general data model (except from identifiers and timestamps to be evaluated in the preservation process). Since Cumulus does not have any facilities for bit preservation, the Library has developed the program "Data & Metadata Bit Preservation Service" (depicted in Fig. 1). This application extracts files and metadata from Cumulus and transform the metadata into the structure described for metadata and the general data model, before sending it to bit preservation.

B. *Postcard Exported from Preservica*

In this section, it is described how Royal Danish Library will extract metadata from the materials placed in Preservica.

The data model used in Preservica (hereafter referred to as the Preservica data model) has some resemblance to the data model described in this paper (called the general data model). Both data models are inspired by the data model from the EU Planets project, which ended in 2010. However, there are differences in terminology and meaning, especially regarding dynamic or static status of the Digital Intellectual Entity. Furthermore, the Preservica data model only takes into account the changes in files – not the changes in metadata. Resemblances and differences in concepts between the Preservica data model and the general data model are:

Files in Preservica contain both a file *and* metadata for the file. Furthermore, these file metadata can be changed. To map Preservica's data model to the general data model, one would have to separate the file from its metadata and have the metadata in a Representation for the file (and with changes to file metadata in new Representations).

Manifestations in Preservica are similar to Representations in the general data model, (at least the preservation Manifestations)[1]. There are, however, two main differences. Firstly, metadata in Manifestations are changeable. Secondly, Manifestations in Preservica assume that there can be only one active preservation Manifestation at any given time. In the case of e.g. preserving a heavily animated PowerPoint presentation, there is a need for several different "preservation Manifestations", e.g. a migration to a PDF version to preserve the look and feel of e.g. colors, and a newer version of PowerPoint to maintain the idea of the animation in the previous version. The general data model does not assume anything about whether Representations are active or not, and can therefore cover such cases.

Deliverable Units in Preservica are similar to Digital Intellectual Entities in the general data model. Again, there is a difference in the fact that a Deliverable Unit has changeable metadata. A Digital Intellectual Entity cannot change at all, since this would require a new identifier, consequently a Digital Intellectual Entity cannot contain metadata. In order to have metadata at this level, it must have a separate layer of Representations.

Preservica also has ***Collections***, which are collections of Deliverable Units, and a Deliverable Unit can belong to one Collection only. If ignoring the latter extra restriction, there is no difference between a Collection of Deliverable Units and a Deliverable Unit consisting of Deliverable Units. Thus, Collections do not need additional comparison.

The above-mentioned postcard example will in the Preservica data model look as depicted in Fig. 6.

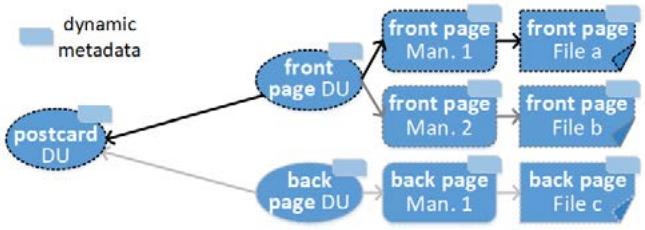


Figure 6 Postcard example implemented in Preservica.

The Preservica data model looks much simpler than the general data model, but this is caused by the fact that the Preservica data model does not support preservation of metadata and changes in these metadata for all Preservica data model entities (Collections, Deliverable Units, Manifestations and Files).

The difference from the general model is that changes are not represented. For instance in Preservica, the adding of the back page cannot be seen in the Manifestation of the postcard, since it is only represented as the back page file pointing to the postcard Representation. In the general model, this relation is explicitly recorded as part of the postcard Representation "postcard repr. 2". Furthermore, the change of metadata for the front page is not visible in the Preservica data model, since the Preservica "front page Man. 1" covers both "front page repr. 1" and "front page repr. 2" representing the metadata before and after the metadata update.

How to map the Preservica data model to the general data model is illustrated in detail in Fig. 7 for the front page (with a file).

In practice, there are two ways of mapping Deliverable Units with files to the general data model. Which method to choose depends on the type of metadata they cover, - or rather, how curators originally decided to place metadata on the Deliverable Unit and Manifestations in Preservica.

[1] Preservica also has presentation Manifestations, which are not concerned with the actual preservation and therefore not part of the general data model.

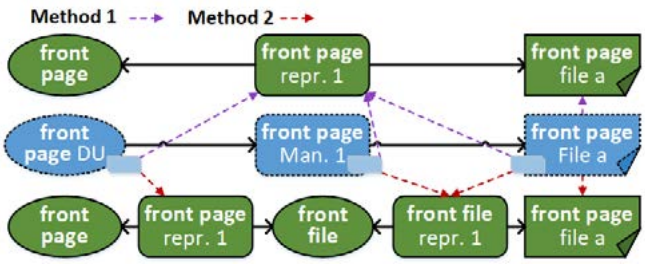


Figure 7 Two ways to map Preservica postcard (in blue) to general data model postcard (in green).

Technical metadata will always be on the File / Manifestation: the File and Manifestation are joint since the Preservica interface does not distinguish between whether metadata comes from a File or its Manifestation. This is also why all the Preservica metadata for the File and Manifestation (at a specific time) are mapped into the Representation of the file for both Method 1 and Method 2.

In most cases, the types of metadata placed on the File / Manifestation and the Deliverable Unit differ (are disjunct). Normally, descriptive metadata are placed on the Deliverable Unit in order not to repeat them for each Manifestation. Therefore, Method 1 is the most likely to be used.

Method 2 is used in cases where there is a need to distinguish between metadata for the File / Manifestation itself and metadata for the logical unit that the file represents (the Deliverable Unit), e.g. descriptive metadata. In such cases, an extra layer (the front file Representation and front file Digital Intellectual Entity) is inserted to distinguish between the two types of descriptive metadata.

The mapping of a Deliverable Unit *without* files (only pointing to other Deliverable Units) is simpler, since all metadata from the Deliverable Unit are mapped into a Representation of the corresponding Digital Intellectual Entity. However, in Preservica, the information about relations to other Deliverable Units comes from the Deliverable Units at a lower level. This means that the postcard Deliverable Unit has e.g. descriptive metadata only, while the structural information about the front and back page Deliverable Units' relation to their parent postcard Deliverable Unit can only be found in these underlying front and back page Deliverable Units. Furthermore, the history information about the adding of the back page has to be found in the Preservica log. Since a Digital Intellectual Entity cannot have metadata, the metadata are placed on the Representations. Again, if there is a need to distinguish between changes e.g. in descriptive metadata and structural information, then an extra layer can be added.

Based on the described mappings, Royal Danish Library will develop a service to transform metadata from Preservica to the standardized metadata format which will then be bit preserved. We are aware that there are challenges to this transformation. One known challenge is calculation of the bit preservation level, which for instance needs information about the Preservica Storage adapter used for the data. Another known challenge is calculation of provenance metadata for tools used in Preservica. However, the challenges so far seem to be solvable, if the right APIs to extract metadata from Preservica are provided.

VIII. Discussion

The way that the Royal Danish Library's digital preservation policy and strategy will ensure metadata preservation is by no means the only way to do it. It will always be a matter of considering which risks the organization is willing to take; number and nature of preservation supporting systems in the organization; and to which degree different materials need to be preserved.

First of all, the decision about transforming metadata before their bit preservation has an inherited risk of losing information during this transformation. On the other hand, a delayed transformation of system dependent metadata (when the system is eventually replaced with another) will also involve risks. The Library regards it as a greater risk to wait with the transformation.

Secondly, there are many ways to choose and structure different metadata, both in interrelations between the different formats and in the way interrelations between data are expressed in a data model. The choices described in this paper will therefore be debatable for other cases.

Concerning the mapping from Preservica's data model to the general data model, there are other

cases than the included ones, e.g. cases where a Deliverable Unit consists of more files without an explicit Manifestation.

Finally, the concept of intellectual entities has been a confusing area for a long time, and it could be worth discussing whether the definition of a Digital Intellectual Entity can assist in the understanding of intellectual entities when we use the term in connection with digital assets.

IX. Conclusion

This paper has provided a case study of one way to ensure proper preservation of metadata, as reflected in the new Royal Danish Library's digital preservation policy and strategy replacing three former and very different sets of policies and strategies.

The case study has included aspects of how to deal with preservation of dynamic metadata along with exit strategies for different current and future systems and access to all metadata for all materials independent of the system from which they originate.

The paper has also provided details of the decisions and reasoning made to ensure that implementation of metadata preservation can fulfil the different requirements to exit strategies, reestablishment after major breakdowns, and support of metadata warehousing.

We hope that this paper can form a basis for feedback and further discussion of metadata preservation strategies.

References

[1] Royal Danish Library Digital Preservation Policy, archived in: netarkivet.dk, archival date: 2019-03-14T14:18:57Z, URI: http://id.kb.dk/documents/DB_Politik_20190109_UK.pdf [part], 2019.

[2] Det Kgl. Biblioteks digital bevaringsstrategi (only in Danish) in: netarkivet.dk, archival date: 2019-03-14T14:18:58Z, URI: http://id.kb.dk/documents/DB_Strategi_20190109.pdf [part], 2019.

[3] Digital Preservation Metadata for Practitioners - Implementing PREMIS, eds. Editors: Dappert, A., Guenther, R. S., Peyrard, S. ISBN 978-3-319-43761-3, 2016.

[4] ISO 14721:2012. "Space data and information transfer systems - Open archival information system (OAIS) – Reference model", 2012.

[5] J. Pearce, D. Pearson, M. Williams, S. Yeadon, "The Australian METS Profile – A Journey about Metadata", D-Lib Magazine, vol. 14, no. 3/4, March/April 2008.

[6] ISO 28500:2017, "Information and documentation -- WARC file format", 2017

[7] B. A. Jurik, A. B. Nielsen, E. Zierau, "Flexible Bit Preservation on a National Basis", *Proceedings of the IS&T Archiving 2012, pp. 2-7, 2012.*

Extensive Extensions

Exploring File Extensions in Library of Congress Collections

Mark Cooper
Library of Congress
United States of America
mcop@loc.gov

Aly DesRochers
Library of Congress
United States of America
alde@loc.gov

Jesse Johnston
Library of Congress
United States of America
jesjohn@loc.gov

Camille Salas
Library of Congress
United States of America
csala@loc.gov

Trevor Owens
Library of Congress
United States of America
trow@loc.gov

Abstract – Through four decades of digital initiatives and collecting programs the U.S. Library of Congress has built up a sizable digital collection. In support of long-term management of this digital content, in 2018 staff worked to review information about file extensions of content in the permanent digital collection through analysis of data in the institution's primary digital content inventory system. This paper reports the results of this analysis and how these findings will inform the development of digital content management policy and practice at the institution.

Keywords – file extensions, digital inventory management, digital collections

Conference Topics – Designing and Delivering Sustainable Digital Preservation

I. Introduction

In 2018, Library of Congress staff began working to identify the major file formats that comprise the institution's permanent digital collection. This paper reports on the initial results of that work, which included an analysis of file extension metadata recorded in the Library's primary digital collections inventory system.

This content is spread across multiple storage architectures and inventorying platforms, and the computing time and ongoing work to process new content makes this unattainable with present tools and systems. Instead of direct analysis, we began work by querying file extension information (e.g., .pdf, .mp3).

This paper describes the methods used to conduct this preliminary analysis and the results. We begin by briefly contextualizing this work within the Library's considerable history of digital initiatives, its digital collecting programs, and the development of standards and practices. We then present the results of our analysis and offer observations and discussion of how those results are informing future planning for digital content management practices and processes

A. Four Decades of Digital Collecting

The Library of Congress has four decades of experience acquiring, managing and preserving digital collection content. In 1982 the Library's Optical Disk Pilot Project began capturing text and images of collections [1]. In 1989 those efforts grew into the American Memory pilot program, which digitized selected collection content for distribution and ultimately became the basis of the institution's first online collections [2].

These efforts have since grown to include a wide range of born-digital collecting programs. Of particular note, the Web Archiving Program launched in 2000, [3] the copyright eDeposit for eSerials launched in 2009, [4] and the Cataloging-in-Publication program began to acquire eBooks in 2012 [5].

These programs, among many others, have resulted in continued growth of the Library's digital holdings. Building on the development of these varied

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

programs and initiatives, the Library of Congress established and published Collecting Digital Content at the Library of Congress, a five-year strategy and plan, which establishes targets in six strategic objectives for growing the digital collections [6]. In 2018, as a companion to its strategic plan, the Library of Congress established its first Digital Strategy which asserts the intention to continue to "exponentially grow" its digital collections and "ensure enduring access" to them [7].

As a result of the early and extensive efforts of the Library of Congress to build robust digitization and digital collecting programs, the institution has developed a digital collection that is, much like the physical collection is varied and expansive. The institution has simultaneously invested in a series of initiatives focused on ensuring enduring access to these parts of the library's collections.

B. Planning for Enduring Access to Content

In parallel to the growth and expansion of the Library of Congress digital collections, the institution has developed a series of resources to support planning for enduring access to these materials. Created and coordinated by a range of stakeholders in the organization, these include the development and maintenance of the Sustainability of Digital Formats [8], the PREMIS Data Dictionary for Preservation Metadata [9], and the Library of Congress Recommended Formats Statement [10]. The institution also convenes stakeholders annually for the Designing Storage Architectures for Digital Collections summit. Along with these areas of work, the Library was a founding convener of the National Digital Stewardship Alliance Levels of Digital Preservation [11].

All of these resources and initiatives have become critical components of international community approaches to digital preservation. Through the development of the Digital Collections Management Compendium (DCMC), the Library is currently working to synthesize and connect these interrelated resources.

C. Connecting Digital Content Theory & Practice

The Library of Congress is now working to develop an integrated DCMC that summarizes current policies and internal guidance. The core concept for DCMC is to integrate and synthesize guidance for policy and practice, clarify roles and responsibilities, and clarify IT business needs. In this way, the effort provides a policy infrastructure that supports improvement of systems that acquire, preserve, and deliver digital content to users now and in the future.

The DCMC is being developed to connect high-level policy to the every day work of digital content management practice. As a result, the Library's Digital Content Management section (DCM) is working to simultaneously integrate these various standards and planning resources into ongoing work to review the extent of existing digital collections and plan and implement processes and workflows to support future growth.

II. Identifying File Extensions in the Collections

In 2018, the DCM section was tasked with analyzing and inventorying file formats in the digital collections. We began working to establish a list of file formats in use in the permanent digital collection of the institution. Ideally, analysis of file formats would involve direct analysis and characterization of the files for signatures. However, at this point it was not feasible to do so across all digital collection content with our diverse digital library infrastructure. Instead, analysis was conducted against file extensions (e.g., .pdf, .mp3), which provide valuable information for understanding the likely formats of files. We discuss the limitations of this approach below.

Using our existing tools and data, DCM staff compiled and analyzed the file extensions of content inventoried in the Content Transfer System (CTS), an internally developed system that serves as the Library's primary inventory system for managing digital collection content. A central value of CTS as an inventory system is that it enables Library staff to maintain inventory control of digital content in the Library's permanent collection. Specifically, CTS provides logs and inventory data for all content managed through the system across a range of distinct storage systems.

CTS is one of two approved inventory systems for managing digital collections content at the Library; the second is the Packard Campus Workflow

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Application (PCWA), which inventories collection content for the Motion Picture, Broadcasting and Recorded Sound (MBRS) Division. Digital collections inventoried in PCWA are managed by MBRS and are not addressed in this analysis. Currently, the collections managed throughout the primary inventory system comprise well over 7 PB of digital content. Notably, this does not include the 9 PB of content stored in the National Audiovisual Conservation Center's separate inventory system.

The results of DCM's analysis illustrate the extensive diversity of digital content in the Library's permanent digital collections. For the "long tail" of collection materials representing the diversity of file extensions, the Library plans to maintain bit-level preservation, as reflected in the DCMC. CTS contains data that was used to report on file extensions of digital content in the collections, but this analysis also illustrates a series of potential next steps for improving inventory data for collection materials en masse. In particular, a significant portion of files in the Library's digital collections is managed within compressed container files (ex. zip and tar), which mask the file extensions contained within them.

A. Scope of File Extension Data

DCM staff used to the Kibana platform to create a custom dashboard to report extension data for Library of Congress Inventoried Collection Content. The Dashboard is not available to external users, however, the development of it may be of general interest to external users.

The Dashboard uses the file extension index generated from CTS inventories to present data on file count and file size by extension. Dashboard users have the ability to filter file extension data by server system, custodial divisions, or content custodians and can download various sets of data based on their categorization.

For both the Dashboard and the present report, we focused on a subset of the content inventoried in CTS that best represents a view of content in the permanent digital collection. This includes digital collection content under inventory control in long-term (tape) storage and access presentation (spinning disk) storage. Digital content in areas used for processing and ingest is excluded from the analysis.

This includes all the files managed in these systems, including information about metadata files and other supporting files. All of this content and its respective file formats are under Library of Congress inventory control and are therefore content the institution is managing as part of the files in its permanent digital collection.

III. Analyzing File Extension Data

As of September 24, 2018, digital content under inventory control in the Library of Congress' primary inventory system include 681 million files, almost 8 petabytes of data, represented in over 16,000 unique file extensions. It should be noted that these numbers include distinct copies of files in presentation and long-term storage. Due to aspects in the Library's digital infrastructure and the difficulty presented in de-duping inventories at the file level, an identical file stored in multiple locations or servers is counted multiple times for each discreet copy on disk. (As noted above, this analysis does not include data from the National Audiovisual Conservation Center, which is managed through a different inventory system.)

The tables below provide detailed information about the top ten file extensions by file count and by file size.

TABLE I

Top 10 File Extensions by Count

Rank	Extension	File Count	File Size
1	jp2	148,242,701	294.83 TB
2	tif	133,388,164	2,516.69 TB
3	jpg	113,952,865	36.40 TB
4	xml	70,117,369	41.98 TB
5	pdf	56,048,470	81.59 TB
6	txt	48,637,417	3.25 TB
7	gif	44,082,987	0.74 TB
8	gz	10,120,046	3,937.79 TB
9	i41	7,425,410	1.96 TB
10		5,091,599	3.16 TB

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

TABLE 2
Top 10 File Extensions by File Size

Rank	Extension	File Count	File Size
1	gz	10,120,046	3,937.79 TB
2	tif	133,388,164	2,516.69 TB
3	mxf	22,920	543.19 TB
4	jp2	148,242,701	294.83 TB
5	mpg	161,656	175.01 TB
6	wav	596,130	127.18 TB
7	mov	29,024	99.03 TB
8	pdf	56,048,470	81.59 TB
9	iso	36,325	63.74 TB
10	dv	5,006	54.48 TB

We were also able to gain insights about the most common file extensions by file count and by file size. Further analysis of these results is presented in the next section, including discussion of the prevalence of image files by file count, the nature of compressed container files, and the unique case of files with no extension appearing to be the tenth most popular file across the Library's digital collections.

Analysis of file extension data suggests the following findings:

1) The majority of Library of Congress digital content—both by file count and by size—has file extensions associated with a relatively small number of file formats.
2) The remainder of file extensions reflect the Library of Congress's wide collecting scope as well as the scale and variation of the digital world.
3) The available set of tools for this analysis produced valuable information on Library of Congress digital collections, but an understanding of the current technical limitations suggests paths forward to improved analysis and monitoring.

A. Most Content has Well Known Extensions

By file size, about 80% of Library of Congress digital content is represented by two file extensions: .gz and .tif. Given the nature of current Library of Congress digital collections, it is not surprising that these two file extensions represent a significant majority by file size. The extension .gz is a compressed container format used, in this case, primarily for holding Web Archives content. This and other compressed container formats are discussed in more detail below. The extension .tif is an image format used extensively across many types of digitized content, including textual, photographic, and map content, as specified in the Recommended Formats Statement [10].

By file count, about 95% of Library of Congress digital content is represented by the following eight file extensions: .jp2, .tif, .jpg, .gif .xml, .txt, .pdf, and .gz. The image formats represent both master files and their corresponding derivative images, largely from digitized collection materials, reflecting the Library's digitization practices. As these files may correspond to a single page of a book or a single-image item such as a photograph, higher numbers are expected compared to, for example, born-digital ebooks, for which a single EPUB or PDF file usually represents a single book, a smaller but growing category of content at the Library. The textual formats may be collection content such as born-digital materials encoded in XML or plain text, but also include derivative content such as transcriptions, or metadata and supporting materials. Files with the extension .pdf may also represent born-digital master files or derivatives of digitized content.

Given the volume of Web Archives, the extension .gz also appears on this list of top file counts. The table below provides a break down of the extent to which these container files are associated with the web archives.

TABLE 3
Container Files by Extension and File Count

Extension	Total	Non Web Arch	Web Arch
gz	10,120,046	1,132,045	8,988,001
zip	410,092	409,878	214
rar	3,314	3,314	0
tar	3,026	3,015	11

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

TABLE 4
Container Files by Extension and File Size

Extension	Total	Non Web Arch	Web Arch
gz	3497.9 TB	560.7 TB	2987.7 TB
zip	15.7 TB	15.7 TB	0 TB
rar	0.2 TB	0.2 TB	0 TB
tar	1.9TB	1.9 TB	0 TB

The Web Archives are stored as WARCs, and then compressed into the .gz format. Display of the web archives content requires its own type of index. Those indexes allow for further analysis into the content of these .gz files. At the time of analysis in 2018, these web archive files contained 6.2 billion unique files [12].

With improvements to the Library's digital inventory systems, DCM will be able to provide more information about Library of Congress digital collections and bring more collection content under inventory control. This initial analysis suggests a large majority of the current state of Library of Congress digital collection content is contained in a small number of file formats, which are widely supported and well understood for purposes of preservation and access. Large portions of the current collections are digitized surrogates of physical items, represented by the prevalence of image formats produced through digitization. As born-digital collecting expands in scope and content areas, we expect to have see a greater variety of file formats in CTS.

B. Long Tail of File Extensions

Beyond the initial list of widely used file extensions exist a considerable long tail of extensions. Most of the 16,689 unique file extensions appear infrequently but provide a window into the diversity of the digital collections.

Across the Library of Congress digital collections, 2,761 file extensions are included only once; most of these likely do not represent a unique file format, but instead represent a single file of an unknown format. In many cases, the "extension" part of the file name may have been used as a note to a user or a system, or a filename may incidentally contain a period, which is parsed in the data as representing an extension.

Furthermore, 14,064 file extensions, or 83% of all extensions, occur less than 100 times. It is very likely that some of these extensions represent collection material of a digital format for which further analysis is needed to determine more information. However, it is also likely that much of this content does not represent individual file formats, but instead are sub-components of data set formats or are simply multiple copies of a file with the issue noted above: an unknown format where the "extension" field was possibly used as a note to a user or a system.

Finally, 3,810 files representing 272 file extensions are 0 byte files, meaning there is no content to the file except for a filename. These files may be purposefully empty to serve as a flag or identifier to a user or a system, or may expose corrupted files where the content was lost due to a system error.

The digital collections include content ranging from formats familiar to the average computer user and for which the preservation risks are well understood, to highly specialized and idiosyncratic formats In some cases, as particularly seen in files created before the mid-1990s, file extensions were used as general purpose indicators or for sorting and may not have any relationship to file format at all.

In order to work with this long tail of file extensions, the DCMC provides guidance on the preservation and maintenance of original file formats. For some born-digital content, such as digital archival manuscript collections, the original content and file formats delivered to the Library may have artifactual qualities that need to be preserved for researchers. Such material will need to be maintained with its original context in addition to any potential migration or emulation for access.

C. Limitations Require System Improvements

Further understanding of the limitations of this analysis provides valuable information for improvements to digital inventory systems.

i. Limitations of Extension Information

The extensions data enabled some analysis and conclusions to be drawn about the Library's digital collections; however, file extensions remain the weakest form of file characterization. In the most striking example, files with no discernable,

conventional extension—no characters following a period within the filename—are the 10th most popular "extension," representing 5,091,599 files and 3.161 TB of data. It is likely that most of this content is related to system functions, scripted operations, or datasets, but more advanced format analysis is required to determine if any of this content represents known file formats that should be managed as collection materials.

Another limitation is highlighted by the large subsets of extensions that can represent groupings of the same file format. In some cases these groupings are clearly defined; otherwise, they require further analysis to determine similarities. In this area, two major groupings emerged: datasets and system files.

Datasets: At least 20 different file extensions such as .i41, .i21, .i22, etc, are different extensions but effectively the same file format related to geospatial datasets. When these are separated out individually by extension they appear to be less prevalent, but combined they represent over 10 million files, or 1.5% of the total Library of Congress collections by file count. There are likely many other file extensions that are broken out as separate elements in this report, but in fact represent a single file format within a dataset.

System files: As with datasets, file formats used by systems in logging or tracking functions can result in misleading counts when analyzed by file extension. For example, .jdb files are stored as 2,388 different extensions, such as ".jdb,9999998", representing 154,411 total files; all of this content represents a single file format related to logging information for a certain type of Web Archives crawl action.

Library of Congress collections contain file extensions that include notes or references created by a user, in particular with digital content created decades ago which is often donated to custodial divisions working with archival collections. These are valuable components of the provenance of the complete collection, but mask the functional file format of the content from current analysis.

ii. CTS Inventory Data Limitations

The CTS inventory data utilized for this analysis must be qualified for its significant limitations. For example, some portion the Library of Congress's digital collections are not inventoried by the CTS system and therefore not included in this analysis. As previously mentioned, the National Audio Visual Conservation Campus content is managed in a separate system. Beyond this, some portion of other digital collection content is in process to be brought under inventory control in CTS. While initial estimates indicate that a large majority of the long-term storage system is inventoried in CTS, DCM estimates that potentially as much as one third of the collection content (by file size) held in presentation server storage is not inventoried in CTS and therefore not represented in this analysis. DCM is thus establishing procedures for inventorying all content and maintaining the health of the inventory system.

In addition to the limitations created by uninventoried content, a potentially significant number of files are counted multiple times due to duplications and overlaps in inventory records. Duplication is expected for files stored in multiple server systems (long-term and presentation), but overlaps in inventory records result in the same discreet file on a server being counted twice. In order to improve the underlying data of CTS inventory records and improve further efforts to characterize the extent of digital content managed through this system, DCM has initiated an ongoing process to establish, improve, and refine use of the functions of the inventory system.

Given current system limitations, DCM is treating all content inventoried in CTS as "collection content." The CTS inventory system cannot currently filter the data into categories such as collection material or supporting files, metadata, or derivatives. CTS treats all files equally, since all are digital content under inventory control and managed in server systems that require monitoring. However, it may be helpful in the future to differentiate between born-digital and digitized collection content; masters from derivative, metadata, or supporting content; as this may determine preservation actions.

IV. Next Steps: Improve Inventory & Infrastructure

The results of this analysis offer substantive opportunities to build on work building and managing digital collections at the Library of Congress. The creation of the internal Kibana Dashboard of file extensions provides an opportunity to both reflect on further enhancements to digital content management tools and infrastructure and a means to track work to improve the inventory control and management of digital collections.

A. Work to Improve Inventories

One of the primary results of this analysis has been to identify and prioritize work necessary to improve the underlying data used to track and monitor content in the digital collections.

In continuing to track this data, we now have a baseline view into the state of the collection, and can plan projects to improve management and inventory control. DCM's future efforts to bring uninventoried collection content under inventory control across all divisions will result in more trustworthy and complete reports and data on the entirety of Library of Congress's digital collections. Furthermore, DCM is in the process of establishing regular data integrity checks of all content managed in CTS, which will ensure the inventory data remains in line with the current state of collections. Additional remediation of metadata fields within CTS inventories will also improve the reportability of the data, such as assisting with filtering by custodial divisions and distinguishing master files from derivatives or supporting materials.

B. Potential Infrastructure Improvements

In addition to opportunities to improve the management of digital content, this analysis suggests a series of potential infrastructure improvements that could result in improved analysis of file formats in the digital collections.

1) Expansion of data available to reporting tools, including data on content in the other inventory system (PCWA), which would enable a wider variety of analyses.
2) Implementation of existing format characterization tools and validators (TIKA, JHOVE2) on a collection-level scale. Within existing systems, these tools can currently be run on individual resources, but to support planning and analysis, it is necessary to and be able to be run these kinds of tools at scale against collection content.
3) Capability to generate and analyze data on contents of container files.
4) Ability to generate reporting across all instances of files, allowing for identification of duplicate files across systems for analysis and counting of unique files. Current deduplication is only possible at the larger inventory record level.

C. Anchoring Policy & Planning in Data

The results of this work serve as direct inputs for development of policy and planning resources for the institution. As we work to create and share an integrated DCMC of practices and guidance, we are working to ensure that this resource bridges the realities of digital collection content in hand and visions for how digital content should work in theory.

Based on this analysis, the DCMC clarifies the intention to maintain content in its original formats and offer levels of planned future support for digital content in alignment with information from the Sustainability of Digital Formats site and the Library of Congress Recommended Formats Statement. As this data and analysis provides a view into all inventoried content in the digital collections, it can be used to inform baseline practices for content management, help shape priorities for work on the Sustainability of Digital Formats and help to inform future revisions to the Library's Recommended Formats Statement.

We opened this paper contextualizing our work in an understanding of the four decades of digital content management practice at the Library of Congress. A significant result of those four decades of work was making the case and then establishing the Digital Content Management unit. As we look to the next four decades and beyond, it will be increasingly essential to engage in the kind of analysis we have engaged in here and put that analysis in dialog with both ongoing practices and the development of policy and planning resources.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

References

[1] M. Flynn and H. Zinkham, "The MARC Format and Electronic Reference Images: Experiences from the Library of Congress Prints and Photographs Division," *Visual Resources*, vol. 11, no. 1, pp. 47–70, Jan. 1995.

[2] C. R. Arms, "Historical collections for the National Digital Library: Lessons and challenges at the Library of Congress," *D-Lib Magazine April: Part*, vol. 1, 1996.

[3] A. Grotke, "Web Archiving at the Library of Congress," *Computers and Libraries*, vol. 31, no. 10, Dec. 2011.

[4] E. Delfino and J. Mandelbaum, "eDeposit for eSerials: Current Work and Plans at the Library of Congress," in *Journal Article Tag Suite Conference (JATS-Con) Proceedings*, 2011.

[5] T. Westervelt, "To Lead to Learning, Not to Madness: E-Books and E-Serials at the Library of Congress," *The Serials Librarian*, vol. 72, no. 1–4, pp. 122–127, May 2017.

[6] J. Puccio, "Collecting Digital Content at the Library of Congress," *The Signal*, 21-Mar-2017 https://blogs.loc.gov/thesignal/2017/03/collecting-digital-content-at-the-library-of-congress/.

[7] K. Zwaard, "Digital Strategy at the Library of Congress," presented at the Federal Depository Library Conference, Washington DC, Oct-2018.

[8] C. Arms and C. Fleischhauer, "Digital formats: Factors for sustainability, functionality, and quality," in *Archiving Conference*, 2005 pp. 222–227.

[9] PREMIS Data Dictionary, http://www.loc.gov/standards/premis

[10] T. Westervelt, "Library of Congress's Recommended Format Specifications," *Serials Review*, vol. 40, no. 3, pp. 217–218, Jul. 2014.M.

[11] Phillips, J. Bailey, A. Goethals, and T. Owens, "The NDSA Levels of Digital Preservation: An Explanation and Uses," *IS&T Archiving, Washington, USA*, 2013.

[12] G. Thomas and C. Dooley, "The Library of Congress Web Archives: Dipping a Toe in a Lake of Data," The Signal, 2019. http://blogs.loc.gov/thesignal/2019/01/the-library-of-congress-web-archives-dipping-a-toe-in-a-lake-of-data/

Aligning the eARK4All Archival Information Package and Oxford Common File Layout Specifications

Complementary rather than competing approaches

Neil Jefferies
Bodleian Libraries
University of Oxford, UK
neil.jefferies@bodleian.ox.ac.uk
orcid.org/0000-0003-3311-3741

Karin Bredenberg
Swedish National Archives
Sweden
karin.bredenberg@riksarkivet.se
orcid.org/0000-0003-1627-2361

Angela Dappert
Springer Nature
UK
angela.dappert@springernature.com
orcid.org/0000-0003-2614-6676

Abstract – The E-ARK4ALL[1] project released an alpha version of its Common Specification for Information Packages (CSIP)[2] to be used in the eArchiving Building Block[3] for review at the end of November 2018. Slightly earlier, the Oxford Common File Layout (OCFL)[4] initiative had also released an alpha version of its software independent preservation file organisation specification. While, at first sight, these would appear to attempt to do similar things, they are in fact, largely complementary approaches. While the eARK specification aims to define the logical structure and content of Open Archival Information Systems (OAIS)[5] Information Packages, the OCFL describes how to map any logical digital object layout onto a physical file system in a preservation-friendly manner, as well as identifying the fundamental operations required to manage such objects. This paper provides a brief introduction to the two specifications and then describes how the OCFL can be applied to an E-ARK IP.

Keywords – Preservation, OAIS, AIP, file system, specification

Conference Topics: The Cutting Edge: Technical Infrastructure and Implementation;

Collaboration: a Necessity, an Opportunity or a Luxury?

I. Introduction

The E-ARK4ALL project released an alpha version of its Common Specification for Information Packages (CSIP) to be used in the eArchiving Building Block for review at the end of November 2018. In September, the Oxford Common File Layout (OCFL) initiative had also released an alpha version of its software independent preservation file organisation specification. While, at first sight, these would appear to attempt to do similar things, they are in fact, largely complementary approaches.

While neither specification is completely finalised at the time of writing, they are largely complete so it is an opportune time to examine how, in practice, they might be aligned. The examination presented here is quite high level since it is based on members of each of the respective communities reading of the other's specification while completing work on their own documents. However, it is possible to usefully identify some basic workable principles and potential areas for further discussion. As always, the fine detail will only emerge when code comes to be written and systems to be built.

II. The Oxford Common File Layout

he Oxford Common File Layout (OCFL) initiative began as a discussion among digital repository practitioners about the ideal layout and characteristics for persisted objects, from a computational and conceptual point of view. It is named, as with

[1] E-ARK4ALL Project, http://e-ark4all.eu/

[2] E-ARK Common Specification for Information Packages, http://earkcsip.dilcis.eu/

[3] eArchving Building Block, https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving

[4] Oxford Common File Layout, https://ocfl.io/

[5] Open Archival Information System, http://www.oais.info/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

a number of other projects[1], for the location of this initial discussion. It has since grown into an open community effort defining an application independent way of storing versioned digital objects with a focus on long term digital preservation.

The approach is informed by three simple observations:

1) Archived objects change relatively slowly compared to archival software, and are rarely deleted.
2) Migration by export and re-ingest is generally slow and error-prone. Data is most at risk of loss or corruption when it is moved or migrated, rather than at-rest.
3) File systems, in particular POSIX-style[2] file systems, have been the most consistently implemented and widely tested Application Programming Interfaces (API's) for accessing storage in any form.

- *Objectives*

The OCFL also builds on practical experience gained from previous work on related initiatives, such as Stanford's MOAB[3] and BagIt[4], both in order to avoid some of their pitfalls and bottlenecks, but also with a view towards interoperability and easy migration.

Consequently, the OCFL is constructed with five main objectives, most of which readily map to the more hardware focussed elements of the emerging Digital Preservation Storage Criteria[5].

- *Completeness*

All the data and metadata required to understand and render or execute an object should be stored within the directory that represents the object on the filesystem. This ensures that a repository can be rebuilt from scratch given just the files on storage. It also aligns very well with the construction of an E-ARK AIP.

- *Parsability*

The structure of content stored using the OCFL should be easy to access and interpret by humans and machines. This ensures that the content can be understood in the absence of the original systems and software. To this end, as with some parts of the E-ARK CSIP, the OCFL allows for embedded documentation. This is crucial since the OCFL does not mandate the internal structure of the objects that it stores.

- *Versioning*

The OCFL is designed with the expectation that digital objects will change over time, even if only as a result of preservation activity. It therefore supports object versioning, provides a mechanism for recording version history and allows access to previous versions.

- *Robustness*

Robustness against errors, corruption, and migration between storage technologies is a basic requirement of any preservation storage system. The OCFL uses SHA256 or SHA512 for content addressing and, consequently, for default fixity provision, which operates at both a file and object version level.

- *Portability*

The ability to store content on different storage infrastructures and migrate between them is essential for maintaining diversity as a hedge against both obsolescence and systemic technological failure. The OCFL requires a minimal set of file system features to operate, and proscribes the use of additional features if they have the potential to affect portability.

To these five criteria we can also add *efficiency* as an additional consideration, which is manifest in several ways. The OCFL is designed to support forward-delta differencing between object versions so that components that do not change between versions are only stored once, reducing the storage overhead, and hence cost, for versioning. It is also constructed to minimise the number of file system

[1] e.g. Dublin Core Metadata Initiative http://dublincore.org/; Portland Common Data Model https://pcdm.org/2016/04/18/models

[2] Posix - The Open Group Library, http://pubs.opengroup.org/onlinepubs/9699919799/

[3] The Moab Design for Digital Object Versioning, https://journal.code4lib.org/articles/8482

[4] The BagIt File Packaging Format (V1.0), https://tools.ietf.org/html/rfc8493

[5] Digital Preservation Storage Criteria (Version 2.0), https://osf.io/sjc6u/

operations that are involved in scanning OCFL structures for validation or rebuilding purposes. This had emerged as a key bottleneck with the design of the MOAB file layout.

Specific Features

Without going into too much technical detail on the specification[1], several features of OCFL should be highlighted with respect to the implementation of E-ARK AIP's.

- *The OCFL Storage Root*

While the CSIP specification deals purely with the internal structure of the object, the OCFL also describes how collections of objects should be managed in a file system, as a necessary feature for building archival systems and repositories. However, this is done in a way that is consistent with the CSIP design principles.

The "OCFL Storage Root" is the top level directory for storing OCFL Objects and should contain both a copy of the OCFL specification and a conformance declaration that indicates which version is implemented - as a check that the correct documentation is present, if nothing else!

In addition, the Storage Root should also contain documentation that describes the scheme used for distributing OCFL Objects across directories on the file system in a balanced way that maintains the efficiency of file system operations. Unfortunately, there is no single scheme that is optimal for all use cases, so various options and their relative merits are discussed in the Implementation Notes[2]. As a general principle, this "File System Disposition" should programmatically derive the path of an OCFL Object from its unique identifier.

- *OCFL Objects*

An OCFL Object is completely contained within one directory termed the "OCFL Object Root". At the top level of the directory there must be an object version conformance declaration, an object inventory, which is discussed further in the next section, and a digest for the inventory for validation purposes. Importantly, the OCFL only requires the version conformance to apply to the top level inventory and the most recent version of the object. This permits legacy versions to be included in an OCFL object without rewriting or otherwise tampering with them.

The content of the object is contained in sequentially numbered version directories within the Object Root, with all but the most recent version considered immutable. No content is stored outside the version directories. An optional *Logs* directory may exist in the Object Root to store information that does not affect the content of the object - for example, records of periodic fixity checks that identify no problems.

- *The OCFL Inventory*

The Inventory is the principal metadata structure mandated by the OCFL specification and is the primary mechanism through which most of its functionality is realised. Its primary function is to map between *content paths*, which point to physical files on storage, and *logical paths*, which indicate where these files appear in the logical representation of a version of an OCFL object. This distinction is important for a number of reasons:

1. *Deduplication* - the OCFL supports deduplication within an object, so that once a file exists in storage, with a given content path, all references to that particular content, regardless of filename are merely different logical paths that reference the single content path.
2. Filesystem Limitations - File systems may have limits on paths (such as length or restricted character sets[3]) that may mean that the object structure cannot be represented accurately on the file system. However, logical paths are not restricted in this way and thus object structure can be preserved regardless of file system restrictions.
3. *Efficiency* - Complex directory structures can be quite inefficient to traverse. For complex objects, the OCFL Inventory allows content paths to exist in a simplified hierarchy while retaining complexity at the logical level.

[1] *Oxford Common File Layout Specification 0.2,* https://ocfl.io/0.2/spec/

[2] *Oxford Common File Layout Implementation Notes 0.2,* https://ocfl.io/0.2/implementation-notes/

[3] *Comparison of file systems,* https://en.wikipedia.org/wiki/Comparison_of_file_systems

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

4. *Future Proofing* - Longer term, storage systems, such as content addressable object stores, are appearing that do not have a hierarchical file system. The OCFL Inventory is designed to be a functional object description even in this case.
5. *Optimisation* - Storage systems that handle large numbers of small files well tend to handle very large files poorly, and vice versa. While it is not the default behaviour, there is no reason content paths cannot point to separate storage locations for problematic files. This is a more robust approach than file segmentation or stressing unsuitable file systems.

The Inventory is formatted using JavaScript Object Notation (JSON)[1] because it is compact and easy to read for computers and humans. It has three main sections:

1. A preamble section that, most importantly, includes at least one unique identifier for the object.
2. A *manifest* section that lists every file in the object along with its digest. These are the *content paths*.
3. One *version* section for each version of the object that exists. Within each version section, a *state* section lists the digests for each of the files in the version (which must exist in the manifest section) alongside the *logical path(s)* for that file.

There is also an optional *fixity* section for additional fixity digests that is formatted in the same way as the manifest section.

Additionally, it is recommended that each version directory holds a copy of the inventory as it was at the time of its creation. This has the effect of the current version providing an additional copy of the inventory and allowing rapid rollback of the entire object state to an earlier version in the event of errors during updates.

Basic Operations

In addition to specifying how files should be organised, the OCFL Implementation Notes go further and define how basic operations on OCFL objects should be implemented with respect to inventory maintenance and the requirement for previous versions of objects to be immutable.

- *Inheritance*

By default, a new version of an OCFL Object inherits all the filenames (logical paths) and file contents (content paths) from the previous version. This serves as the basis against which changes are applied to create a new version. A newly created OCFL Object, obviously, inherits nothing and is populated by file additions.

- *Addition*

Adds a new logical path and new content with a content path in the current version. The logical path cannot exist in the previous version of the object, and the content cannot have existed in any earlier versions of the object.

- *Updating*

Changes the content pointed to by a logical path, which must exist in the previous version of the OCFL Object. A new content path is created in the current version of the object. The content cannot have existed in any earlier versions of the object.

- *Renaming*

Changes the logical path of existing content. The logical path cannot exist in the previous version of the OCFL Object.

- *Deletion*

Removes a logical path from the current version of an OCFL Object. The logical path and content remain available in earlier versions of the object.

- *Reinstatement*

Makes content from a version earlier than the previous version available in the current version of an OCFL Object. The content must exist in an earlier version, and not the previous version. The logical path may exist in the previous version, effectively updating the file path with older content, or it may not, effectively adding the older content as a new file.

- *Purging*

Purging, as distinct from deletion, covers the complete removal of content from all versions of an OCFL Object. This is a special case that is not

[1] *Introducing JSON,* https://www.json.org/

supported as part of regular OCFL versioning operations since it breaks the previous version immutability requirement. Ideally, a new OCFL Object with an amended version history should be created.

Community

The OCFL Community Google Group[1] is where discussion takes place and meeting announcements are made. At the time of writing, community conference calls are scheduled monthly. The specification and implementation notes are managed on Github2 and everyone is welcome to raise issues or even submit pull requests.

III. The e-ARK Common Specification for Information Packages

In 2017 the European Archival Records and Knowledge Preservation Project (E-ARK project[3]) delivered its draft common specifications for information packages to the Digital information LifeCycle Interoperability Standards Board (DILCIS Board[4]) . The board is responsible for the enhancement, maintenance, continuous development and endorsement of specifications. Specifications concern information packages as well as Content Information Types. The information package specifications describe OAIS reference model packages for archival transfer, but can also be used for other types of transfer. Content Information Type Specifications (CITS) describe the content itself as well as its structure within the package in order to facilitate easier content validation.

The DILCIS Board specifications are the core specifications in the Connecting Europe Facility Building Block eArchiving[5].

The drafts for information packages are:

1. Common Specification for Information Packages
2. Specification for SIP
3. Specification for AIP
4. Specification for DIP

These draft specifications have been updated, enhanced and published in version 2.0 during spring 2019. The specifications are available as pdf at the DILCIS board's webpage and as markdown in GitHub[6] accompanied with METS profiles and XML-schemas. Questions and issues are handled in each specification's GitHub repository issue tracker. GitHub has been chosen as the transparent platform in which users can follow progress, see notes and comment on the current work.

- *Common Specification for Information Packages (CSIP)*[7]

The core package specification describes general principles and requirements for an information package, that are shared by all types of information package in the OAIS reference model.

The principles present a conceptual view of an Information Package, including an overall IP data model, and use of data and metadata. These principles could be implemented with a physical directory structure and the requirements are expressed with the Metadata Encoding and Transmission Standard (METS)[8].

The principles describe:

- General requirements for the use of the specification;
- Identification requirements ranging from identification of the package to identification of the transferred digital files;
- Structural requirements for the content in the package, for example how different kinds of metadata should be structured and added;
- Metadata requirements outlining the use of standards for describing data.

[1] Oxford Common File Layout Community, https://groups.google.com/forum/#!forum/ocfl-community

[2] The OCFL Specifications (WIP), https://github.com/OCFL/spec

[3] European Archival Records and Knowledge Preservation, http://eark-project.com/

[4] The Digital Information LifeCycle Interoperability Standards Board, http://dilcis.eu/

[5] CEF eArchiving BB, https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving

[6] DILCIS Board in GitHub https://github.com/DILCISBoard

[7] E-ARK CSIP http://earkcsip.dilcis.eu/

[8] Metadata Encoding and Transmission Standards http://www.loc.gov/standards/mets/

The requirements are expressed using METS and PREservation Metadata Implementation Strategies (PREMIS)[1]. METS describes the requirements on the package level and PREMIS defines the preservation metadata needed, especially those for the AIP. The METS specification available both as a METS Profile and in text form in the specification expresses the requirements for how each part of METS is to be used and how it fulfills the CSIP principles. A validation tool has been created to support automatic metadata validation.

In summary, the requirements specify:

- how to identify the package ;
- how to describe the type of content;
- how to link descriptive, technical and provenance metadata;
- which files are to be contained in the package, where each file is described by its :
  - File name
  - Path
  - Mime Type
  - File size
  - Checksum
  - Creation date
- The mandatory METS structural map which describes the package structure on a high level.

Sometimes an IP is large, reaching tera bytes in size. This is cumbersome to handle, both for the submitter and the receiver of the IP. An example is an IP that contains a whole database from an electronic records management system with records comprising over a year. Therefore, CSIP contains a section that discusses how to split large IP's. In a coming version of CSIP this section will be extended and give guidance on splitting large packages. Splitting leads to more than one IP being created. The full IP is established by creating package referencing connections between the split IP packages. Draft specification texts are currently being written that describe how to carry out the splitting, as well as to how to describe the different parts and their relationships. These updates will be published after a review period.

[1] PREservation Metadata Implementation Strategies, http://www.loc.gov/standards/premis/

E-ARK profiles building upon CSIP

E-ARK SIP[2], E-ARK AIP[3] and E-ARK DIP[4] profiles all use CSIP as their basis and extend the CSIP requirements with requirements for their specific type of information package. Some requirements extend existing specific CSIP elements, for example by requiring the value for the element describing the OAIS Reference Model type of the package being set to the appropriate value (SIP/AIP/DIP). The focus in this paper is the AIP. To learn more about the E-ARK SIP and E-ARK DIP please refer to their available specifications.

E-ARK AIP

- The objectives for the E-ARK AIP are as follows:
- To define a generic structure of the AIP format in a way that it is suitable for a wide variety of data types, such as document and image collections, archival records, databases or geographical data.
- To recommend a set of metadata standards related to the structural and the preservation aspects of the AIP.
- To ensure the format is suitable for storing large quantities of data.
- To mitigate the potential preservation risk of repository obsolescence by implementing a repository succession strategy.

The purpose of defining a standard format for the archival information package is to pave the way for simplified repository migration. Given the increasing amount of digital content archives need to safeguard, changing the repository solution should be based on a standard exchange format. This is to say that a data repository solution provider does not necessarily have to implement this format as the internal storage format, but it should at least support exporting AIPs. By this way, the costly procedure of exporting AIP data as Dissemination Information Packages (DIPs), producing SIPs for the new repository solution, and ingesting them again in the new repository can be simplified. Data repository solution providers know what kind of data they can expect if they choose to replace an existing

[2] E-ARK SIP, https://earksip.dilcis.eu/

[3] E-ARK AIP, https://earkaip.dilcis.eu/

[4] E-ARK DIP, https://earkdip.dilcis.eu/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

repository solution. An E-ARK compliant repository solution should be able to immediately analyse and incorporate existing data in the form of AIPs without the need of applying data transformations or having to fulfil varying SIP creation requirements.

Generally, a variety of repository systems are being developed by different providers. The way the AIP is stored often depends on specific requirements which have been addressed according to the needs of their respective customers. For this reason, the purpose of the E-ARK AIP format is not to impose a common storage format that all repository systems need to implement. While it can be used as an archival storage format, it can also be seen as a format that makes system migration easier.

IV. Alignment of The E-ARK AIP and The OCFL

The OCFL is engineered from the viewpoint that a digital object should be considered a greater whole, comprising several streams of information that can be arbitrarily labelled data or metadata but all of which contribute to the intellectual content of the object. Consequently, it does not make any assumptions about the internal structure or composition of a digital object, which is key to the alignment between the E-ARK and OCFL specifications. In this respect, the CSIP specification and the extension profile for E-ARK AIP can be considered as filling this intentional gap in the OCFL for a number of use cases, to provide a more complete approach.

A very simplistic implementation could therefore just encapsulate an entire eARK AIP structure within an OCFL object. However, since the OCFL provides file mechanisms for fixity, versioning, deduplication and logging that are optimised for simplicity and computational efficiency, a more nuanced and functional approach would be to consider where these could interoperate with corresponding elements within the AIP structure.

The essential part of the alignment of the two approaches is that the AIP structure is implemented at the *logical* level within the OCFL. The OCFL client software can then handle versioning, deduplication and other features transparently but present the AIP structure when queried by other software.

Fixity

As stated previously, the OCFL supports SHA512 or SHA256 as the default digests for its content addressability, however other algorithms are permitted and the *fixity* section of the inventory allows storage of hashes generated by additional algorithms. These could be extracted from an AIP (by parsing METS files or manifest.txt, if it exists) as part of object creation or version updating. Using a SHA256-based implementation of OCFL obviously aligns well with the E-ARK AIP since these values can be shared.

The OCFL can technically support the use of other hash functions for *manifest* content addressing, but validity checks will generate errors for fixity algorithms that are considered broken/deprecated (e.g. MD5[1]). As a result, using the other fixity algorithms in place of SHA512 or SHA256 is not advised.

Copying digests from the OCFL inventory into the AIP is also possible but requires a little more care, since OCFL includes digests for every part of the AIP. Such a process would therefore need to exclude the METS and/or manifest files that would be updated.

Versioning and Deduplication

The OCFL differs from the AIP specification in the way that versions are treated, since it makes no assumptions about the types of changes that may occur. It also makes the version history explicit in the manifest with *state sections for every version.*

The E-ARK AIP versioning model is, in some respects, analogous to the OCFL model, in that the parent AIP can be seen as equivalent to the OCFL Object Root with child AIP's equivalent to OCFL versions. However, the AIP model is somewhat encumbered by the requirement for the parent to be compliant with the CSIP which results in additional complexity. In addition, using the AIP model can require multiple file parsing operations to determine version differences whereas the OCFL requires minimal processing.

This can become a significant overhead when objects are referenced externally, since, for

[1] G. Ramirez, MD5: The broken algorithm, https://blog.avira.com/md5-the-broken-algorithm/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

referential integrity, the version of an object current at the time of citation should be readily accessible, using a protocol such as Memento[1]. Being able to easily identify the difference between any two versions is also essential for the efficient synchronisation of distributed storage systems that are maintained asynchronously. This was a bottleneck encountered, in practice, with systems that use the MOAB layout.

Thus, while it is perfectly possible to implement the parent-child AIP versioning model as distinct AIP's in the OCFL, a more efficient approach would be to create new versions of an AIP within a single OCFL Object, allowing the OCFL client to deduplicate the common elements between versions and providing quick access to the version history. This also eliminates the duplication of information between parent and child IP's, along with the consequent maintenance overheads.

The OCFL is constructed so that all changes to an OCFL Object are additions to its contents. This allows AIP's to be updatable but, at the same time, forces each version to be immutable but without incurring undue storage overheads. Using the *reinstatement* mechanism described earlier, it also allows rollback of failed DP actions such as migration at any point after the event[2].

Logging

The OCFL expects new versions to be created when a meaningful change is made to an object. A periodic virus scan or fixity check with a null result thus does not automatically result in the generation of a new version. In practice, there are a wide variety of events that may impact storage but which are largely invisible to preservation systems without explicit action. Examples would include operating system file system maintenance, and hard drive replacement and subsequent array rebuilding operations.

In practice, then, these can be potentially numerous and creating a new AIP each time would not make sense either. However, there is merit in capturing this information for recovery and audit purposes. In the OCFL, these can be captured in the *logs* directory which is outside the object version structure. PREMIS is suggested, but not mandated for this purpose in the OCFL, but it would be sensible to do so if using E-ARK AIP's. If desired, these logs could then be periodically added to a new version of the AIP to embed this audit trail without undue AIP version proliferation.

In the E-ARK AIP the use of PREMIS is mandatory, including the use of events. The full description of the PREMIS use in the specifications and the eArchiving Building Block is not ready at the time of the writing. The work is ongoing and the use of the semantic units of PREMIS will be described in its own document to allow it to be used in all the different IP's easily.

Pathname Mappings

Complex objects can contain paths that are not supported by the file system being used for preservation, especially if they have been imported from another system. This can be for reasons of length, number of directory levels or character encodings, amongst others. The OCFL handles this by allowing long Unicode *logical paths* while implementing *content paths* on storage that may be shortened or use different character encodings. No specific algorithm is mandated since the mappings are explicit in the OCFL inventory.

If the AIP is implemented at the OCFL logical level, then complex AIP structures need not be subject to such file system limits.

V. Conclusion

The choices that can be made when creating a digital archiving approach are numerous, starting with what you consider to be the first AIP. Should it be the SIP that has just been transferred and put directly into preservation storage so you can go back if everything gets demolished through a "bad decision in migration" further down the preservation journey?

Are we concerned more with the preservation of bitstreams as standalone entities or with the

[1] HTTP Framework for Time-Based Access to Resource States -- Memento, https://mementoweb.org/guide/rfc/

[2] You will thanks us for this, believe me! (Neil Jefferies)

preservation of *knowledge*, where the meaning of an object can be largely determined by its relationships to other objects - relationships that necessarily change over time as a result of human discourse? Thus we need to consider how to design systems to capture and preserve this metadata and when and how often to capture this in new AIP versions.

At a technical level, we need to create systems that support the curatorial requirements of digital preservation yet also address the unavoidable limitations of the underlying computational and storage technologies.

Both OCFL and the E-ARK AIP standards go some way to addressing these issues, whilst accepting that not everyone will necessarily make the same decisions about their approaches, for entirely logical reasons. This preliminary analysis shows that, in many respects, the standards are largely complementary in that their primary foci are differ

ent aspects of the broader digital preservation problem space - the structure of preserved digital objects, and the efficient storage and management of them, respectively. This, somewhat fortuitous, "separation of concerns" is considered good practice in terms of systems design.

It can be seen that abstracting the logical structure of an object from the storage structure with the simple logic embodied in the OCFL inventory permits the E-ARK AIP to be realised over a broader range of platforms, very much in keeping with its purpose. It even has the potential to allow the use of object stores which do not implement hierarchical path systems at all.

Both efforts are still in the development phase and more work is required to bring them to fruition. However, this paper shows that there is value in working together, learning and contributing to each other. One early recommendation from the OCFL community to the E-ARK CSIP community is to look further into the selection of checksum algorithms. An area that probably requires further discussion on both sides is the issue of object/AIP dependencies - where one object, such as a collection, depends on the existence of others.

More recommendations and comments will most certainly pass between the groups as we move forward, particularly once we begin to write code and develop systems. Collaboration between efforts can only be beneficial!

...Diversity and choice is always good for Digital Preservation - as is discourse and alignment between concerns and communities.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

A Pragmatic Application Of PREMIS

Mapping the key concepts to a real-world system

Jack O'Sullivan
Preservica Ltd, United Kingdom
jack.osullivan@preservica.com
0000-0002-0306-761X

Richard Smith
Preservica Ltd, United Kingdom
richard.smith@preservica.com

Alan Gairey
Preservica Ltd, United Kingdom
alan.gairey@preservica,com

Kevin O'Farrelly
Preservica Ltd, United Kingdom
kevin.ofarrelly@preservica.com

Abstract – a data model is an expression of how a system is intended to be used, and a statement of how it should interact with other systems. As part of the development of the latest version of Preservica, the underlying data model was significantly altered, informed by what went before and by reference to the experiences and best practices distilled into the PREMIS Data Dictionary. This paper reports on some of the key decisions made in the application of the PREMIS concepts to an extant digital preservation system.

Keywords – PREMIS, Data Model, Digital Preservation Systems

Conference Topics – The Cutting Edge: Technical Infrastructure and Implementation; Designing and Delivering Sustainable Digital Preservation.

I. Introduction

This paper describes some of the decisions made in the course of making improvements to the underlying data model upon which the Preservica product is built. In section II we describe the role and importance of the data model in a system, section III describes why this work has been undertaken and sections IV and V describe some of the key features and how they correspond to the PREMIS Data Dictionary.

II. The Role Of The Data Model

The data model of a system codifies how it views and understands the wider ecosystem in which it operates. It is a means of describing external realities, and of describing and enabling the processes performed by the system. As part of the system interface, it can shape how the system interoperates with other systems. Since it is generally defined to enable the processes the system is designed to perform, it can act as limitation on what the system is able to do.

The data model is also the lens through which the system is viewed. As such, it shapes the way in which users and developers think about the system, its capabilities and its limitations. This means that even if the data model can support an operation, users may be reluctant to perform it, or even unaware that they can, and developers may seek to artificially restrict it. Conversely if something seems like it should be possible, users and developers can be encouraged to use the system in ways it was never intended, often with less than optimal results.

The data model also reflects an understanding of what the system is, or should be, at a particular moment in time. As the world around the system changes, the definition of what the system can, should or should not do is likely to change.

To allow a system to be used flexibly, and to perform functions that could not have been anticipated at the time of the original design, a data model must itself be flexible and extensible. There are however pitfalls of an overly flexible model. If the model does not clearly describe the required functionality, or if it is overly permissive in interpretation, then actual implementations will tend to diverge to the point that managing system changes becomes problematic. A good data model is thus always based on a trade-off between allowing users the freedom to do what they need and constraining

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

them to behave in a way that is consistent with the intended use.

III. The Preservica Data Model Challenge

Preservica has been built on an eXtensible Information Package-based data model (XIP) since the initial end-to-end preservation system was developed in 2006. Minor extensions to this model have been made with each subsequent release through to v5.10 (July 2018), but the fundamentals of the model have not been significantly altered in this time. This model was built with the intention of describing processes that would be generally applicable to anyone performing digital preservation, in the light of the original PREMIS definition published slightly earlier [1]; however, since the original users of Preservica were archives, this model was initially tested and validated specifically in the context of the requirements of an archival setting. It is a testament to this original design that as its user base has swelled to include libraries, museums, and business records repositories, across multiple sectors, each bringing different, sometimes competing requirements, the model itself has held firm, able to satisfy most of these requirements. The XIP model has proven itself to be flexible enough to describe several complicated use cases, some of which are referred to within this paper.

Increasingly however, it has become a constraint on the functionality Preservica can offer, or at best a complication, making the development of new features slower and more complex than they might seem at face value. It has also become increasingly clear that the way in which the model has been implemented by many users has diverged from the intention of the original design in some areas. This often leads to having to decide whether a new feature should interpret the model in its "pure" form, making the behavior unintuitive for users, or in the way it is more commonly used, gradually breaking the assumptions that can be safely made. As such, version 6.0 of Preservica is based on a new data model that seeks to address these limitations in the existing XIP.

Whilst a "new data model" sounds like a large green field opportunity, the development has been shaped from the start by several constraints. Firstly, there is the decade plus of data amassed in Preservica systems world-wide in the older XIP format, which dictate that clear mapping from XIP to the new model should be available; whilst it is desirable to be able to perform the reverse mapping, it was not a requirement to ensure a "round-trip" would return exactly the same data in exactly the same structure. Secondly, the longevity of XIP indicates that it has been a successful model in many ways, losing sight of what it does well would risk throwing the baby out with the bathwater. Thirdly, Preservica does not exist in a vacuum. It must interact with other systems, from standard operating and file systems, through widely used content management, catalogue, and discovery systems, through to bespoke access, workflow, and storage systems, all of which means mapping to other formats and standards must be possible. Perhaps most importantly, Preservica is intended to be a long term preservation system, and its users have responsibility for long term planning, so being able to describe the model in terms of a recognized and accepted industry standard is paramount, hence all stages of the design have informed by version 3.0 of PREMIS [2], and the process has been, in effect, an exercise in an implementation of PREMIS in the context of an active preservation system.

Finally, Preservica will need to interact with systems that use the operating system/file system model of folders/directories and files. This is also the default mental-model that end users tend to bring to any hierarchical presentation of information. Whilst being able to map to this model of the world may seem so obvious that it doesn't need to be stated, the concepts involved in digital preservation mean that this is not necessarily a straightforward process.

The high-level conception around how the existing data model maps to PREMIS entities and file system models is given in Figure 1.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

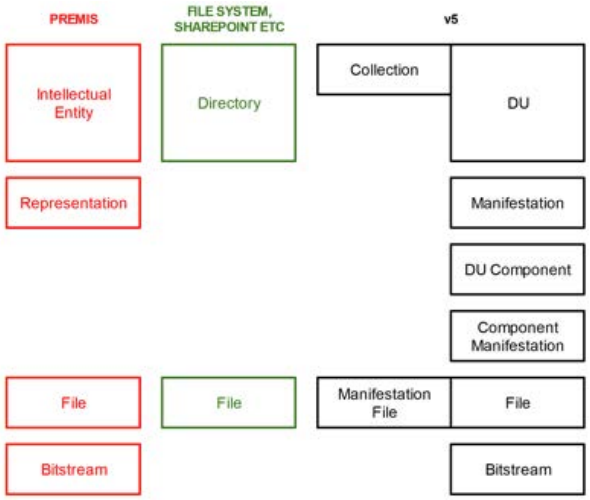


Figure 1 - A mapping comparison of high level entities in PREMIS, a standard File System, and Preservica, up to version 5 (XIP data model). Entities at the same vertical level have the same responsibilities.

The "DU" in this mapping stands for "Deliverable Unit". These are described as "information objects that are the conceptual entities (records) that are delivered by archives and libraries"; they are hierarchical in that a DU may contain nested "child" DUs, but are also considered, confusingly in many cases, atomic units of information. A "top-level DU", that is a DU with no parent DU, is taken as the implementation of an Archival Information Package as described in the OAIS reference model.

The "Collection" is described as "a hierarchical grouping of Deliverable Units". They are hierarchical in that a Collection may contain nested "child" Collections, but they are effectively just metadata records for organizing DUs.

In mapping this existing model to PREMIS, both Collections and DUs have historically been regarded as Intellectual Entities.

The "Manifestation File" is a conceptual entity creating a link between a "File" (which follows the PREMIS definition of File) and the "Manifestation" (a close analogue of the PREMIS Representation). This allows a single digital file to simultaneously be part of multiple Manifestations (and indeed multiple separate DUs).

This abstraction between files as "physical" entities and the construction of higher level "conceptual" entities allows for some of the most complex use cases of the data model. For example, one user of the system stores television broadcast video. The "physical" files represent distinct time units of broadcast (for example the hour of video stream broadcast between 00:00:00 and 00:59:59 on a given day). The "conceptual" entities, the DUs, represent actual programmed content, (e.g. an episode of the daily news). The programmed content may be wholly contained within a single time unit, may approximately equal a single time unit, or may require content from two (or more) consecutive time units. The same "File" is thus potentially a component of multiple DUs. Secondary "Manifestations" can then be created which splice the relevant parts of the underlying streams into a single video file representing only the programmed content.

Similarly, when performing web-crawls to preserve a website, the content of the crawl is stored across multiple "physical" WARC files (so that no individual file is too large), but creates a single Manifestation of a single conceptual "DU". For the most part, users interact with the DU, particularly for rendering, where the whole web crawl is played back rather than individual files.

The "DU Component" and "Component Manifestation" are elements attempting to record and reconcile how a single DU is composed of multiple elements. In most cases each component is effectively comprised of a single "File", but conceptually, a component such as an email may require multiple files (the message itself plus any discrete attachments). This recognizes the idea that a single "record" may comprise multiple pieces of otherwise distinct content.

The rest of this paper will describe Preservica's new data model in terms of the external realities we are modelling, how these are modelled in PREMIS, and the decisions, trade-offs and compromises made. References to "XIP" should be read as the data model of Preservica versions up to 5.10, whereas "XIPv6" refers to the new data model.

The exercise of performing an in-place update of a V5.X system to a V6 system has been developed in parallel to the development of the new model to ensure that existing customers will be able to upgrade and that information will be presented in the new model in the way that best represents what it actually is. This process itself merits further discussion but is beyond the scope of this paper.

IV. Entities And Objects

The goal of digital preservation has long been described in terms of preserving digital *information* [3]. In a general sense, information is obtained when we use some software to interact in some way with some digital data. All digital data can be thought of as streams of bits, binary ones and zeroes. The data only become information when the software used to interact with it can interpret those ones and zeroes as higher-level abstractions. The lowest abstraction above the stream of bits, is a stream of bytes, each byte consisting of 8 bits. This is generally the lowest level aggregation of digital data that software systems address and process. Above this, bytes may be interpreted either individually or in aggregation as simple entities such as numbers, characters, words, or more complex entities such as arrays.

In order to consider the information we can abstract from these data to be "preserved", we need to be confident that we have some piece of software that is able to correctly interpret a particular stream of bytes as the intended series of more complex entities. In some cases, it is sufficient that the software can render the correct "human-readable" entities such as strings, numbers and images, in other cases the "machine-readable" entities manifest as some form of behavior that must be correctly replicated.

Whilst processing these data, the software builds an internal state map of how certain bytes or groups of bytes should be interpreted, potentially updating as user inputs modify the data. These data need to be persisted between sessions in such a way as allows the software to re-build this internal state map; to do this, the data is written out, or serialized, according to a set of rules, typically referred to as a format. It is this format that allows software to interpret data, and it is changes to the list of formats that a given piece of software is aware of that leaves digital data vulnerable to becoming unusable.

A. *Aggregations of Data*

Operating systems and storage systems group the formatted bytes that need to be persisted into units called files. Just as the byte is the lowest aggregation of data that software interacts with, the file is the lowest aggregation that the operating and storage systems will address, as evidenced in the PREMIS definition: *"A named and ordered sequence of bytes that is known to an operating system*" [2]. It is often assumed that the file is an atomic aggregation of data that can be regarded as an intellectually complete unit of information, whereas in fact it is an artifact of the implementation of data persistence, part of the data model for an operating system or file system.

However, files are not necessarily atomic. PREMIS specifically makes provision for files to be considered as containers or aggregations of other units of information, drawing a distinction between two different types of potentially embedded units, a Bitstream and a filestream.

Bitstreams are first class Object entities within the schema, defined as "*Contiguous or non-contiguous data within a file that has meaningful properties for preservation purposes" [2].*

It is further clarified that these are *"only those embedded bitstreams that cannot be transformed into standalone files without the addition of file structure (e.g. headers) or other reformatting to comply with some particular file format*" [2]. An example would be the video and audio streams within a video file; although encoded according to well defined schemes, these are not necessarily directly extractable as standalone files.

If an embedded stream can be transformed into a standalone file without the addition of extra information, for example a TIFF image embedded within a zip File, PREMIS considers it to be a filestream. The filestream is not a first-class Object entity within the schema, although all properties of files apply to filestreams as well.

In either case, a bitstream is only significant if it can a) be identified and characterized independently of any overall file, and b) carries a separate preservation risk. This describes both the video and audio streams within an AV file, and individual documents within a zip file.

The logical separation of files and filestreams makes sense in the context of relating to external systems, where a file is a physical reality, whereas a filestream is an abstract concept. Internally however, to perform any preservation action, Preservica will have to retrieve either a file or a filestream from a long-term storage location, creating a temporary local copy and as such the distinction between the two becomes rather academic. It is also clear that the definition ”contiguous or non-contiguous data that has meaningful properties for preservation purposes” fits both files and filestreams, as well as describing why they are interesting to a preservation system. As such, both files and filestreams are represented internally in XIPv6 as “bitstreams”.

It should also be recognized that an individual file can be a “sub-atomic” unit as well as a “super-atomic” unit. In this context, we should understand that atomicity can mean different things and whilst there are several senses in which this may occur, these can generally each be considered as one of two categories.

1) *Cases where the file is technically atomic:* there exists software that is capable of interpreting the file in isolation, and where examining the file provides a self-consistent and interpretable set of data to a consumer, but where the true information content can only be understood in the context of other files. There are two important sub-cases here, firstly where the other files are unknowable from a technical interrogation of the files, and secondly where the other files are explicitly referenced. In the first case, the context will have to be defined and recorded explicitly by a user.

In the second, this context can be inferred and recorded automatically in software. This case requires higher-level entities to be introduced to the data model and will be addressed in later sections.

Examples of technically, but not informationally atomic files include:

- A series of word processing documents, each of which represent the minutes for a particular meeting, but where the meetings all relate to a single project and the complete set of minutes should be considered the atomic unit of information to be preserved.
- A Digital Video package where video, audio, and additional data such as subtitles are stored in separate files. Each file might be usable in its own right, but the true information context is revealed only by software that interprets all the files together. Such an example is given in [4].
- A web-page, where a single HTML file can be well structured and readable by a browser, but where images are referenced rather than embedded, or links to documents exist; in this case, all the “linked” files must be available and returned by following the links, before the full data is provided to user and before the information content can be understood.

2) *Cases where the file is not technically atomic*: there is, by design, no software capable of correctly interpreting this file in isolation. In this case, there may well be alternative means of storing the same information that would be a single, technically atomic file.

Examples of technically non-atomic files include:

- Certain types of disk images (e.g. BIN/CUE) or scientific data sets (e.g. ENVI [5]), where the raw binary data is contained in one file, but the header information that instructs the software in how to read that data is contained in a separate file. In this case, the “data” file without the headers is a meaningless binary dump, and the header file without the data is better understood as metadata than actual data. In some cases, these could be considered PREMIS Bitstreams, with the additional “file structure (e.g. headers)” coming from a different file, but they are already, by definition, a “stand alone file”.

- A case where a single large file has been partitioned into many smaller files for the purposes of by-passing storage or transport limits, and where only by recombining the files into a single large file can the data be understood by any software. Again, they are Files by

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

definition, but in these cases, it is not clear that the Bitstream definition applies at all. They are not just missing "file structure" or in need of re-formatting, but rather are missing actual bytes of content, often with boundaries in the middle of a complex entity such as a word or array.

It is clear from this that the risks associated with format obsolescence should be properly thought of as being an "atomic content" issue rather than strictly a file issue. In this light, XIPv6 uses a "Content" entity to track content with a specific "format". This entity can contain multiple bitstreams, although in practice, in most cases a "Content" entity has a direct mapping to a single bitstream.

B. Higher Level Entities

Having established that even in the case of technical atomicity, the file is not necessarily an atomic unit of information that should be preserved, we are left with the need for a higher-level entity to describe that unit. This is an Intellectual Entity in the PREMIS model.

Intellectual Entities can include other intellectual entities and may have one or more digital or non-digital instantiations. Digital instantiations are Representations in the PREMIS model.

There is an implication here that an Intellectual Entity has two responsibilities. The first, as a hierarchical aggregation of material, and the second, as an expression of an individual piece of information.

XIP's "Deliverable Unit" satisfied this definition, however combining these two responsibilities in a single entity type made the system complex and allowed different customers' usage of the same system to diverge.

For our earlier "Archives" customers, the DU was considered to be an immutable record, something that could be defined and laid out during ingest, and never altered. In this interpretation of the model, the hierarchical nature of the DU was fairly abstract and secondary to its "immutable" nature.

The fact that DUs were hierarchical became more of a primary concern for two main reasons. Firstly, a growth in the user base of "non-Archival" customers, for whom records are not necessarily fixed for all time, but may require restructuring. Secondly, a growth in the number of less technically expert users, for whom a hierarchical entity like a DU mapped in their mental model to a folder/directory. The second of these was probably not helped by a choice to represent DUs in the GUI with a "folder" icon, whilst Collections were represented as "filing cabinets".

Requests from both the non-Archival and less expert users to be able to "re-arrange their folders" increased in both volume and legitimacy of use case. The implementation of post-ingest restructuring was a good example of a change that appears "trivial" to users, but was complex to implement. It is also an example of a change that broke several assumptions implicit in the way the data model was envisioned, but not explicitly codified. For example, the XIP model had an "IngestedFileSet" entity, which broadly represented the set of a physical content ingested as part of the same transfer. A working assumption that files in the same FileSet were directly related by something other than coincidence of timing was broken by allowing Files to be spread at will through the logical hierarchy of the repository.

For XIPv6 therefore, it was determined that a separation of these responsibilities should be undertaken.

A top level intellectual entity should be created as a means of aggregating related content and providing a primary hierarchical structure.

This "Structural" entity can be part of another structural entity and can be a parent for any other intellectual entity, giving rise to hierarchical structure. It does not have any direct instantiations itself, and whilst the metadata describing it may need to be retained for the long term, it is not considered to have long term preservation risks beyond that of metadata storage. This entity is analogous to the "Collection" from the XIP model, however as the details of the metadata fields available has changed between the two models, and to ensure that the term "collection" does not have confused meaning in internal communication (and communication with users), a different name ("Structural Object") is used in XIPv6.

The second intellectual entity that XIPv6 defines exists as a means of defining the unit of information

that should be considered atomic for the purposes of preservation. This is most closely related to the DU in XIP but differs in important respects. By definition, as an atomic unit, this "Information" entity cannot be part of another "information" entity (it will exist in the hierarchy within a "Structural" entity). This is probably the most important respect in which it differs from the DU. It does allow for multiple instantiations, as discussed further below.

The assumption that files are atomic information units (as mentioned in A above) manifests as requests from users to be able to perform operations such as movements and deletions on individual files. In most cases, a file is a complete set of data required for an atomic entity, and such requests are entirely reasonable, however a long-term preservation system must protect against those cases where that is not true, and changes to an individual file risk corrupting the entity.

Defining a higher-level Information entity such as the "Information" entity allows us to define this as the level at which actions such as deletion or movement are performed. As the "atomic" entity of information, it is also the natural level at which preservation actions such as normalization, migration and rendering should be performed, and is thus the "natural" level to define as an AIP. In most cases, this Information entity is the base unit that should appear to the user, thus in the user's mental model it will likely equate to "file", solving the issue of allowing "individual file deletion" and allowing the "Structural" entity to map more correctly to "folder/directory".

The conceptual structure of entities within the repository means that another PREMIS concept, the Relationship, is also required in the model. There are clearly inherent parent-child relationships within the structure, but an explicit Link between two or more entities can also be defined, allowing the model to record connections such as "derivedFrom", "inResponseTo", that may not be otherwise apparent.

C. Instantiations

The PREMIS definition of a Representation provides for it being "*The set of files, including structural metadata, needed for a complete rendition of an Intellectual Entity". Within the XIPv6 model, this is applicable to "Information" entities, but will rely on the more generic "Content" rather than files.*

The need for more than one instantiation of an entity can arise from multiple requirements. There are however, two obvious cases.

In the first, there is considered to be a long-term preservation risk due to format obsolescence, i.e. it is feared that there is no available software capable of correctly interpreting the content, or that access to such software is not feasible or possible for the repository, and as such long-term management should occur within the context of an institutionally supported format. In such cases, the transformation of content is at issue. A standard example of this concern is word processing documents in an outdated or proprietary format, such as WordPerfect or Microsoft Word. In this case, having an instantiation in OpenDocument Text (ODT) might be the desired outcome.

In the second, there is considered to be an issue with the presentability of the content as is. In such cases the holding institution may not consider the original content to be at risk but may consider that the designated community will not have the means to interact with the information in its original form, or simply that there is a more convenient alternative means of dissemination. Examples of this might include a desire to have lower resolution/lower bit rate images/videos for access to conserve bandwidth; or a digitised manuscript, book or journal where the digital information is in the form of a series of high resolution images, but a PDF or eBook is the more natural digital surrogate.

In the first case, the format of the content may need to be changed, but the fundamental form of information remains unaltered, WordPerfect and ODT are merely different encoding of the same fundamental "word processing document" type. In the second case, the form of the information may be altered, but the content still represents the same intellectual entity, a book is still a book whether presented as a series of TIFFs or a single PDF.

If the first type of migration can be considered as happening at the Content entity level, then the Representation of the Information entity level can always refer to the same Content entity. In our example, the Information entity might be a report, with a single Representation that is comprised of a

single “word processing document”. After migrating from WordPerfect to ODT, the Representation simply provides a later “Generation” of the document. It is in these cases that serious consideration of significant properties, and significant significant properties [6] must be made in validation of the migration.

The migration from a series of TIFFs to a PDF creates an entirely new Representation that leans more naturally to the idea that “this book can be represented as a collection of pages, or as a single complete document”. In this migration, it is more problematic to rely on significant properties as a validation as the significant properties for an image are likely to look very different to those for a document; similarly, certain properties like image size might deliberately be changed in the creation of a presentation version.

It is in the second sense that XIPv6 uses the term “Representation”, using “Generation” to reflect the set of files needed for a complete rendition of an atomically complete piece of content. This allows the use of “Normalization” [7] to describe the process of creating new generations of content, and the more generic “Migration” [7] to describe the process of creating new representations of information, allowing different policies to be specified for each and different levels of validation criteria to be applied to each.

In XIP, both forms of “instantiation” were represented by “Manifestations” of a DU. This made it difficult to assess whether pieces of content in different manifestations should be considered to be intrinsically linked for the purposes of validating transformations. By separating Generations from Representations it is clear that all Generations of a piece of content should ultimately derive from the same source and have some shared invariant “properties”, but content in different Representations may be entirely unconnected (except insofar as the combined content of one Representation should convey the same “information” as the combined content of another).

A simplified summary of the entities in the XIPv6 data model, alongside the levels at which they will relate to other logical data models in the system, is depicted in Figure 2.

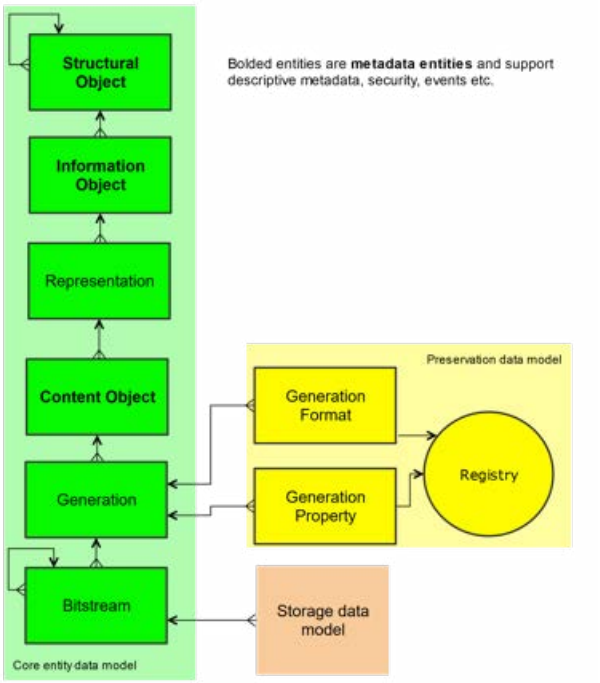


Figure 2 - Simplified summary of the XIPv6 Data Model

Figure 3 - Screenshot of an Asset in a Preservica V6 system shows a screenshot of the details of an Asset in a Preservica V6 system, this highlights several of the concepts. The “breadcrumbs” in the light blue bar towards the top show the hierarchical structure, where in this case “Test SO” and “videos” represent Structural Entities and “WindowsMedia” represents the Asset in Question.

The tree in the centre of the image shows the Preservation Representation, which consists of a single piece of content. The first Generation of this content was a Windows Media Video (wmv) file. At some point in its history, this content has undergone normalisation to create a second Generation; in this case to a Matroska (mkv) file. There is also an Access Representation of this asset, which also consists of a single piece of content. In this case, the content is a MP4 video, which is more usable for streaming video access than either the original wmv or the mkv.

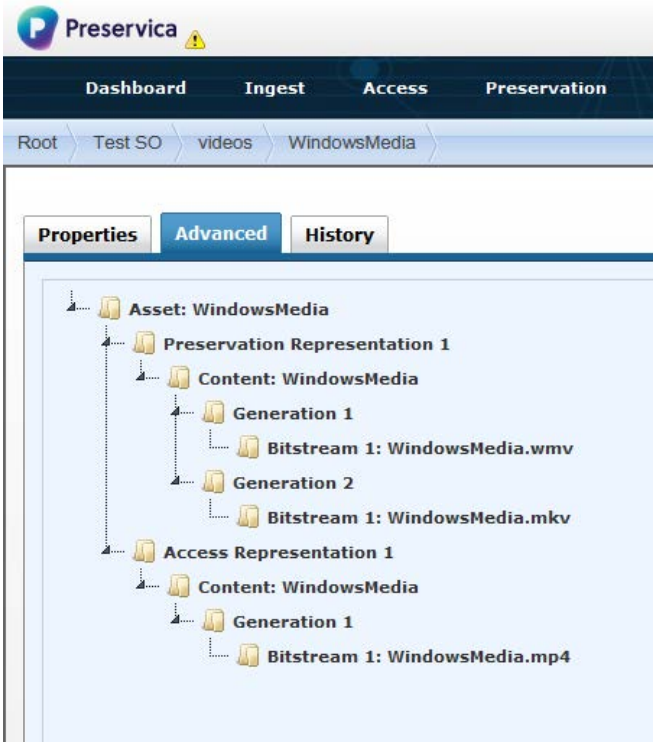


Figure 3 - Screenshot of an Asset in a Preservica V6 system

V. Events And Agents

As a preservation system, Preservica undertakes preservation actions that will need to be recorded as part of a chain of custody for the materials stored within it. These actions may be performed directly by the system or by deferral to third party tools and applications. Again, PREMIS defines the relevant entities to enable this record to be made, with Events and Agents.

Events in PREMIS are defined as "*an action that involves or affects at least one Object or Agent associated with or known by the preservation repository*". What it does not specify is the granularity of record required, or the actions that should be audited.

Reference [8] discusses what event metadata needs to be recorded, specifically with reference to idea of "metadata bloat". That discussion is helpful in guiding a decision but does phrase the discussion in terms of "the organization". As a platform used by many organizations, who may come to different conclusions in this regard, Preservica needs to ensure that enough metadata is recorded to satisfy any of its users.

Metadata bloat is unlikely to be a major concern in terms of the sheer volume of storage required as it is likely that this will always be marginal compared to the size of the repository as a whole, but it can impact on the performance of the system, particularly in reference to any process that attempts to extract the audit trail of a particular object, and so in designing this model, care was taken to try to avoid "event explosion".

To take the example of the characterization of content during submission to the repository, this requires running a format identification tool on each incoming bitstream. As a result, validation of the format may be required, followed by the measurement of technical properties, potentially by multiple tools. However, at a high level, all that has happened is that the Information has been characterized.

The fact that these actions happened should be surfaced to the user through the User Interface (UI), otherwise there is little point in recording it. From the earlier statement that the Information entity is the atomic unit of preservation, and the base unit presented to the user, it follows that this user presentation must happen at this entity level.

For an Information entity consisting of 10 bitstreams, the single process of "characterizing" an entity could result in at least 30 events being presented to the user. This is likely to be an overwhelming to a user, and risks burying pertinent information in a sea of detail.

This "event explosion" is avoided by asserting that Events in Preservica must be recorded against a high-level entity (generally the Information entity, but potentially also the Content itself), but that individual actions can be recorded against an event at a more granular level. This achieves the compromise of allowing the high-level events to be examined free from the minutiae of what happened, whilst ensuring those details are available if required. In this model, the Events in XIPv6 do not store results directly. These are associated with each individual action.

The Event history of the Asset from Figure 3 - Screenshot of an Asset in a Preservica V6 system is show in Figure 4 - The Event History of an Asset in a Preservica V6 system. This shows "Characterise" events connected to each of the original "Ingest",

"Preserve" and "Create New Representation" workflows, each of which also records an event. The details of the characterization events are recorded against lower level objects.

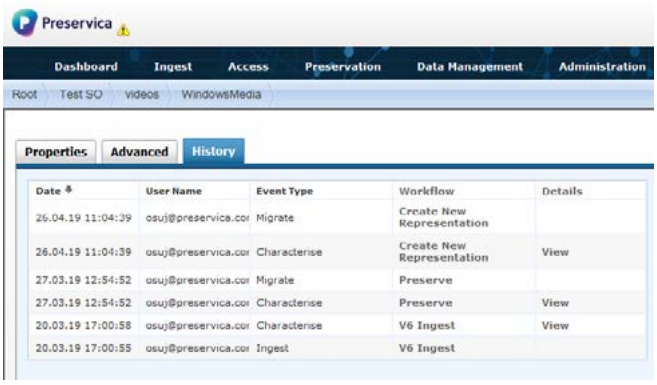


Figure 4 - The Event History of an Asset in a Preservica V6 system

Agents have not been formally implemented as an entity in the model itself, but each event action is recorded as having been performed by a specific tool, or piece of software (generally also including details of the version of the software used). In cases where actions were triggered by a specific user, details of the user are also recorded. In such a way, there is a "soft" implementation of agents as meaning the user and/or the software performing an action.

This definition of Events formalizes the way in which events and actions were recorded in XIP, which allowed for the recording of some events as part of the metadata of the specific entities on which they acted, others to be recorded in specific audit tables in the database, and others to be use the process metadata of the workflow to record them.

A UML diagram of the combined "Entity" and "Event" models is given in Figure 5.

VI. Discussion

Part of Preservica's aim is to provide a system that enables non-specialist, non-technical end users to ensure they can preserve their digital assets. To do this, the system must mask some of the complexity of the digital preservation process.

Having to balance our aim to conform to PREMIS with the need to continue operating the product with over a hundred live users, and millions of live objects amounting to hundreds of terabytes of live content, and to provide a system that masks the complexity of the digital preservation processes it provides has meant that the strict definitions of the Data Dictionary have not always quite met the exact need of this model.

As such conformance to Level 3 [9] (through internal implementation) has not been possible.

However, we have attempted to define our entities to be as close as is possible to those in the Data Dictionary, with the intention that Level 2 (through export) should be possible, and even straightforward. In this way we can provide our users with a system-independent exchange format version of their data.

The basis of our entity model has been the Intellectual Entity, albeit we have specifically sub-classed this into three distinct types (meaning no XIPv6 entity explicitly implements the PREMIS Intellectual Entity). The first, the "structural entity", to provide means of aggregation and to allow multiple levels of description; the second, the "information entity", to define the unit of information to preserve; and the third, the "content" entity to describe the base level at which data is formatted in a way that carries some long-term preservation risk.

To attempt to break a reliance on assumptions of file-level atomicity, and because conceptually to the system it doesn't matter whether a particular stream of bytes is recognized on its own terms by the underlying storage and operating system, we have effectively consolidated the PREMIS file and filestream into a single "bitstream" entity.

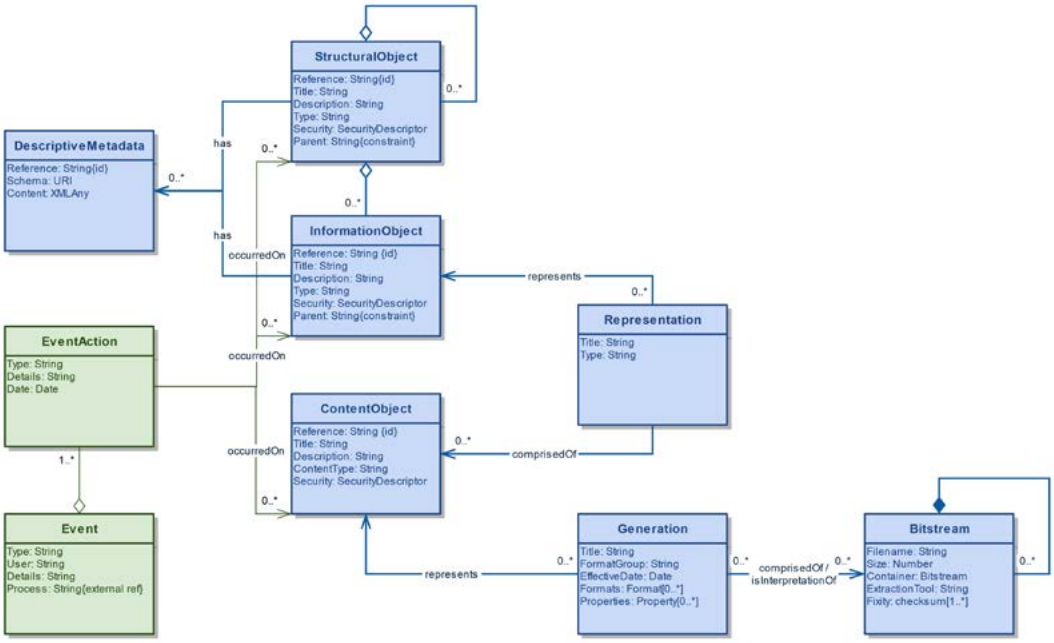


Figure 5 - UML Diagram of the Entity and Event conceptual model

We have modelled the record of specific preservation actions in the system around the PREMIS event, with a controlled vocabulary of event types, each with actions performed by specific named users or tools (i.e. agents). By doing this, we believe that our mappings and exports should conform to level B of the PREMIS conformance statement.

We have not changed our existing security model, and so whilst our users may model their own access and permissions around the PREMIS Rights model, this is not done explicitly in the system itself.

From the inception of this project, conformance to PREMIS has been both a goal and a requirement, and one which we believe we have achieved whilst delivering a data model that will enable Preservica to continue to improve its digital preservation functionality.

References

[1] PREMIS Data Dictionary v1, https://www.loc.gov/standards/premis/v1/index.html

[2] PREMIS Data Dictionary v3, https://www.loc.gov/standards/premis/v3/index.html

[3] D. Waters and J. Garrett, "Preserving Digital Information, Report of the Task Force on Archiving of Digital Information", Council on Library and Information Resources, 1996

[4] K. Van Malssen, "Digital Preservation Metadata Practice for Audio-Visual Materials", in Digital Preservation Metadata for Practitioners, Springer Press, 2016, pp 45-58

[5] ENVI Image Files, Harris Geospatial, https://www.harrisgeospatial.com/docs/enviimagefiles.html

[6] P. Lucker, C. Sijtsma, R. van Veenendaal, "Significant Significant Properties", *iPres 2018*

[7] Library of Congress Event Type controlled vocabulary, http://id.loc.gov/vocabulary/preservation/eventType.html

[8] E. Cochrane, "Implementing Event and Agent Metadata for Digital Preservation", in Digital Preservation Metadata for Practitioners, Springer Press, 2016, pp139-150

[9] P. McKinney, "Conformance With PREMIS", in Digital Preservation Metadata for Practitioners, Springer Press, 2016, pp247-257

CoreTrustSeal Certified Repositories

Enabling Findable, Accessible, Interoperable, and Reusable (FAIR) Data

Mustapha Mokrane
Data Archiving and Networked Services/CoreTrustSeal Standards & CertificationBoard
Netherlands
mustapha.mokrane@dans.knaw.nl
0000-0002-0925-7983

Jonas Recker
GESIS Data Archive for the Social Sciences/CoreTrustSeal Standards & CertificationBoard
Germany
jonas.recker@gesis.org
0000-0001-9562-3339

Abstract – The preservation of research data to enable replication and reuse is critically dependent on efficient, effective and sustainable data stewardship by the research communities. The certification of trustworthy data repositories (TDRs) custodian organizations that ensure data stewardship and long-term preservation by means of a standard such as the CoreTrustSeal is an established and recognized procedure to support long-term access to reusable data. Likewise, the FAIR Guiding Principles and the developing FAIR metrics have largely codified the contemporary discourse and policies on research data management and stewardship. The proximity of objectives between the CoreTrustSeal certification of TDRs and the implementation of FAIR Principles calls for a close examination of their overlaps and complementarities. In particular, the concept of FAIR data cannot be detached from the characteristics of the data infrastructure, the environment in which FAIR data objects reside. It is therefore necessary to examine, under which circumstances the assessment of FAIRness should be carried out at collectionor repository-level, and to what extent CoreTrustSeal certification can be considered positioning TDRs as enabling FAIR data.

Keywords – FAIR Data, Data Preservation, Trustworthy Data Repositories, Certification.

Conference Topics – Exploring New Horizons; Building Capacity, Capability and Community.

I. Introduction

Data repositories are key research infrastructures entrusted with the mission of managing research data assets and preserving their usefulness by ensuring accessibility, understandability and reusability over time. By deploying both human and technical capacities for data stewardship, repositories play a critical role in enabling reproducibility of research and data reuse for future discoveries. The mission of research data repositories is thus strongly aligned with the FAIR Guiding Principles for Scientific Data Management and Stewardship [1].

The FAIR Principles were published as a set of high level aspirational principles describing four characteristics that data assets, tools, vocabularies and infrastructures should exhibit throughout their entire lifecycle: Findability, Accessibility, Interoperability, and Reusability. However, the principles do not explicitly describe how to achieve compliance or how to measure 'FAIRness'. While some FAIR principles address characteristics which are dependent on the (technical) environment in which a data object is stored and accessed (e.g. F1 on data identifiers or A1 on communication protocols) and thus can be usefully (and sometimes only) assessed at the level of the repository, other principles require a more detailed assessment at the level of the dataset. To support the measuring of FAIRness of data objects several ongoing initiatives have begun to explore the definition of FAIR metrics [2]. Increasingly, the FAIR Principles are also recognized and used as a benchmark to develop and improve research data infrastructure for maximizing the reuse of scholarly data.

In the Turning FAIR into Reality report [3] the European Commission Expert Group on FAIR Data states that "[t]he FAIR principles focus on access

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

to the data and do not explicitly address the long-term preservation needed to ensure that this access endures. Data should be stored in a trusted and sustainable digital repository to provide reassurances about the standard of stewardship and the commitment to preserve" (p. 22). Accordingly, an important role of trustworthy repositories in this ecosystem is the provision of long-term stewardship of FAIR data objects, including curation activities to ensure that the data objects remain FAIR. This entails support and guidance for the data producers e.g. advice on which metadata standards should be used as well as for the data users. For the latter, a trustworthy repository serves as a guarantor that the data they download remain citable, accessible, and usable for the long term.

In this light, the importance of certification of repositories as trustworthy is twofold with regard to FAIR: Firstly, it can demonstrate to users that the repository enables FAIR data; secondly, certification of repositories as trustworthy may serve as a baseline for the evaluation of the FAIRness of datasets namely if we assume that a correlation exists between a repositorys trustworthiness,i.e. its demonstrated sustainability and capacity to perform data management appropriately, and its capacity to enable FAIR data.

The *Core Trustworthy Data Repositories Requirements* [4] were published as a universal catalog defining the minimum capacities research data repositories should achieve and the characteristics they must exhibit to be recognized as trustworthy. Because they were designed to assess the quality and performance of data management practices and compliance with internationally recognized and community-adopted standards, these requirements share a common objective and spirit with the FAIR Guiding Principles. The CoreTrustSeal certification based on these requirements provides a structured assessment of data repositories' trustworthiness. It is both a measure of sustainability and soundness of a data repository as an organization and a measure of its technical and technological reliability. The assessment also covers the management of digital objects in the repository and therefore sheds light on the overall quality of the data holdings. This certification scheme supported by a community effort (i.e. the CoreTrustSeal Foundation) is operational and open to data repositories worldwide. Over 140 data repositories have already been successfully certified as trustworthy by CoreTrustSeal and its precursors, the Data Seal of Approval and the World Data System Regular Members Certification [5].

It can be expected that if a data repository fulfills the CoreTrustSeal requirements, the data it holds will also meet a number of the FAIR criteria. Thus the CoreTrustSeal certification may provide a good basis to assess FAIR compliance of datasets, at least for the FAIR principles directly linked to characteristics or capacities exhibited by the repositories holding the data. It may also provide a good proxy to assess compliance with other FAIR principles. Moreover, the CoreTrustSeal Requirements address other aspects such as maintaining the understandability and reusability of datasets over time (data curation and stewardship) which are not covered by the FAIR Guiding Principles but are extremely important to maintain the FAIRness of a data object.

While there is clearly some overlap between CoreTrustSeal requirements and FAIR Principles (see section IV. below), there is not yet a full understanding of the extent to which we can assume that data held by a CoreTrustSeal-certified repository comply with the FAIR principles. In this paper we therefore explore the extent to which the CoreTrustSeal certification can serve as a baseline to assess FAIR compliance of datasets and infrastructure. For this purpose, we will look at the FAIR Guiding Principles and their assessment in particular with relation to the sustainability and long-term preservation of datasets. We will then analyze CoreTrustSeal Requirements for Trustworthy Data Repositories (TDRs), their associated certification procedure and the relationship with the FAIR Principles. Subsequently a mapping between CoreTrustSeal Requirements and the FAIR Principles will be presented to discuss where and how they overlap.1[1] Based on this we will investigate the extent to which the CoreTrustSeal certification conducted at the level of the data repository can be used to assess the implementation of FAIR Principles at the level of datasets. We will conclude with

[1] A high-level mapping of the CoreTrustSeal Requirements against the FAIR categories was done by [6]. A more granular mapping is presented by [7]. Comprehensive work towards linking FAIR and CoreTrustSeal Requirements is done by the WDS/RDA Assessment of Data Fitness for Use WG (see below).

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

considerations on the review of the CoreTrustSeal Requirements and how this process can incorporate the outcomes of relevant FAIR initiatives.

II. FAIR Guiding Principles and their assessment

The FAIR Guiding Principles for scientific data management and stewardship define the characteristics that data resources on the one hand, and the tools, vocabularies, and infrastructures used for data management and sharing on the other should exhibit to assist discovery and reuse by third parties. As they were by design defined at a high level, the principles do not include an implementation framework nor an approach to assess datasets FAIRness. The principles have a datacentric approach and the main focus in the current discussions about FAIR metrics and FAIR assessment is on the data objects without sufficient attention given to the characteristics of the environment in which the data are held, in particular data repositories. Thus, several initiatives have begun to develop tools to assess the FAIRness of datasets.

The World Data System (WDS)/Research Data Alliance (RDA) Assessment of Data Fitness for Use WG created a checklist for the Evaluation of Dataset Fitness for Use whose categories of dataset fitness (Metadata completeness, Accessibility, Data completeness and correctness, Findability & interoperability, Curation) are mapped to the FAIR principles and which is "meant to supplement the CoreTrustSeal Repository Certification process"2[1].

The RDA FAIR Data Maturity Model WG develops "a common set of core assessment criteria for FAIRness and a generic and expandable self-assessment model for measuring the maturity level of a dataset". The WG seeks to build on existing approaches and to identify common elements in these approaches. The initial step in this work was to classify the different approaches currently taken, leading to a landscape overview of FAIR assessment tools and approaches.3[2] The overview illustrates the different approaches currently taken to the question of FAIR assessments. These differ, for example, with regard to the subject of assessment (e.g. datasets, data management plans, repositories), the evaluating entity (e.g. researchers, repository managers), and the mode of assessment (manually or automatically).

Table 1: The FAIR Guiding Principles. Source: [1]

	To be Findable:
F1	(meta)data are assigned a glob-ally unique and persistent iden-tifier
F2	data are described with rich metadata (defined by R1 below)
F3.	metadata clearly and explicitly in-clude the identifier of the data it describes
F3.	(meta)data are registered or in-dexed in a searchable resource To be Accessible:
A1.	(meta)data are retrievable by their identifier using a standard-ized communications protocol
A1.1.	(meta)data are retrievable by their identifier using a standard-ized communications protocol
A1.2	the protocol allows for an au-thentication and authorization procedure, where necessary
A2.	metadata are accessible, even when the data are no longer available To be Interoperable:
I1.	(meta)data use a formal, accessi-ble, shared, and broadly applica-ble language for knowledge rep-resentation
I2.	(meta)data use vocabularies that follow FAIR principles
I3.	(meta)data use vocabularies that follow FAIR principles To be Reusable:
R1.	(meta)data are richly described with a plurality of accurate and relevant attributes
R1.1.	(meta)data are released with a clear and accessible data usage license
R1.2.	(meta)data are associated with detailed provenance
R1.3.	(meta)data meet domain-relevant community standards

[1] WDS/RDA Assessment of Data Fitness for Use WG Outputs and Recommendations: https://doi.org/10.15497/rda00034

[2] The WGs results page (https://doi.org/10.15497/RDA00035 provides a link to the GitHub repository with more information on the landscape overview.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The FAIRmetrics Group seeks to create a set of univer-sal FAIR metrics valid across all scholarly domains and for all digital object types, measuring how FAIR a dataset is "for machines i.e., the degree to which a digital resource is findable, accessible, interoperable, and reusable with-out human intervention" [8]. The group explicitly states that these generic metrics need to be supplemented by "domain-specific or community-specific metrics" [8].

The work begun in these and other initiatives has to be continued to define an agreed core set of metrics as well as to come to an agreement about community specific metrics suitable to measure the FAIRness of specific data types.

Characteristics that determine the FAIRness of a data object and which thus become the focus of an assessment can be intrinsic or extrinsic to the object, i.e. they can either be an integral part of the object, or the object derives these characteristics from the infrastructure in which it resides. For example, to check compliance with the FAIR Principle F1 (see Table 1) we can either look at the data object itself to see if there is a unique and resolvable PID attached to it (e.g. as part of the metadata). However, it can also be verified globally for an entire class or collection of data objects by checking if the infrastructure holding the dataset assigns PIDs to its data assets. By contrast, FAIR Principle F4 refers to a characteristic extrinsic to the data object. It can only be verified by ascertaining that the infrastructure holding the data object implements such a registry or index. It follows that for some of the principles the FAIRness of a data object can be assessed at repository-level by looking at the policies and standards employed by the infrastructure holding the data object.

Assuming that the FAIRness of a data object can be assessed based on the FAIR Guiding Principles, if the data object meets these principles it is assigned a high score at the time of assessment. However, as Mari Kleemola points out, "[r]esearch data will not become nor stay FAIR by magic. We need skilled people, transparent pro-cesses, interoperable technologies and collaboration to build, operate and maintain research data infrastruc-tures" [9]. That said, the current FAIR Principles neither cover data stewardship activities such as curation and long-term preservation nor strategies and procedures to promote the sustainability of the data repository, all of which ensure that the data objects remain FAIR over time. Consequently, the FAIRness score of a data object could decay over time and should be time stamped and updated regularly. An example is a data object to which a Digital Object Identifier (DOI) a persistent, unique and resolvable identifier (PID) is assigned, resulting in a pos-itive assessment for FAIR Principle F1 (see Table 1). If the data object is not managed and preserved in a TDR, the odds that the DOI no longer resolves to the data object are high because the persistence of the DOI depends entirely on the data custodian, i.e the data repository in most cases, to update the URL for the landing page. Hence global services for PID minting and resolution on their own can only enable persistence, but not guaran-tee it. Similarly, a data object that meets community and domain approved standards today because it uses a preferred file format may be assessed as FAIR regarding Principle R1.3 (see Table 1). Yet if the data object is not pre-served in a TDR which ensures file formats migration in adherence with the needs of its designated community, it could become unreadable in the future, e.g. because the file format becomes obsolete or is deprecated in the research community.

Accordingly, a FAIR assessment that considers only the data object at a given point in time but which does not take into account the characteristics of the infrastructure in which the object is stored is not sufficient to predict whether a data object will remain accessible and usable over time.

III. CoreTrustSeal requirements and certification

The examples provided earlier illustrate the risks of limiting the assessment of FAIRness to characteristics of data objects thus highlighting the importance of including an assessment of the environment in which the objects reside. In particular this concerns the quality and trustworthiness of the data repositories providing the key infrastructure for the dissemination and preservation of scholarly data.

The Core Trustworthy Data Repositories Requirements [4] define universal and essential ("core") characteristics of trustworthy data repositories. The CoreTrustSeal Requirements are the result

of integrating and improving two predecessor catalogs of criteria the Data Seal of Approval (DSA) and the World Data System (WDS) Regular Members certifications, already used by many domain repositories in the natural and social sciences and humanities across the globe into a set of universal requirements that can be applied to research data repositories generally. The CoreTrustSeal Requirements were developed by a DSA and WDS Partnership Working Group established under the umbrella of the Research Data Alliance[1]. In addition to unifying DSA and WDS catalogs of criteria, CoreTrustSeal requirements are aligned with the concepts defined in the Reference Model for an Open Archival Information System (OAIS), an international standard for data repositories also known as ISO 14721:2012 [10]. The requirements are also mapped to ISO 16363:2012, the international standard for Audit and Certification of Trustworthy Digital Repositories [11] as well as the German nestor Seal (based on DIN 31644) [12]. This facilitates the transition from a CoreTrustSeal certification to a DIN or ISO certification as agreed in the Memorandum of Understanding to create a European Framework for Audit and Certification of Repositories.5[2]

In the CoreTrustSeal framework, the trustworthiness of a data repository is assessed through a formal certification process which starts with the submission of a self-assessment against the 16 CoreTrustSeal requirements via the Application Management Tool. This selfassessment is then peer-reviewed by two independent experts to verify that the repository meets the requirements and that there is sufficient public evidence supporting the claims made in the self-assessment. In the case of missing evidence or open questions the assessment is returned to the applicant with comments in an iterative process. A successful review results in the award of the CoreTrustSeal by the Board, signaling that the repository can be considered as trustworthy for a period of three years.

As indicated, the CoreTrustSeal certification is conducted at the repository level and the requirements are organized in three main categories addressing the context, structure and activities of a data repository in alignment with ISO 16363. In their self-assessment repositories provide evidence that

1. the organizational infrastructure is sound to ensure sustainability: this includes requirements on the mission and scope, licenses, continuity of access, confidentiality and ethics, funding, and expert guidance.
2. Digital objects management is performed according to standards to ensure understandability and reusability for the long term of datasets by the designated community: this includes requirements on data integrity and authenticity, appraisal, documented storage procedures, preservation plan, data quality, workflows, data discovery and identification, and data reuse.
3. Technical infrastructure and security measures are adequate to protect the data against loss and unauthorized and/or undocumented manipulation.

The CoreTrustSeal requirements were also heavily influenced by the discussions in the data management and data sharing communities, including the emerging consensus and momentum surrounding the FAIR Guiding Principles, which they consequently incorporate although with a different focus and slightly different terminology.

IV. CoreTrustSeal alignment with FAIR

As discussed, the FAIR Guiding Principles follow a data and metadata-centric approach with a focus on data discovery, reuse and machine readability, whereas the CoreTrustSeal Requirements are formulated following an infrastructure-centric perspective which incorporates the aspects addressed in the FAIR Guiding Principles but shifts the focus towards data preservation and organizational sustainability. In addition, unlike current approaches to determining FAIR metrics, the CoreTrustSeal criteria are not designed as a checklist of mandatory requirements that repositories and reviewers "tick off" to determine a repositorys trustworthiness. While the accompanying guidance and extended guidance contain hints and suggestions as to what kind

[1] https://rd-alliance.org/groups/repository-audit-and-certification-dsa%E2%80%93wds-partnership-wg.html

[2] http://www.trusteddigitalrepository.eu

of technical implementations and evidence applicants are expected to provide, these largely do not lend themselves to semi-automated assessment. Rather, reviewers and the CoreTrustSeal Board consider whether the evidence provided sufficiently demonstrates that the repository can be considered trustworthy in relation to its goals (e.g. the level of curation offered) and the context in which it operates (for example, the needs of the designated community for which the data are preserved): "Reviewers are looking for clear, open statements of evidence specific to the applicant. Not necessarily all bullet points in all requirements are mandatory; final judgment depends on the completeness and quality of the answer in the self-assessment of a specific Requirement" [13](p. 5).

These differences in approach also explain to a certain extent the different terminology between the CoreTrustSeal Requirements and the FAIR Principles and more importantly why they do not map in a one to one relationship.

A mapping between the FAIR Guiding Principles and the CoreTrustSeal Requirements is presented in Fig. 1 and will be discussed in more detail in the following sections.

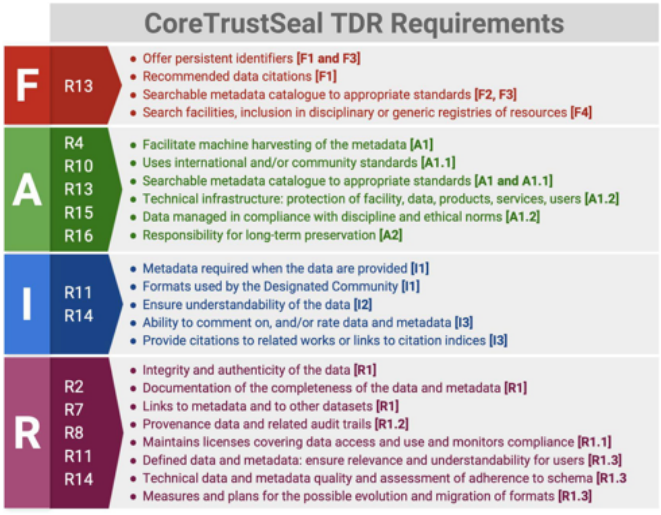


Figure 1: Mapping between FAIR and CoreTrustSeal

A. *Findability*

The FAIR Findability principle has four sub-principles (FAIR-F1 toF4; see Table 1) covering persistent identifiers, rich metadata including a reference to the identifier, and registration in a searchable resource. In the CoreTrustSeal (CTS) Requirements, this is mostly covered by CTS-R13 requesting evidence that the repository "enables users to discover the data and refer to them in a persistent way" [4]. The additional guidance provided to data repositories for this requirement makes it clear that data discovery is key to data sharing, and that datasets should be citable including with persistent identifiers to ensure that data can be accessed into the future. The CoreTrustSeal reviewers will particularly look for evidence that the repository offers search facilities, which covers FAIR-F4, and maintains a searchable metadata catalog to appropriate (internationally agreed) standards addressing FAIR-F2 andF3. The reviewers will confirm if the catalogue is registered in one or more disciplinary or generic registries of resources (FAIR-F4), offers recommended data citations (FAIR-F1) and makes use of persistent identifiers (FAIR-F1 andF3).

It is worth noting that FAIR-F2 explicitly requires rich metadata describing the data (as defined in FAIR-R1: "(meta)data are richly described with a plurality of accurate and relevant attributes"). The CoreTrustSeal Requirements focus on the availability of metadata for discovery and the use of domain-specific and international standards without going (much like the FAIR Principles) into the details of necessary attributes. Such attributes to enrich the (meta)data which include controlled vocabularies and ontologies are not necessarily included in discovery or domain-specific standards. What exactly constitutes "rich" metadata requires further discussion in the FAIR context as well [2], in particular as "richness" of metadata will mean different things in different scientific communities. Thus, while the FAIR-Findability principles can be assessed at repository level and would therefore lend themselves to using CoreTrustSeal certification as a baseline for measuring FAIRness, further discussion is required to determine what "richness" means in different contexts and disciplines and how to measure it.

B. *Accessibility*

The FAIR Accessibility principles prescribe the use of standardized, open, free, and universally implementable communication protocols allowing authentication and authorization where necessary to retrieve data and metadata. They also require that metadata remain accessible even when the data

are no longer available. At the repository level these principles are either explicitly or implicitly covered by several CoreTrustSeal Requirements.

The guidance for CTS-R13 explicitly covers FAIR-A1 by indicating for instance that the repository should facilitate machine harvesting of the metadata. The CoreTrustSeal guidance also mentions that the repository should maintain a searchable metadata catalogue to appropriate (internationally agreed) standards, which implicitly covers FAIR-A1.1. Similarly, CTS-R15 addresses the technical infrastructure of data repositories to ensure that it is appropriate and in particular that the standards used are relevant to their designated community of users. Although free, open and universally implementable communication protocols are not mentioned explicitly, they are implicitly required to implement a searchable and machine-harvestable metadata catalogue. Thus HTTP/HTTPS and the Open Archives Initiative Protocol for Metadata Harvesting OAI-PMH [14] are among the international standards widely used by data repositories.

For CTS-R4 (and CTS-R9 to some extent) reviewers will check that the repositorys data in particular personal data with a disclosure risk are created, curated, accessed, and used in accordance with disciplinary and ethical norms. Evidence must include availability of human expertise and technical capacities, for example for anonymization and secure access. Similarly, CTS-R16 stipulates that the technical infrastructure of the data repository must provide for protection of the facility and its data, products, services, and users. Both of these requirements thus mirror FAIR-A1.2.

FAIR-A2 requires that metadata should remain accessible, even when the data are no longer available. This is an area where CoreTrustSeal-certified repositories excel by definition because they commit to preserving data and metadata for the long term. CTS-R10 addresses the responsibility for long-term preservation and reviewers will look for evidence that the repository manages this function well. As a consequence data in general and metadata in particular can be expected to continue to be accessible in the case of TDRs (within the boundaries of the data retention policies the TDR set itself).

C. *Interoperability and Reusability*

Reference [3] stipulates that Interoperability and Reusability depend on the FAIR Digital Objects being "represented in common and ideally open formats, and [being] richly documented using metadata standards and vocabularies adopted by the related research community" (p. 12). With regard to interoperability, in many disciplines the necessary frameworks already exist. However, due to the increasingly interdisciplinary nature of research, "attention needs to be paid to the extremely challenging task of developing FAIR data frameworks across disciplines and for interdisciplinary research" [3] (p. 11).

The same is true for Reusability, which strongly depends on the use of communityagreed file formats and software as well as on the description of the data objects with standardized metadata and documentation. While community-specific standards and agreements exist in this regard, which formats and metadata should be used for a given class of data objects to facilitate crossdisciplinary reuse depends on the context and thus has to be decided with regard to specific use scenarios.

It follows that the assessment of both Interoperability and Reusability is impossible without taking into account the purpose for which a given community seeks to use a data object. It does not seem feasible to assess this at the level of individual objects but on the level of collections and repository-level, making the CoreTrustSeal certification a potential tool to support this assessment.

Two CoreTrustSeal Requirements particularly relevant to Interoperability and Reusability CTS-R8, "Appraisal,"R11 "Data quality," andR14 "Data reuse" will always be assessed in relation to the repository's scope, preservation goals, and the needs of the designated community.[1] Reviewers will particularly focus on the question of if and how the repository seeking certification ensures that the data objects deposited can be rendered by and are understandable to the intended user community, and that all metadata deemed necessary for this purpose are of sufficient quality.

[1] Further relevant CoreTrustSeal requirements include CTS-R2 "Licenses," mapping to FAIR-R1.1 and CTS-R7 "Authenticity and Integrity" mapping to FAIR-R1.2

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

V. Discussion

Datasets in a TDR meeting the CoreTrustSeal Requirements are managed, curated, and archived in such a way that they are useful and meaningful (FAIR enough) for the repository's designated community and remain so in the future.

We therefore assert that CoreTrustSeal certification of data repositories facilitates the FAIRness assessment of data objects by providing proxy information to evaluate compliance with many FAIR Principles. Certified trustworthy repositories enable a baseline FAIRness level to the datasets they hold and contribute to maintain or even increase the level of FAIRness over time through appropriate data curation and stewardship services.

An automated assessment of the FAIR Findability and Accessibility Principles can rely on machine-readable metadata and testable data services. In contrast, an assessment of Interoperability and Reusability (for human users) is more difficult as it requires domain expertise to evaluate for example conformance with community standards or the completeness of data and metadata content. For sensitive data in particular, automatic assessment of anonymization is hardly possible which means that a data curator will always be required. In these cases a possible FAIR assessment procedure cannot rely on comparably simple metrics that could be assessed semi-automatically; instead, the evaluation of an infrastructures ingest and quality control procedures, for example, has to take into account a complex set of community-specific conditions that does not lend itself to automated assessment easily.

CoreTrustSeal TDR certification addresses the FAIR Interoperability and Reusability requirements at the level of the repository for example by ensuring that sufficient levels of data curation are applied and that procedures for checking the quality of the data and metadata are in place, in accordance with the needs of the repositorys designated community. Therefore, a FAIR assessment for data objects could usefully build upon the certification status of the repository holding the object to make assumptions on its interoperability and reusability.

FAIR Guiding Principles are still being assimilated in the various research communities and their implementation will affect data infrastructures at large. Like many other certification processes, the CoreTrustSeal Requirements are reviewed regularly to incorporate feedback received from certified data repositories and to account for the evolution of practices of the data community. As part of this evolution process the CoreTrustSeal Board has to reflect on how to incorporate references to FAIR Principles as well as to FAIR-enabling standards and technologies (e.g. ontologies and controlled vocabularies) and their implementation in the (extended) CoreTrustSeal guidance.

In March 2019, the CoreTrustSeal Board initiated an open review of the CoreTrustSeal Requirements to define the requirements for the period 20202022.7[1] The Board anticipates a certain stability of the requirements, yet it makes a commitment to consider the requirements in the light of FAIR Principles implementation.

The CoreTrustSeal Requirements will also gain from the work and outputs of many initiatives worldwide aim contribute to the adoption FAIR Principles in practice and will cater mainly to the European Open Science Cloud project which brings together European research communities, infrastructure providers and practitioners.[2] It is expected that the outcomes of the FAIRsFAIR project will be directly relevant to the CoreTrustSeal Requirements and will most certainly be considered in the next scheduled review of CoreTrustSeal Requirements.

The CoreTrustSeal, unlike other certification frameworks, emerged directly out of the community of research data repositories. As a "core" certification it provides an entry-level procedure to help data repositories continuously improve and demonstrate their trustworthiness. To be able to continue fulfilling this role, the CoreTrustSeal Board considers it an important task to take into account the current developments around FAIR: to ensure that data repositories are recognized by researchers,

[1] https://www.coretrustseal.org/why-certification/review-of-requirements//

[2] https://www.fairsfair.eu/, https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

publishers, and funders as both trustworthy and as enabling FAIR data, now and in the future.

Acknowledgment

The authors would like to acknowledge the contributions to this paper from Dr Peter Doorn, Dr Eliane Fankhauser as well as members of the CoreTrustSeal Standards and Certification Board.

References

[1] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, vol. 3, no. 1, Dec. 2016,ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18.

[2] M. D. Wilkinson, S.-A. Sansone, E. Schultes, P. Doorn, L. O. Bonino da Silva Santos, and M. Dumontier, "A design framework and exemplar metrics for FAIRness," Scientific Data, vol. 5, p. 180 118, Jun. 2018, ISSN: 2052-4463. DOI: 10.1038/sdata.2018.118.

[3] European Commission Expert Group on FAIR Data, Turning FAIR into reality: Final report and action plan from the European Commission Expert Group on FAIR data, European Commission, Ed. 2018. DOI: 10.2777/1524.

[4] Rorie Edmunds, Hervé L'Hours, Lesley Rickards, Paul Trilsbeek, Mary Vardigan, and Mustapha Mokrane, Core Trustworthy Data Repositories Requirements. 2016. DOI: 10.5281/zenodo.168411.

[5] CoreTrustSeal, CoreTrustSeal Certified Reposito-ries. [Online]. Available: https://www.coretrustseal.org/why-certification/certified-repositories/.

[6] N. Harrower, How the CoreTrustSeal (CTS) Enables FAIR Data. 2019. [Online]. Available: https://www.slideshare.net/dri_ireland/how-the-core-trust-seal-cts-enables-fair-data.

[7] M. White, L. Winfree, Mehndiratta, Payal, Robasky, Kimberly, and Champieux, Robin, "Modeling the FAIR Rubrics Landscape (Poster)," Coral Gables, FL, 2019. [Online]. Available: https://osf.io/rthgx/.

[8] M. Wilkinson, E. Schultes, Luiz Olavo Bonino, Susanna-Assunta Sansone, P. Doorn, and M. Dumontier, Fairmetrics/Metrics: Fair Metrics, Evaluation Results, And Initial Release Of Automated Evaluator Code, en. Zenodo, Jul. 2018. DOI: 10.5281/ zenodo.1305060.

[9] M. Kleemola, Being trustworthy and FAIR requires people, processes, technologies and collaboration, ser. Tietarkisto Blogi. Nov. 2018. [Online]. Available: https://tietoarkistoblogi.blogspot.com/2018/11/being-trustworthy-and-fair.html.

[10] CCSDS, Reference Model for an Open Archival Information System (OAIS). Jun. 2012. [Online]. Available: https://public.ccsds.org/Pubs/650x0m2.pdf.

[11] International Organization for Standardization, Space data and information transfer systems Audit and certification of trustworthy digital repositories. Feb. 2012. [Online]. Available: https://www.iso.org/standard/56510.html.

[12] nestor Working Group Certification, nestor Seal for Trustworthy Digital Archives. 2013. [Online]. Available: https://www.langzeitarchivierung.de/Webs/nestor/EN/Services/nestor_Siegel/nestor_siegel_node.html

[13] CoreTrustSeal, Core Trustworthy Data Repositories Extended Guidance: Core Trustworthy Data Repositories Requirements for 20172019 - Extended Guidance (Version 1.1). Jun. 2018. [Online]. Available: https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf.

[14] The Open Archives Initiative Protocol for Metadata Harvesting: Protocol Version 2.0 of 2002-06-14. 2015. [Online]. Available: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm.

Setting Up Open Access Repositories

Challenges and Lessons from Palestine

Rawia Awadallah
IUG
Palestine rradi@iugaza.edu.ps
0000-0001-6317-9747

Iyad ALAgha
IUG
Palestine ialagha@iugaza.edu.ps
0000-0002-7516-4902

Tomasz Miksa, Andreas Rauber
TU Wien Austria
miksa@ifs.tuwien.ac.at
0000-0002-4929-7875

Joy Davidson
UoG UK
Joy.Davidson@glasgow.ac.uk
0000-0003-3484-7675

Adnan Yahya
BZU
*Palestine yahya@@*birzeit.edu
0000-0002-8769-1961

Anna Maria Tammaro, Stefano Caselli
University of Parma Italy
annamaria.tammaro@unipr.it
0000-0002-9205-2435

Janet Anderson, David Anderson
UoB UK
J.Delve@brighton.ac.uk
0000-0003-2673-4830

Nael Salman
PTUK
Palestine N.Salman@ptuk.edu.ps
0000-0002-5190-3889

Yousef AbuZir
QOU
Palestine yabuzir@qou.edu
0000-0002-1220-1411

Abstract – Research outputs produced by developing countries lack visibility. Common reasons are high costs of publishing, reluctance to share, as well as lack of journals focusing on research challenges specific to developing countries. The ongoing change towards open access and the rising popularity of institutional repositories allows for bridging the gap to developed countries. However, this still requires cultural, organizational and technical changes. In this paper, we describe a holistic approach for deploying open access repositories and building research data management services and increasing data curation skills. We describe how we identified users' needs and necessary supporting systems and services. We also explain the rationale and challenges faced when implementing popular repository systems and share experiences in developing institutional data management policies. Finally, we provide common goals for a national roadmap. All these actions are the first step towards the preservation of both research outputs and cultural assets. The paper is based on our experiences from ROMOR project that helped in establishing policies, systems, and organizational workflows at four Palestinian universities.

Keywords – Open Access, Institutional Repositories, RDM, Policy, Developing Countries.

Conference topics – 1. Collaboration: a Necessity, an Opportunity or a Luxury? 4. Building Capacity, Capability and Community

I. Introduction

In the Digital Age, more and more people are accessing and contributing digital content which has the potential to improve people's lives by making information better available, increasing opportunities for political and economic engagement, and making government more transparent

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

and responsive. However, global participation in the Digital Age still remains uneven for many and thus can lead to a growing knowledge divide among research communities or whole countries.

Research outputs of Palestine researchers are often essential for the development of appropriate programs to solve local problems, e.g. estimating underground water reserves in Gaza strip. Their research is often essential for policymakers and development practitioners to identify the exact needs of the local community and to optimize their policies and investments according to these needs.

However, developing countries such as Palestine are at low rank in terms of research output and its visibility (same applies for digital preservation of cultural heritage,

e.g. digitization of rare manuscripts). This not only can be explained by the low budget allocated to research and the fluctuating political climate but also by lack of proper practices, policies, infrastructure and culture of sharing research outputs.

To close the gap in sharing and access to scientific data and publications, developing countries need to participate in developing knowledge exchange anisms and systems that benefit them by making their outputs Findable, Accessible, Interoperable and Reusable (FAIR)[1]; universities should encourage publishing through cooperative, peer-reviewed open access platforms; and the governments and development agencies should invest in developing more effective knowledge sharing systems and digital repositories.

mech-Table 1: OAIRs Developed within ROMOR

IUGSpace:	iugspace.iugaza.edu.ps
FADA:	fada.birzeit.edu
OSOL:	dspace.qou.edu
PTUK Repository:	scholar.ptuk.edu.ps

This paper aims to share the experience of ROMOR project in building and improving Open Access Institutional Repositories (OAIRs) of four universities in Palestine (see Table 1. In particular, the paper describes the development of OAIRs, and their planning, implementation/improvement, and evaluation. These include the repository, scope and coverage, metadata standards and interoperability, policies, and business models. These repositories include not only staff publications and theses, but also hundreds of scanned rare books that are freely available to the public[1], and will soon include environmental research datasets including geological, hydrological, and meteorological datasets. The paper also identifies next steps needed for introducing a national roadmap to include further universities and provide common services.

Despite the focus on Palestine, this paper can be seen as a blueprint for other developing countries that share similar challenges and consider establishing OAIRs.

The paper is organized as follows. Section 2 describes related work on open access and research visibility in Palestine. Section 3 presents requirements identification and training material development that laid the foundation for OAIRs development. Section 4 deals with business modeling and preparation of OAIR deployment plans. Section 5 presents the process of installing OAIRs, challenges, and lessons learned. Section 6 focuses on data management policy development. Section 7 presents goals for a national roadmap. Conclusion and future work are provided in Section 8.

II. Related work

A. *The Visibility of Research in Palestine*

Developing countries such as Middle Eastern Arab countries are at low rank in terms of research output. However, these countries have witnessed a considerable growth in research size and impact in the last decade due to the growing investments in research ([2]; [3]).

Palestine is a developing country located at the south- eastern side of the Mediterranean coast. It is only since the 1970s that universities came into existence in Palestine. According to the latest publications and statistics of the Palestinian Ministry of Higher Education (MOHE), there are 16 universities. Most of the universities started as 2-year College institutions and then were developed to provide university-level

[1] fada.birzeit.edu/handle/20.500.11889/2836

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

education. Currently, many of these universities are involved in research and offer graduate programs in various scientific disciplines including engineering, health, basic science, economics, humanities, and other fields [4].

Despite the affordance of research activities in Palestine, access to the results of implemented research has been severely limited due to publishing and access inabilities ([5]; [6]; [7]). Because of the common trend among scholars to publish their research findings in highly reputed journals, most of the Palestinian scholarship is either unpublished or delayed. In addition, the traditional system of scholarly publishing and the high costs of publications have contributed to the limited growth of the Palestinian scholarship ([8]). Apart from the research, the visibility of research outcomes is also kept to the minimum, due to financial restrictions limiting the publication and distribution of national research outcomes. In addition, information important for the resolution of problems specific to developing countries is not often published in journals from the developed world.

As a result, the research outcomes, which most of is imperative to address local and regional developmental issues, die at the institutional level as they remain invisible to those who may need them. Although some results may eventually get published in local journals, the outcomes may not be widely disseminated due to the poor distribution and recognition of these journals. This disappointment comes despite the so much commitment of efforts and resources that may be devoted to undertaking research in Palestine.

B. Open Access in Palestine

Little attention has been paid to the development of Open Access Institutional Repositories in the Arab academic institutions. OAIRs can be defined as digital archives of intellectual outputs created by faculty, research staff and students of an institution, and accessible to end-users both within and outside the institution, with few if any barriers to access ([9]; [10]). One of the main benefits of OAIRs is to maximize the availability, accessibility, discoverability, and functionality of scholarly outputs at no cost to the user ([11]; [12]). In addition, OAIRs often provide technologies and methods to support the preservation of research outputs and ensure that they remain accessible and reusable over time ([13]; [14]).

A number of academic institutions in Arab countries have taken the initiative to build OAIRs to foster the dissemination of their research outputs ([15]; [16]). Few studies have tried to track and assess the progress of institutional repositories in the Arab world ('[17]; [16]; [18]). All these studies agreed that OAIR developments in the Arab region are still at early stages. Carlson [17] assessed a sample of repositories from the Arab world in terms of accessibility and transparency and reported that they were notably lacking in explicit policies, metadata, and preservation. Studies also reported that repositories have not been growing consistently due to the lack of mandating policies or the lack of awareness of OAIR benefits.

There is a positive response toward OA in Palestine. This is indispensable in order to raise the profile of research performed in the Palestinian higher education institution and integrate it into the international knowledge pool. As of January 2019, there are 23 OA journals indexed in the Directory of Open Access Scholarly Resources. The Open Access initiative has further strengthened in Palestine through embarking digital projects funded by international entities. Until recently, only two OAIRs from Palestine were oZcially registered with the Registry of Open Access Repositories[2]) or the Open Directory of Open Access Repositories (OpenDOAR[3]).

C. ROMOR Project

One recent, ongoing and important initiative for OA in Palestine is through a project titled "Research Output Management through Open Access Institutional Repositories in Palestinian Higher Education" ROMOR. ROMOR was kicked off on January 2017 by four Palestinian universities and four EU universities with EU funding from Erasmus+. ROMOR aims to improve the management, visibility, and accessibility of scientific research outputs in Palestinian HEIs by establishing new or enhancing existing OAIRs, improving institutional capacity for

[2] http://roar.eprints.org

[3] http://v2.sherpa.ac.uk/opendoar

the management and sharing of research outputs, and developing and/or refining curricula to ensure that emerging researchers are better able to manage their work across the entire research lifecycle.

ROMOR is the first project that seeks to build capacity for developing OAIRs at Palestinian universities. Without the input of knowledge from research that tackles national challenges and priorities, the development stakeholders and policymakers in Palestine and beyond may not be able to act effectively, and development initiatives may suffer from inappropriate programs. Considering the mobility restrictions between the West Bank and Gaza, the OAIRs is providing a unique opportunity for cross-border coordination and exchange of research findings between Palestinian universities, thus leading to more complementary and less duplicated research. In addition, the project is innovative in that it embraces a variety of capacity building activities that include: conducting needs assessments, holding workshops, building OAIRs for both publications and research data, providing vocational and academic teaching, implementing policies and exploring sustainable solutions for digital curation and preservation. The overall activities aim to bridge theory with practice. The Project seeks to build OAIRs not only by using existing solutions but also by exploring innovative solutions that are tailored to the local needs such as the need to handle the Arabic content.

III. Identifying Needs and Developing Training Material

To help scope the OAIR service development and/or refinement, user requirements were gathered at the start of the ROMOR project through two surveys targeted to both researchers and support staff. The first [19] aimed to assess researchers' current practices, and the second explored institutional support staff capacity. The four participating institutions include: Islamic University of Gaza, Al-Quds Open University, Birzeit University, and Palestine Technical University-Khadoori.

We used the DCC[4] research data management service model [21] which helps break down RDM infrastructure into specific activities to assist with planning and implementation. 'Soft' infrastructure aspects, including policies, business planning, and training, underpin the more technical infrastructure requirements in the center which is based around the data lifecycle. It is important to note here that this is an idealized view of an RDM service model. In reality, most institutions will have some services in place as well as many gaps.

The gaps were identified through analysis of the ROMOR academic staff and managerial staff surveys. The findings are mapped to the DCC's RDM Service model and results are compared to the RDM landscape in the UK. Overall, results revealed the lack or complete absence of RDM policies at the institution's level. In addition, there is a lack of business and sustainability plans for data repositories and RDM services. As a result, adopted data archiving procedures were mostly immature and not consistent with known good RDM practices. Most researchers still have to store and manage their research outputs on their own, with limited support provided by the institution. Results also showed a lack of awareness of the potential of OA publishing and OAIRs to promote the visibility of research outputs.

These results were used to make decisions on the training structure and the educational material to be produced. The areas of competencies related to Data Management and Data Engineering, developed by the EDISON Project [22], were matched to the objectives/needs of the ROMOR Project. The EDISON competencies of these two areas have been compared with the educational and training needs of the Palestinian Universities. ROMOR partners have agreed to employ the DCC's RDM service model to structure the training including an emphasis on softer infrastructure aspects including policies, business planning and training; and also on more technical infrastructure requirements which are based around the data lifecycle.

The training included each of the model components, considering aspects of EU partners' good practices along with Palestinian universities specific needs and expectations. Following the DCC's RDM service model, three areas of training and education are highlighted in the ROMOR RDM process in Palestinian universities:

[4] http://www.dcc.ac.uk

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

1. *Data Governance:* a set of activities including strategic planning, supervision, and enforcement that governs the process and methodologies that are carried out to ensure and improve the quality of RDM, including organizational structure and business planning, policy, and advocacy awareness to data stakeholders.
2. *Data Management:* consists of data selection and handover, data storage in repositories, data catalogues, data sharing, data reuse, and data appreciation.
3. *Data literacy:* embedded in R&D data flow and related to research data life-cycle. R&D management and data utilization are main perspectives in data literacy; practical skills such as data analysis, data description (metadata and methods) and tools are the main focus in data literacy.

IV. Developing Business Models

Once a general picture of user needs was identified, the Palestinian partners began to scope out the OAIR services that would be delivered at each institution. A series of practical exercises were run as part of dedicated training workshops and continued at each Palestinian institution. Each training workshop aimed to build upon the work carried out during the previous session.

In the first workshop participants were introduced to the Business Model Canvas (BMC)[5]. The BMC helps an organization to create a business model by considering among others: what value the service will deliver to customers, how to maintain a relationship with customers, what are the key activities and operational units involved in delivering the service, etc. The key value of the BMC approach is that it allows a group of stakeholders to work collaboratively to develop a high-level view of the business model for a given service. The graphical presentation of the BMC makes it easier to identify any gaps or disconnects that may hinder the value proposition(s) being realized.

Some Palestinian partners were just starting out with scoping RDM services while others are further down the line and are currently rolling out pilot services or are delivering more established RDM services. Through the short exercises, each group developed a BMC for their RDM service. The BMCs were completed collaboratively within working groups involving main stakeholders to allow different viewpoints to be captured. It was agreed on that in the Palestinian context, the structural funding of the institution may be the best option for sustainability. Resource sharing or even shared operation and shifting the data entry work burden to material producers and integrating the effort into the management process may work well to reduce the costs to sustainable levels. What makes such an approach reasonable is that the archiving volumes are quite low and are most likely to remain so for several more years. Potential funding bodies (e.g. MoHE) is supportive of structuring the costs of archiving into the research costs that can be supported. We also believe that the pioneering institutions, in terms of OAIR development and adoption, can offer services to latecomers that can help offset some of the costs. This includes: (1) Consultancy services both technical, policy drafting and needs assessment. Such services can be offered by Library, IT and maybe legal staff. (2) Development services by offering a running system that suits the needs of the client institution against a fee, possibly with maintenance and training provisions. (3) Hosting services for the OAIR of institutions who find it too costly to run their independent OAIR facilities.

The Palestinian partners started their BMC during the first workshop and then further refined it with input from relevant stakeholders back at their own institutions.Below we outline high-level recommendations for developing a BMC, highlighting the main recommendations provided by EU partners on each of the Palestinian partners' BMCs:

- Create an inventory of your current infrastructure (hardware and technical solutions but also to the range of policies, guidance, training, and support that are provided) before you start.
- Aim to keep the number of potential value propositions included in the BMC to a minimum.
- With OAIRs, end-users do not usually provide a resource to support activities, so try to make sure that those who will provide the resource for carrying out activities (i.e., who pays for these) can see a value in doing this such as making internal

[5] https://drive.google.com/open?id=122M11PC-CRW2YXq-ArKiQ6ap-46dyWUbO

reporting easier or more eZcient.
- Bear in mind that value can be realized indirectly as well directly.
- Do not underestimate the importance of good communication channels and establishing relationships with end-users (customers). Remember that communication does not always mean person-to- person, it can also be realized through automated systems.
- Consider how best to leverage key partnerships – both within your own institution but also with other universities to deliver value.

Building upon the BMC and policy development work, representatives of the Palestinian universities developed an Institutional Implementation Roadmap.

Its goal was to move from broad aims to a defined program of activities that would help to realize the value proposition(s) of the OAIR and associated services. Participants were asked in particular to consider time frames for delivering each activity and to define mechanisms for assessing progress towards completing each activity (milestones).

The process of developing implementation roadmap was continued locally. The implementation plans were reviewed by the EU partners and feedback was shared with individual partners.

General recommendations for preparing an implementation plan based on the individual feedback reports:

- Scope a pilot implementation before rolling out wider implementation across the entire institution to see if the approach is viable and to identify potential problem areas are likely to scale.
- Provide a specific deadline for each activity to be completed and include relevant milestones to help monitor progress.
- Be realistic about time frames associated with each activity.
- Be clear about who is involved in carrying out each activity (e.g., IT, Library, Research OZce) and where possible, include named staff.
- Keeping track of time, effort and resource required to carry out each activity in the roadmap is a good idea. This will help with assessing the feasibility of scaling up the pilot with the resources available.
- Define metrics relating to each of the roadmap activities (i.e., provide targets for content uploads; number of users).

V. Setting Up OAIRs

In this section we discuss and justify our choice of the repository framework, and present our detailed implementation plan. The main challenges faced during the implementation process are also discussed, focusing on how we managed to approach and resolve these problems by automating services where possible. We believe that our experience and adopted solutions can be informative to other institutions who plan to establish OAIRs.

A. Setting up the repository software (DSpace)

There is plenty of off-the-shield repository software frameworks that are widely used to develop OAIRs around the globe. These frameworks can be either commercial or free to use, and often have different capabilities and limitations. Examples of these frameworks include DSpace[6] and EPrints[7], which are of the most popular repository frameworks. Other frameworks provide only back-end solutions for archiving and data preservation, but the front-end component needs to be developed based on the institution's needs. Examples of these frameworks include Fedora[8] and Islandora[9].

The partner Palestinian universities have decided to go for open source and free-to-use repository frameworks due to the limited financial resources. We believe that it may be risky to start the transition towards OA publishing by investing large budgets on commercial repository frameworks, considering the limited budgets of Palestinian universities.

DSpace was chosen to be used by the fours Palestinian universities due to its popularity, well-documentation, and the huge technical support available online. In addition, the lack of skilled software

[6] https://duraspace.org/dspace

[7] https://www.eprints.org/uk

[8] https://duraspace.org/fedora

[9] https://islandora.ca

developers who have experience with developing repository services has made us compelled to use a "turn-the-handle" solution such as DSpace.

Having decided to use DSpace, it important to report the limitations that DSpace has, and how these limitations conflict with the needs of Palestinian institutions in particular. First, the customization of the user interface and back-end services of DSpace is diZcult. In addition, DSpace, like most open-access frameworks, has limited support to the Arabic language. DSpace only supports the localization of control buttons, menus, and toolbars, but it does not support the adaptation of metadata presentation and repository structure based on the user's language of interest.

The setup of the repository software was assigned to the IT teams in all universities. The teams were guided to online resources about the implementation steps. To coordinate activities between teams at different universities, a special online forum was launched so that IT staff can discuss implementation details, share experiences and troubleshoot technical problems.

B. Define and set up metadata catalogues

When planning for metadata catalogues, we had two objectives in mind: First, standard metadata catalogues should be adopted so that the OAIRs contents can be easily discovered and indexed by web crawlers and harvesters. Dublin Core metadata was chosen for building metadata catalogues, but was slightly extended to cope with the specific needs of Palestinian research. In fact, large volumes of Palestinian research outputs, including staff publications and theses, are published in Arabic, especially from Art and humanities departments. Therefore, it is important to have metadata for capturing both Arabic and English bibliographic data. We decided that each Palestinian research output, which can be of any type, should have metadata values in both Arabic and English. For example, a journal article to be deposited into the OAIR should have titles and abstracts in both Arabic and English. This is important to maximize the reach to OAIR contents by both the Arab and wider communities.

Second, one of the objectives of ROMOR is not only to build standalone OAIRs, but also to establish the infrastructure needed to link these repositories in the future. The ultimate goal is to build a national metadata repository that aggregates and enables access to the collective Palestinian research volumes through a single user interface. This goal cannot be achieved without unifying the metadata catalogues to achieve the desired interoperability, and assure that these catalogues are compliant with the FAIR principles. Several meetings have been conducted at both institution's level and between institutions to identify and standardize, where possible, the required metadata catalogues.

C. Customization of user interface to address Arabic language needs

The DSpace front-end theme was modified so that metadata values relevant to the language selected by the user are retrieved and presented. Metadata values of the same item can be rendered in either English or Arabic by simply switching the preferred language from the top bar of the DSpace interface.

Another common problem with repository systems is the inability to distinguish alternative names used by the same author. Authors may write their names in Arabic or English in their publications or may have multiple names due to variations of spelling across different papers. As a result, the same author registered with DSpace may end up with two or more collections of publications, each of which is associated with a different name.

Solutions have been proposed to approach this problem by adopting unique identifiers for authors, such as ORCID, and then grouping repository items by using these identifiers instead of author names. However, Palestinian partners decided not to adopt ORCIDs at this stage because the use of ORCID is still not common in Palestine, and there is a lack of awareness about it among the research community. In addition, many Palestinian researchers publish in local and regional journals not indexed by popular digital libraries, and thus their publications cannot be retrieved by using ORCIDs. We decided at this stage to use the oZcial university email addresses of authors as identifiers in DSpace. Variant names of the same author will be mapped to one oZcial email address, and publications will be grouped by the email address rather than by the author's name.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

D. Populating repository through batch import

Batch import is a service offered by most repository systems to allow a fast population of repositories. Multiple items with associated metadata can be imported and published in one process rather than submitting items one by one through an input form.

We decided to use batch import only in the first stage in order not to launch empty repositories. In later stages when repositories stabilize and users familiarize with OA practices, relevant policies will be applied to mandate researchers to deposit their research papers. We believe that using batch import is necessary at the beginning of the repository life because it will be frustrating to ask individual researchers to submit dozens of publications, which perhaps were published over decades, one by one. In contrast, if the researcher can find most of his/her publications already deposited into the OAIR, he/she will become more motivated to keep his/her research records up to date.

Above all, the main obstacle for performing batch import into Palestinian OAIRs is to find collections of research outputs to be imported. Only academic theses and staff publications published in university journals are archived and published through the library and journal websites. In addition, it is very diZcult to contact the whole research community at Palestinian universities to collect their research belongings.

Therefore, we decided to collect staff publications and university theses by means of automatic data scraping. The project team developed tools to automatically download theses and staff publications along with associated metadata from different university sites, and deposit them into the OAIR. As the library and university journals websites are not compatible with DSpace in terms of metadata schemes and database structures, the developed tools can restructure collected data into the format required by the DSpace. As a result, thousands of publications could be imported into DSpace from different university's legacy systems. It is worth mentioning that the developed tools included solutions to avoid duplicate submissions of items, to exclude items not published by university staff, and to resolve the problem of variant, multi-lingual and misspelled author names by using text similarity algorithms.

E. Fix and Complete Missing Metadata

Publications retrieved from the web or from the university's legacy systems may have many missing or incorrect metadata. Also, they are unlikely to have metadata values in both Arabic and English. Most importantly, keywords were of the most missing metadata fields because they often exist inside documents but are not embedded in webpages accessible to web crawlers. Keywords are very important for OAIRs to enable subjective classification and browsing of contents. Additional effort is still needed to fix or complete missing metadata of deposited items. As we were dealing with thousands of repository items, we decided to automate this process as possible. To do so, we tried to extract missing metadata from documents by using a machine learning library called Grobid[10]. Grobid is trained to extract bibliographic data from PDF documents automatically. It takes a well-structured publication as input and gives bibliographic data, such as the title, abstract, author names, keywords, journal or conference details, as output. We developed a pipeline to which we inputted thousands of publications into Grobid, and then repository items were directly updated with the extracted metadata. Overall, we found that our metadata correction steps updated or corrected about 23% of OAIR items. However, Grobid is not trained on Arabic text, thus were not applied to Arabic publications.

F. Integrate OAIR with RDM process and other university systems

After being deployed and populated, it is very important for the OAIRs to be part of the research workflow currently implemented at Palestinian universities. In other words, the research oZce, library, and different faculties should start using and contributing to the repository. At the same time, responsibilities should be clearly defined to avoid conflict between university entities and assure a seamless flow of deposited items from the submission step until they get published. After planning workflows on paper, the repository software of OAIRs has been configured to apply the planned workflows. This involves representing the workflow in a machine-understandable format, creating users groups associated with workflow tasks, and assigning responsibilities.

[10] https://grobid.readthedocs.io/en/latest/Introduction

Another important service is to link repository content with the institutional and personal pages of academic and research staff. It may be advantageous for researchers if they can retrieve the list of their publications from the OAIRs and present them on their pages. Linking the repository content with staff pages presents the following benefits: 1) it will release the staff from the burden of creating and editing publication lists manually. 2) It will keep the staff motivated to deposit content regularly to keep their publication lists. 3) Most importantly, it will increase access to the repository content since the visitors of staff pages will eventually land on the repository pages. Therefore, we developed a service that shows on each researcher's site the list of publications he/she deposited into the IR. All that the researcher needs to do is to insert a short script inside his/her site, and the list of publications will be rendered automatically (see Example [11]).

One challenge we identified from the needs assessment studies was the inability of the research oZces and faculties to monitor the research impact and progress. OAIRs provide valuable metrics that can be used to track accessibility and reach to deposited items such as the number of views, the number of downloads and the countries from which visits originate. DSpace provides item-based metrics, but there are no similar metrics for faculties or communities. In addition, there is currently no support for detailed analysis such as tracking the progress of faculties or individuals over a period of time or classifying deposited items based on types or publishers. Therefore, we extended OAIRs by building a reporting service that enables to submit queries similar to the aforementioned scenarios and to present results in both tabular and visual representations. An example of the reporting service can be found here . This reporting service can be easily extended in the future to handle different queries upon request.

VI. Policies

Another aspect of data management being addressed by ROMOR is the development of an institutional policy. The success of ROMOR initiative at Palestinian universities and at the national level in the future will depend strongly on the existing policies and regulations related to the deposition of the different research outputs types, especially on the OA mandate policy at a national and an institutional level. No policy for research data management has been in place before ROMOR. In this section, we describe our efforts in the development of policies in ROMOR Palestinian institutions striving for OA.[12]

A. Development Phase

During the training workshops, participants were introduced to the range of policies that may need to be in place to support their OAIR and associated services. They have been introduced to RCUK Common Principles on Data Policy[13]. The DCC and the Jisc MRD projects guidance and support materials[14] have also been introduced to the participants. The participants have been also trained using the guidance for developing a research data management Policy provided by LEARN Toolkit of Best Practice for Research Data Management.[15]

The process of developing a data management policy in Palestinian universities required extensive consultation with stakeholders and the approval of the relevant committees. In Palestinian universities, the management of research outputs is usually led by deans of scientific research and in collaboration with the libraries, and the IT centers. Natural they took part in developing the data management policy with ROMOR team in each university. The experiences of the local universities showed that such policies need to be in Arabic language and need to be accompanied both by activities to raise awareness of what is being advised, and institutional support services to enable researchers to actually implement the recommendations.

Participants have been working on the development and refinement of their policies iteratively over the life of the ROMOR project. A policy document has been produced and is currently is either pending

[11] http://site.iugaza.edu.ps/ialagha/dspace_pub

[12] https://iugspace.iugaza.edu.ps/DSpaceDBProject/testreport.jsp

[13] http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx

[14] http://www.dcc.ac.uk/resources

[15] http://learn-rdm.eu/wp-content/uploads/RDMToolkit.pdf

approval or approved in each partner Palestinian university. In addition, a policy model in Arabic language has been approved by Palestinian partners. It should serve as a template that can be used by other universities who would like to start the process of developing a data management policy[16].

General recommendations for developing policies have been distilled through aggregation from the individual feedback reports received form EU partners on each of the Palestinian partners' policy:

- It is wise to get senior management buy-in from the outset when developing or revising policies in relation to the OAIR.
- The scope of the policy should be very clearly stated (e.g., the policy covers the deposit of journal articles, research data, and/or theses).
- Be very clear about who is expected to comply with the policy (staff, students).
- Avoid drafting very long or complex policies with legal sounding language. Aim to produce concise policies written in plain language.
- Make a clear delineation between the content of the policy itself and specific details that should be shared as separate guidance.
- It is advisable to avoid referring to named tools or products in the policy itself as these can change over time. Again, this sort of specific information should be provided as guidance.
- Test the policy out before seeking ratification. This will help to ensure that the scope of the policy is realistically achievable with the resources available. Be sure to involve researchers as well as operational units.
- Be sure to consider workflows and institutional infrastructures (hard and soft) that will be required and/or impacted by the policy and be sure to consult with all relevant stakeholders (e.g., operational units that may need to support services to enable compliance with the policy).
- Be sure to spell out any terms that may lack clarity (i.e., if the policy covers 'active' researchers, specify what this term includes).
- Consider who will be responsible for monitoring compliance with the policy and how this will be carried out.

[16] https://drive.google.com/open?id=1a08etl1MTabC2VP3vb-N7QNyKSR5QuaQg

B. Challenges

There are some factors leading to diZculties in effectively implementing the policies:

Compliance with the policy: If OA is only recommended in the developed policies, some researchers will deposit their publications and research data, and some will not. It is not only about personal preferences, but there are also significant differences in OA practices among disciplines too. From the Palestinian universities experience and point of views, if the institution wants to store and preserve its intellectual contribution to human knowledge, the best way is to declare deposition manory. It would be much easier for an institution if the framework for such a mandate exists on a national level, in-laws, codes, regulations, and rules. In the next section of this paper, we will describe ROMOR activities in order to realize all these efforts at the national level. However individual Palestinian universities, for example, mandated that the repository is the main and the only means for offcially recording the researchers' publications activities. These universities generate automatic reports about the research activities of its researchers and consider these reports for evaluations and academic promotions. Researchers are also encouraged to publish their publications in OA journals, however, the preprint, the final version of the accepted or published paper must be deposited, and OA provided in the accordance with the right holders' provisions. If the rights are partially or totally transferred to the publishers, the author is obliged to deposit an eligible version of the paper and provide OA to it as soon as possible. On the other hand, the mandate for depositing the master's and the doctoral thesis is stated in the developed policies.

The capacity of the university to implement the policy: for policy implementation to be successful, it is important for institutions to have the necessary administrative and other abilities required to carry out the implementation process. These necessary skills and resources include the availability of suZcient numbers of qualified staff within the different responsible units. However, the lack of suZcient staff was found to be of particular concern during the implementation process and was seen as a burning issue and a major stumbling block. To overcome this problem, some partner Palestinian

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

universities are decentralizing the RDM efforts. They are including graduate students, researchers, research representatives, and support staff from the different faculties and other research units in the RDM process by involving them first in a series of RDM training workshops. Second, this decentralization of is RDM is reflected in the workflow of the repositories. Through the repository, the students or the researchers first deposit their publications, which are then received by the research representatives or support staff in their faculties, who in return check the deposited item and forward them to the responsible person in the library or in the dean of scientific research, before actually being deposited to the repository.

Copyright issues: there is a lack or almost complete absence of awareness about copyright literacy among most Palestinian researchers. There seems to be a general misunderstanding amongst academics about copyright and the rights they have to re-use their publications after they have signed copyright transfer agreements. Explaining these policies and talking through these issues is often the most diZcult and frustrating part of encouraging people to deposit material. Offering training on copyright literacy and the use of tools for copyright checking such as SHERPA/RoMEO[17] can help indeed. Also during the submission process, the support staff can do any copyright checking on behalf of authors which edge, the best way is to declare deposition mandatory. It would be much easier for an institution if the framework can be done easily using the SHERPA/RoMEO or by contacting publishers directly with a standard e-mail. However, the problem becomes more severe when it comes to publications in Arabic journals. Most of these journals have no clear OA statement and are not included in SHERPA/RoMEO. These diZculties have highlighted the importance of having an institutional statement of copyright, which supports and encourages authors to retain control over their work.

Quality control: Peer-review, is outside the scope of the repository itself. In order to ensure a certain level of quality control, Palestinian universities recommend peer-reviewed submissions to their repositories. However, many Arabic language publications are not peer- reviewed. On the other hand, these publications focus on local problems on most of the time, and therefore there is an essential need to have them visible. Restricting depositing to peer-reviewed publications will leave many important publications invisible.

Cultural issues: some researchers tend to be conservative in their publications habits. They have the impression that OA publishing reflects low-quality research. Some researchers also, have technophobia problem especially those from humanitarian fields. We approach such researchers most of the time personally and using the word-of mouth method. We also offer to deposit in the repository on their behalf until they feel confident they can do it themselves.

VII. National Roadmap

While institutions and their libraries are moving ahead to establish OAIRs to maximize the visibility of their academic output and make it as widely available as possible, there are still considerable obstacles that inhibit academic community from participating in this initiative. Among challenges are low bandwidth, technophobia, technical support and security, finance and legal aspects, lack of skilled personnel to manage the repository, and lack of adequate power supply (as the electric power sector of Gaza is poor and it really affects the servers that house the repository).

A need for a national road map for research output management and a national repository in Palestine was clearly articulated by stakeholders of workshops organized by ROMOR. The participants represented different stakeholders such as the institutions' libraries, research centers, Ministry of Higher Education (MoHE), the Accreditation and Quality Assurance Commission (AQAC), and numerous higher education institutions.

The institutions that already have OAIRs stress the need to identify ways to automatically ingest objects into the repository to minimize the workload on researchers and supporting services. They also mentioned the need to have better external visibility (through integration with hubs like OpenAIRE, etc.), and to support research data that include types beyond theses and publications (such as Databases,

[17] http://www.sherpa.ac.uk/romeo.php

datasets, source code, etc.). However, research data requires special policies and procedures that go beyond the capabilities of current OAIRs. Research data comes in different formats that require special metadata for interpretation and demand distinct formats and description as required by different research communities. In addition, special solutions for digital preservation are required as research data are often stored in formats that rapidly become technologically obsolete. The sensitivity of research data may also influence the security needs that require special consideration in policies. The workshops participants have also discussed the fragmented environment of OAIRs, in which effort and costs are duplicated, numerous software platforms and versions are managed individually by each institution, metadata are applied inconsistently, users are served poorly, and the ministry and the universities are unable to take advantage of collective data about content and users. Besides, digital works come with preservation and storage issues that are only just being realized in Palestine. Institutions need to begin tackling the preservation of digitally produced and recorded material. But the enormity of the task is only just beyond the capacity of many institutions.

A committee was established that comprises representatives from ROMOR team, the Palestinian MoHE, and the AQAC. The committee was responsible for shaping the national roadmap with inputs from the discussions within ROMOR working groups and workshops. The committee hoped by this to foster future efforts towards the organizational, informational and technical development of the national infrastructure.

The national roadmap defines the following objectives:

- identify the needs to integrate scholarly publishing and dissemination of the universities
- identify the requirements to build a shared platform for research outputs management
- build capacity needed for building, managing and sustaining the shared platform
- adopt a common metadata schemes that would be essential for interoperability and for collecting and comparing data across institutions
- establish, maintain and manage a national network of repositories to allow discovery and reuse
- establish, maintain and manage a central infrastructures to enable reliable and interoperable OAIRs
- establish a national center for research data management that will provide infrastructure and services for archiving and sharing research data of different types
- develop research output management and OA policies at the national level
- deploy preservation technology framework
- promote collaboration and partnerships among a large number of institutions and individuals from the academic community in Palestine and in the world
- improve the criteria used for assessment and accreditation of HEIs and academic programs to be research-driven and transparent. AQAC/MoHE can benefit from the metrics gained from OAIRs to assess the research performance and activity of HEIs.

VIII. Conclusion

In this paper, we discussed methodology, solutions, and challenges when developing Open Access Institutional Repositories and research data management services at four universities in Palestine. We have highlighted specific technical challenges, e.g. customization of repository software to support Arabic metadata, as well as institutional challenges relating to research data management policy definition, ratification and adoption. For each of the thematic areas, we have provided specific lessons learned that can be of high value to other developing countries which aim at increasing visibility of their research outputs and strengthening competences of their researchers in data management and preservation.

The ongoing and future work focuses on implementing a national roadmap for integrated services that allow not only to promote publications but also to share and preserve data. This will enable to reduce costs and facilitate the exchange of expertise and highly skilled human resources. We are also working towards better integration with initiatives such as the European Open Science Cloud or PlanS.

References

[1] M. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. O. Bonino da Silva Santos, P. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, R. Finkers, and B. Mons, "he FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, 2016. DOI: 10.1038/sdata.2016.18.

[2] S. Zyoud, S. Al-Jabi, and P. W. Sweileh, "Scientific publications from Arab world in leading journals of Integrative and Complementary Medicine: A bibliometric analysis," *BMC Complementary and Alternative Medicine*, vol. 15, 2015. DOI: 10.1186/s12906- 015-0840-z.

[3] S. H. Zyoud, "Dengue research: a bibliometric analysis of worldwide and Arab publications during 1872–2015," *VirologyJournal*, vol. 13, pp. 1–10, 2016.

[4] M. of Higher Education, *Annual Statistical Directory of Palestinian Higher Education Institutions* 2017-2018, 2018. [Online]. Available: https://www.mohe.pna.ps/services/statistics.

[5] IAP, "Scientific Research in Palestine," *Palestine Academy for Science & Technology*, 2002.

[6] W. M. Sweileh, S. H. Zyoud, S. W. Al-Jabi, A. F. Sawalha, and S. A. Khalil, "Research Output from Palestine (1995–2012): A Bibliometric Study," *International Information & Library Review*, vol. 46, no. 3- 4, pp. 99–112, 2014. DOI: 10.1080/10572317.2014. 943070.

[7] W. M. Sweileh, S. H. Zyoud, S. Al-Khalil, S. W. AlJabi, and A. F. Sawalha, "Assessing the Scientific Research Productivity of the Palestinian Higher Education Institutions: A Case Study at An-Najah National University, Palestine," SAGE Open, vol. 4, no. 3, 2014. DOI: 10.1177/2158244014544287.

[8] M. Qumsiyeh and J. Isaac, "RESEARCH AND DEVELOPMENT IN THE OCCUPIED PALESTINIAN TERRITORIES: CHALLENGES AND OPPORTUNITIES," *Arab Studies Quarterly*, vol. 34, no. 3, pp. 158–172, 2012. [Online]. Available: http://www.jstor.org/stable/41858700.

[9] X. Y. C. W. Yongchao ZHAO, "Academic institutional repositories in China: A survey of CALIS member libraries," *Journal of Data and Information Science*, vol. 5, no. 2, 2012.

[10] S. Ali, S. Jan, and I. Amin, "Status of Open Access Repositories: a Global Perspective," *International Journal Of Knowledge Management And Practices*, vol. 1, no. 1, pp. 35–42, 2013.

[11] D. C. Prosser, "Institutional repositories and Open Access: The future of scholarly communication," *Inf Serv Use*, vol. 23, 2003. DOI: 10.3233/ISU- 2003- 232-327.

[12] R. Cullen and B. Chawner, "Institutional Repositories, Open Access, and Scholarly Communication: A Study of Conflicting Paradigms," *The Journal of Academic Librarianship*, vol. 37, no. 6, pp. 460 –470, 2011. DOI: https://doi.org/10.1016/j.acalib.2011.07.002.

[13] H. Hockx-Yu, "Digital preservation in the con text of institutional repositories," *Program*, vol. 40, pp. 232–243, 2006.

[14] O. Saini, "Understanding the Role of Institutional Repository in Digital Preservation in Academic Libraries: A Review of Literature," *Library Philosophy and Practice,* vol. 2018, 2018.

[15] J. Mavodza, "A review of the open access concept in the UAE," *New Library World*, vol. 114, 2013. DOI: 10.1108/03074801311326885.

[16] S. S. Ahmed and S. Al-Baridi, "An overview of institutional repository developments in the Arabian Gulf Region," *OCLC Systems & Services: International digital library perspectives*, vol. 28, pp. 79–89, 2012. DOI: 10.1108/10650751211236613.

[17] S. Carlson, "An Assessment of Institutional Repositories in the Arab World," *D-Lib Magazine*, vol. 21, no. 5/6, 2015. DOI: 10.1045/may2015-carlson.

[18] A. aalYateem and N. Bn-Hameed, "Digital Repositories in the Arab Universities: A Comparative Analytical Study," *Procedia Computer Science*, vol. 65, pp. 768 –777, 2015. DOI: https :// doi . org / 10 . 1016/j.procs.2015.09.022.

[19] I. AlAgha and R. Awadallah, *Assessing Current Open Access and Research Data Management Practices and Services in Palestinian HEIs,* Project number: 573700- EPP-1-2016-1-PS-EPPKA2 -CBHE-JP. Co-funded by the Erasmus+ programme of the European Union., 2017. DOI: 10.5281/zenodo.801735.

[20] R. Awadallah and I. AlAgha, *Assessing Research Data Output Management at the Managerial Levels at Partner PS HEIs*, Project number:573700-EPP-1-2016-1- PS-EPPKA2-CBHE-JP. Co-funded by the Erasmus+ programme of the European Union, 2017. DOI: 10. 5281/zenodo.801773.

[21] J. Rans and A. Whyte, *Using RISE, the Research Infrastructure Self-Evaluation Framework*, Edinburgh: Digital Curation Centre, 2017. [Online]. Available: www.dcc.ac.uk/resources/how-guides.

[22] W. Los, "EDISON Project: Building Data Science Profession for European Research and Industry," 2016.

Supporting Virtual Reality and 3D in Academic Libraries

Defining Preservation and Curation Challenges

Zack Lischer-Katz
University of Oklahoma, USA
zlkatz@ou.edu
https://orcid.org/0000-0002-4688-1275

Matt Cook
Harvard University, USA
matt_cook@harvard.edu
https://orcid.org/0000-0002-1513-0444

Nathan Hall
Virginia Tech, USA
nfhall@vt.edu
https://orcid.org/0000-0002-0676-9916

Juliet Hardesty
Indiana University, USA
jlhardes@iu.edu
https://orcid.org/0000-0002-7705-5937

Jamie Wittenberg
Indiana University,USA
jvwitten@indiana.edu
https://orcid.org/0000-0002-5266-0508

Jennifer Johnson
Indiana University - Purdue University Indianapolis, USA
jennajoh@iupui.edu
https://orcid.org/0000-0003-3994-4538

Robert McDonald
University of Colorado Boulder, USA
rhmcdonald@colorado.edu
https://orcid.org/0000-0003-4258-0982

Tara Carlisle
University of Oklahoma, USA
tara.carlisle@ou.edu

Abstract[1] – Academic libraries are increasingly adopting virtual reality (VR) technologies for a variety of research and teaching purposes, yet there is a lack of clear guidance in the community on how to manage these technologies in effective and sustainable ways. In June 2018, the University of Oklahoma hosted the second of three forums on the topic of using 3D and VR for visualization and analysis in academic libraries, as part of the IMLS-funded project, Developing Library Strategy for 3D and Virtual Reality Collection Development and Reuse (LIB3DVR). This project uses nominal group technique to collect data from the invited experts (from diverse academic disciplines and institutional contexts) at the Forum to identify common preservation and curation challenges in the visualization and analysis of 3D data and the management of VR programs. This paper describes the findings of this project and outlines strategies and techniques for curating and preserving 3D/VR.

Keywords – Virtual Reality, 3D Data, Libraries, Preservation, Curation

Conference Topics – Exploring New Horizons; Building Capacity, Capability and Community.

[1] This project was made possible in part by the Institute of Museum and Library Services (LG-73-17-0141-17). The views, findings, conclusions, and recommendations expressed in this program do not necessarily represent those of the Institute of Museum and Library Services.

I. Introduction

Virtual reality (VR) has resurfaced as an engaging and innovative technology, with a surge in the availability of low-cost hardware. Academic libraries are increasingly adopting VR as a means of providing enhanced access to collections of 3D models, new research tools, and new immersive learning environments for students [1]. VR is useful for enhancing visualization and analysis for big data applications [2, 3] and scientific research, and for contributing to increased engagement in the classroom [4, 5]. The demonstrated efficacy of VR for research and teaching purposes, and the increasing affordability of hardware, has inspired library administrators and technologists to introduce VR to makerspaces and other sites across university campuses, as well as to provide for the checkout of VR equipment by library patrons [6, 7].

The adoption of VR is part of a trend towards technological innovation now taking place in academic libraries; however, there is a clear lack of guidance in the library community on how to introduce, integrate, and sustain these technologies in ways that serve all library stakeholders. A multitude of institutions

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

are tackling the same issues, oftentimes replicating similar work, indicating a need for leadership on the part of early-adopters, including academic libraries, to determine best practices for supporting VR technologies across different types of institutions. Practical considerations, such as designing systems to reduce motion sickness and increase accessibility, for example, have been tackled on an ad hoc basis, making it difficult to scale up VR services for widespread adoption. Similarly, preservation issues have not been adequately addressed for 3D/VR. The fundamental problem is that best practices have not been systematically collated, analyzed, and packaged for widespread dissemination and adoption across the community.

To address these challenges and aid in the maturation of 3D and VR as learning and research technologies, an interdisciplinary group of librarians and researchers from Virginia Tech, Indiana University, and the University of Oklahoma convened to develop a series of three national forums on this topic, funded by the Institute for Museum and Library Services (IMLS), as a project titled Developing Library Strategy for 3D and Virtual Reality Collection Development and Reuse (LIB3DVR) [8]. Each forum was designed to cover a particular phase of the lifecycle of 3D and VR within academic contexts: The first forum looked at 3D/VR creation and publishing; the second forum looked at 3D/VR visualization and analysis; and the third forum looked at 3D/VR repository practice and standards. This paper presents findings from the second forum, held in June 2018 at the University of Oklahoma, Norman, OK.

II. Literature Review

3D and VR technologies offer new potential for interactive engagement with and analysis of spatially complex artifacts, spaces, and data, which enables the possibility of new insights [9]. They are also being used as immersive learning environments for a range of fields, from anthropology to biochemistry [10] to architecture and design [11]. Researchers in a range of fields are already incorporating 3D technologies into their scholarly practice in order to enhance their methods of analysis [12, 13, 14, 15]. Research has shown that scientists are able to make more inferences from 3D digital models than from photos, while humanists can visually represent texts, images, and material artifacts in VR spaces for detailed analysis and to better understand their cultural and historical context [16, 17, 18, 19]. In addition, the 3D representation of fragile or otherwise inaccessible artifacts opens up access to a host of archived objects for a wider audience of researchers, students, and the general public [20, 21].

Studies on the preservation and curation of 3D/VR have pointed out that as researchers increasingly use these technologies, there will be a greater need for archiving and preservation services [22, 23, 24, 25]. As emerging technologies, however, there is still a lack of knowledge about how best to create and curate the scholarly products of 3D/VR projects. 3D data is being valued for its potential to be reused beyond the original context of creation [26, 27], which makes it important to ensure adequate data curation procedures are in place. In recent years, a handful of domain-specific research groups have attempted to develop 3D data creation workflows and repository structures [28, 29], and metadata guidelines (e.g., the Archaeology Data Service's Guide to Best Practices [30]). Technical groups (e.g. the Khronos Group's COLLADA and OpenXR initiatives [31]) have been working towards interoperability standards for 3D/VR file formats, software, and hardware. Early 3D/VR metadata projects, such as Mourkoussis, et al. [32], have not seen their guidelines widely adopted. At the same time, more recent projects have identified a lack of suitable metadata standards, particularly with regards to preservation metadata [33, 34], as a serious challenge to working with 3D data. Bennet (2015) suggests, "3D data archiving remains a multifaceted web of decision points on file formats, their relational organization, packaging, and final storage" [35].

There have been some attempts to develop digital repositories and common metadata guidelines for 3D data [36], when it is framed as "cultural heritage" data. For example, the goal of the European-based, 3D-ICONS Project [37] was "to provide Europeana with 3D models of architectural and archaeological monuments." This project also developed metadata guidelines [38], which were released in 2013. It was announced in early 2019 that Europeana [39], the European Union's digital library platform for cultural heritage, would start introducing 3D materials into its collections [40] with guidance from 3D-ICONS.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

In 2017, the project team who established the IMLS-funded project, Community Standards for 3D Data Preservation (CS3DP) surveyed an international community of researchers and librarians involved in digital curation and 3D data acquisition and research (112 total participants). They reported that "72% said that they did not use best practices or standards for preservation, documentation, and dissemination of 3D data. Of those not using standards/best practices, 69% said that they did not use them because they were unaware of such standards" [41]. Cook and Lischer-Katz (2019) have defined three important preservation areas in which libraries can take the lead: Managing VR hardware and software obsolescence; establishing file formats for archiving 3D content; and developing metadata standards [42]. Moore and Skates Kettler (2018) also point to the importance of metadata: "Creating a standard for metadata and a set of best practice recommendations would have immense impact on the overall preservation and interoperability of 3D research" [43]. One of the critical challenges to common standards is the diversity of approaches being carried out as part of 3D and VR creation methodologies [44].

At the Coalition for Networked Information (CNI), Fall 2017 Plenary, Clifford Lynch noted that while in many cases it is now possible to support the full research lifecycle of a significant range of 3D objects at reasonable quality and cost, there remains a significant and important challenge to implement a whole library apparatus, including the development of good standards for storage and description; good provenance metadata to tell us where 3D objects came from; and suitable documentation specifying whether they are produced by scanning real objects in the world or are designed entirely on a computer (e.g., architectural CAD designs) [45].

A report published in February 2019 by the Council on Library and Information Resources argued that libraries need to take the lead in supporting "new and complex technical workflows, scholarly practices, and data curation and digital preservation requirements," if 3D/VR technology is to be widely used as a set of scholarly and pedagogical tools [46]. In many cases, academic libraries are already taking the lead in adopting these technologies, providing support and developing effective course integrations and research support. However, Cook and Lischer-Katz (2019) note, "the sustainability of VR as a legitimate library resource depends on managing VR-related data and digital tools throughout the research life cycle" [47]. Meyerson (2019) has suggested that establishing preservation guidelines for 3D and VR can follow existing guidelines for other types of software, with some modification [48].

The findings and analyses in these reports and articles unanimously point to a critical need to establish 3D/VR creation and curation best practices and standards, and they emphasize the essential role played by community engagement in establishing those best practices and standards. They also acknowledge that because of the diversity of approaches and contexts it is necessary to look closely at how a broad cross section of stakeholders is approaching this problem space in order to establish guidelines that will be useful for all involved.

III. Research Objective

The main research objective of this phase of the LIB3DVR project is to determine how academic libraries and other institutions with 3D/VR programs are planning for the reuse and long-term sustainability of 3D and VR resources. Identifying the challenges and strategies in current practice will help establish a foundation for community-generated best practices and standards.

IV. Methods

The project team assembled a two-and-a-half day forum in Norman, Oklahoma with fifteen expert participants, including academic librarians, researchers from a variety of disciplines, and commercial game designers and software engineers. Participants were selected by identifying national experts in representative fields, with an eye towards achieving institutional, disciplinary, racial, and gender diversity. The project team shared the participant list with an advisory board that provided further input on the selection of participants. In addition to the meeting of invited experts, a half-day public forum was held in which local stakeholders were invited to attend and discuss their experiences working with 3D/VR.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The project team used a nominal group technique to generate research data for this study [49]. Nominal group technique is a consensus-building method for achieving general agreement on a topic through face-to-face small group discussions. This method was adopted in order to reveal key challenges related to the visualization and analysis of 3D and VR data and the design and management of library programs to support those activities. The agenda for the forum was divided into special sessions on specific topics. Data were generated through methods of community note taking, facilitated using shared Google Drive documents for each forum session. At the end of each discussion session, a group note taker summarized and presented the views of each small group to the wider forum. Both the raw community notes and the summarized facilitator notes were collected and analyzed. Notes produced from the smaller groups and from the larger group form the basis of the findings. We validated these findings by disseminating an early draft of this paper to participants, asking them to correct, clarify, or elaborate on the paper's findings. The authors incorporated all participant feedback into a subsequent draft.

Data analysis consisted of grouping data from the community note taking documents into higher level categories based on the research objectives and emergent themes, following an inductive analysis approach [50]. A central part of the data analysis process involved moving from grouping specific examples of institutional practices and personal perspectives in order to link them to more general, community-wide phenomena. In this way, a set of shared challenges and strategies could be identified at the community level of analysis. One of the limitations of this methodology is that it is limited to a small group of experts, which could potentially leave out other perspectives. Including a public forum, which was open to more participants from a greater range of institutions, helped to mitigate this limitation.

V. Findings

Participants were primarily concerned about the practical implementation of VR in their institutions, particularly the costs of maintaining VR equipment over time. Beyond the ongoing costs of maintaining and upgrading VR hardware and software, there are a number of other issues identified by forum participants that impact the management, use, and reuse of valuable VR content. These include the development of suitable documentation practices and tools for tracking the 3D content creation process; legal and ethical concerns, especially in the context of cultural heritage content; and preservation and curation concerns related to research transparency and reproducibility.

A. The Importance of Documentation

Documentation was seen as essential by participants because it can impact the accuracy and reliability of the 3D models and the structure and behaviors of the VR environment. Ensuring transparency in the creation process of VR is essential so that future users can interpret the accuracy of the VR content, which impacts the types of inferences that they can make from it. Strategies suggested by participants for addressing documentation concerns included using project management tools that can document processes over time (i.e., producing process-based project documentation), documenting overall workflows, and using journaling and lab notebooks during the course of a project in order to keep track of decisions made in the production process. Another part of developing good documentation practices is getting into the habit of storing VR project files in open, well documented and widely supported formats, which would enable future users to be able to access the original source files that went into creating the VR project file. If original software is no longer supported, it becomes difficult or nearly impossible to open up VR projects and see how they were assembled, which makes it important to document which software packages were used for a given project.

B. Ethical and Legal Issues

The need for documentation is also related to important legal and ethical questions raised by the use of VR content. Working with cultural heritage content in particular raises a number of concerns in this area. While historical materials are often in the public domain and not encumbered by copyright, there were concerns raised by participants about companies or organizations doing scanning projects and then claiming copyright on the resulting digital products. In addition, participants raised questions about how 3D scanning of a cultural heritage

object might impact the integrity of the object, with consequences for the ownership and value of the original object [51]. This is particularly important when models are produced from culturally-sensitive materials. In other cases, in order to prevent "digital colonization," a term brought up by one participant, it is important for 3D and VR content creators to respect the cultural protocols of indigenous communities. In contexts where personal information may be captured via 3D or VR, data privacy was also seen as a potential issue.

In addition to these ethical and legal concerns, there are also intellectual property issues identified that could impact how VR content is used in the future. VR projects that employ plugins, interactions, models, or other components that involve third-party licenses are at risk of having limited options for reuse, or not being reusable at all if the underlying licenses or digital rights management (DRM) technologies place burdensome restrictions on users. One participant was concerned that DRM could restrict how 3D and VR content are reused, for instance, that DRM might one day limit which 3D-printed models could be printed (the participant was concerned that 3D printers might someday be designed to only print certain models that were authorized via restrictive DRM systems). Increasing use of "software as a service" models, which are built on "black box" systems and cloud storage, also complicates how legacy VR content can be sustained over time and how it can be used for research. This issue is an example of a legal issue that has implications for research transparency and reproducibility, which will be addressed in the next section.

C. *Research Transparency & Reproducibility Concerns*

Being able to access research data and digital scholarly products over time has become an important aspect of research transparency and reproducibility. Participants voiced concern that if software relies on external servers and those servers are shut down some day, then access to the software may be lost. It may be nearly impossible to replicate research findings that relied on server-based software or proprietary software built with non-transparent processes and algorithms. Researchers in the forum were also concerned about how 3D and VR scholarly outputs could be cited as persistent scholarly objects if they rely on "software as a service"-based systems. One participant suggested that blockchain technology might be useful as a means of keeping track of provenance and the intellectual property chain. This is an area that deserves further investigation, as it may help to address some of the documentation and transparency challenges of managing 3D and VR over time, but with the caveat that other research has shown that blockchain may have limited utility as a preservation tool [52, 53].

Grant-funding agencies, such as the National Science Foundation, are making data management plans (DMPs) a required component of grant applications and preservation of research data is an important component of a DMP. Being able to preserve and access 3D/VR into the future is important for a range of stakeholders in the research community. The discussion around preserving 3D and VR content revolved around questions of 1) defining what elements of 3D/VR projects to save, 2) identifying the range of technological obsolescence and interoperability challenges that are typically encountered, and 3) defining strategies for preservation.

1. *Defining the Objects of Preservation*

Participants tried to come to some consensus on what should be considered the most basic unit of a 3D or VR asset that might be reused in the future. Some answers included preserving 3D models along with the VR behaviors and "physics" of those models, including the structures and interactions between elements in the VR environment. In terms of preserving 3D models, one challenge is to identify how much quality is necessary. There was some discussion of preserving low-resolution models in a VR environment for re-use and some participants argued that preservation efforts should focus on the high-resolution models that are produced through 3D capture processes such as LiDAR and photogrammetry. One participant introduced the concept of the "smallest preservable unit" (i.e., the smallest unit that can be exported and used to reconstruct the VR environment or build new environments in the future), which could be defined depending on the particular use-case or the organization's preservation intent. One example of this is the concern over preserving the behaviors and interactions of

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

objects in the virtual space. In terms of behaviors and physics, one participant pointed out that these elements could not be saved separately from the models or the VR environment because they are generated by the game engine that was used to create the VR environment, which can change as the game engine is updated over time. For instance, the Unity game engine, which is a commonly used platform for producing VR content, is constantly being updated and it is difficult to ensure that the behaviors of elements in a VR environment at one point in time will interact consistently in the future. Only the game engine design company has complete control over how those elements will change. Because of this, participants pointed out that this makes preserving the actual performance of a VR environment very difficult. While many of the elements may be preservable individually, this still does not capture how that VR environment behaves when in use. For that reason, forum participants emphasized the importance of documenting behaviors of VR environments using video recordings of users engaging with them.

2. Obsolescence & Interoperability Challenges

Participants pointed out that because of the complexity of VR technologies, the risk posed by obsolescence to the long-term accessibility of VR is very high. They pointed to the updating of firmware, dependencies on third-party software, and upgrading hardware as activities that could impact how the VR system behaves and whether or not older VR projects can be accessed in the future. Older VR projects may need to be migrated over time to new systems. For instance, one participant presented a case study on the preservation initiative to preserve the Virtual Harlem project (a project developed by Dr. Bryan Carter, Associate Professor in Africana Studies at the University of Arizona) [54]. This involved moving the project to a new VR platform every few years, which typically required recreating most of the VR environment from scratch because the different VR systems were not compatible and did not have import or export capabilities. This shows how even active and ongoing migration of a project from one VR system to another is challenging and requires significant resources. Another preservation challenge of current VR technologies is their dependency on server-based resources. The software packages that run VR headsets also rely on external servers for accessing user accounts. If VR headsets do not have "offline" modes, users will no longer be able to operate the VR hardware if the company's server (e.g., Oculus) goes down or the company ceases operation.

Interoperability was also identified as an important issue that had implications for preservation and reuse of VR content. One of the biggest challenges identified was the lack of concerted effort at the level of university campuses to communicate about VR projects and promote VR adoption in ways that would mandate interoperability. Different units on campus are creating VR content that may be useful for other units, but lack common interoperability standards and use an array of software and hardware configurations that may not be compatible. Thus, VR content may not be easily shared across campus units, not to mention between different institutions, if there is no coordination of interoperable VR solutions. One suggestion provided by participants to address this problem was to develop a database that would help identify who is using particular hardware/software configurations across campus, which would make it easier to adopt similar configurations and share content. The use of containerization tools (e.g., Docker), which bundle dependencies and system configurations together, could be useful for ensuring that VR projects are interoperable between units and institutions.

From these discussions, a set of considerations emerged that need to be taken into account when planning for preservation of 3D/VR. First, it was acknowledged that involving a range of stakeholder groups in preservation planning initiatives is essential for tackling preservation problems. In particular, software engineers should be involved in preservation planning in order to address the technical preservation issues. Second, standardization and adopting standards is critical. Developing common, sustainable preservation practices requires the standardization of preservation and access formats for VR and 3D. Third, preservation is closely connected with questions of interoperability and the ability to network and connect different virtual worlds. One participant pointed out that preservation is not enough and that virtual worlds also need to be interconnectable (i.e., use interoperable standards so that content can be shared and reused, and users are

able to move between different worlds in VR), otherwise they will remain isolated and unused, inevitably becoming inaccessible. Finally, participants pointed out that other fields have been tackling similar issues around preserving complex configurations of visual information and computer software and hardware, such as audiovisual/moving image preservation and software preservation communities. Looking to strategies from these fields could also help the 3D/VR preservation community.

3. Defining Strategies for 3D/VR Preservation

Participants identified a range of possible preservation strategies for dealing with these challenges and preservation considerations. Selection and documentation were seen as important activities for ensuring the long-term preservation of 3D/VR content. Selection criteria for 3D/VR content, particularly in terms of making decisions about archiving low- versus high-resolution content was seen as essential, and participants saw an urgent need for best practices for appraising 3D/VR for archiving. This is complicated by the earlier point about identifying the "smallest preservable unit," because it is not always clear what needs to be saved throughout the lifecycle of 3D data (from capture to processing, editing, etc.). Participants agreed that preserving the "raw data" from the earliest phase of the 3D/VR project is important for future-proofing them, because even if the finished projects become inaccessible due to system obsolescence, they can still be rebuilt from their constituent elements.

What counts as "raw data" in research is still an open question being debated in many fields [55], and in the case of 3D data creation, participants pointed out that some "raw data," such as scanner data are typically in proprietary formats that have significant long-term sustainability issues. Documentation practices are also important throughout the lifecycle of curating 3D/VR content and they complement selection practices because they both provide information about the processes that created the 3D/VR content, how they interconnect, and the decision making process underlying their archival appraisal. Recent software released by Cultural Heritage Imaging for documenting the creation process of photogrammetry-based 3D projects was offered as a model for how documentation systems can be built into 3D/VR workflows in order to seamlessly capture key moments in the creation process [56].

In addition to these preservation strategies that are particular to 3D/VR content, preservation of 3D/VR can also draw on more general digital preservation approaches, such as emulation and migration. The use of virtual machines was suggested for running obsolete operating systems and VR software, but participants cautioned that one of the challenges would be supporting all of the drivers for the complex network of VR peripherals (e.g., head and hand tracking sensors, head mounted displays, etc.). Migration was seen as a potentially sustainable strategy for moving files out of obsolete systems to more sustainable ones. Planning for migration involves selecting VR systems that have the range of import and export functions necessary for moving files out of that system and into a new one at some point in the future. Based on the case studies considered in this forum (e.g., the Virtual Harlem project), if systems do not have export functions, migration will require rebuilding the virtual environment from scratch. This strategy also connects with the "smallest preservable unit" concept. As discussed earlier, there are issues with behaviors and physics tied to the game engine that limit the effectiveness of a migration strategy.

One final strategy discussed by participants was maintaining hardware and software in a fixed state (i.e., preventing automatic updates and hardware upgrades). While maintaining systems in a fixed state is only a short-term solution that is difficult to maintain in the long-term or scale-up for wider use, preserving software and hardware in this way could provide important examples for future emulation and migration projects. Because the hardware and software configurations for VR systems are typically very complex, having examples of running systems (for instance, in a computer museum context) is essential for understanding how they originally behaved through user interaction, which is necessary for developing future systems that accurately emulate earlier ones.

IV. Summary and Discussion

The findings drawn from the discussions and presentations at this forum offer a broad view of the current concerns of this diverse community. The

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

range of stakeholder groups is expansive and demonstrates a growing interest in immersive visualization technology across many fields and institution types. From the findings, we can identify and summarize a set of common challenges facing libraries and other information institutions that are implementing 3D/VR technologies.

Participants engaged in a lengthy discussion on issues associated with managing, using, and reusing VR content. The main challenges participants identified in this area included the need to develop reproducible workflows and documentation tools and procedures; concerns over research transparency and reproducibility, which are related to documentation concerns; and a complex array of ethical and legal issues that require further investigation.

For supporting documentation efforts, participants recommended the use of project management tools; keeping a journal or lab notebook to keep track of decisions made throughout the creation process; and storing data in open, well-documented file formats. Documentation is an essential component of ensuring research transparency and reproducibility for all forms of research data, but it is only part of the picture for 3D/VR. Participants also identified a set of practices that could help address the challenges of 3D/VR research data curation and preservation:

- Specify which elements of 3D/VR projects to save and at which levels of granularity.
- Define the level of quality at which types of 3D/VR elements should be saved.
- Identify the range of technological obsolescence and interoperability challenges, including: updating firmware and hardware; dependencies on third-party software; dependency on server-based resources or credentialing mechanisms; migrating older VR projects to newer systems; and Interoperability between VR systems and game engines.

From discussions on these challenges, participants defined a set of strategies and recommendations to address them:

- Involve diverse stakeholders in preservation planning to ensure that preservation plans will support the range of future uses.
- Involve software engineers in preservation planning to advise on the technical aspects of preserving hardware/software.
- Design and/or purchase systems with interoperability in mind to increase chance of long-term use.
- Actively monitor other fields, such as moving image preservation, that also preserve complex digital media.
- Adopt a lifecycle approach to managing and preserving 3D/VR content.

Techniques for preservation defined by participants to support these strategies include: planning from the beginning of a project for eventual migration and emulation; maintaining hardware and software in a fixed state, as documentation to guide migration and emulation; and recording videos in order to document fully-functioning VR environments. Beyond documentation and preservation concerns, management of 3D/VR content also involves negotiating ethical and legal issues. Some of the key areas identified that need additional work include: concerns about scanning cultural heritage sites and artifacts and claiming ownership of resulting files; understanding the impact of 3D scanning on original artifacts, and the owners or custodians of original artifacts; and establishing protocols for protecting culturally sensitive materials.

Many of these techniques and challenges are not unique to 3D/VR, but overlap with digital preservation concerns for preserving other types of complex digital objects. The field of video game preservation is also concerned with exploring emulation and migration as preservation strategies [57, 58, 59, 60]. For 3D/VR, emulation and migration strategies appear to be more difficult because of the complex array of hardware peripherals and drivers that constitute a VR system. Migration to new software/hardware environments may be particularly difficult. For instance, the case of the Virtual Harlem project, discussed earlier, suggests that migration may be so difficult that researchers will choose to entirely rebuild the virtual environment on a new VR platform. One hope is that the building blocks of VR environments will be interchangeable, following shared technical standards, so that even if the virtual environments need to be recreated in the future, much of the underlying content will be reusable.

3D/VR also shares similar concerns with the field of *time-based media conservation* [61] in terms of the need to document complex configurations of audio-visual media technologies as they function in their original context, via photography, video recordings, diagrams, etc. In time-based media conservation, the resulting documentation can help guide conservators as they take steps to conserve and prepare the work for exhibition in the future, as part of a critical discourse and investigation of the meaning of media and performance-based art [62] or to match emulated or migrated versions to the documented originals [63]. Similarly, documentation can help emulation or migration efforts for VR environments to make them renderable in the future. Others in the digital preservation field have suggested that the documentation of complex digital objects may be more valuable to future archivists and historians than preserving working versions of the original software [64]. The nascent field of 3D/VR preservation should look to these established fields for guidance and collaboration.

V. Conclusion

The overriding theme across the findings from the Forum is the importance of interinstitutional and interdisciplinary collaboration. Confirming what we had assumed going into this project, it is clear that many of the challenges of 3D/VR can only be solved through systematic and concerted effort across multiple stakeholder groups and existing subfields of preservation research and practice. Furthermore, 3D/VR is not limited to a niche area. As we can see from the range of participants and the diversity of uses they identified, there are wide applications and growing mainstream acceptance in many contexts. Further collaboration through future forums and working groups could and should generate standards and best practices for application across the broad 3D/VR community. These need to be specific enough that they can offer real guidance to stakeholders of varying capacities, but flexible enough to be useful for a range of applications and disciplinary practices.

While the findings from the Forum suggest a variety of techniques and strategies for addressing the challenges identified, there is still much more work that needs to be done to establish standards and best practices. In addition, developing tools for supporting 3D/VR throughout the research or educational lifecycle is critical. Such tools should include:

- Project management and documentation tools.
- Universal 3D viewers that are able to integrate with diverse VR equipment and 3D repositories.
- Sustainable, preservation-quality file formats for 3D/VR.
- Open platforms for hosting and preserving 3D/VR content.

There are a number of other projects that are addressing some of the most pressing challenges in the field of 3D and VR research and teaching, including Community Standards for 3D Data Preservation (CS3DP), discussed earlier; Building for Tomorrow, an additional IMLS-funded project that is developing guidelines for preserving 3D models in the fields of architecture, design, architectural archives, and architectural history; the 3D Digitization Project at the Smithsonian Institution's Digitization Program Office, which is developing software, workflows, and metadata guidelines for a variety of 3D creation processes; and the Library of Congress's Born to Be 3D initiative, which has started convening experts in the field to look at the preservation challenges of "born digital" 3D data (e.g., CAD models, GIS data, etc.). The LIB3DVR project team plans to continue to collaborate with members of these projects, and is confident that through these initiatives, useful standards and best practices will emerge to help librarians, digital curators, and other information professionals address the complex challenges of preserving and curating 3D/VR for academic use.

References

[1] M. Cook and Z. Lischer-Katz, "Integrating 3D and VR into research and pedagogy in higher education," in *Beyond Reality: Augmented, Virtual, and Mixed Reality in the Library*, K. Varnum, Ed. Chicago: ALA Editions, 2019, pp. 69-85.

[2] C. Donalek, et al., "Immersive and collaborative data visualization using virtual reality platforms," in *Proceedings of 2014 IEEE International Conference on Big Data*, pp. 609–14, 2014.

[3] G. Evagorou and T. Heinis, "Visual exploration and interaction with scientific data in virtual reality," in *Proceedings of Knowledge Discovery and Data Mining Conference*, London, United Kingdom, August 20th, 2018.

[4] J. Milovanovic, "Virtual and augmented reality in architectural design and education" in *Proceedings of the 17th International Conference, CAAD Futures 2017*, Istanbul, Turkey, July 2017,
[5] https://www.researchgate.net/publication/319665970_Virtual_and_Augmented_Reality_in_Architectural_Design_and_Education_An_Immersive_Multimodal_Platform_to_Support_Architectural_Pedagogy
[6] M. Thompson, "Making virtual reality a reality in today's classroom," *THE Journal*, November 1, 2018,
[7] https://thejournal.com/Articles/2018/01/11/Making-Virtual-Reality-a-Reality-in-Todays-Classrooms.aspx?Page=1
[8] J.F. Hahn, "Virtual reality library environments," in *Mobile Technology and Academic Libraries: Innovative Services for Research and Learning*, Chicago: Association of College and Research Libraries, 2017, pp. 237-248.
[9] R. Smith and O. Bridle, "Using virtual reality to create real world collaborations," in *Proceedings of the IATUL Conferences*, 2018,
[10] https://docs.lib.purdue.edu/iatul/2018/collaboration/5
[11] Grant proposal for Project LG-73-17-0141-17, Funded by the Institute for Museum and Library Services (IMLS), https://www.imls.gov/grants/awarded/lg-73-17-0141-17
[12] Donalek, et al., 2014.
[13] Z. Lischer-Katz, M., Cook, and K. Boulden, "Evaluating the impact of a virtual reality workstation in an academic library: Methodology and preliminary findings," In *Proceedings of the Association for Information Science and Technology Annual Conference*, Vancouver, Canada, Nov. 9-14, 2018,
[14] https://shareok.org/handle/11244/317112
[15] E. Pober, and M. Cook, "The design and development of an immersive learning system for spatial analysis and visual cognition," Conference of the Design Communication Association, Bozeman MT, 2016, https://bit.ly/2TRziP6
[16] S. Zhang, et al., "An immersive virtual environment for DT-MRI volume visualization applications: a case study," In *Visualization 2001 Proceedings,* IEEE, pp. 437-584. IEEE,
[17] http://dx.doi.org/10.1109/VISUAL.2001.964545.
[18] A. Seth, J.M. Vance, and J. H. Oliver, "Virtual reality for assembly methods prototyping: a review," *Virtual Reality*, vol. 15, no. 1, pp. 5-20, 2011, http://dx.doi.org/doi:10.1007/s10055-009-0153-y
[19] C. Helbig, et al., "Concept and workflow for 3D visualization of atmospheric data in a virtual reality environment for analytical approaches." *Environmental Earth Sciences,* vol. 72, no. 10, pp. 3767-3780, 2014,
[20] http://dx.doi.org/doi:10.1007/s12665-014-3136-6
[21] S. Jang, J.M. Vitale, R.W. Jyung, and J.B. Black, "Direct manipulation is better than passive viewing for learning anatomy in a three-dimensional virtual reality environment," *Computers & Education*, vol. 106, pp. 150-165, 2017, https://doi.org/10.1016/j.compedu.2016.12.009
[22] T. Rowe, and L.R. Frank, "The disappearing third dimension," *Science*, vol. 331, issue, 6018, pp. 712-714, 2011,
[23] http://doi.org/10.1126/science.1202828
[24] D. Viggiano, et al., "A low-cost system to acquire 3D surface data from anatomical sample," *Eur. J. Anat*, vol. 19, no. 4, pp. 343-349, 2015.
[25] E.D. Ragan, R. Kopper, P. Schuchardt, and D.A. Bowman, "Studying the effects of stereo, head tracking, and field of regard on a small-scale spatial judgment task," IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 5, pp. 886-896, 2013,
[26] http://dx.doi.org/10.1109/TVCG.2012.163
[27] B. Laha, D.A. Bowman, and J. J. Socha, "Effects of VR system fidelity on analyzing isosurface visualization of volume datasets," *IEEE Transactions on Visualization and Computer Graphics* vol. 20, no. 4, pp 513-522, 2014, 513-522,
[28] https://www.ncbi.nlm.nih.gov/pubmed/24650978
[29] F. Limp, A. Payne, K. Simon, S. Winters, and J. Cothren, "Developing a 3-D digital heritage ecosystem: From object to representation and the role of a virtual museum in the 21st century," *Internet Archaeology,* vol. 30, 2011, http://intarch.ac.uk/journal/issue30/limp_index.html
[30] B.J. Fernández-Palacios, D. Morabito, and F. Remondino, "Access to complex reality-based 3D models using virtual reality solutions," *Journal of Cultural Heritage,* no. 23, pp. 40–48, 2017.
[31] D. Koller, B. Frischer, and G. Humphreys, "Research challenges for digital archives of 3D cultural Heritage Models," *ACM Journal on Computing and Cultural Heritage*, vol. 2, no. 3, pp. 7:1-7:17, 2009.
[32] Z. Lischer-Katz, K. Golubiewski-Davis, J. Grayburn, V. Ikeshoji-Orlati, "Introduction – 3D/VR creation and curation: An emerging field of inquiry." In *3D/VR in the Academic Library: Emerging Practices and Trend*s, CLIR Report 176, February 2019,
[33] https://www.clir.org/wp-content/uploads/sites/6/2019/02/Pub-176.pdf
[34] J. Moore, H.S. Kettler, and A. Rountrey, "CS3DP: Developing agreement for 3D standards and practices based on community needs and values," in *3D/VR in the Academic Library: Emerging Practices and Trend*s, J. Grayburn, Z. Lischer-Katz K. Golubiewski-Davis, V. Ikeshoji-Orlati, Eds., CLIR Report 176, February 2019, pp.114-121,
[35] https://www.clir.org/wp-content/uploads/sites/6/2019/02/Pub-176.pdf
[36] J. Moore, and H.S. Kettler, (2018) "Who cares about 3D preservation?," *IASSIST Quarterly*, vol. 42, no. 1, pp. 1-14, 2018,

[37] https://doi.org/10.29173/iq20
[38] H. Chapman, E. Baldwin, H. Moulden, and M. Lobb, "More than just a sum of the points: re-thinking the value of laser scanning data. In *Visual Heritage in the Digital Age* (pp. 15–31), London: Springer-Verlag, 2013,
[39] http://dx.doi.org/10.1007/978-1-4471-5535-5_2
[40] H. Richards-Rissetto, J. von Schwerin, "A catch 22 of 3D data sustainability: Lessons in 3D archaeological data management & accessibility," in *Digital Applications in Archaeology and Cultural Heritage*, Volume 6, pp. 38-48 2017,
[41] https://doi.org/10.1016/j.daach.2017.04.005
[42] M. Doerr, et al., "A repository for 3D model production and interpretation in culture and beyond," in *Proceedings of the 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST*, 2010.
[43] M. Doerr, et al., "A framework for maintaining provenance information of cultural heritage 3D-models," in *Proceedings of the Electronic Visualisation and the Arts (EVA 2014)*, London, UK, pp. 267–274, 2014.
[44] Archaeology Data Service / Digital Antiquity, "Guides to good practice," accessed June 25, 2019, http://guides.archaeologydataservice.ac.uk
[45] Khronos Group, "OpenXR initiative," accessed June 25, 2019, https://www.khronos.org/openxr
[46] N. Mourkoussis, M. White, M. Patel, J. Chmielewski, and K. Walczak, "AMS - Metadata for cultural exhibitions using virtual reality," presented at the International Conference on Dublin Core and Metadata Applications, Seattle, 2003.
[47] J. Doyle, H. Viktor, and E. Paquet, "Preservation metadata: A framework for 3D data based on the semantic web," in *Proceedings of 2008 Third International Conference on Digital Information Management*, pp. 925–927, Nov. 2008.
[48] J. Doyle, H. Viktor, and E. Paquet, "Long-term digital preservation: Preserving authenticity and usability of 3-D data," *Int J Digit Libr*, vol. 10, pp. 33–47, 2009.
[49] M.J. Bennett, "Evaluating the creation and preservation challenges of photogrammetry-based 3D models," UConn Published Works, no. 52, 2015, http://digitalcommons.uconn.edu/libr_pubs/52
[50] A. Felicetti and L. Matteo, "Metadata and tools for integration and preservation of cultural heritage 3D information," *Geoinformatics FCE CTU Journal*, vol. 6, 2011.
[51] A. Corns, *3D-ICONS: D7.3-Guidelines and Case Studies (Version Final).* Zenodo, Nov. 30, 2013, p. 6, http://doi.org/10.5281/zenodo.1311797
[52] A. D'Andrea, and K. Fernie, *3D-ICONS: D6.1-Report on Metadata and Thesaurii (Version Final)*, Zenodo, March 10, 2013, http://doi.org/10.5281/zenodo.1311780
[53] Europeana Foundation, "Europeana collections," accessed June 25, 2019, https://www.europeana.eu/portal/en
[54] Europeana Foundation, "3D content in Europeana," accessed June 25, 2019, https://pro.europeana.eu/project/3d-content-in-europeana
[55] Community Standards for 3D Data Preservation (CS3DP) Project, "Background," 2017,
[56] https://osf.io/ewt2h/wiki/Background/
[57] Cook and Lischer-Katz, 2019.
[58] Moore and Kettler, 2018, p. 11.
[59] P. Alliez et al., "Digital 3D Objects in Art and Humanities: challenges of creation, interoperability and preservation," PARTHENOS, White Paper, May 2017,
[60] https://hal.inria.fr/hal-01526713v2/document
[61] C. Lynch, "Resilience and engagement in an era of uncertainty," Presentation at Coalition for Networked Information, Fall 2017 Meeting, (37:35 - 41:09),
[62] https://www.cni.org/events/membership-meetings/past-meetings/fall-2017/plenary-sessions-f17-2 .
[63] J. Grayburn, Z. Lischer-Katz, K. Golubiewski-Davis, and V. Ikeshoji-Orlati. *3D/VR in the Academic Library: Emerging Practices and Trends*, CLIR Report 176 (2019 February), p. 122, https://www.clir.org/wp-content/uploads/sites/6/2019/02/Pub-176.pdf
[64] J.-P. von Arnhem, C. Elliott, and M. Rose, *Augmented and Virtual Reality in Librari*es Lanham, Maryland: Rowman & Littlefield, 2018.
[65] J. Meyerson, "3D/VR preservation: Drawing on a common agenda for collective impact," in *3D/VR in the Academic Library: Emerging Practices and Trends,* J. Grayburn, Z. Lischer-Katz, K. Golubiewski-Davis, and V. Ikeshoji-Orlati, Eds., CLIR Report 176, 2019 February, p. 100.
[66] S.S. McMillan, M. A. King, and M. P. Tully, "How to use the Nominal Group and Delphi Techniques," *International Journal of Clinical Pharmacy*, vol. 38, 2016, 655–662
[67] https://doi.org/10.1007/s11096-016-0257-x
[68] M.D. LeCompte, and J.J. Schensul, *Analyzing and Interpreting Ethnographic Data,* Walnut Creek, CA: AltaMira Press, 1999, p. 46.
[69] E. Shein, "Who owns 3D scans of historic sites"? *Communications of the ACM*, January 2019, vol. 62, no. 1, pp.15-17,
[70] https://cacm.acm.org/magazines/2019/1/233513-who-owns-3d-scans-of-historic-sites/fulltext
[71] M. Nelson, "Blockchain can not be used to verify replayed archived web pages," Coalition for Networked Information, Fall 2018 Meeting, https://www.cni.org/wp-content/uploads/2019/01/CNI_Block_nelson.pdf
[72] D. Rosenthal, "Blockchain: What's not To like," Coalition for Networked Information, Fall 2018 Meeting,
[73] https://www.cni.org/topics/information-access-retrieval/

blockchain-whats-not-to-like

[74] A. Johnson, J. Leigh, B. Carter, J. Sosnoski, and S. Jones, "Virtual Harlem," *IEEE Computer Graphics and Applications*, vol. 22, no. 5, 2002, pp. 61-67.

[75] C.L. Borgman. *Big Data, Little Data, No Data: Scholarship in the Networked World*, MIT Press, 2015.

[76] Cultural Heritage Imaging, "Digital Lab Notebook," accessed June 25, 2019, http://culturalheritageimaging.org/Technologies/Digital_Lab_Notebook/

[77] P. Gooding and Melissa Terras, "'Grand theft archive': a quantitative analysis of the state of computer game preservation," *International Journal of Digital Curation*, vol. 3, no. 2, 2008, http://www.ijdc.net/index.php/ijdc/article/download/85/56

[78] M. Hedstrom and C. Lampe, "Emulation vs. migration: Do users care?" *RLG Diginews,* vol. 5 no. 6, 2001, http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070511/viewer/file2448.html#feature1

[79] M. Hedstrom, C. Lee, J. Olson, and C. Lampe, "'The old version flickers more': Digital preservation from the user's perspective." *The American Archivist* vol. 69, no. 1, pp. 159-187, 2006. https://americanarchivist.org/doi/pdf/10.17723/aarc.69.1.1765364485n41800

[80] J. McDonough, et al., *Preserving Virtual Worlds - Final Report*, 2010, http://www.ideals.illinois.edu/bitstream/handle/2142/17097/PVW.FinalReport.pdf

[81] P. Laurenson. "The management of display equipment in time-based media installations," *Studies in Conservation*, vol. 49, no. sup2, pp. 49-53, 2004, https://www.tandfonline.com/doi/abs/10.1179/sic.2004.49.s2.011

[82] J. Santone. "Marina Abramović's Seven Easy Pieces: Critical documentation strategies for preserving art's history," *Leonardo* vol. 41, no. 2, pp. 147-152, 2008, https://www.mitpressjournals.org/doi/pdf/10.1162/leon.2008.41.2.147

[83] P. Laurenson, "Old media, new media? Significant difference and the conservation of software-based art," in *New Collecting: Exhibiting and Audiences After New Media Art*, pp. 73-96, Routledge, 2016, http://www.academia.edu/download/44565452/Laurenson_OldMediaNewMedia.pdf

[84] T. Owens, "Preservation intent and collection development," in *The Theory and Craft of Digital Preservation*, Baltimore, MD: Johns Hopkins University Press, 2018, pp. 81-102.

Archivematica as a Case Study for Sustained Digital Preservation

Ashley Blewer
Artefactual Systems, Canada
ablewer@artefactual.com

Sarah Romkey
Artefactual Systems, Canada
sromkey@artefactual.com
https://orcid.org/0000-0003-3833-7648

Ross Spencer
Artefactual Systems, Canada
rspencer@artefactual.com

Abstract – Archivematica is an open source software platform that incorporates the community's experiences and skills to create a robust digital preservation processing system. This paper describes Archivematica as a case study for sustained digital-preservation software. This paper briefly covers the software's history before exploring in more detail the components that make Archivematica what it is today, taking in some of the benefits and some of the current limitations of the approach. Looking to the future these limitations are also discussed, as well as other considerations about how the system, and how the software can move forward alongside the entire digital preservation community.

Keywords – Digital preservation; open source software; maintenance

Conference Topics – Designing and Delivering Sustainable Digital Preservation; Building Capacity; Capability and Community

I. History

In 2008, Peter Van Garderen had an idea for a "simple" filesystem-based digital preservation system. Archivematica has since grown into a comprehensive digital preservation system.

Archivematica's goal was introduced at the iPres 2010 conference [1], which was stated as to "reduce the cost and technical complexity of deploying a comprehensive, interoperable digital curation solution that is compliant with standards and best practices." In a few years, the software grew to encompass "system scalability, customization, digital repository interfaces, format policy implementation, and a business plan that stays true to the ideals of the free software community" [2].

Archivematica uses a microservices design pattern to provide an integrated suite of software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model. Users monitor and control the digital preservation pipeline via a web-based dashboard. Archivematica uses METS, PREMIS, Dublin Core, the Library of Congress BagIt specification and other best practice standards and practices to provide trustworthy and interoperable archival information packages (AIPs) for storage in your preferred repository. As a product, Archivematica seeks to "not re-invent the wheel" but rather leverage existing utilities by wrapping them in a fluid digital preservation system. Archivematica is made up of services that any institution could seek to run on their own, but one of the goals of the product is to lower the barrier to entry for digital preservation, and create a system where all of these services work in congruence with each other.

Archivematica's sustainability model was addressed in the 2012 iPRES proceedings by Peter Van Garderen and Courtney Mumma, in their paper "The Community driven Evolution of the Archivematica Project": "The free and open-source, community-driven model provides the best avenue for institutions to pool their technology budgets and to attract external funding to continue to develop core application features as requirements evolve. This means the community pays only once to have features developed, either by in-house technical staff or by third-party contractors such as Artefactual Systems" [3].

II. Dependencies

Archivematica as a software has been a work continually in progress of development for over a decade, and has grown from a small grant-funded

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

"backend for AtoM (Access to Memory)"[1] into a robust, standards-based digital preservation system. This paper seeks to disclose the current practices for development and long-term support for Archivematica by its primary caretaker, Artefactual Systems, as a transparent case study for sustainable open source digital preservation systems.

Supporting software requires much and many different kinds of work. There are dependencies both in and around software systems. These dependencies in software form a functional system much like cogs in a machine or groups of people -- everything must function together and thusly depend on each other for success.

Software systems are written in one or more specific programming languages and sometimes a framework for those languages. The core structures of Archivematica is written in Django, which is a framework for writing web applications in Python. Just like Archivematica as a software requires regular maintenance and support, the languages and frameworks of which software is composed of also require a certain level of maintenance. Updating to the latest versions of programming languages and associated frameworks within a software system are essential maintenance tasks.

Archivematica typically runs by itself on a dedicated server or virtual machine, and between Archivematica and hardware is the Linux operating system. This operating system also requires regular system updates for maintaining the security and integrity of the platform on which Archivematica and all of its dependencies run. Falling behind on these updates can affect how the software performs in the greater computing environment, whether on the web or in a locally installed application. Software, programming languages, and operating systems all often have an "end of life" or "end of support" date for given versions, and it is important to heed those dates and perform appropriate updates. The farther along a software, language, or system develops, the harder it can be to perform an update from a software's current version to its latest version

Archivematica as a software is particularly dependency-heavy due to its primary design philosophy. Archivematica's intention is not to reinvent any wheels, so it utilizes existing tools to perform most of its core preservation tasks (such as file identification, virus scanning, characterization, normalization, et al). Archivematica as a software intends to provide a wrapper around these preservation services and operates as a workflow engine that allows users to run preservation microservices in a logical, standards-based, and recommended sequence. Each of these core preservation tools follow their own development cycles, and in addition to the maintenance needed to upkeep itself, Archivematica must consider the upkeep of the tools that run within it.

These technical considerations are all part of what can be described as "maintenance," which has gained traction as a worthwhile topic of discussion and reflection in technical communities[2] All of the non-technical components that go into enhancing software (writing and documenting requirements, training, writing documentation) are also a crucial part of the maintenance and long-term sustainability of software projects and initiatives.

[1] An open source software for archival discovery.

[2] See, the Maintainers Research Network: https://perma.cc/68WW-M9L8

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

III. Artefactual's Role as Software Steward

As stewards of the software, Artefactual tries to guide sponsoring institutions in software development projects towards more generalizable workflows for both ease of maintenance and community benefit. Because it is not feasible to support all features forever, sometimes tough decisions have to be made. The project team may decide to depreciate a feature if it has fallen into disuse and removing it is easier than maintaining it. For example, the project team is currently depreciating the integration with Archivists' Toolkit.

One major aspect of this is Artefactual's core business model, the traditional "bounty model." This bounty model means that institutions can sponsor development of new features or fixes that they require for their workflows, and these features are then rolled into the software in a future public release. The bounty model allows institutions to asynchronously share resources, and allows smaller, less funded organisations the ability to benefit from features requested by larger, better funded organisations. The bounty model funds what features and functionality people want in Archivematica, so it leaves feature prioritization development in the client's hands. The tension of innovation versus maintenance is a prominent part of the management of funding the software and keeping it healthy and maintained.

Artefactual incorporates a fee of 10% as a "Community Support fee", added on to most development contracts specifically to support the ongoing code maintenance: activities such as resolving merge conflicts in the code, documentation and regression testing. The fee also emphasizes to development sponsors that they are part of the commitment for the maintenance of the feature Other than this maintenance fee, fundraising to pay for code maintenance or technical debt derived from the bounty model is a non-trivial task. It is difficult not just for Artefactual, but also for clients to convince their administrators to approve funding maintenance work in this context.

Artefactual serves as the stewards of the Archivematica software but the community stretches far beyond the reaches of Artefactual. Many Archivematica users have never interacted with Artefactual. This is one of the nice things about open source software, but it can be difficult to gauge community interest broadly. This creates a tricky dynamic that is difficult to balance for Artefactual -- wanting to be contributing members of the community but not be the sole drivers of the future of the software.

Recently, Artefactual has begun developing a more open and organized road map for Archivematica. The hope is to achieve better management of maintenance tasks through better planning. A roadmap can take many forms and Archivematica's remains a work in progress. A Trello board[1] is used to track enhancements and fixes which either have a milestone for an upcoming release of the software, or remain on the "wishlist". Roadmap items are intended to be a broader description than an individual issue; in agile methodology they would be referred to as "epics" [4]. The purpose of describing epics on the roadmap is to allow the user community a window into what they can expect from the software in the future at a high level, and also for them to see, and possibly contribute to gaps in functionality.

IV. Funding Software

A. The Traditional Bounty Model

As mentioned, Artefactual's primary method of developing the Archivematica software has been the bounty model -- one institution (or several in collaboration) pays for a feature which is then incorporated into the project for general use. Taking a look at a typical estimate is broken down for a very small feature or "bugfix" in Archivematica[2], the complexity of any change to the system is apparent:

[1] See, the Archivematica Roadmap: https://perma.cc/6K-KT-BREV

[2] Numbers are only representative of a small-feature or fix. As an example of what might be out of scope in these estimates might be the development of a new microservice. Numbers can also vary from case-to-case. As an example, documentation might already be largely complete for a feature for which a bug-fix is being submitted, so might not need adding to or enhancing.

TABLE I
Sample Estimate for Archivematica Development Project

	Step	Who	Estimate (Hours)
1	Analyze the requirements of the feature or the fix.	Analyst	6
2	Create a pull-request to satisfy those requirements.	Developer	12
3	Create appropriate unit tests.	Developer	6
4	Code review	Development team	2
5	Seek approval on product milestone	Product Owner	0
6	Rebase and merge with QA branch	Developer	0.5
7	Update or create new documentation	Analyst, developer and development team	6
8	Update sample data for testing	Developer	2
9	Release-candidate made.	Development team	Part of product release
10	Verification testing	Analyst team	Part of product release
11	Regression testing	Analyst team	Part of product release
12	Additional feature documentation, e.g. delivery of a screencast to users.	Analyst or project team member	4
Total			38.5 hours

As demonstrated above, there is rarely any part of an estimate that might be characterised as a quick fix. The number of people involved in the creation of a software change is also quite high. Not represented in the tasks above include the work done by Artefactual and a client required to get to the stage of creating an estimate, internal and external project meetings along the way, and other administrative overhead associated with work. The tasks that fall under "Part of product release" are, in theory, funded by the 10% Community Support fee that Artefactual adds to each development contract.

B. An increasingly agile bounty model

A number of projects in recent years have adopted more agile approaches for Archivematica development, such as buying hours to be used toward a final goal, but without that goal being 'fixed', as in the "waterfall model" of project management alluded to in the above example. Bentley Historical Library sponsored a number of features to support their inhouse workflows and took an iterative approach to the development, allowing developers and users to collaborate on requirements as the project progressed. In another model, Wellcome Collection has improved testing in Archivematica by sponsoring development work around its automated test suite. Wellcome and Artefactual have worked together to define goals in roughly bi-weekly sprints based on what was felt was needed and could be achieved in that time with the resources available -- in contrast to setting a strict goal up-front. In the future it is hoped that organisations will continue to sponsor development in iterative ways that may contribute to sustainability of the project through better testing, dependency upgrades and the like.

C. Alternative approaches?

With the bounty approach Artefactual runs the risk of developing features most pertinent to larger, well-funded organisations and neglecting the input of smaller organisations. With the agile approach there are risks too. One such risk is that the final output might not be 'all that was envisioned at first'. That being said, one of the benefits of an agile approach is that it is better able to manage the uncertainty involved in any development process. An example of this might be having to react to develop a new programming library to deal with a type of data that was previously unanticipated. Perhaps, the implication for an organisation working with Artefactual this way is that they are just happy to push Archivematica in a forward-direction. Benefits

of agile for the clients that Artefactual work with are:

- The creation of burn-down data in support of decision-making to create a greater trust around estimation in future.
- Greater freedom to affect the small changes along the way, bug-fixes, patches, documentation, release-packaging, etc.
- An incremental, but systematic approach to feature development, that without a fixed end-point, allows the feature to evolve as its use is further understood. Something that can be lost when a feature has to travel a fixed path from point inception to implementation.

D. Contributions

Receiving contributions of code, documentation and other community participation is a marker of health for any open source project. In this respect Archivematica has been less than healthy in the sense that the vast majority of code for the project has been written by employees of Artefactual Systems. Artefactual has begun to address this problem through collaborations with partners; namely, an Memorandum of Understanding (MOU) and a project sponsored by Wellcome Collection has opened the doors to code contributions through collaborative code review and increased automated test coverage for the code (thereby making it easier to accept code from an external contributor).

A common question faced by Artefactual relates to the chosen open source license for Archivematica. All Archivematica code is released under a GNU Affero General Public License (A-GPL 3.0) and associated documentation is also released under a Creative Commons Share-alike license. The decision-making around this choice is asserted in the Archivematica Contributor Agreement, which all contributors are required to sign:

"One of the key challenges for open source software is to support a collaborative development environment while protecting the rights of contributors and users over the long-term. Unifying Archivematica copyrights through contributor agreements is the best way to protect the availability and sustainability of Archivematica over the long-term as free and open-source software. In all cases, contributors who sign the Contributor's Agreement retain full rights to use their original contributions for any other purpose outside of Archivematica, while enabling Artefactual Systems, any successor Foundation which may eventually take over responsibility for Archivematica, and the wider Archivematica community to benefit from their collaboration and contributions in this open source project."

V. Maintenance

Beyond new feature development, software projects needs to be patched, fixed, upgraded, debugged and monitored. In addition to that, processes and regulations for taking these actions need to be addressed, maintained, and supported [5].

From the Maintainers conference in 2016, Nathan Ensmenger presented on the "unexpected durability of digital technologies," and found that in studies very little of maintenance goes into what we would think of as bug fixing (e.g., making sure the software works the way we expect it to). He wrote that, "The majority of software maintenance involve what are vaguely referred to in the literature as "enhancements"... This included the introduction of new functionality, as dictated by market, organisational, or legislative developments" [6].

A. Addressing technical debt

Increasingly estimates for Archivematica projects look to manage technical debt upfront -- making it as clear as possible to clients that a feature is only a small percentage of what is involved. Alluded to also is the maintenance cost built into development contracts on-top of what is already estimated.

Contributors engaging with Archivematica via the GitHub organisation are presented with a list of acceptance criteria designed to tackle technical debt. These criteria need to be satisfied to see a contribution accepted by Artefactual. The practice is followed inside Artefactual as well.

The Archivematica Contributing guidelines [7] describe coding standards that should be adopted when developing a fix or a feature which in part, aims to avoid technical debt. Artefactual could go further such as in these examples:

Maintaining calendars of dependency end-of-life dates and building the time and financial dedication required for updates into the software release cycle

- Allowing space in the development schedule to remove "dead" code and simplify the codebase
- Paying attention when a developer says "I could make our lives so much easier if I had time to do X." Often the most vocalized needs are user facing but if an adjustment to the code makes the developers' work easier it ultimately results in a better, more maintainable product.
- Balancing new feature development with known maintenance cycles. For example, by leveraging agile rituals, a project could devote a sprint (or more) to focus on upgrades and maintenance work.

B. Accept some depreciation

Decision-making around the maintenance of software means accepting a level of depreciation over time. As mentioned previously, depreciating features is possible and does happen. Questions around deprecation are around how to perform this thoughtfully and with appropriate consultation of the community. As an example, Artefactual has put forth to the Archivematica community the decision to depreciate Archivists Toolkit support in Archivematica, based on the perception that most Archivists Toolkit users have moved along to using ArchivesSpace or other tools.

VI. Recent developments

In 2018, Artefactual opened a single Issues repository[1] to track all known issues or bugs, whereas previously they had been spread amongst many code repositories. Artefactual invited members of the digital preservation community to join the GitHub organisation and the Issues team with the goal of allowing the larger digital preservation community to take part in the process changes too, opening up the conversation. Labels are used to provide consistency and guide movement through Artefactual's public kanban. Labels help determine the milestone (release timeline) of a project. For the community, organisations are invited to add their own label so that they may signify interest in an issue and see how that work progresses.

Within the Issues tracker, Artefactual asks all issues be described as 'problems'. This started as an internal regulation but stems from parts of the broader open-source community, to help focus the attention of the writer on the problem that needs to be solved and not just adding what they believe to be the solution. The hope is that this increases engagement amongst everyone involved or interested in a particular issue.

VII. Keeping focused

It is fair to say that the interest in improving Archivematica comes from a few different disciplines. The ability to deploy with ease might come from a systems administrator's perspective. Asking to use the most up-to-date programming libraries might make a developers life easier; but the original use-case shouldn't be forgotten while many different streams of work might happen parallel to each other.

As a demonstration, Artefactual is continuing to seek out ways to increase Archivematica's ability to scale horizontally and vertically [8]. The end result of any software optimization would mean very little to a community of digital preservationists if the outputs (system-agnostic archival information objects [AIPs]), are somehow rendered unhelpful, or worse, incorrect. The hypothesis of digital-preservation is that the AIP is the object that will be taken into the future. As such, when all the optimizations that make Archivematica as fast as possible, or as efficient as possible, are complete, then the details of the process by which the material was processed and the AIP was made still need to be reflected in the preservation metadata.. Users must still be able to trace the original-order, and order-of-activities that happened on transfer objects so that their chain of custody remains unbroken.

Tackling this in software, a collaborative, multi-company community project such as the Preservation Action Registries[2] (PAR), alongside

[1] See, Archivematica's collected issues repository: https://perma.cc/TC24-PHT8

[2] See, the RDSS-PAR Project: https://perma.cc/A2AM-ZRTD

OPF, Preservica, and JISC, represents an important mechanism for interfacing with external tools. As more preservation, or preservation adjacent systems interact with a common interface such as PAR, then there are more eyes on a consistent digital preservation community's ecosystem of tools and utilities. Through the PAR more organisations can take responsibility for maintaining the integrity of the output of digital preservation tools, their performance and their reliability, thus improving them for the benefit of a whole.

VIII. Looking forward

As Archivematica's primary maintainers, Artefactual believes that the sustained way forward for the Archivematica project is through community involvement. Artefactual wants to see as many people preserving digital heritage in the most sustainable way possible; in order to achieve this end Artefactual needs to know its team is focused on the features and workflows needed the most. Meaningful community involvement also means lifting the veil on software development practices and challenges, asking for help when needed, and being transparent about what they can and cannot do.

Some ways in which Artefactual have been working to improve the practices around community development in Archivematica not included in this paper include:

- Developing better practices around release cadences e.g. by being more deliberate in managing our release scope, and looking back on what was achieved, will make releases more predictable and build trust with the user community.
- Understanding more about how users want to use Archivematica and what they want to use it for, challenging assumptions encoded in the system and understand how to rewire it to improve on those.
- Identifying and filling missing community roles. Identifying who is doing the work, how this work is being organized, and who is the advocate for maintenance -- Is it Artefactual, other companies, clients, non-clients, or end users?
- Figuring out how to talk about community and maintenance, such as what kinds of language and in what situations (public or private, at conferences or online, et al).
- Negotiating points of conflict, such as in a pull request to the codebase or architectural decisions.
- Developing better automated testing practices. Automated integration looks beyond what an individual script might do in the context of a workflow to the output of the system as a whole.
- Building 'maintainer' capabilities across a broader number of community members and companies so that aspects of a release such as code-review are not automatically a bottleneck for community submissions.
- To foster greater developer contributions Artefactual must improve the consistency of its software development practices which, in-part, will come out of testing, and developer-documentation; but it will also rely on standardizing interfaces to key parts of the system, where currently there are the many idiosyncrasies of Archivematica's many past developers to be seen.

These are themes that are appearing when Artefactual talks to clients or reviews the output of user groups and forums, such as the 'Archivematica User Forum'[1]. The notes taken in this type of forum are invaluable to identifying where opaqueness exists and where Artefactual can improve, in communicating intent, in engaging with contributions, and in directing the roadmap.

IX. Conclusion

Like the field of digital preservation, Archivematica is still young. And like digital preservation practices, Archivematica's are still evolving. As many assertions as this paper makes, it is hoped that the spirit that comes across is one of 'community'. In the presentation created to support Mumma and van Garderen (2012) [4] one of the more humorous slides might be considered as the one of 'Mr-T' making it very clear that 'WE NEED YOU'. And this hasn't changed. In 2012, Mumma and van Garderen list ways to contribute:

[1] See, Archivematica User Forum, Call for Participation: https://perma.cc/GNB7-WD5Q

- Discussion: ask questions and provide opinions/suggestions on project list
- Support: answer questions on the discussion list
- Dissemination: blog, tweet, present, train
- Documentation: manual updates, wiki gardening
- Testing: report bugs, request features
- Development: fix bugs, contribute patches, write Plugins
- Maintenance: provide technical services

And seven years later, the entire digital preservation community and these methods are still very much at the core of what makes Archivematica a sustainable digital preservation project.

Archivematica has and will continue to be a resource driven by the digital preservation community. The aforementioned projects in 2018 and 2019, featuring collaborations with the organisations of JISC, Wellcome, and PIQL, will allow Archivematica to move beyond the current development bounty-model and support contract models performed primarily by Artefactual Systems to something larger, healthier, and more robust.

Acknowledgements

Thanks to everyone who currently or previously has worked at Artefactual. Hillel Arnold and Erin O'Meara for their thoughts organized during their DigiPres presentation "Acts of Maintenance". To the various organisations that have sponsored features development in Archivematica. To those who provide technical services. Finally, to those who blog, or present, or train, about using Archivematica; supporting the ecosystem through whatever resources are available to them.

References

[1] P. Van Garderen, "Archivematica: Using Micro-services and Open-source Software to Deliver a Comprehensive Digital Curation Solution," iPRES 2010. https://perma.cc/2Z8V-RBYX

[2] C. Mumma, P. van Garderen, "Realizing the Archivematica vision: delivering a comprehensive and free OAIS implementation," iPRES 2013.

[3] C. Mumma, P. van Garderen, "The Community-Driven Evolution of the Archivematica Project," iPRES 2012. https://perma.cc/RWN9-QVU8

[4] Agile Alliance, "Epic," https://perma.cc/LM5X-S437

[5] H. Arnold, E. O'Meara, S. Romkey, "Acts of Maintenance," DigiPres 2018. https://www.slideshare.net/Archivematica/acts-of-maintenance

[6] N. Engsmenger, "When Good Software Goes Bad, The Surprising Durability of an Ephemeral Technology," The Maintainers 2016. https://perma.cc/H3R2-8FEL

[7] Artefactaul Systems, "Contributing." https://perma.cc/2TCP-6JPK

[8] N. Shalom, "What is Horizontal and Veritcal Scaling?" https://perma.cc/979P-674B

Preservation Planning, Beacons for a TDR

Three Cases on Archiving Emerging Media

Marjolein Steeman
Netherlands Institute for Sound and Vision
The Netherlands
msteeman@beeldengeluid.nl

Abstract – In order to preserve digital objects for the long term repositories need to choose a preservation strategy. For new emerging types of media this is a challenge. This paper describes how various cases occurred at the Netherlands Institute for Sound and Vision. It shows how preservation planning helps management in putting these matters in the right context and taking informed decisions based on knowing what we know now. It concludes with an overview of the content of a Preservation Plan, as has been implemented in practice.

Keywords – preservation planning, trusted digital repositories, new media, audiovisual heritage, AV archives

Conference Topics – Designing and Delivering Sustainable Digital Preservation; Exploring New Horizons.

I. Introduction

The oldest existing lighthouse of the Netherlands is called the Brandaris. It stands at the eastpoint of one of the isles in the North and dates from 1592. Many years it served as a beacon to guide ships from the far east, west and from the northeast to Amsterdam. And also to guide them on the way out.

This paper is called Preservation Planning, beacons for a TDR (trusted digital repository). It shares recent experiences on preservation planning at The Netherlands Institute for Sound and Vision (NISV). Especially with the emergent new media formats in our modern information society. It will present how preservation planning is put into practice in this institute and how it serves as a beacon that helps guiding the ingest of and access to media works in our repository.

II. Preservable formats

Sound and Vision is an independent media institute that holds a heterogeneous collection including the public broadcast archives, education and science collections as well as amateur and independent works. The archive stores more than 1 million hours of digital AV material and also at least 20,000 objects and over 2.5 million photos. The institute is a museum, an archive and a knowledge institute. In 2016 the Data Seal of Approval was granted: a certificate for trustworthiness of repositories.

At Sound and Vision the complete archival storage contains 34 petabyte of files. These 34 petabyte are used by only a few different file types. Dpx and wav files (40% of the used capacity) are used to store our digitised film. A tiny part of the storage consist of tiff files, representing the photos. Wav-files (4% of the storage) are used for audio and mxf-files (55% of the storage) for video. Overall, Sound and Vision has only four preservable file formats in its repository.

Only content that is presented in or will be digitised to one of these preservable formats, qualifies for full preservation. Other formats are not accepted because the longevity can not be guaranteed. This is called a “just in case” policy.

The preservable formats have been described in detail in a Preservation Metadata Dictionary (PMD). This PMD is the first product of our preservation planning activities. It is used as a reference for new ingest: what technical metadata must be provided and what characteristics are allowed. Also: via a systematic mapping it records where the characteristics are documented in the repository systems. The PMD is conformant to level 1A of PREMIS and it

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

is recognised as essential information for NISV as a trusted repository [1]..

All principles and choices for execution of the business of sustainable digital preservation have been outlined in a policy document [2]. By documenting the current policy and the standards employed, it is possible to account to all parties that entrust their digital collections to Sound and Vision, and to offer the staff of Sound and Vision transparency and clarity on the rules and procedures that apply.

III. Emerging media

But what if new media, new formats, new requirements come into play? To answer this question the following case is exemplary.

A. Webvideo

In 2004 Vimeo, the first big webvideo platform arrived. Soon followed by YouTube in 2005. A few years later Sound and Vision did research on the options for archiving webvideo, followed by some internal projects and an exhibition in 2016. In 2018, the institute decided it was ready to store webvideo in its trusted archive, as the following terms had been met:

1. Our mission is comprehensive: "Sound and Vision wants to improve everyone's life in and through media by archiving, exploring and clarifying that media". Webvideo is definitely within scope.
2. We recognised the Internet is great for sharing, but it is not an archive: we sure must take on our role here.
3. We developed selection criteria for webvideo that should cover the new Dutch media landscape of webvideo.
4. Agreements were made with rightsholders of the videos on archival services and on publishing in specific context.
5. Tooling had been implemented to gather metadata from the web.
6. And last but not least a proposal was made for a new preservable format.

This was when preservation planning was allerted. To get a full understanding of this proposal, let us first give some context.

The current preservable format for video is an MXF-wrapper with D10-30 or D10-50 videocodec. D10 is an implemented MPEG-2 codec used in production workflows for digital television. It is an industry standard, well documented and widely supported. The MXF/D10 is transcoded to a proxy for viewing or dissemination.

But the codec uses a bitrate of 30 or even 50 Mbps. Where the webvideo comes in max. 2,5 Mbps. This means that transcoding all webvideo to MXF/D10 would inflate the size of the files. An unwelcome effect. Also: the MXF/D10 isn't lossless; it is lossy. Transcoding a lossy compressed file (webvideo) to another lossy codec is far from ideal for preservation.

Therefore webvideo team proposed the introduction of a new preservable format: an MP4-wrapper with an H264 codec. This seemed a plausible proposal. A lot of webvideo nowadays has exactly this format, so transcoding would then often not be needed.

However from a preservation point of view, one might question this option. The MP4/H264 might be widespread at the moment, but for how long? H265 with even better compression is coming up. Also: H264 defines the codec, but there are a range of other file-characteristics that may have implications for access or playout. And on top of that: again it is a lossy compression. Transcoding may have impact on the quality of the file, which is ofcourse undesired.

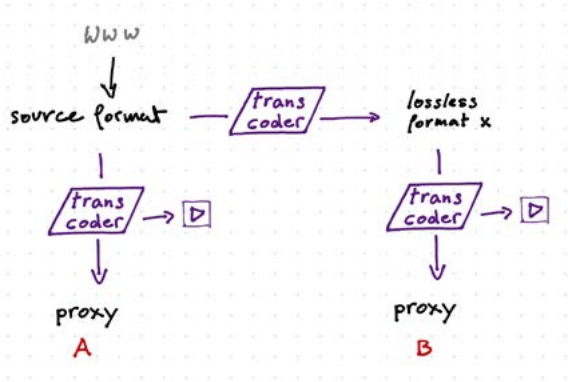


Figure 1 New scenarios A and B for webvideo.[1]

So instead a scenario for a lossless format was

[1] all illustrations by M. Steeman/NISV licensed under CC BY-SA

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

made. A new preservable format that is trusted to stand the test of time. That is: to live long and, or migrate well. The suggestion (fig.1) is to archive the source format "as is", provided that our systems can create a proxy and support playout. If not, the source format is transcoded to a lossless format "x". From this a proxy can be created infinitely.

Once the source format has been accepted (Scenario A), there is no immediate need to transcode to a lossless format. The source format can be disseminated or the proxy itself as a standard derivation.

There will be a need to keep monitoring though. If the source format threatens to become obsolete then still a lossless archival master must be created (Scenario B). This will in fact depend on evolutions in the playout environments. As a starting point the internal transcoding software will act as a reference for the playout environment. It has been provisionally agreed to that when new versions of this software cease to support certain outdated formats, this calls for action.

Transcoding to a lossless format will probably also inflate the size of the file. But instead of inflating all files, this will only happen when it is relevant to do so. In other words: a "just in time" policy is applied instead of "just in case".

B. Getting our Bearing

From just in case, to just in time. This is an essential addition to the NISV preservation policies. It opens up the archive for new media that so far were put aside on separate disks, where the risks of not being properly looked after are eminent. It also introduces a new operational practice, following scenario A or B. And it sheds light upon the issue of obsolescence. In particular how this risk must be monitored.

The scenarios were documented in a preservation plan, that was presented to the NISV preservation board. It was important to have their consent, before the consequences of the policy were worked out. Even more important: making this preservation plan, together with all internal stakeholders, indeed helped Sound and Vision to retrieve its bearing with respect to preservable formats. Preservation planning operated as a true beacon and put us back on track.

VII. Preservation planning

The case ends with drawing up a "preservation plan" to underpin the new policy on preserving webvideo content. How does this relate to the latest standards on preservation planning?

A. Planets and OAIS

Becker c.s. [3] make an important distinction between concrete preservation plans and high-level policies. It is claimed that a preservation plan is seen on a more specific and concrete level and Becker refers to the definition as was adopted by the Planets project: "A preservation plan defines a series of preservation actions to be taken by a responsible institution due to an identified risk for a given set of digital objects or records (called collection)" [4]. The preservation actions are specified, along with responsibilities and rules and conditions for execution on the collection.

The Planets preservation workflow consists of four phases:

1. Define requirements
2. Evaluate alternatives
3. Analyse results
4. Build preservation plan

In this view the preservation plan is right at the end of the process of working out all details. The definition speaks of 'preservation actions'. The preservation plan contains an executable workflow definition to perform a specific migration on a specific set of records or files. However, in the case of webvideo the preservation plan documented the recommendations to the board on how to approach this new preservation case. This implies a more generic plan, proposing new policy guidelines.

The Open Archival Information System (OAIS) [5] is a widely accepted reference model to become a so called Trusted Digital Repository. Preservation Planning is one of the entities of the OAIS functional model.

Preservation Planning is linked to the entity of Administration, that contains the services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis. Administration functions include maintaining

configuration management of system hardware and software. It is also responsible for establishing and maintaining Archive standards and policies.

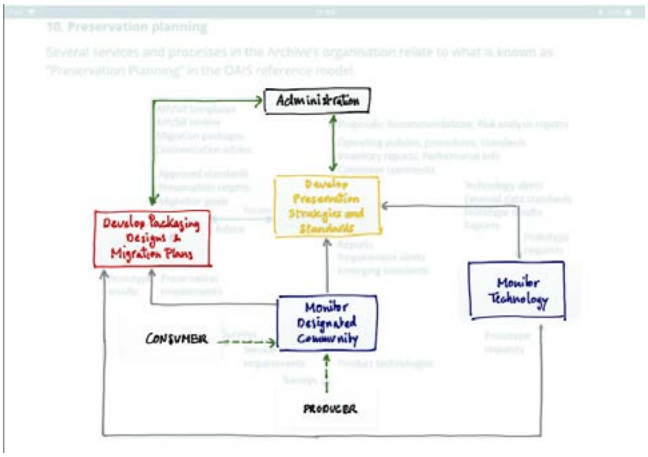


Figure 2 Functions within Preservation Planning according to OAIS.

Fig. 2 shows this relation between Preservation Planning and Administration more in depth, by unfolding Preservation Planning in the composite functions. Preservation Planning consists of four functions. The webvideo case seems to fit very well in one of these functions: developing preservation strategies and standards (yellow).

To develop packaging designs and migration plans (red) refers to more operational planning. This function is more in line with the concept of Planets. It delivers a detailed timetable of actions.

Both have a relationship with Administration but in a very different way. Developing strategies and standards relates to management that establish the policies and make decisions on scenarios or options. Where packaging designs and migration plans are input for System Configuration, the operational level of Administration.

In the workflow presented by Planets, the first three steps are said to be compliant with Develop Preservation Strategies and Standards. The outcome is provided to the Develop Packaging Designs and Migration Plans function as advice to create a detailed migration plan.

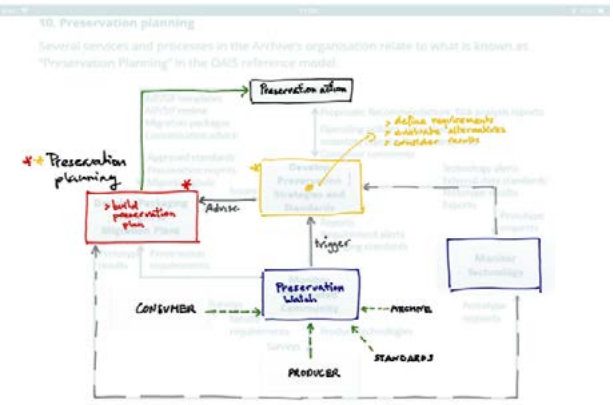


Figure 3 Planets workflow within OAIS.

It is evident that the Planets workflow is very straightforward. Preservation watch leads to testing and evaluating, resulting in an advise. A detailed plan is built and carried out by Preservation Action. Management is not involved explicitly. Policies seem already set and covered.

Given the experiences at Sound and Vision both planning functions are not necessarily part of the same workflow. The “Preservation Plan” that documented the additional policies on webvideo is the outcome of Develop Preservation Strategies and Standards. This plan is explicitly presented to Administration. A detailed action plan on a given set of digital objects would rather be referred to as “Migration plan”.

Using the metaphor of the beacon, “developing preservation strategies and standards” can very well be the lighthouse that guides the ships at the horizon. Where “packaging designs and migration plans” are like the mooring buoys that are placed to navigate between shallows or along the fairway at a particular location.

B. Triggering a Preservation Plan

The two other functions of OAIS preservation planning, are monitoring functions (blue). First of the designated community (consumer, producer) and secondly of technology (file formats, standards, tooling etc). Both give input to the yellow Strategies and Standards and to the red Develop packaging designs and migration plans.

The difference between the two can be illustrated by an example. A topical issue right now is the fact that production technology in broadcast

environments is changing gears towards 4K. Makers create files in 4K. MXF/D10 might not be adequate for those producers. Our consumers might no longer be happy with an - in this respect - inferior standard.The following questions arise:

- what do we know about the production context?
- how widespread and fast is this change?
- will the broadcasters come up with a high resolution standard broadcast format?
- what formats will be conventional among makers?
- and who's deciding about new standards

These questions are addressed by the monitoring function of the designated community.

From a technology point of view the new emerging formats and codecs are studied by the OAIS-technology monitoring function. They ask questions like:

- open source? how is versioning done? what about backward compatibility?
- proprietary? are there licensing issues?
- how do new codecs perform in terms of transcoding speed?
- will our own infrastructure and tooling be able to adopt the new format?

With these two monitoring functions the repository builds up knowledge. The aim is that this knowledge is adequate to give a timely and substantiated advise on which preservable format to choose. The urgency of the issue in combination with the comprehensiveness of the knowledge, will trigger the preparation of a preservation plan to introduce this new format to management.

The two monitoring functions can trigger a preservation plan in several ways:

- Producer: new production technology, new collections
- Consumer: new requirements for playout
- Archive: new collection profile, priorities in budgets, outcomes of self assessment
- Standards: new opportunities or risk alert (obsoletion)

Monitoring implies an ongoing activity. The outcome is always temporary; based on current findings. But in terms of risk management the outcome must be assessed and sometimes calls for action. Then preservation planning must document the options and give advise, thus presenting a preservation plan. In some cases this will give rise to a specific migration, but certainly not necessarily.

V. Requirements for preservation

Making the effort of drawing a preservation plan offers the opportunity to think through the preservation challenge as it emerges as exemplified in the webvideo case. This will be further illustrated by the following two other cases.

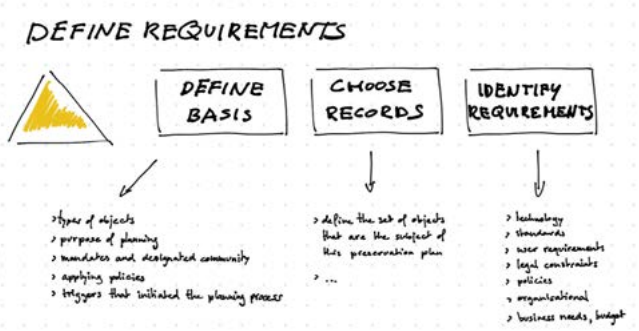


Figure 4 Steps within the first phase of the Planets workflow.

In terms of scope it will turn out that drawing a preservation plan has much similarity with very first phase of the Planets workflow, "define requirements". This will become apparent when the outline of the NISV preservation plan will be given, at the end of this paper. It is interesting to note that this phase is followed by the definition of alternatives and a Go/No-Go. Perhaps this is the parallel with presenting the plan to the NISV preservation board.

A. GIF - Graphical Interchange Format

Recently it was decided that Sound and Vision wants to include GIF images to the collection. This triggered preparing a preservation plan.

First the technical aspects of the GIF-format were investigated. GIF was introduced in 1987 by CompuServe; it's history goes back to the start of the internet. It became popular because it used a very efficient compression technique. Many pictures could be downloaded rapidly, even on slow connections.

The extreme limitations of the GIF format and the restrictions of websites that display them played a vital role in the way GIFs were made, with makers tweaking the size and color palette as well as editing frame-by-frame to make the best-looking, smallest possible file. All that nuance can disappear if the archive is not careful to preserve both the GIF itself, and the context of its creation.

For instance: rendering a GIF on current browsers might not give the same result as the original. Some users (like exhibition curators) might even go so far as fully emulate old hardware to ensure that variables like CPU speeds or screen technology don't mess up the visual representation the artist intended.

To avoid this, the GIF may be transformed into a video file. But there is a significant risk that this may change the way the GIF appears, caused by misinterpretation of instructional metadata, or by the introduction of color shifts or even potentially compression artifacts through the process of encoding as video.

One final point of consideration when rendering GIFs from the early web: it is often the case that these GIFs play back at a faster rate today, as they were limited by the slow CPUs of the time of their creation. Employing emulation to view historic Animated GIFs in something close to a period specific CPU, operating system, and web browser is therefore often recommended.

This short introduction illustrates there are at least two options on preserving GIFs in a repository. First: one could add GIF as a preservable format. This would imply that the minimum set of metadata for GIF would be documented in the Preservation Metadata Dictionary (PMD) , together with a mapping to the NISV systems and table columns, where this metadata will be stored. There would be some research needed to define what technical metadata can assure that all the specifications to render the GIF properly, are covered. This will include some specifications of the suitable environment for rendering the GIF. Also some more insight must be given on possibilities (or necessity) of emulation.

Or, the other option is to ingest the GIF as a reference file and to create (or acquire) an MP4 that resembles the original GIF. For this option no additional preservable format is needed; the MP4 is treated as the archive master and will be preserved as any web video, as presented in the first case.

In both options, the main question is: how can we establish whether rendering the master file represents the original work? The only difference between both options is: do we assess this later, given the requirements at that point in time, or do we make this assessment now, at the moment we accept the MP4 as peer. Either way, the archive must define what significant properties it wants to preserve, for whom and with what costs.

These scenarios and their implications must (and will) be addressed in a preservation plan. As a basis the context of the plan will be described (triggered by collection policies, typology of the main designated community). The GIF-object will be explained followed by the requirements that must be met like the extension to the PMD or the procedure of consent to the acquired or created MP4.

B. Games

For GIF, emulation was introduced as a way to render the original GIF, provided you simulate the original environment. For Games emulation is the only option, as there is no working substitute for the interactive feature of the game. After all a single standard format that can represent all possible interactive user experience does not exist.

In the NISV preservation plan on games the following three requirements are included, because these will have to be met in order to preserve games in the NISV repository.

Firstly the PMD should be extended with the new preservation format for Games (disk images that hold the original game-software). Find a way to document additional content like instruction videos. In a PREMIS-schema (fig. 5) is shown how this should be done. Several rights have to be managed too. The environment is added as a separate object.

For now Sound and Vision chooses not to archive environments but it must document the characteristics, to be able to create or emulate the environment when needed.

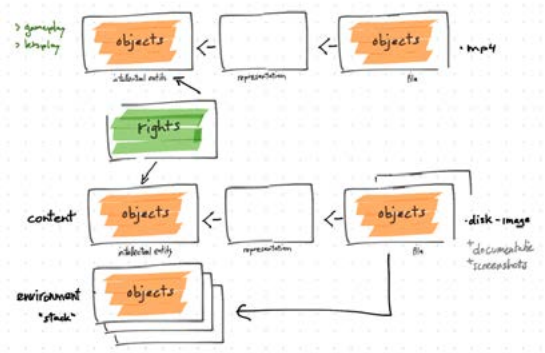


Figure 5 Schema of object categories for preservation metadata on games; based on PREMIS.

Secondly, consider that the policy on digital preservation doesn't support emulation yet. This should be added in the next update. This implies that:

- emulation must be added, next to migration
- the preservable format *disk image* should be added
- and preservation service levels as in what do we promise to preserve, should be redefined

Thirdly, to monitor the longevity of the games, NISV will organise a 5-year monitoring cycle. Once every 5 years it will check the rendering of disk-images. Is NISV still able to configure the hard and software that runs the game? Will new versions of emulators still do the job?

This brings to mind our just in time policy for webvideo: it is the same challenge: will new versions of the transcoder still be able to transcode the source files to the standard proxy? And: will this be an acceptable norm for our designated community? In a "just in time" policy NISV must somehow organise a trigger not to be too late!

VI. The preservation plan

By creating a standard table of contents for the NISV Preservation Plan, better informed decisions by management are ensured. The Preservation Plan at NISV has 4 sections.

First the outline of the context of the plan. What triggered the plan. What risks are to be mitigated; for instance legal issues, legacy or increasing backlogs. Then specific goals of the plan and the foreseen impact on digital assets already in the Archive are to be addressed.

Secondly the collection itself is described. Which Designated Community is leading, and what will be the designated use; the nature and scale of the expected ingest, the 'significant properties' of the material, and notes on selection criteria or demarcation in agreement with other archives in the Netherlands.

Third it defines what requirements are to be met. Special attention is paid to preconditions or assumptions regarding technical issues, planning, internal users (availability, competences), and internal procedures to be redesigned, implemented or just applied.

Then, at the heart of the Preservation Plan are the scenarios, followed by a recommendation. The scenarios may differ in the outline of the preservation strategy, chosen preservation formats, implications for the metadata dictionary, technical requirements, and so on.

These Preservation Plans are discussed by the NISV preservation board and as a result may lead to assignments to implement tooling, prepare specific upgrades to IT infrastructure or start prototyping a new format. Also, the outcome may be the formulation of add-ons to the preservation metadata dictionary, or even to current preservation policies themselves.

VII. Conclusion

With three cases it is shown how preservation planning at NISV plays a role in checking preparations for new ingest to standing preservation policy. And how it suggests updates to this policy. Preservation planning gives NISV archival management the opportunity to make deliberate choices on preservation. And the documentation makes these choices transparent.

The way NISV adopted preservation planning is consistent with OAIS. It differs from the implementation by the Planets project, although it certainly has corresponding elements. Especially the outline of the NISV Preservation Plan owes to the work done by this working group.

Also the way the two monitoring functions can trigger a preservation plan is very similar to Planets.

NISV has combined the two functions accordingly. This "preservation watch" is in reality an abstract state of mind and sense of responsibility of all colleagues that have knowledge of audio visual technology, in house, outside at our DC's or in the field in general. Given the topicality of preservation issues Sound and Vision will mobilise this implicit knowledge by organising meetups on these issues.

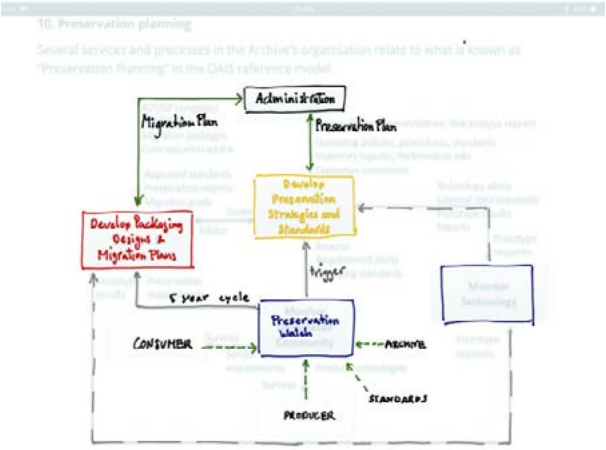


Figure 6 NISV adoption of Preservation Planning within OAIS.

The schema (fig. 6) shows how NISV has adopted preservation planning. The NISV preservation plans are triggered by risk alerts from Preservation watch. The plans, together with risk assessments, standards (like the PMD) and policies add up to the knowledge base of Administration. All operational preservation actions by Administration build on this knowledge base. Parallel on the drawing of preservation plans is the set up of migration plans. Preservation watch fosters this function by a cyclical process, like the five year cycle for the "just in time" policy.

Preservation planning is not the equivalent of a once every five year general policy on preservation. Neither it is reduced to the preparation of preservation actions on a specific set of objects. It stretches out over adjustments or add-ons to preservation policies on one side and the set up of concrete migrations on the other side. As some beacons will guide our main course with a reassuring light on the horizon, while other beacons will set out a strict direction that must be followed. Each will help us reach our preservation goals, even in poor weather or heading for unknown shores.

Acknowledgment

NISV thanks Annemieke de Jong for her work on making digital preservation a leading principle for our institute. Her outstanding work on preservation policies and TDR certification could pave the way for a organisation-wide tackling of preservation issues and inspires many colleagues, at NISV, as well as elsewhere to take up the challenge of designing longevity for our collections.

References

[1] M. Steeman, J. Schuurman, "Preservation metadata for DPX files: A case study" Journal of Digital Media Management, vol. 7, no. 1, pp. 33-45, Autumn 2018

[2] A. de Jong, "Digital Preservation Sound and Vision: Policy, Standards and Procedures," Netherlands Institute for Sound and Vision, 2016

[3] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, H. Hofman, "Systematic planning for digital preservation: evaluating potential strategies and building preservation plans", International journal on digital libraries, vol. 10, issue 4, pp 133-175, 2009, http://www.ifs.tuwien.ac.at/~becker/pubs/becker-ijdl2009.pdf

[4] H. Hofman, Planets-PP subproject, C. Becker, S. Strodl, H. Kulovits, A. Rauber, "Preservation plan template. Technical report", The Planets project, 2008, http://www.ifs.tuwien.ac.at/dp/plato/docs/plan-template.pdf

[5] OAIS2012, CCSDS, "Reference Model for an Open Archival Information System (OAIS)", Magenta book, http://public.ccsds.org/publications/archive/650x0m2.pdf, 2012.

The Integrated Preservation Suite

Scaled and automated preservation planning for highly diverse digital collections

Peter May
British Library
London, UK
Peter.May@bl.uk
https://orcid.org/0000-0001-8625-9176

Maureen Pennock
British Library
Boston Spa, Wetherby, UK
Maureen.Pennock@bl.uk
https://orcid.org/0000-0002-7521-8536

David A. Russo
British Library
London, UK
David.Russo@bl.uk
https://orcid.org/0000-0003-2829-3936

Abstract — The Integrated Preservation Suite is an internally funded project at the British Library to develop and enhance the Library's preservation planning capability, largely focussed on automation and addressing the Library's heterogeneous collections. Through agile development practices, the project is iteratively designing and implementing the technical infrastructure for the suite as well as populating it with the content required for the infrastructure to work in a business environment. This paper provides an initial description of the suite's architecture and supporting data model.

Keywords — Knowledge base, preservation planning, software preservation, preservation watch

Conference Topics — The Cutting Edge: Technical Infrastructure and Implementation.

I. INTRODUCTION

A digital format becomes obsolete because it is data that does not have the required digital environment in which to interpret and render it for human consumption. Assuming bit-level preservation is accounted for, then as Ryan [1] states, the "file format is not threatened with extinction or a discontinued existence; rather the threat is to the ability to access information from a file that is encoded in that format". The challenge lies in the availability of a *suitable environment* that is able to render a *suitable representation* of our digital object to a reader. And, as we know, digital environments — hardware and software — evolve over time.

This has led to the two common format-focussed digital preservation approaches: migration and emulation. Migration aims to provide a *suitable representation* of a digital object that can be rendered in a modern environment; as the environment landscape evolves, so must the migrated representation. Emulation, on the other hand, aims to create a *suitable environment* in which the original digital object can be rendered.

But what is the most suitable strategy to use in any given circumstance? How should one best migrate a digital object to a suitable representation? What is needed to create a suitable emulation environment?

These questions are not straightforward to answer in their own right. Simply obtaining the knowledge about the set of available migration tools for current environments can be challenging, let alone considering how to keep this knowledge up to date as environments evolve. On top of this we need to consider the sheer variation in *circumstances* for which we are trying to define our strategy. The British Library collect large amounts of heterogeneous digital content — eBooks, geospatial data, websites, audio and visual content, digitised images, eTheses, electoral register data, digital sheet music, and personal digital archives, to name a few broad categories. And this set expands as new technologies — new environments — become available.

How do we best deine suitable preservation strategies for all these collections? Does each one require a separate strategy? Or more than one strategy? When should we create strategies? When should we re-evaluate our strategies, and how often?

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Preservation planning is a core function of an Open Archival Information System (OAIS) (ISO 14721:2012), "responsible for mapping out the OAIS's preservation strategy, as well as recommending appropriate revisions to this strategy in response to evolving conditions in the OAIS environment" [2]. It encompasses a wide range of activities including monitoring the wider environment in which preservation is taking place to identify risks and opportunities which may affect the long-term accessibility of digital objects, such as new technologies or standards, as well as developing strategies for addressing these. It is the "OAIS's safeguard against a constantly evolving user and technology environment" [2]. Becker *et al.* [3] have contrasted this relatively high-level definition with the practical need for plans that could be used "for preserving a specific set of objects for a given purpose." With this approach, alternative preservation approaches are empirically tested to identify the most suitable option for the given circumstances.

But addressing these kinds of activities at scale across large heterogeneous collections, such as held by the British Library, is difficult and time consuming. As Becker *et al.* note [4] "as content grows in volume and becomes increasingly heterogeneous, the aspects of technologies that need to be monitored are by far outgrowing any organisation's manual capabilities." We need to streamline preservation planning activities and turn to more automated solutions to help minimise the burden of identifying, monitoring and addressing the risks and opportunities.

The Integrated Preservation Suite is an internally funded project at the British Library that builds upon several years of preservation activities to develop and enhance the Library's preservation planning capability, largely focussed on automation and addressing the risks and opportunities specific to the Library's heterogeneous collections. It aims to achieve this through the development and integration of several components — a knowledge base, a software repository, a policy and planning repository, and a web-based workbench — designed to meet separate but complementary goals (such as the gathering and curation of technical knowledge about formats, or the preservation of institutionally relevant access software), combined with the population of these components with content required for the infrastructure to work in a business environment. This paper provides an initial description of the suite's currently defined architecture and knowledge base data model, which will be used to help us preserve the Library's digital collections.

II. Background and Related Work

A. *Preservation Activities at the British Library*

Preservation work undertaken by the Digital Preservation Team (DPT) at the British Library encompasses many different activities. Our collection profiles, developed for all types of digital content held, were an initial exploration of what might be needed to preserve the different collection types (web archives, eJournals, eBooks, audio-visual content, digitized content, *etc.*), specifying at a high level for each collection type: the constituent formats, the preservation intent, and the known issues that should be addressed [5]. These have all recently undergone a periodic review to ensure they remain upto-date and continue to reflect the on-going evolution of the collections themselves, our curator's understanding of the collections, as well as our readers' evolving needs. From a planning perspective, such work and the resulting profiles provide useful information to contextualise a plan, guidance on what the plan should achieve (the intent), and potential issues that need to be taken into consideration (for example, colour profile considerations when converting from TIFF to JP2).

Companion and complimentary work to this included our format sustainability assessments, designed to provide a nuanced understanding of preservation risks that could feed into a preservation planning exercise alongside other business requirements such as storage costs and access needs [6]. Fed into preservation planning, such assessments could provide a useful source of preservation related risks, and when combined with format information in our collection profiles, enable further depth to collection-based risk assessments.

Wider analysis is underway to explore the threat model for our digital preservation infrastructure and to explore the relationship between these relatively highlevel threats, our understanding of digital preservation risks, the risk assessment process and

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

the preservation planning process. This work is still at an early stage and so is not elaborated upon here but will be shared at a later date as our thinking develops.

The team is also called upon at various points to assist with collection-specific preservation and access challenges. With this in mind we run a helpdesk system for colleagues in other areas of the Library to request help. Tasks vary from helping architect ingest workflows, giving guidance on the operation or debugging of validation tools such as JHOVE, performing in-depth research into suitable validation approaches, to more subjective visual assessments of content rendering (*e.g.*, EPUBs [7]). These activities typically result in new knowledge generation which can be used, or built upon, to serve subsequent helpdesk requests. Capturing this knowledge — and the evidence for it — in a way that could be used for risk assessment and preservation planning would facilitate such activities and improve transparency, and therefore trust, in the outcomes.

This wide range of preservation planning activities complements and supports the automated and formatbased preservation planning process that IPS has been designed to address.

B. Related Work Elsewhere

Several initiatives have worked to create reasonably automated systems which help monitor the preservation environment and provide means to instigate some form of preservation planning, such as the Automated Obsolescence Notification System (AONS) [8] and its successor, AONS II [9], the DiPRec system and its associated File Format Metadata Aggregator (FFMA) [10][11], and the SCAPE project's Planning and Watch suite [12] which comprises three *independent* tools to characterise a repository (c3po[1]), monitor the wider environment (Scout[2]), and develop preservation plans (Plato 4[3]).

Largely, these approaches follow the same broad concepts: external information is aggregated into a knowledge base; an organisation's repository is profiled to determine characteristics of its contents (*e.g.*, formats); all this information is compared and used to notify an administrator of potential risks or opportunities; which leads to preservation planning being initiated.

AONS I used information from PRONOM and the Library of Congress' sustainability of digital formats registry to help identify when objects in a user's repository were in danger of obsolescence and notified repository administrators. AONS II refactored the system to work with an adapter based architecture, facilitating the import of data from other file-format information sources [9]. Similarly, FFMA links together knowledge from different publicly available data repositories (initially: Freebase, DBPedia, and PRONOM) and uses this to make recommendations about preservation actions based on risk scores and institutional risk profiles [11][13]. SCAPE's Scout tool also uses an adapter-based architecture, but its approach is broader than AONS enabling it to import other data such as repository events and institutional policy information, and use this for generating notifications to initiate preservation planning [4].

Such knowledge bases form the backbone for more automated means of monitoring the wider preservation environment, forming a central place for collecting information useful for preserving digital objects, and allowing gaps in one source's knowledge to (potentially) be filled. Graf and Gordea [10] found the approach of aggregating linked open data in FFMA increased the amount of information available, with "~10% more file formats, about 13 times more software and with 60% more vendors than PRONOM" alone, demonstrating the potential for aggregated knowledge. The usefulness of a knowledge base, though, really depends on the quality, accessibility, scope and reliability of the incoming data; Becker *et al.* [4] note that "sources that focus on digital preservation have a generally very reduced coverage (registries) or machine readability (reports), while general purpose sources normally cover very limited facets of the information relevant for digital preservation."

More recently, Yale University Library have taken a slightly different approach to developing a knowledge base of technical metadata about computing resources (file formats, software, *etc.*) — they are

[1] https://c3po.openpreservation.org/

[2] https://scout.openpreservation.org/

[3] https://plato.openpreservation.org/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

driving a community effort to enhance the information in Wikidata with the view that Wikidata's "infrastructure will enable the long term continued access to the data digital preservation practitioners collate and capture" [14] [1s]. To support this, they are developing a web portal[4] which acts as a layer over the Wikidata infrastructure, allowing users to browse and easily contribute knowledge to the Wikidata knowledge base. They are effectively championing the improvement of source data through a community effort. Providing a domain-specific web interface will certainly help contributions, but effective additions are perhaps more likely to come from alignment and integration with business workflows[5].

Notification of risks is intended to initiate some form of preservation planning to devise an appropriate mitigation strategy. The SCAPE suite uses (and has enhanced) the Plato tool specifically for this. Plato guides users through a preservation planning workflow enabling users to evaluate alternative preservation strategies (*e.g.*, alternative migration software), review the results, and make an informed decision about the most appropriate preservation action plan. Plans need to include preservation requirements (*e.g.*, significant properties) for fair evaluation of preservation actions, and evidence of the preservation strategy decision (*e.g.*, approaches tested, results, and decisions made) [3]. Trust is therefore promoted through transparency of the process undertaken, potential for reproducing the evaluations, and openness of the options considered and the decision taken.

One of the key challenges with such a planning approach is the efficiency of the process, particularly when trying to do this at scale across large heterogeneous collections. Becker *et al.* note that these challenges can often be lessened through better automation and improved preservation-related business documentation, however a large proportion of time can still be spent discussing preservation requirements, particularly formats, significant properties, and technical encodings. To aid with this, the SCAPE suite defines a controlled vocabulary[6] which could be used when defining policies and collection profiles to enable more automated import of information into the planning process.

III. IPS Architecture

Our Integrated Preservation Suite is intended to help us with risk mitigation at scale and across all of our collections, primarily through development and implementation of preservation plans. Functionality, trust, and ease of use are critical factors, which has led us down an avenue of integrating functionality behind a single, managed web interface. The ability to enhance functionality as our needs evolve is also important; one area we see this will be vital is in realising the outcomes from our risk assessment and preservation planning explorations.

We have developed the architecture and associated data models in a recursive manner in line with our learning as the project has proceeded, building components from the ground up to meet our needs where necessary. The project is a three-year initiative, due to complete in late 2019, however the intention would be to maintain and expand (where necessary) the suite to meet our continued and developing requirements. The work presented here reflects our thinking (at the time of writing).

An overview of the architecture is shown in Figure 1, highlighting the main components of the suite:

- *Knowledge Base* (KB) — a graph-based curated knowledge base with information, initially, about formats, software, and wider technical environments relevant to the Library's digital collections;
- *Preservation Software Repository* (SR) — a digital repository containing requisite current and legacy software for rendering files and implementing preservation plans;
- *Policy and Planning Repository* (PPR) — a document repository for storing collection-specific data including collection profiles, preservation policies, and collection-specific preservation plans;

[4] http://wikidp.org/

[5] One suggestion mooted was the use of 'bots' to push data directly into Wikidata from other registries, for example, PRONOM. More generally, though, effective contributions are likely to require a user to have a business motivation.

[6] https://github.com/openpreserve/policies

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

- *Preservation Workbench* — a web-based graphical user interface providing unifying functionality: for searching and curating the knowledge base, the Software Repository, and the Policy and Planning Repository; monitoring the preservation environment to provide notifications to users about potential preservation risks; as well as for managing and developing format-specific preservation plans;
- *Execution Platform* — a platform for testing preservation actions on.

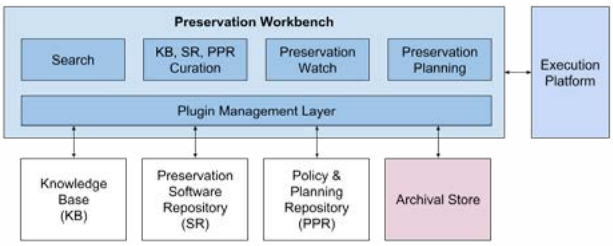


Figure 1: Overview of the IPS Architecture

These components are designed to integrate with any repository system through a modular, API-based architecture. The Workbench defines a standardised API for interacting with the various components, with bespoke plug-ins written to target technology-specific implementations of each component. For example, a graph-database-specific plug-in implements the Knowledge Base API. To interface with an organisation's repository system, an appropriate plug-in will need be written to translate between the IPS Archival Store API and the repository's own API.

To date, most effort has focussed on the Knowledge Base and the Workbench for querying it, curating the data going in to it, and developing preservation plans. The Software Repository and the Policy and Planning Repository make use of open-source software for their implementations to quickly develop against; longer term, our preservation repository system would make a good home for the data these components store. Preservation Watch functionality (part of the Workbench) and the Execution Platform are part of our next steps.

A. Preservation Workbench

This is the main entry point to IPS and provides a webbased user interface for digital preservation practitioners. Functionally, the current implementation provides three main tasks: it enables a user to search for information from the Knowledge Base using a single-search-box interface;it allows users to curate incoming data in the Knowledge Base; and, it walks the users through a process for developing a preservation plan. Over time, this existing functionality will be enhanced and new functionality will be added (for example, to support preservation watch).

The interface is an Angular web application[7] currently running in an Ubuntu virtual machine on a HP Proliant departmental server. API calls to the other IPS components, *e.g.*, the Knowledge Base, are currently made directly from within the web application, however this has been coded in such a way that it can be easily replaced with a call to the IPS API once that has been implemented. Working in this way is intentional as it allows us to design the Workbench functionality we need without having to define the IPS API upfront. Once we understand the needs of the API layer, we can implement that and refactor the Workbench to use it.

1. Searching the Knowledge Base

Usability has been a key consideration for the interface's overall design. We have purposefully kept the interface clean, affording only a single search box to search the Knowledge Base. Keyword searching is supported, *e.g.*, a user can search for "PDF", "Adobe", or any other term. This matches on key properties within the Knowledge Base, such as the (file format/ software) name or extension.

To facilitate more in-depth queries, such as for identifying software that can migrate file formats, we provide a set of search *labels* with which to tailor queries:

- **"type:"** — enables the user's search to be filtered by the type of result, such as 'software' or 'format', *e.g.*, "Adobe type:software"
- **"extension:"** — enables the user to search specifically for information based on the file extension value, *e.g.*, "extension:pdf"
- **"create:"** — enables a user to search for software that can create a specific file format, *e.g.*, "create:pdf"
- **"render:"** — enables a user to search for software that can render a specific file format, *e.g.*,

[7] https://angular.io/

"render:pdf"

- **"migrate-from:"** — enables a user to search for software that can migrate from a particular file format, *e.g.*, "migrate-from:tiff"
- **"migrate-to:"** — enables a user to search for software that can migrate to a particular file format, *e.g.*, "migrate-to:jp2"

These last two could be used in combination, for example a search of "migrate-from:tiff migrate-to:jp2" would allow a user to search for software that can migrate from TIFF to JP2.

The set of labels listed here have evolved to their current state. It is fully anticipated that new labels will be added as they are deemed useful.

2. Curating Incoming Data for the Knowledge Base

Data curation is described in further detail in section B.3, after the data model has been described. Chiefly, though, the Workbench provides a web-based interface to allow an appropriate user to compare incoming data with existing data and make decisions about how to proceed with each incoming piece of data.

3. Preservation Watch

The suite's preservation watch element relies largely on the integration with the other IPS components and Archival Store, along with findings from our exploration of preservation threats and risks. In terms of development, the other IPS components have been our focus to date, so one of our next steps is to design and implement this functionality. Broadly though, it is envisaged that key data within the other components will be monitored on a routine, scheduled, or event-driven (*e.g.*, new software added to the Software Repository) basis, initiating user notifications of interest to specific risks.

4. Preservation Planning

Currently, our preservation planning approach is broadly following a SCAPE/Plato planning methodology [3] bringing together various facets of information about a collection at risk to define the plan requirements, evaluating different strategies to mitigate any risks, analysing the results, making a recommendation, and constructing an executable plan.

Our current implementation is in its infancy. The web page allows an offline preservation plan template to be downloaded, walks the user through the necessary steps to complete the plan, and allows them to upload their completed plan into the PPR. However, this will be modified in future releases to allow the definition and execution of the plan directly from the Workbench.

We have begun to experiment with improving the effectiveness of the guiding steps by incorporating embedded search boxes into the page at relevant points for a planner to search for specific information, such as finding collection profile documents in the Policy and Planning Repository. We expect this functionality to improve as we evolve the Knowledge Base, and make improvements to the content within the PPR to better support machine-interpretation.

Evaluating different preservation strategies, and developing executable preservation plans has only loosely been considered, again broadly in line with SCAPE approaches. Executable scripts will most likely be stored in the IPS Software Repository alongside their required applications.

5. Integration with Other Components

To facilitate technology-agnostic connectivity to the various IPS components and existing Library archival store, the Workbench provides a standardised API allowing plug-ins to be written to meet each component's underlying technology.

B. Knowledge Base

The Knowledge Base is intended as the fundamental, curated knowledge base upon which to search and reason over key information to establish preservation actions and base decisions on. It was initially conceived as a database of technical information and relationships about file formats and software, with a view to enabling digital preservation practitioners within the Library to produce, contextualise, and validate preservation plans. By searching through this knowledge base practitioners should be able to get a set of information to help them make judgements about questions, such as:

- What software applications can be used to open or edit files of this particular format? (query relation: format > software)
- What formats can this software import? (query

relation: software > format)
- What software can I use to migrate from format A to format B? (query relation: format, format software)

The focus of such queries is on the relationships between information points, *e.g.*, the software that can *open* a particular format, or the software that can *read* one format and *write* out a second. This led us to orientate towards graph-based databases, in particular Neo4J[8], for which relationships are first-class entities. On top of this we constructed a data model based around file format and software information, with a view to addressing the above questions. Further details about the data model are given below.

The data model supporting this knowledge base is not static and is expected to evolve over time. Indeed, as the project has progressed we are beginning to see the scope of the Knowledge Base gradually expand to cover broader information sets, such as hardware, licensing information, and detailing software we have in our Software Repository. We envisage that this expansion could continue to include collection profile details, policies, and risks, allowing greater depth to the reasoning capabilities of the system, for example:

- What hardware were these type of floppy disk typically used with? (query relation: disk > computing equipment)
- What risks are associated with this file format? (query relation: format > risks)
- What mitigation strategies are needed with this file format? (query relation: format > risks > mitigation strategies)
- What are the known problems with using this software? (query relation: software > problems)

Of course, as has been hinted at and highlighted in previous work [4][9], such knowledge bases are only as useful as the data contained within them. Information within our Knowledge Base is thus a mixture of data from outside sources — web pages, databases, registries, *etc.*and manual contributions from domain experts within the Library.

This presents a couple of challenges. Firstly, the variable nature of all this information needs to be aggregated together in a standardised way to ensure that it can be reasoned over. Broadly, this means that data from any given source needs to be translated into our IPS data model. To do this, we use an adapter approach, as has been used in other projects [9][12]. Data import is combined with a curation stage to ensure that newly arriving data is effectively merged with existing data; this requires the use of a *staging* instance of the Knowledge Base.

A second challenge is establishing and maintaining trust in the data to ensure that preservation actions/decisions are based upon sound reasoning. We see a number of key aspects here. One is that it will be important to maintain knowledge of the source of each piece of information. Relatedly, given sources of information could disappear (or simply become inaccessible to us), preserving a snapshot of those sources is also essential.

1. Data Model

The data model needed to allow the aggregation and association of information from various sources, both internal and external, while also keeping track of the provenance of all incoming information. To that end we devised a model comprising a backbone of high-level *canonical* nodes, nodes whose properties and organisational relationships could be curated by ourselves, associated with any number of *informational* nodes, which provide related information extracted from specific sources of data. This allowed us to organise file format and software information into a structure that would suit our needs, while also allowing the addition of externally generated information.

Informational nodes currently contain a set of predefined properties (such as *name, description*, or *aliases*) which are normalised between sources, where possible, so that they can be easily compared or queried alongside nodes of the same type (*e.g.*, file formats) from other sources. The set of normalized properties is expected to increase as more sources of information are added to the Knowledge Base and more properties worth capturing are discovered.

[8] https://neo4j.com/

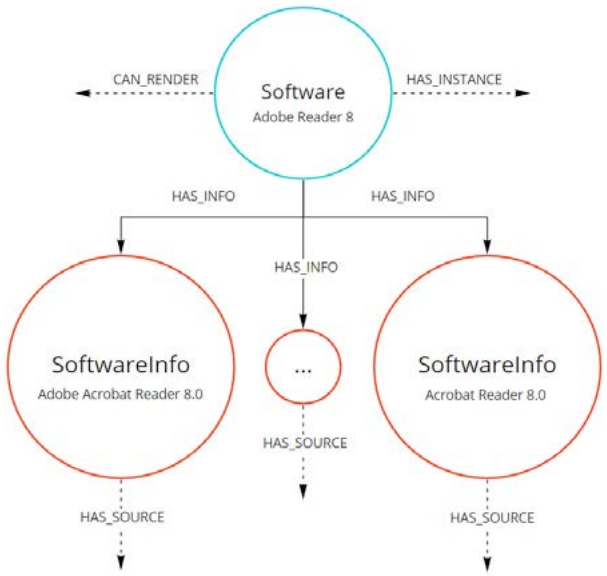


Figure 2: An example of a canonical Software node ("Adobe Reader 8") and its relationships to informational nodes with differing names and sources.

Source information is kept for every informational node and relationship extracted from a data source, allowing us to judge the trustworthiness of specific information by its source, or track down and correct an erroneous source after it's been ingested. The current data model also supports a degree of versioning (not shown in Figure 3), allowing us to search and investigate informational nodes and relationships ingested from previous source snapshots. This could assist in determining what information the Knowledge Base could have provided at a given point in time, allowing a certain amount of traceability.

To build on previous work done by the preservation community and simplify integration with external data sources, certain relationships and vocabularies were adopted, where possible, from existing registries, such as PRONOM, and augmented with additional items where it was thought necessary to fulfill certain preservation queries. For example, while the preservation vocabularies we initially adopted could easily describe a software's ability to 'render' a file format, they were unable to capture the simpler ability to understand, or 'read', a format. This became an issue when we wanted to more precisely discover software with the potential for migrating formats.

While one could easily argue that conversion software is technically *rendering* one file format into another, failing to differentiate between that and the more conventional sense of rendering for consumption (*e.g.*, visually or aurally) meant that we were unable to discover only those pieces of software which could 'read' one format and 'create' another without the results also being muddied by conventional rendering software. Results for software which could render a format for consumption would have been similarly muddied by software only capable of reading the format for conversion purposes.

The current data model has undergone extensive evolution and expansion since its initial version, growing as we discover new information we wish to extract, and changing to accommodate better graph design principles as our experience with the underlying technology has grown.

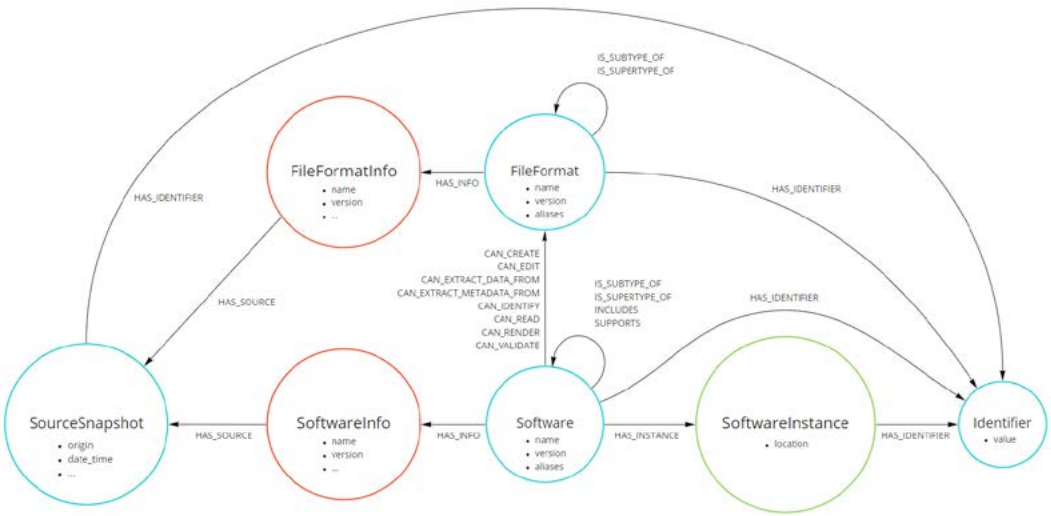


Figure 3: A simplified extract of the KB data model, showing the relationships between software, file formats, and their data sources.

2. Data Source Adapters

A Data Source Adapter is defined for each source, performing source-specific ETL (Extract, Transform, Load) functions to load the data into the curation area of our database, as shown in Figure 4.

Each adapter will eventually implement a standardised Adapter API which will enable a Data Source Management component of the Workbench to control it, such as to start or stop an on-demand import or to configure automated polling of a data source. Adapters are written in Python and make use of a *Data Management Library* module developed to act as an IPS Data Model-aware wrapper around our Neo4J databases.

Adapters are also responsible for capturing snapshots of the source information for preservation[9]. This ensures that we will always have a copy of the raw data we parsed and imported into the Knowledge Base. It also means that if there is a problem transforming the data, we can modify the adapter and rerun the process without needing to reacquire the data.

To date we have defined adapters for PRONOM, FileExtensions.org[10] (website), and an Excel spreadsheet provided by the National Library of Australia containing file formats and software information. In the immediate pipeline we will be developing adapters for the SPDX License List[11], COPTR[12] and Wikidata[13].

3. Data Curation

The data curation process is still largely in development. Broadly, it needs to allow curators to take data incoming from a source and merge it, in a managed way, into the existing Knowledge Base of information. An incoming record (*e.g.*, file format, software, *etc.*) could represent completely new information (*i.e.*, a new file format not held in our existing Knowledge Base), existing or otherwise overlapping information, or information it would be unhelpful to retain at all.

To cope with these scenarios and allow managed and documented contributions into the main Knowledge Base, we make use of a *staging area* in which to prepare the incoming data before it is pushed into the main Knowledge Base. The staging area is currently a separate instance of our Neo4J database and operates with largely the same data model but with the addition of information to record individual curatorial decisions (as described below).

A source's incoming information is initially imported into this staging area for curation. A curator is then able to see, via the Workbench, the list of incoming records side-by-side the list of existing records in the main Knowledge Base. Any items previously curated are marked with icons signifying those past decisions.

Individual records can be chosen for closer inspection, or two can be chosen for side-by-side comparison, whereupon the curator is shown each records' contents.The curator can then decide whether to keep the incoming record, have the two merged into one, or have the incoming record discarded entirely. A level of editing is allowed on the canonical nodes when either retaining the incoming data or merging (*e.g.*, editing the name, aliases, or identifiers). The curator's decision is captured as a *decision* node within the staging area.

Once the curation of the incoming data is complete, the curator can initiate a push from the staging area to the main Knowledge Base. Decision nodes are processed to determine what needs to happen to each incoming record and the action itself is captured in a log. Once complete, the staging area is wiped clean in preparation for importing data from another source.

[9] Although for some websites we could make use of our web archived content.

[10] With permission.

[11] https://spdx.org/license-list

[12] http://coptr.digipres.org/Main_Page

[13] https://www.wikidata.org/wiki/Wikidata:WikiProject_Informatics

Curating data is likely to be a laborious process, particularly for the initial import of new data sources, where aligning existing and incoming information needs to be thoroughly considered. Over time, however, we expect the workload to decrease as we begin to apply rules and heuristics to improve the process. For example, each informational node imported from a data source also has a unique, source-dependant, external identifier (such as a PRONOM ID or scraped URL) which can be used to automatically link it to any newer versions of that same node on subsequent imports.

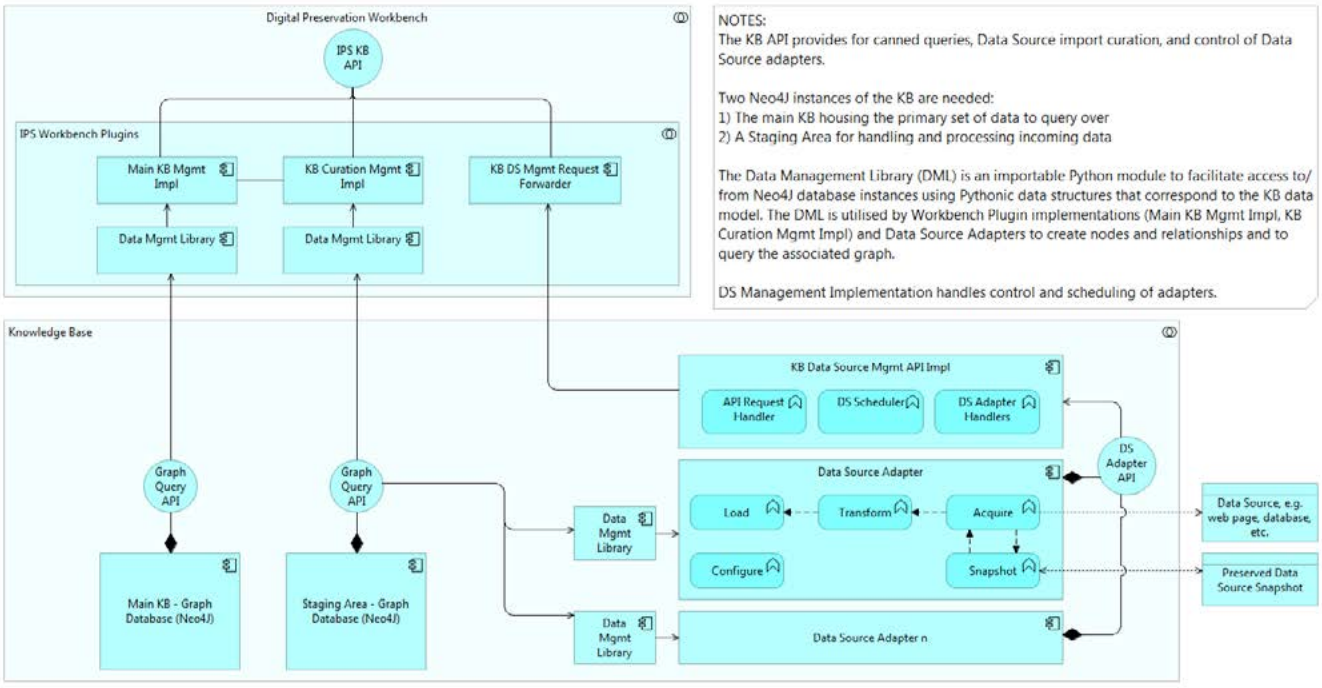


Figure 4: Knowledge Base Architec

Another avenue could entail leveraging each node's alias information (*e.g.*, alternative names for file formats) to automatically suggest links between incoming and existing nodes, reducing the curator's job to one of confirmation. Whilst burdensome, we felt that the value added by curation — allowing an organisable information structure, and the removal of misleading or erroneous information — was important for productive use of the Knowledge Base.

Curation of data that has already made its way into the main Knowledge Base are less developed at this stage, but current intentions are to allow editing of canonical nodes and structures through the Workbench, while keeping all external informational nodes as uneditable views on their source information.

C. *Software Repository*

We initially stated that a digital object becomes inaccessible because it lacks an appropriate rendering environment (wholly or in part). Within the context of the Library (or any other organisation), we can refine this to say that a digital object becomes inaccessible because that rendering environment does not exist within the context of the Library (organisation); the format is institutionally obsolete.

Our approach to solving this is to retain the software needed to access our digital objects. That includes: the software required to open the file directly on current institutional computing technology; the migration and rendering software for such a preservation strategy; and emulators, base operating systems, and any other dependencies necessary to render the digital objects in question. This is the purpose of the *software repository*; to preserve the software necessary to maintain access to all our digital collections. Licensing details are noted and the project is engaging with software providers (such as Microsoft) and the Library's in-house legal team to address concerns around long term use of the software for preservation purposes.

At its heart, the software repository is simply an archival store. The British Library currently has its homebuilt Digital Library System which would serve for keeping such software safe. From a development perspective however, we have opted to run our own instance of the open-source repository system RODA[14], backed by network storage, and use the RODA-in[15] tool to create SIPs for ingestion into this repository.

1. *Populating the Software Repository*

Considerable effort has been placed into identifying software of relevance to the Library, and subsequently locating installation files for it. Discussions with IT, Architecture, and collection-specific colleagues have led to capturing a list of software the Library uses (or has used) in ingest and access workflows, including on reading room PCs. The selection of software is based on analysis of formats in the current archival store (excluding web archive content), with at least five viable software options ingested for each format and format version in the repository to date[16].

Most software has been acquired from our IT department's existing and legacy application library. In addition to this we have been downloading software from the software's official web sources, or an archived version of that source.

A Microsoft Access database is currently used to capture information about the software. This is split into two main parts, information relating to media (*e.g.*, media from IT), and information about the software itself (which may be on physical media, a digital download, or simply knowledge one has about software without actually having acquired it). Software information is of most relevance for discussion, and includes the name, version, developer, release date, technical information (e.g., requirements), licensing information, and whether we have a copy of the actual software.

In time we expect a subset of this information to make its way into the Knowledge Base, and other more descriptive information to be included as AIP metadata within the Software Repository for cataloguing purposes.

D. *Policy and Planning Repository*

Risk identification and mitigation, including preservation planning, is based on and influenced by a variety of factors including organisational policies. Through bitlevel preservation we may be able to preserve the raw digital objects themselves, and through preserving software we're able to maintain access, but our approaches will be influenced by our overall risk appetite. Without an understanding of the factors influencing our risk appetite, we will not be able to completely and unambiguously demonstrate the rationale behind any preservation decisions that have been made. This is especially important in order to retain knowledge due to the turnover of staff (whether short term, or eventual).

The Policy and Planning Repository acts as a document store for all this supporting information. It is the place where all known documentation relevant to preservation of digital collections within the repository is centralised. This includes, but is not limited to: preservation plans, policies, collection profiles, architectural documents or diagrams, and workflow documents or diagrams.

For development purposes we are currently using an open-source electronic document management system — Mayan EDMS[17] — installed and running on our own server, to store documents. This provides functionality for organising and tagging documents, performing optical character recognition, and even developing bespoke workflows to manage documents through a lifecycle (*e.g.*, for editing and review).

IV. Conclusions and Further Work

This paper has presented a description of the current status and thinking of the British Library's internally funded Integrated Preservation Suite project. The suite comprises a web-based Workbench providing the central, overarching interface for digital preservation users, a Knowledge

[14] https://github.com/keeps/roda

[15] https://rodain.roda-community.org/

[16] This is, in some ways, slightly circular as a fully working IPS solution should help us do this task.

[17] https://www.mayan-edms.com/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Base of information (initially) about file formats and software, a repository for preserving software, and a further repository for storing Library-specific preservation information, such as policies, preservation plans, and collection profiles.

At the time of writing the project has the majority of the year left to run. Development is still in progress and work will continue with a focus on producing a more robust release of the suite's components.

Our understanding of preservation risk management and subsequent preservation planning is also developing and so work around improving the Workbench to support this will undoubtedly be needed. As mentioned in prior work, supporting any form of automated risk identification largely depends on the availability and quality of underlying information. Enhancing risk identification within IPS will require making more of the Library's preservation policies and collection profiles, amenable to machine-reading and information processing. Improving the Workbench to aid development of such preservation documentation may be useful.

Building on this, Preservation Watch functionality will also need to be developed and integrated into the main IPS Workbench interface to support a unified approach to risk management and subsequent planning actions.

Similarly, the IPS Execution Platform needs development. In particular, the Library are in the process of procuring and implementing a new digital repository system. Functional overlap between IPS and this new system will need to be considered, and integration between the two will need to happen. Ideally, the IPS Software Repository and Policy and Planning Repository implementations would be removed in favour of implementation by our digital repository system.

Finally, trust is vital for such preservation planning endeavours, and one key aspect will be to ensure that user logins, and where necessary user roles, are implemented to ensure appropriate access to functionality. Relatedly, a logging system would be necessary to ensure user actions are auditable; the beginnings of this functionality exists in the logging provided by Knowledge Base data curation.

Acknowledgments

We would like to thank the other IPS team members whose dedicated work has helped shape and deliver the project: Michael Whitmore, Kevin Davies, Michael Day, Akiko Kimura, Simon Whibley, and John Beaman. We would also like to thank the National Library of Australia for kindly supplying us with a spreadsheet of their collected data, and the maintainers of File-Extensions.org for giving us permission to scrape their website.

References

[1] H. M. Ryan, "Occam's razor and file format endangerment factors," in *Proceedings of the 11th International Conference on Digital Preservation, iPRES 2014, Melbourne, Australia, October 6–10, 2014,* 2014. [Online]. Available: http://hdl.handle.net/11353/10.378114.

[2] B. Lavoie, "The open archival information system (oais) reference model: Introductory guide (2nd edition)," DPC Technology Watch Report 14-02, 2014. DOI: https://doi.org/10.7207/twr14-02.

[3] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman, "Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans," *Int. J. on Digital Libraries*, vol. 10, no. 4, pp. 133–1s7, 2009. DOI: https://doi.org/10.1007/s00799-009-0057-1.

[4] C. Becker, J. C. Ramalho, M. Ferreira, K. Duretec, P. Petrov, and L. Faria, "Preservation watch: What to monitor and how," in *Proceedings of the 9th International Conference on Digital Preservation, iPRES 2012, Toronto, Canada, October 1–5*, 2012, 2012. [Online]. Available: http://hdl.handle.net/11353/10.293864.

[5] M. J. Day, M. Pennock, A. Kimura, and A. MacDonald, "Identifying digital preservation requirements: Digital preservation strategy and collection profiling at the british library," *in Proceedings of the 11th International Conference on Digital Preservation, iPRES 2014, Melbourne, Australia, October 6–10*, 2014, 2014. [Online]. Available: http://hdl.handle.net/11353/ 10.378119.

[6] M. Pennock, P. May, and P. Wheatley, "Sustainability assessments at the british library: Formats, frameworks, & findings," in *Proceedings of the 11th International Conference on Digital Preservation, iPRES 2014, Melbourne, Australia, October 6–10, 2014*, 2014. [Online]. Available: http://hdl.handle.net/11353/ 10.378110.

[7] M. Pennock and M. Day, "Adventures with epub3: When rendering goes wrong," in *Proceedings of the 15th International Conference on Digital Preservation, iPRES 2018, Boston, Massachusetts, United States, September 24–27, 2018*, 2018. DOI: 10.17605/OSF.IO/94TEB.

[8] J. Curtis, P. Koerbin, P. Raftos, D. Berriman, and J. Hunter, "AONS an obsolescence detection and notification service for web archives and digital repositories," *The New Review of Hypermedia and Multimedia*, vol. 13, no. 1, pp. 39–s3, 2007. DOI: https://doi.org/10.1080/13614560701423711.

[9] D. Pearson, "AONS II: continuing the trend towards preservation software 'nirvana'," in *Proceedings of the 4th International Conference on Digital Preservation, iPRES 2007, Beijing, China, October 11–12, 2007*, 2007. [Online]. Available: http://hdl.handle.net/11353/10.294518.

[10] R. Graf and S. Gordea, "Aggregating a knowledge base of file formats from linked open data," in *Proceedings of the 9th International Conference on Digital Preservation, iPRES 2012, Toronto, Canada, October 1–5*, 2012, 2012. [Online]. Available: http://hdl.handle.net/11353/10.293868.

[11] S. Gordea, A. Lindley, and R. Graf, "Computing recommendations for long term data accessibility basing on open knowledge and linked data," in *Joint proceedings of the RecSys 2011 Workshops Decisions@RecSys'11 and UCERSTI 2,, vol. 811, 2011*. [Online]. Available: http://ceur-ws.org/Vol-811/paper8.pdf.

[12] M. Kraxner, M. Plangg, K. Duretec, C. Becker, and L. Faria, "The SCAPE planning and watch suite," in *Proceedings of the 10th International Conference on Digital Preservation, iPRES 2013, Lisbon, Portugal, September 2–6, 2013*, 2013. [Online]. Available: http://hdl.handle.net/11353/10.378091.

[13] R. Graf, S. Gordea, H. M. Ryan, and T. Houzanme, "A decision support system to facilitate file format selection for digital preservation," *Libellarium: journal for the research of writing, books, and cultural heritage institutions*, vol. 9, Mar. 2017. DOI: 10.15291/libellarium.v9i2.274.

[14] K. Thornton, E. Cochrane, T. Ledoux, B. Caron, and C. Wilson, "Modeling the domain of digital preservation in wikidata," in *Proceedings of the 14th International Conference on Digital Preservation, iPRES 2017, Kyoto, Japan, September 25–29, 2017*, 2017. [Online]. Available: https://www.wikidata.org/wiki/Q41533080.

[15] K. Thornton, K. Seals-Nutt, E. Cochrane, and C. Wilson, "Wikidata for digital preservation," 2018, [Online]. Available: https://doi.org/10.5281/zenodo.1214318, unpublished.

Digital preservation and enterprise architecture collaboration at the University of Melbourne

A meeting of mindsets

Jaye Weatherburn
University of Melbourne Australia
jaye.weatherburn@unimelb.edu.au
orcid.org/0000-0002-2325-0331

Lyle Winton
University of Melbourne Australia
l.winton@unimelb.edu.au
orcid.org/0000-0002-3049-1221

Sean Turner
University of Melbourne Australia
sean.turner@unimelb.edu.au
orcid.org/0000-0002-9497-9576

Abstract – This paper examines a meeting of mindsets at the University of Melbourne, and how digital preservation and enterprise architecture have joined forces to meet goals of the university's Digital Preservation Strategy 2015-2025. We identify the points of connection and similarities between digital preservation and enterprise architecture. We explore how a core foundation for university-wide infrastructure implementation is being achieved through effective collaboration between the Scholarly Services-based Digital Scholarship team at the university (the team responsible for driving digital preservation project work) and the Infrastructure Services-based Enterprise Architecture unit. We investigate the similarities and differences in approach for these two diverse business units within the university context and identify how collaboration between digital preservation and enterprise architecture can continue to drive mutually beneficial digital preservation operations at the university.

Keywords – Digital preservation, Enterprise Architecture, Collaboration

Conference Topics – Collaboration: a Necessity, and Opportunity, or a Luxury?; Designing and Delivering Sustainable Digital Preservation

I. Introduction

Scholarly Services and the Digital Scholarship team at the University of Melbourne has been faced with a continual challenge throughout its three years of active digital preservation project work: how to achieve university-wide buy-in and investment for the people, infrastructure, policy, and processes required to protect and maximize the long- term value of digital assets in a complex environment with many demands on funding and resources. This challenge drives the need to find effective allies within the organization to achieve the required level of implementation and change, including long-term sustainable operations.

With the arrival of a newly-employed Enterprise Architect at the university in 2018, and a forthcoming revamp of the university's Enterprise Architecture Roadmap, the Digital Scholarship team recognised the opportunity to explore how digital preservation infrastructure could emerge as a core university foundation, by working collaboratively with Enterprise Architecture in their capacity as trusted university business and technology design experts. The resulting collaboration has facilitated faster progress for the design and first steps of implementing core digital preservation system infrastructure for the university, and has paved an easier path for future progress to meet the other goals (beyond the technology solutions) of the university's Digital Preservation Strategy 2015-2025 (Culture, Policy, and Organisation goals) [1].

Over the three years of digital preservation project work, the Digital Scholarship team has gathered substantial knowledge about organizational readiness to support digital preservation, including

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

analysis of current technology, resourcing, and skills gaps. One of the key questions driving digital preservation implementation at the university is: how best to provide long-term digital preservation storage and services for digital research data, research outputs, university records, and cultural collections?

An internal review of research support services in 2016 analysed many of the elements inhibiting university-wide implementation of digital preservation across these varied domains of research outputs, records, and collections. The review identified a "siloed" way of working at the university as a major cultural barrier inhibiting the development of shared messaging of the value of and need for digital preservation. The wide range of stakeholders who must be involved in the implementation of digital preservation at the university are spread across multiple different business units. Successfully engaging these stakeholders (for whom digital preservation is not currently part of their job description) requires time consuming consultations and engagements.

Driven by the need to work iteratively and simultaneously across all four goals of the Digital Preservation Strategy (Culture, Infrastructure, Policy, and Organisation), and to address the major cultural barrier of siloed ways of working, a majority of the digital preservation project work has been focussed on drawing in the necessary stakeholders through targeted working groups, ensuring key outcomes are clearly defined in advance. These working groups have included key individuals from varied business units (e.g. the Preservation Storage Working Group, tasked with developing key preservation storage criteria for the university, involved central IT, Records, Archives, Library, and research service and infrastructure staff; the Appraisal Working Group, tasked with scrutinizing current collection policies and procedures to inform digital preservation policy development brought together Records, Archives, Library, Collections, and Digital and Data leadership staff.)

These working groups have produced essential foundational work that is progressing the Culture, Policy, and Organisation goals of the Strategy, but the Digital Scholarship team was aware of the lack of major progress for university-wide infrastructure implementation to support digital preservation. Small-scale "infrastructure blueprints" (small iterative technology-focussed projects) undertaken over the three years of project work generated learnings for the project staff in relation to the technology and skills required for digital preservation processes, but these projects did not significantly progress to the point of holistic infrastructure implementation. Compelled by the goals of the Digital Preservation Strategy to implement and sustain core university-wide platforms and services for digital preservation infrastructure, the Digital Scholarship team recognised the need to engage initially with the university's central Infrastructure Services, and then subsequently with the Enterprise Architecture team sitting within Infrastructure Services, in order to drive greater visibility of digital preservation project work across the organization. This approach would also help to ensure that digital preservation appeared (and remained) on the emerging Enterprise Architecture Roadmap as a foundational and crucial component of the socio-technical ecosystem of the university.

During the writing of this paper, we searched for examples of information professionals working collaboratively with enterprise architecture to meet digital stewardship goals and can validate Sam Searle's findings that "there is little discussion in the literature about the EA [enterprise architecture] process as a collaborative effort"; and "there are few documented examples of librarians working closely with enterprise architects in higher education or elsewhere" [2].

There are some examples investigating the use of enterprise architecture for application in records and archives contexts [3]; [4], and there is documented experience of the use of enterprise architecture approaches in a library context [5]. However, we have found limited documented evidence of collaborations between digital preservation practitioners and enterprise architects.

This paper explores the mindsets of enterprise architecture and digital preservation, how these mindsets work in the University of Melbourne context, and how best we can collaboratively work together to deliver the Infrastructure goals of the university's Digital Preservation Strategy. We describe how this collaboration has accelerated the

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

design and the first steps to delivering the required technology for implementing digital preservation as a core university-wide function.

We are documenting our collaborative efforts in designing and delivering core digital preservation infrastructure in order to share our learnings with others in similar contexts who are striving to implement digital preservation operations across a distributed organisation.

II. Meeting of mindsets: enterprise architecture and digital preservation

A. *Enterprise Architecture Mindset*

As defined by Gartner, enterprise architecture is "the process of translating business vision and strategy into effective enterprise change by creating, communicating and improving the key requirements, principles and models that describe the enterprise's future state and enable its evolution" [6]. Key aspects of enterprise architecture aim to deliver effective enterprise-wide change through holistic solutions that address various organizational challenges, with a scope extending beyond just technology implementation to include governance, people, and processes.

Enterprise architecture emerged in response to a clear disconnect in many organisations between business strategy and the change initiatives that were being funded and implemented. This was particularly evident in IT-heavy projects and often led to large investments that did not assist in achieving the goals of organizations. Numerous audits of project portfolios in larger organizations found that there was significant duplication, and that different solutions were being applied to the same problems in an uncoordinated way. In addition, many organizations were not managing the lifecycles of their technology investments and were not effectively leveraging new technology opportunities. Enterprise architecture thus focuses on optimizing various portfolios and projects that are focused on technology and change, to ensure alignment with organizational strategy and to effectively leverage technology trends. Given that strategy and particularly technology can change quickly, enterprise architecture horizons are often relatively short (3-5 years) and focused on prioritizing investment.

This Enterprise Architecture focus on optimization and standardization is similar to that described by Scott Prater for general IT practitioners: "Most IT shops strive to standardize processes and infrastructure for all their users, to increase efficiency, reduce duplication of effort, and free up time and resources to improve their suite of services and introduce new ones" [7]. As digital preservation is often not one of the standard services offered by IT, issues can arise when attempting to explain to IT staff the need for implementing long-term digital preservation processes, particularly when IT environments have been setup with mandates to respond to the here-and-now challenges of data storage, management, and dissemination, without consideration of long-term requirements.

In order to provide alignment and optimization holistically, enterprise architecture teams need to apply a "continuous practice of describing the essential elements of a socio-technical organization, their relationships to each other and to the environment, in order to understand complexity and manage change" [8]. In addition to this focus on description and documentation, enterprise architecture teams apply standardization and integration as key approaches to the optimization of portfolios and investment. Thus fewer solutions are better, and reuse of information and processes is preferred.

B. *Digital Preservation Mindset*

The various information professions that have a stake in, and that are involved in digital preservation activity (e.g. libraries, archives, museums, records management) have time-honoured mandates for preservation and custodianship of the scholarly, scientific, and cultural record, and bring a long-term mindset to their various organizations and institutions.

Information professionals working within the digital preservation field have a wide range of valuable knowledge and skills: how to ascertain the value of digital assets, and how to ensure good practice for preservation of and access to digital assets over time, being two key areas. These skills are becoming integral to organizations dependent on digital materials, "for now it is librarians and archivists, the custodians of the past, that are the advance-guards of the future. They have worked with open approaches to

software development, data practices and scholarly communication for years. These communities, networks and processes are a vital part of the solution" [9].

Scott Prater provides a useful perspective of the difference between the mindset, needs, and priorities of the archivist as opposed to the priorities of IT practitioners, particularly when talking about digital preservation: "when the discussion turns to digital preservation, there may be a divergence in priorities and understanding. While the archivist's priority is on making sure their digital assets are preserved and accessible forever, the IT personnel's focus may be on making current data publicly accessible, making sure that systems are running smoothly right now" [7]. The fundamental differences in approach that result from the mandate to preserve and access digital materials forever and the imperative to ensure things are working smoothly in the now are some of the elements that make it so difficult to meld the mindsets of digital preservation practitioners and IT practitioners for the benefit of the organization's long- term digital stewardship goals.

Digital preservation practitioners charged with implementing digital preservation capabilities bring an understanding of complex information management, the ability to manage change over time, and knowledge of robust standards and frameworks that enable long-term management and preservation of information. Models in the digital preservation field such as the Three Legged Stool [10] usefully demonstrate how to develop sustainable digital preservation programs that adequately address the socio- technical complexities involved in long-term preservation of information.

C. The Two Mindsets: Similarities and Differences

Both the enterprise architecture and digital preservation fields of practice are concerned with socio- technical complexity and managing change, and both have a holistic mandate across an entire organisation. Both fields use frameworks, including models, standards, and reproducible methods, and have a key approach of documenting and visualizing complex domains. Experts in both domains have used various approaches to "divide and conquer" the breadth and complexity, including in the University of Melbourne context the use of frameworks that separate "concerns" (for example into People, Process, Data, and Technology for Enterprise Architecture [11], and Culture, Infrastructure, Policy, and Organisation for Digital Preservation) [12].

There are, however, some key differences between the domains, with enterprise architecture struggling to achieve true long-term focus, often restricted to effective planning with horizons of less than a decade. Enterprise architecture teams also tend to focus on data and information more as something to be dealt with rather than as an asset, with older data often seen as something to be "archived off" and devalued. This is evident in the naming of a widely used cloud archive as "Glacier". These aspects are driven by an underlying imperative to optimize outcomes from limited resources, which tends to be linked to a focus on cost and financial risk. This drives the "automate as much as possible" approach that assumes people are an expense to be removed.

In contrast, from a preservation perspective, information can gain greater value over time, with the risk focus dominated by emphasis on "held in trust" and the impact on resources (especially the ongoing financial resources required for long-term digital preservation) as a consequence. In addition, digital preservation practitioners have inherited an understanding from archival mandates and principles that the "human" aspects are critical, and believe that the identification of the items to be preserved will require judgement and may never be fully automatable.

A good example of the differences we have discovered in the University of Melbourne context between the two "tribes" is the use of the phrase "long-term". For Enterprise Architecture this may be as short as 5 years, where for digital preservation this may be centuries.

III. Digital preservation and enterprise architecture at the University of Melbourne

As a world-class research institution, the University of Melbourne generates considerable digital materials of enduring value that will be relied upon into the future to support the university's functions, accountability, and legacy. Management

of digital information is now increasingly recognized as a critical capability for the university, and the ability to find and leverage the "gold" is particularly important.

Scholarly Services and Enterprise Architecture are two distinct and distributed business units at the University of Melbourne. Both have university-wide goals to improve the way the university functions by streamlining business processes and offering sustainable services for teaching and research.

Scholarly Services is a business unit that includes the university's librarian and archivist "information managers", and it is the custodian and champion of the university's ten- year Digital Preservation Strategy. The Digital Scholarship team within Scholarly Services is the responsible division for driving the goals of the Strategy through iterative digital preservation project work (ongoing since March 2016). A key objective of the Digital Scholarship team is to leverage state- of-the-art technologies to guide the ways in which teaching, research, and engagement are performed to ensure that scholarly information resources are sustained through time.

The Enterprise Architecture team within Infrastructure Services works with business and technology stakeholders to develop roadmaps and plans which optimize investment in technology across the university. This is achieved by:

- Defining agreed target states which can be used to better coordinate and align multiple initiatives
- Helping ensure that the trade-offs between longer- term operational efficiency and short-term value generation are identified, debated and resolved
- Identifying focus areas from an operational/ IT perspective which allows better targeting technology investment
- Assessing new demand for alignment to roadmaps and recommending technology options
- Developing solution architectures to support the delivery of business initiatives.

A. Analysing Terminology Differences

As an exercise to better understand the terminology, topical, and conceptual differences between digital preservation and enterprise architecture at the university, we applied qualitative analysis techniques to compare two sets of core documentation. One set of documents was the internal digital preservation policy framework, currently under development, and the other the internal Enterprise Architecture Handbook.

The internal digital preservation policy framework covers all of the elements considered essential for providing a common understanding of how digital preservation activities are undertaken at the University of Melbourne. The framework serves as an initial point of reference for understanding:

- How the university approaches digital preservation
- The current state of digital preservation at the university
- Who is responsible for various digital preservation activities
- How sustained funding is ensured
- How to make decisions about what to preserve
- How to develop operational procedures for digital preservation work across varied and multiple domains that are responsible for digital content generation and management

The internal Enterprise Architecture Handbook documentation contains material that is used by architects to help describe the structure and activities of the Enterprise Architecture team at the university. It contains the operating model for enterprise architecture, the roles and responsibilities of enterprise architects, and a clear description of university stakeholders, the range of services provided, and the architecture governance structure that ensures the most effective management and use of IT.

An initial analysis of term frequencies within these documents indicated significant differences in the terminology of each practice. (Textual analysis was undertaken using the tools Voyant [13] and NVivo [14]). Word clouds of the most frequent terms are a simple demonstration of this apparent difference. (Figure 1.)

Figure 1 Digital Preservation term frequency cloud top, Enterprise Architecture bottom

A Correspondence Analysis [15] was applied to the combined set of documents to further investigate terms and document similarities. (Figure 2.) The analysis confirmed that both the associated term distribution, especially the most frequently used terms, and the corresponding document sets are at opposite ends of the most significant dimension. However, the analysis also revealed a cluster of common lower frequency terms shared by both sets of documentation: ("management", "standards", "support", "required", "process", "strategies", "decision".)

An approach was then adopted to begin exploring topics and more abstract concepts. The aim was to determine if the apparent document separations were simply "language" differences between the two practices. Computational topic modelling was used to extract the ten most probable topics in both of the sets. The Latent Dirichlet allocation technique was used for topic modelling with over 20,000 iterations performed on each set [16].

Topics resulting from this technique are essential groups of commonly associated and likely related terms found throughout the texts. This revealed that the majority of the common lower frequency terms (found in the Correspondence Analysis) also formed part of the ten most probable topics in both of the sets. This could indicate an overlap in the key topics, or it could represent a change in

the areas to begin appropriating language from the other (a result of collaboration).

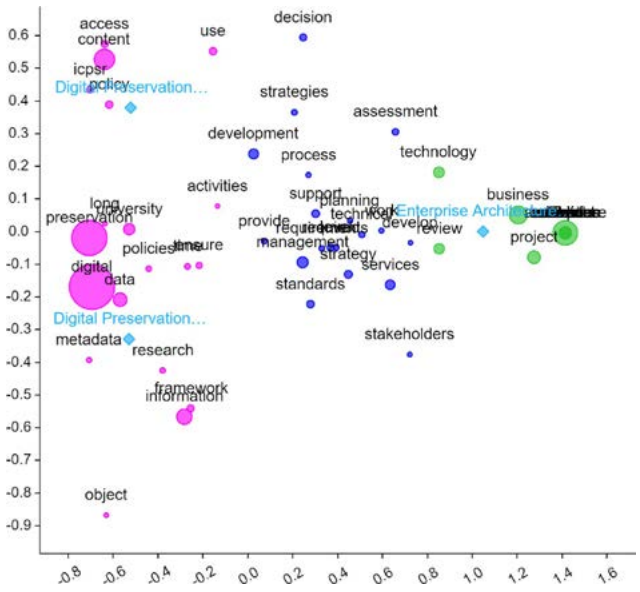


Figure 2 Correspondence Analysis of Documents. Light blue points and labels indicate documents and relative distance (Digital Preservation found left, Enterprise Architecture found right). Other colours represent clusters of terms, and each point/circle is a term with relative distance the association between terms. The point/circle size indicates relative frequency of each term.

An open coding approach was then applied to the digital preservation policy framework documents to identify the initial abstract concepts with closer analysis, and a number of core responsibilities were identified within the Enterprise Architecture Handbook documentation (see Table 1).

Digital preservation concepts and enterprise architecture responsibilities

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Digital preservation concepts	Enterprise architecture responsibilities
Active management	Active management for change
Awareness	Business vision
Managing risk, meaningful access	Business strategy
Changing technologies	Information and technology
Consistent practice, continued authenticity	Consistent practice
Organizational direction	Organizational direction
Roles and responsibilities	People, processes
Documenting	Documenting/describing
Sustainable funding	
Community engagement	
Value	

TABLE I

The digital preservation concepts identified involved: active management, awareness, changing technologies, consistent practice, continued authenticity (of digital content), managing risk, meaningful access (ensuring this for digital content), organizational direction, roles and responsibilities, sustainable funding, wider community engagement, documenting, and the concept of value.

The Enterprise Architecture documentation identified a number of overlapping responsibilities: "consistent practice", "documenting" and describing, active "management" for "change", "organisational direction" which involves business vision and strategy. Enterprise architecture explicitly deals with layers of the business – people, processes, information and technology – as does digital preservation, which is also clear from the range of concepts identified.

Digital preservation can bring a different perspective to an organisation that complements enterprise architecture, as suggested by the differences in concepts and focus areas, and further confirmed by discussions between the teams. We recognise the need for ongoing, sustainable funding as a core digital preservation concept, as well as investing in community engagement (e.g. developing guidance for content creators to better implement measures for sustainable digital materials; encouraging community sharing of digital preservation knowledge and expertise). Also highlighted though the digital preservation documentation is the importance of the concept of long- term "value," which drives many digital preservation methods.

The assumed point of connection between both practices would typically be the technology systems or IT requirements, and initial engagements with Enterprise Architecture by the Digital Scholarship team at the University of Melbourne were actually triggered by this need. However, the qualitative analysis of alignment we have undertaken indicates a broader range of connections and similarities between the two practices.

This analysis was useful in demonstrating to us that it is essential to find ways to work with this "meeting of the mindsets", as we both approached this collaboration with different methods and approaches. There were also the initial assumptions to overcome, including that digital preservation was looking only for a technical solution, whereas the Digital Scholarship team was attempting to communicate to infrastructure staff the need for holistic inclusion of the cultural, policy, process, and governance elements that must be developed and enhanced in conjunction with any technology "solution".

Our combined experience was that the initial struggle between Enterprise Architecture and the Digital Scholarship team at the university in understanding each other was, in fact, related to the framing and the use of language. For many in Enterprise Architecture there is a "corporate" mindset, which is focused on cost, whereas digital preservation is about "trust", where the organisation is responsible for the long-term management of digital assets.

In addition, because of the University drive to develop "business cases" that quantify the value of the investment in adding new capabilities (such as digital preservation), the monetary component often dominates. We have also found a need to distinguish digital preservation requirements from a more records management approach that focuses on compliance (which enterprise architecture teams are often more familiar with than they are with digital preservation).

With non-digital preservation staff there is still a need to clarify why digital preservation is required. We have found that "stories" are useful to overcome language and conceptual differences: for example, reference to significant investments that the university has made in particular collections and the consequent need to effectively "preserve" that investment. Another "story" that resonates widely is the example of climate change research, which needs to identify longitudinal datasets and to leverage diverse sources such as ice cores and tree ring data.

By identifying and appreciating our differences in terminology, and recognising common areas of connection and the extensive range of organizational and socio- technical levels on which we need to engage with each other not just for the initial infrastructure design and solution, but also into the future – we are better placed to realise the broader goals of the university's ten-year Digital Preservation Strategy.

IV. The collaborative journey of digital preservation and enterprise architecture

The collaborative journey has not been easy, and it has taken over a year of continual engagement to arrive at the constructive point we now find ourselves. This was due to a range of factors, including issues caused by:

- Changeable funding cycles for university projects
- Fluctuating rules around the development of business cases for new projects
- The issues involved in identifying and subsequently engaging with the right technology experts who understand drivers and needs for digital preservation
- The challenge of having a range of key stakeholders not fully understanding how broad and wide-ranging digital preservation requirements are across the university.

A. The Catalyst and Way Forward

The Digital Scholarship team began seeking collaborative ways of working with the university's central Infrastructure Services unit in late 2017, in order to drive implementation of a university-wide core digital preservation platform and service. At this early stage, the Digital Scholarship team was actually not aware of a new iteration of the role and function of Enterprise Architecture that had recently been put in place within the broader Infrastructure Services unit.

Initial engagements with Infrastructure Services and Solution Architects (who focus on delivering specific and well-defined technology solutions) were not ideal, in that the skills needed to drive the overall platform design for digital preservation at a university-wide level are more aligned with the role of an Enterprise Architect. This appreciation took time to emerge and was the most significant barrier for the Digital Scholarship team to overcome. Solution Architecture in the university context is concerned with a smaller subset of the functions that an Enterprise Architect performs. A Solution Architect was able to help the Digital Scholarship team to identify potential solutions to defined problems, assist with implementation planning for those solutions, and help put in place governance and change management to help embed the solutions. However, what digital preservation infrastructure implementation at a university level for various different domains of digital content requires in the first instance is an architecture vision, an overall design, and a plan for how that vision could best fit within the business and information systems architecture of the university. These areas are the remit of Enterprise Architecture.

The positive outcome from the considerable time and effort put into this first engagement is the knowledge sharing we were able to do: both the Digital Scholarship team and the university technology experts were able to share their expertise with each other, creating greater awareness of each others' roles and challenges, thus enabling both cohorts to arrive at greater tolerance and understanding – a progressive place to be, for future work together. Having learned a lot from this initial engagement, in June 2018 the Digital Scholarship team engaged with a newly employed Enterprise Architect tasked with enhancing the research domains of the university (e.g. the business systems and processes for supporting research practice). A common language was established by utilizing the OAIS framework [17] and we began quite quickly to understand the value in working together.

The skills that the Enterprise Architect brought (high- level platform design thinking and an understanding of how the university-wide technology landscape functions and interacts), combined with digital preservation expertise, enabled us to collaboratively formulate a design for a core foundation platform and the associated processes suitable to the university context. The digital preservation project governance endorsed this enterprise design plan in July 2018, and the latter half of 2018 delivered an extensive evaluation process for procurement of a digital preservation system (Figure 3.)

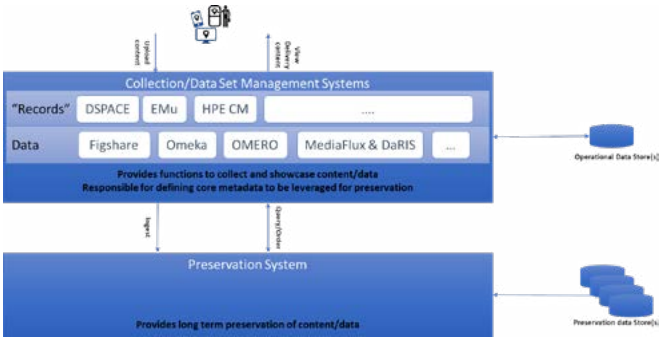


Figure 3 Design for a core foundation platform suitable to the university context, with planning and funding activities underway to implement and begin an ongoing process of ingestion.

V. Learnings

With this collaborative approach, the Digital Scholarship team has achieved much greater success with current digital preservation business cases, and we have greater hope for our future business cases as we seek the next phase of funding for project work. The partnership with Enterprise Architecture at the university has directly supported and strengthened our joint success, and we are now aiming to embed a more sustainable funding stream for digital preservation than the current situation of two-year funding cycles. The level of respect typically afforded to enterprise architects by an organization can make them powerful allies for driving organizational change, as Sam Searle articulates: "In our organisational context, enterprise architects are trusted by very senior executives; they are perceived as neutral and objective, and the processes that they use are understood to be systematic and data-driven" [2].

Digital preservation project work at the university has greatly benefited from the two-way knowledge exchange and the benefits of the enterprise architecture mindset to drive the change of business processes to support long-term preservation and access for digital materials. Digital preservation knowledge, concepts, goals, and processes can be disseminated and communicated more easily throughout the organization, in conjunction with Enterprise Architecture improvement work across multidisciplinary teams at the university. Through this collaboration, the Digital Scholarship team now fully recognizes the essential role of Enterprise Architecture in the university ecosystem, how we can best align our strengths, and how we can overcome differences in approach towards a common goal.

A concrete outcome of the collaboration for the university is that digital preservation and related aspects such as digital collection management have now been added to the internal Enterprise Capability Models and Roadmap to ensure they are addressed with the same importance as other core capabilities such as Enterprise Resource Planning (ERP) and Student Management. In addition, the Enterprise Storage Strategy now includes digital preservation as a "first class" capability that is to be supported for all university data domains, along with current, backup, and archive storage tiers (Figure 4.)

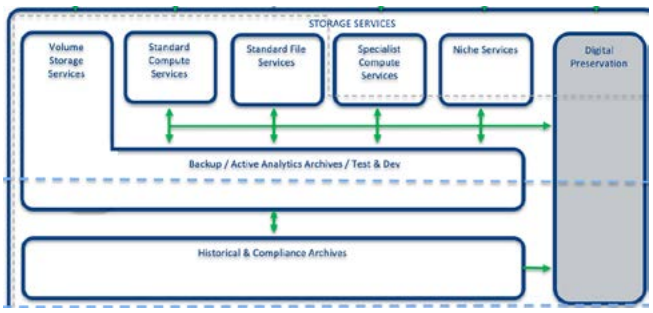


Figure 4 Enterprise Storage Strategy including Digital Preservation (far right) as a "first class" capability underpinning all university data domains

Because of this meeting of the mindsets of digital preservation and enterprise architecture, there is now wider understanding that digital preservation is an entirely new capability being added to the organization, one that underpins and helps enable many of the university's strategic goals. There is

more understanding throughout the wider central Infrastructure Services team that digital preservation is a capability that requires a technology foundation, but also one that requires ongoing work around resourcing, policy, process, and governance, in order to make it function effectively and enhance the organization's ability to maximise the long-term value of its digital assets. Now, there is a collegial two-way appreciation: digital preservation needs to engage with Enterprise Architecture regarding holistic layers and an enterprise approach; and Enterprise Architecture has discovered an understanding of digital preservation as a holistic platform and program of organizational change, enhancing technology solutions.

We acknowledge that not all organisations embarking on the implementation of digital preservation will have an Enterprise Architecture function or equivalent role with which to collaborate. However, many of the terms, concepts and in particular fundamental differences, that were explored in our "meeting of mindsets" are not exclusive to enterprise architects and are common to other technology specialists. The holistic mindset common to both digital preservation and enterprise architecture has provided the opportunity to better place digital preservation within the enterprise at the University of Melbourne. But through sharing our findings we seek to enable others to recognise differences and find common ground more easily with IT practitioners in general.

The broader message we have to share from our particular case study is that the influence of good interpersonal relationships cannot be overstated in the quest for successful collaboration to realise digital preservation goals. Collegiality, an openness to change, preparedness to listen, and general willingness and determination to challenge existing organizational structures and operations that do not facilitate effective digital preservation implementation have all been major factors in this particular collaboration. Even the act of collaboratively writing this paper has had immediate cultural impact within the university – the development of the paper has generated wider awareness of digital preservation goals and their importance, and it has enabled further collegial connections to be built across our previously siloed and unconnected business units, paving the way for collaboration into the future.

The advice we have to share with other organizations faced with the need to implement digital preservation is to continually seek and maintain relationships with those who need to be key stakeholders in managing digital materials, wherever they sit within the organization. Although this can be a challenging path to follow, it can also be a powerful way to progress with digital preservation goals. Forging and maintaining such relationships can be achieved in part by persistently aiming for clear terminology understandings, and clear communication of needs and mindsets. We offer a message of hope and persistence: clarifying the why and what of the organization's digital preservation goals so that everyone understands the purpose and the benefits can help to forge the relationships that enable action. Our experience reflects the advice offered by Scott Prater in that "what is true in life is also true in digital preservation discussions: generosity, mutual respect, and patience win the day...frame requests as interesting problems...not as demands to be met" [7]. By collaboratively tackling the "interesting problems" that digital preservation presents and ensuring that all the required mindsets are included early on when embarking on the journey, it is possible to find common ground to move forward together, utilising different skillsets to meet shared goals.

VI. NEXT STEPS

The work that has been undertaken through this collaboration thus far is highly focused on the technological aspects of implementing a university-wide digital preservation technology foundation. But to continue evolving, we need to explore further how the collaboration and relationship building with Enterprise Architecture at the university can enable the other priorities (Culture, Policy, and Organisation goals, not just Infrastructure) within the university's Digital Preservation Strategy [1]. Through the three years of project work for digital preservation implementation that has been undertaken so far, including outreach and advocacy (e.g. workshops, presentations, consultations), policy, procedure, and processes analysis and reviews, and pre-ingest and ingest workflow developments with various existing expertise at the university, we have a solid foundation for further engagement.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Our analysis comparing the terminology, topical, and conceptual differences between digital preservation and enterprise architecture can be taken further, particularly by investigating a comparison of the standards and frameworks utilized by each field of practice. By examining how these standards complement each other, and how they diverge, we may better understand how best to work together to achieve our various aims and to implement digital preservation functionality across a large distributed organization. We have also started thinking about possible applications of emerging technologies, such as machine learning, within the business systems of the university, and how the logging practices of digital preservation metadata could be a key enabler for this planning and implementation. Analysing what such processes would require, and how especially digital preservation metadata could be used, could prove a valuable next step for the university.

We find ourselves having a similar broad aim for the future as Searle's own case study goal: "to encourage other librarians [digital preservationists] to learn more about architects' work practices and to seek opportunities to apply EA methods...for the benefit [of] the organisation as a whole" [2]. We also have a keen interest in sharing the results of this collaboration with the wider enterprise architecture community (e.g. at enterprise architecture conferences and events). Broader sharing of the benefits of collaboration between enterprise architecture and digital preservation could help pave the way for establishing value in long-term thinking within other increasingly corporate- focussed organizations.

Acknowledgments

We would like to thank Ross Harvey and Michelle Borzi for providing editorial guidance for this paper. We would also like to thank the University of Melbourne leadership that has invested in a robust Enterprise Architecture function, and that continues to support digital preservation program development.

References

[1] The University of Melbourne, "Digital Preservation Strategy 2015-2025 Vision Mandate and Principles" http://hdl.handle.net/11343/45135

[2] S. Searle, "The Benefits of Enterprise Architecture for Library Technology Management: An Exploratory Case Study," Information Technology and Libraries, vol. 37, no 4, p. 28; p.44, p.43, 2018 https://doi.org/10.6017/ital.v37i4.10437

[3] S. Katuu, "The Utility of Enterprise Architecture to Records and Archives Specialists," December 2018 [2018 IEEE International Conference on Big Data] https://www.researchgate.net/publication/329235405_The_Utility_of_Enterprise_Architecture_to_Records_and_Archives_Specialists

[4] NSW Government State Records, "Recordkeeping in Brief 59 An Introduction to Enterprise Architecture for Records Managers," https://web.archive.org/web/20120502184420/https:/www.records.nsw.gov.au/recordkeeping/government-recordkeeping- manual/guidance/record-keeping-in-brief/recordkeeping-in-brief-59- an-introduction-to-enterprise-architecture-for-records-managers

[5] S. Anthony, "Implementing Service Oriented Architecture at the Canada Institute for Scientific and Technical Information," The Serials Librarian, vol 55, issue 1-2, pp. 235-253, October 2008 https://doi.org/10.1080/03615260801970907

[6] P. Allega, "Defining EA: Low Barriers to Entry" https://web.archive.org/web/20100901131058/http://blogs.gartner.com/philip-allega/2010/08/11/defining-ea-low-barriers-to-entry-my- mother-has-an-ea-definition-too, 2010

[7] S. Prater, "How to Talk to IT About Digital Preservation," Journal of Archival Organization, vol 14, issue 1-2, pp.90-101, https://doi.org/10.1080/15332748.2018.1528827

[8] Enterprise Architecture Research Forum, "Defining Enterprise Architecture," https://web.archive.org/web/20100820165859/http://earf.meraka.org.za/earfhome/definining-enterprise-arcitecture

[9] R. Ovenden, "We Must Fight to Preserve Digital Information," The Economist, 21 February 2019, https://www.economist.com/open- future/2019/02/21/we-must-fight-to-preserve-digital-information

[10] Digital Preservation Management Tutorial, "Where to Begin?" 2014, http://dpworkshop.org/dpm-eng/conclusion.html

[11] The University of Melbourne, "Enterprise Architecture Handbook," unpublished

[12] J. Weatherburn, "Establishing Digital Preservation at the University of Melbourne," September 2016 [Proceedings of

the 13th International Conference on Digital Preservation] http://hdl.handle.net/11343/143364

[13] S. Sinclair, and G. Rockwell, “Cirrus”, Voyant Tools, Version VoyantServer 2.4 M17, 2018 https://voyant-tools.org/

[14] QSR International Pty Ltd, NVivo qualitative data analysis software, Version 12, 2018

[15] S. Sinclair, and G. Rockwell, “ScatterPlot”, Voyant Tools, Version VoyantServer 2.4 M17, 2018 https://voyant-tools.org/

[16] S. Sinclair, and G. Rockwell, “Topics (jsLDA by David Mimno”, Voyant Tools, Version VoyantServer 2.4 M17, 2018 https://voyant-tools.org/

[17] Consultative Committee for Space Data Systems, “Reference Model for an Open Archival Information System (OAIS),” 2012 https://public.ccsds.org/pubs/650x0m2.pdf

FAIR DATA WORK @ DANS

Implementing Metrics, Reviews, and More

Peter Doorn
DANS
The Netherlands
peter.doorn@dans.knaw.nl
ORCID 0000-0002-8246-675

Marjan Grootveld
DANS
The Netherlands
marjan.grootveld@dans.knaw.nl
ORCID 0000-0002-2789-322X

Francesca Morselli
DANS
The Netherlands
Francesca.Morselli@dans.knaw.nl
ORCID 0000-0003-05010-9811

Abstract – Ever since the origin of the FAIR data guiding principles, various members of the DANS staff have been involved in a variety of activities on thinking about their implications and implementing them. This paper presents an overview of the fruits of our work so far and sketches our ideas for the years to come. We were involved as co-authors of the original publication on the FAIR principles, developed and tested FAIR metrics, worked on tools to rate the FAIRness of datasets, on a FAIR checklist for researchers, we evaluated how our own data archives score on FAIRness, we compared the principles to the requirements of the Data Seal of Approval and the CoreTrustSeal, explored the applicability of the FAIR principles to Software Sustainability, prepared guidelines for FAIR data management, and we lead the prominent Horizon 2020 FAIRsFAIR project in the context of the European Open Science Cloud.

Keywords – FAIR data, repositories, CoreTrustSeal, research data management

Conference Topics – 3. Exploring New Horizons; 5. The Cutting Edge

I. Introduction

At a Lorentz workshop in Leiden in January 2014, the idea of the FAIR guiding principles was born [1]. They were formulated as a minimal set of community-agreed guiding principles to make data more easily discoverable, accessible, appropriately integrated and re-usable, and adequately citable [2]. In the FAIR Data approach, data should be:

Findable – Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets;

Accessible – Stored for long term such that they can be easily accessed and/or downloaded with well-defined license and access conditions (Open Access when possible), whether at the level of metadata, or at the level of the actual data content;

Interoperable – Ready to be combined with other datasets by humans as well as computer systems;

Reusable – Ready to be used for future research and to be processed further using computational methods.

DANS vice-director Ingrid Dillo was one of the 53 co-authors of the paper published in Nature Scientific Data [3]. Since that very beginning, various members of staff at DANS contributed to the FAIR success story, and to putting the principles into practice. This proves to be a highly stimulating but by no means straightforward process. To support others on their road to FAIR data in this paper we will outline our activities and their results so far, and we will end by outlining the next steps.

II. FAIR and Data Seal of Approval

When DANS was set up in 2005 by the Royal

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO), one of our tasks was to think of criteria for trustworthy digital repositories (TDR). At that time, there were other initiatives to formulate an ISO standard, which resulted in ISO 16363 in 2012, and the German nestor seal, which was codified as DIN 31644 in 2013. We considered these initiatives as laudable, but also as ambitious and difficult to apply in a scientific context, and therefore aimed at a limited number of core criteria, which became the Data Seal of Approval (DSA) in 2010. The DSA relies on peer review for the assessment of repositories aspiring to comply. We started their development by defining five principles, which bear a remarkable resemblance to the FAIR principles. Data in a DSA-certified repository:

1. can be found on the internet
2. are accessible
3. are in a usable format
4. are reliable
5. can be referred to

These five DSA principles are refined into 16 guidelines [4], whereas the four core FAIR elements are expanded into 15 (sub) principles [3].

Despite the similarity, there are some differences as well. First, the DSA principles were designed to be applied to repositories, whereas the FAIR principles refer to data and metadata at an unspecified level of aggregation. In addition to that, for the DSA the *long-term* preservation of and access to data is central, a viewpoint that is not explicit in FAIR.

Second, the FAIR principles explicitly aspire to be applicable both by humans and by machines or automated processes, whereas the DSA does not mention this.

Third, "a usable format" can be considered as an aspect of either interoperability or reusability (or both). The idea was that data in a DSA certified repository should be in a format that would make data usable across software platforms and time.

Fourth, reliability (DSA) is not the same as reusability (FAIR), but there is similarity in intention: the core idea is that re-users of data can rely on their integrity, precision and validity, aspects that are not explicitly mentioned in the FAIR principles.

Finally, citability was seen as such a fundamental criterion, that it was treated separately in the DSA principles; in the FAIR principles, citability is equally important, but it is treated as an aspect of Findability (F1). Both DSA and FAIR stress the importance of data citation and referencing via persistent identifiers.

On the basis of this comparison, it was our conviction that the DSA and FAIR principles are not only similar, but also complement each other. DSA focusing on long term preservation and access of data in trustworthy repositories, FAIR stressing interoperability, applicability for machines, and focusing on data and metadata in whatever context, either in a curated repository or otherwise.

III. FAIR Data Assessment Tool

It is exactly this complementarity that led us to explore how the FAIR principles do apply to datasets within TDRs. In other words, rather than focusing on *making* data FAIR, we focused on measuring the FAIRness of *existing* data. For this purpose, we used our former data review approach and operationalized the FAIR principles in the form of an online questionnaire [5]. The idea was that answering the questions would result in a FAIRness rating of datasets stored in any DSA-certified archive. However, operationalizing the FAIR principles and making them independent of one another proved to be no easy task. Some of the principles address the same or overlapping aspects of a data set, i.e. are interdependent, and some of them are hard to interpret, even with the help of the available explanations [6]. Moreover, some of the principles include qualitative or subjective elements, such as "broadly applicable language for knowledge representation" (I1); "richly described with a plurality of accurate and relevant attributes" (R1). Sometimes the specification of a principle can even be challenged, e.g. "(Meta)data use vocabularies that follow the FAIR principles" (I2), which results in an endless loop.

We had especially great difficulty to find valid operationalizations for Reuse. For practical reasons, we decided to move some of the principles from the R to other FAI-letters under the assumption: if data is findable, accessible and

interoperable, it is by definition reusable (or: F+A+I=R). We also decided to include a measure of openness, not just of metadata, but also of datasets themselves, whereas the FAIR principles explicitly state that open data per se is not a FAIR requirement: there can be valid reasons for data not to be openly accessible (e.g. for privacy reasons), and in such cases the metadata (which should always be open) should explain why and specify the access conditions. In spite of this, we thought that researchers wanting to reuse data would like to see immediately if a dataset is open or not on the basis of the FAIR data assessment tool FAIRdat. For this purpose, we devised a FAIR badge with scores from 1 to 5 stars on the F-A-I, and decided to use the R as an average reusability score of the dataset under assessment.

Figure 1. Example of a FAIR badge designed for FAIRdat

A prototype was set up in SurveyMonkey (R) in 2017 [7] and tested by colleagues on a couple of datasets in the DANS archive, and next in an international setting by four sister archives (see Table I).

TABLE I.

Test assessments with FAIRdat tool of datasets in other repositories than DANS EASY.

Name of Repository	Number of Datasets reviewed	Number of Reviewers	Number of reviews
VirginiaTech	5	1	5
Mendeley Data	10	3 (for 8 datasets) 2 (for 2 datasets)	28
Dryad	9	3 (for 2 datasets) 2 (for 3 datasets)	16
CCDC	11	? (no names) 2 (for 1 dataset)	12

Source [8].

Finally, we presented and tested the prototype at the Open Science FAIR in Athens [9] and received feedback from 17 participants in the workshop we organized on that occasion.

All in all, the feedback was mixed, which corresponded to our own experiences. Most people liked the idea of the FAIRdat approach a lot, but there were critical voices as well. The simplicity of the well-documented tool was a pro, and most people found the star-ratings useful. Criticisms concerned: the subjectivity in some questions; misunderstandings about the intention of some questions; worry that data that could not be open for valid reasons would never receive a high overall FAIR score; the (perceived) absence of questions under Reusability; unclarity about the target audience for the questionnaire (researchers, data specialists, data archive staff?).

One other element we found difficult to resolve was that many research datasets consist of multiple data files/objects, and that the questions targeted the whole dataset, whereas sometimes different individual files would be assessed differently. Moreover, some FAIRdat questions asked were already solved at the level of the repository, and were actually not necessary.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

In short, it was necessary to take one step back to rethink our approach and how to improve the prototype [10].

IV. FAIR metrics

In the meantime, in the summer of 2017 a small group of FAIR insiders of the first hour came together in Leiden to operationalize the principles into metrics. Peter Doorn was invited to participate [11]. The group had a number of intense discussions on how to measure FAIRness. The ambition was also that the metrics should not be implemented by asking humans, as in the DANS FAIRdat approach, but by machine processes, as automatic as feasible. The group started by formulating points of departure for the metrics, which were to be clear, realistic, discriminating, measurable and universal [12]. A metric form was devised, stating for each metric:

- Metric Identifier
- Metric Name
- To which principle does it apply?
- What is being measured?
- Why should we measure it?
- What must be provided?
- How do we measure it?
- What is a valid result?
- For which digital resource(s) is this relevant?
- Examples of application across types of digital resources
- Comments

The group ended up with 14 exemplar metrics, which were published on GitHub to stimulate debate, to invite submissions of further metrics, and to test them [13]. A paper describing the metrics was published in Nature Scientific Data [14] after the preprint was first uploaded to bioRxiv [15]. Although the FAIR metrics exercise is both intellectually challenging and stimulating, the applicability of most metrics to concrete data sets seems as yet an aspiration rather than a reality.

As DANS did not see a short-term opportunity to implement the FAIR metrics, we decided to move into new directions which would provide more immediate practical results. These are:

- a FAIR checklist for researchers (section V)
- a revision of the FAIRdat tool (section VI).
- a self-audit of compliance of the DANS-EASY archives with the FAIR principles (section VII).
- an evaluation of the correspondence between the CoreTrustSeal (CTS) and the FAIR principles, including a gap-analysis (section VIII).

V. "FAIR Enough" Checklist

Even though the FAIR principles at the level of the four letters have an immediate appeal to many, partly perhaps because of the clever acronym, the more detailed principles are not easily understood by the non-initiated or by non-data specialists, that is: by most researchers. Therefore, we decided to create a simple questionnaire form explaining the core ideas behind FAIR, enabling researchers to do a self-check on the data they are working with, aiming more at awareness raising and educating researchers about the principles than following them to the letter.

The "FAIR Enough" checklist is a short and concise questionnaire for researchers who are planning to deposit their data in a repository. It covers FAIRness at different levels: the repository, the metadata, the dataset, and files or objects within a dataset. It is set up using Google Forms and is currently in beta [16].

The questions are formulated in terms that aim to be understandable for non-data specialists. There is no one-on-one correspondence to the more detailed FAIR principles, although the four core elements are covered. The checklist offers brief explanations of terms and concepts, including reference to trustworthy repositories and CoreTrustSeal. At the end, an indication of the FAIRness is obtained as an overall score. Recommendations are provided for questions resulting in negative answers with respect to FAIR, so that researchers can take measures to make their data more FAIR.

VI. FAIR Data Review Tool (2nd prototype)

On the basis of the testing and evaluation of our first prototype of the FAIR data assessment tool, we worked on a second version, that is now also available as a Google Form. It explicitly focuses on data in a CTS-certified TDR. We first devised a table with all 15 FAIR (sub)principles, and determined at which level they can be best assessed [17].

As will be argued in section VIII, TDRs compliant with CTS take care of virtually all FAIR principles. Only for a minority of the principles, especially those with subjective elements, there are variations for data within a repository. For example, all datasets in a CTS repository "are assigned a globally unique and persistent identifier" (F1). But the degree to which "(meta)data are richly described with a plurality of accurate and relevant attributes" (R1) can vary from dataset to dataset: some data in a TDR are better documented than other.

Moreover, most FAIR principles adhere to both data and metadata, and hence for those principles that may have different scores within a repository, we formulated questions for data and metadata separately. Therefore, in the new data review tool, the questions cover the FAIR data principles, in so far as they are not already guaranteed by a CTS-certified data repository.

One of the criticisms often voiced about the FAIR principles is that they do neither cover aspects of data quality such as completeness, precision/accuracy and validity, nor the logic of the data organization. Therefore, we added some questions covering these aspects as well.

The new FAIR Data Review Form is aimed at both data specialists and at researchers reusing data from a CTS-certified trustworthy data repository. It is also used for reviewing data belonging to data papers submitted to the Research Data Journal (RDJ) for the Arts and Humanities [18].

Reviewers are requested to answer 18 questions about how they rate the data and accompanying documentation or metadata. The questions are divided over four sections:

- General quality and FAIRness of the data: Q1 - Q8
- Quality of the metadata (i.e. description and documentation of the data): Q9 - Q14
- Further FAIR characteristics of the data (set and objects): Q15 - Q16
- Further FAIR details on the metadata: Q17 - Q18

In addition to providing a rating, reviewers can add qualifying remarks. The answers will result in an overall rating of the data reviewed, but the reviewer is also asked to give a summary verdict. With the consent of the reviewer, the data review can be published. In this way, we aim to give the author of the data paper and data users an impression of the fitness for reuse of the data.

VII. FAIRness of DANS Archives

Soon after the publication of the FAIR principles, three staff members of 4TU.Researchdata evaluated the compliance of a sample of 37 repositories, online databases, and research infrastructures with their interpretation of the FAIR principles [19]. The DANS EASY archive was one of them. In the interpretation of the authors, DANS EASY complied with 11 principles, did not comply with two (I2: (meta)data use vocabularies that follow FAIR principles; and R1.2: (meta)data are associated with their provenance), whereas compliance was not clear for two more (A2: metadata are accessible, even when the data are no longer available; and R1.3: (meta)data meet domain-relevant community standards); see the underlying dataset [20].

We carried out a self-assessment on the basis of the FAIR principles as well, and also found some room for improvement [21]. In our self-audit, we distinguished data from metadata, as they have different FAIR characteristics (or rather: there are different implications with respect to their compliance with the FAIR guiding principles). With respect to metadata, in our evaluation DANS EASY complies with the FAIR principles, except R1.2: "(meta)data are associated with detailed provenance". Which metadata and documentation was entered and edited by whom and when is something to be taken into consideration.

Three FAIR principles explicitly apply to metadata and not to actual research data (F2, F3 and A2). However, it is not legally permitted nor ethically responsible to demand that all data in the archive comply with principle F4: "data are registered or indexed in a searchable resource": data that are legally protected cannot be indexed in a searchable resource, because that would violate their protection! Therefore, we claim that this principle needs to be more precisely specified.

Another problematic principle is A1: (meta)data are retrievable by their identifier using a standardized communications protocol. In the case of DANS EASY, the data are retrievable via the metadata, because the identifier is part of the metadata, not of the data. In our opinion, this demand should also be more precisely formulated in the principles. Finally, there are four principles that only partially apply, because of variations within datasets or/and within the archive:

F1: "(meta)data are assigned a globally unique and persistent identifier": Individual files do not get a persistent identifier; the PID resolves to the data description page (i.e. the metadata).

I2: "(meta)data use vocabularies that follow FAIR principles": Controlled vocabularies used for metadata are few; whether or not vocabularies are used on the level of data depends on the dataset.

I3: "(meta)data include qualified references to other (meta)data": whether or not datasets have references depends on the dataset.

R1.3: "(meta)data meet domain relevant community standards": whether or not such standards are used depends on the dataset (same as with principle I2) and whether such community standards exist at all.

To conclude, the (self-)assessment of data archives in the light of the FAIR principles is a useful exercise: it provides guidance on the improvement of archival systems and procedures, and in some cases it brings to light where the specification of the principles themselves can be improved.

VIII. FAIRness of CoreTrustSeal

As mentioned above, the principles on which the Data Seal of Approval (DSA) was based bear a striking resemblance to the FAIR guiding principles, and the same is true for the successor of DSA, the CoreTrustSeal). We already discussed the subtle differences, and the same applies to CTS. Mokrane and Recker made a detailed mapping of the alignment of the 15 FAIR principles to the 16 CTS requirements. They write: "the proximity of objectives between the certification of trustworthy data repositories and the implementation of FAIR Principles calls for a close examination of their overlaps and complementarities. The characteristics of the data repositories assessed by means of the CoreTrustSeal requirements position TDRs as enabling FAIR data. In particular, the concept of FAIR data cannot be decoupled from the characteristics of the data infrastructure necessary for inferring compliance of the data object with some of the FAIR Principles (mostly under F and A). In addition, CoreTrustSeal TDRs can usefully be considered as having reached baseline FAIR compliance at the data collection level for other principles (mostly under I and R)." [22]. Their comparison is summarized in Figure 2 below. They conclude that in the next revision of the CTS requirements, the FAIR principles should be reflected even more explicitly.

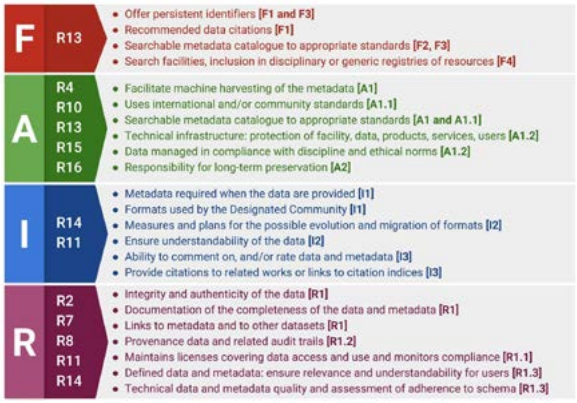


Figure 2. A mapping between the FAIR Guiding Principles and the CoreTrustSeal Requirements. Source: [22].

IX. Guidelines for FAIR Data

Although the FAIR principles have attracted a substantial following among many groups of stakeholders, a lot of explanation, training and advocacy is still needed. DANS contributed to this expertise development in a variety of ways and in various national and international projects, in conference contributions, workshops, webinars, training courses, brochures, etc. Here we restrict ourselves to three examples: first, the "Data Management Expert Guide", an online training module offered by the Consortium of European Social Science Data Archives (CESSDA) contains extensive information for social scientists wanting to make their data FAIR [23]; second and similarly the "Guidelines to FAIRify data management and make data reusable" developed in the context of the Parthenos Project focus on humanities scholars [24].

X. Applying FAIR to Software

Although the FAIR guiding principles originated with research data in mind, it is worthwhile to investigate their applicability to other research outputs, especially software [25]. Software sustainability is an emerging topic in digital preservation. DANS is not yet equipped to support software up to par with data archiving, but it is devoting an increasing amount of attention to it [26, 27, 28], and we think there is an urgent need for a European or International Software Sustainability Infrastructure. We are collaborating with organizations such as the Software Heritage Archive at INRIA in France and the Software Sustainability Institute in the UK to raise the awareness about this need.

Preserving software obviously involves more than just archiving the code. In order to keep software running, much more is required. In 2010 NASA published a report proposing Reuse Readiness Levels (RRLs) for software[29]. In 2011 SSI [30] and in 2016 CLARIAH [31] defined sets of criteria for assessing software sustainability, maintainability, (re)uasability and overall "quality". Both organisations also provide an online (self) assessment tool.

In the Table at the end of this paper a comparison of the good practices proposed by CLARIAH and SSI are presented and mapped to the FAIR principles. The table shows that many of the recommendations already in use today to assess software for readiness and sustainability can be matched to the FAIR principles.

The SSI and CLARIAH criteria under the respective categories of "Usability" and "Sustainability & Manageability" are the most relevant ones in the context of FAIR. A substantial difference, however, is that the FAIR principles are formulated at a more general and less operational level than the SSI/ CLARIAH criteria. It is therefore obvious that their numbers of (sub)criteria are much higher: SSI has 73 criteria for Usability and CLARIAH has 42; for Sustainability & Manageability they have respectively 130 and 45 criteria. Hereby we should remark that CLARIAH distinguishes different support situations, in which the applicability of the criteria varies: software for end users and "experimental" software, which can be either actively supported or be unsupported. Anyhow, the number of software criteria is a multiple of the number of FAIR principles (and also of the FAIR metrics).

Grouped into categories, it is nevertheless fairly well possible to map most main criteria (reflected as questions is Table II) to the main FAIR principles. The FAIR principles do not cover aspects of project management, buildability and installation/ deployment, and the FAIR principles do not cover security and privacy, elements that seem to be most related to Reusability. In this way, we would arrive at four rather operational principles for the Findability of software, 3 for Accessibility, 4 for Interoperability and 8 for Reusability; perhaps we would need an extra "principle" to take care of project management, buildability and installation/deployment of software.

TABLE II.
Mapping of main categories for software sustainability to the FAIR principles (summarized)

Main SSI/CLARIAH criteria for software sustainability	FAIR principle
Is the software easily understood?	F
Is there comprehensive well-structured documentation?	F
Is it straightforward to build from source on a supported system?	Not covered
Is it straightforward to install and deploy on a supported system?	Not covered
Is it easy/intuitive to learn how to use its functions?	R
Does the software perform well?	R
Is the project/software identity clear and unique?	F
Is it easy to see who owns the project/ software?	A
Is an appropriate licence adopted?	A
Is it easy to understand how the project is run and the development of the software managed?	Not covered
Is there evidence of a current/future user community?	R
Is there evidence of current/future ability to download?	A
Is it easy to test the correctness of the source code?	R
Is the software usable on multiple platforms?	I
Is there evidence of current/future developer support?	R
Is the software easy to understand at the source level?	F
Is the software easy to modify and to contribute changes to developers?	I
Is there evidence of current/future development?	R
Is the software interoperable with other required/related software?	I
Does the software comply to requirements for integration into the community (CLARIAH) infrastructure?	I
To what extent is the software reusable?	R
Are security and privacy dealt with adequately?	R? (Not covered)

In order not to make the number of requirements too high, it seems useful to rank them in terms of their importance, using the "MoSCoW criteria" (Must/Should/Could/Won't or Would have). If we do this, we could concentrate on a limited number of requirements: 9 core requirements and about 10 additional ones, with a few decisions left open at this stage.

One of the directions we took, in collaboration with the Netherlands eScience Center (NLeSC), was setting up a "FAIR software route", advising what steps researchers can take to make sure that their research software is "FAIRly" sustained. The NLeSC Research Software Directory and the DANS NARCIS portal will be adapted and expanded to serve this purpose [32, 33].

XI. Outlook: FAIRsFAIR Project

The European project FAIRsFAIR [34], which is led by DANS, aims to supply practical solutions for the use of the FAIR data principles throughout the research data life cycle. FAIRsFAIR addresses the development and realization of a knowledge infrastructure on academic quality data management, procedures, standards, metrics and related matters, which are all based on the FAIR principles. This knowledge infrastructure will be an essential part of the European Open Science Cloud or EOSC. The EOSC itself is envisaged as a research data commons, including all disciplines and member states, associated countries and global initiatives. It is to be sustainable in the long-term, based on sound and transparent data stewardship, in which re-use of scientific outputs is the default. Hence the importance of FAIR principles and practices.

The FAIRsFAIR project addresses stakeholders ranging from research communities, research infrastructures and SMEs to research funders and publishers, analyzing current data policies and making recommendations for FAIR-enhancing policies, practices and data services. It plays a key role in the development of global standards for FAIR certification of repositories, strengthening certification schemas such as CTS, and will also support repositories in implementing these, recognizing the fact of different maturity levels. Data stewards and data scientists will be trained, also with a view to

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

providing training themselves ("train the trainer"). A FAIR competence framework for higher education and a virtual competence center with experts in disciplinary communities will neatly combine formal learning with everyday learning-on-the-job. With its 22 partners and a duration of three years (2019-2022), FAIRsFAIR will foster a FAIR data culture including the uptake of good practices in making and keeping data FAIR. For project leader DANS this is a great platform to contribute our earlier products and expertise to.

REFERENCES

[1] Jointly designing a data FAIRPORT, Leiden, 13-16 January 2014. Available: https://www.lorentzcenter.nl/lc/web/2014/602/info.php3?wsid=602 [Accessed March 19, 2019].

[2] I. Dillo, "FAIR Data in Trustworthy Data Repositories", keynote, in iPRES2017 - 14th International Conference on Digital Preservation, Kyoto, Japan, 25-29 September 2017. Full abstract. Available: https://ipres2017.jp/wp-content/uploads/Keynote-ingrid-edited-by-Nakayama.pdf [Accessed March 14, 2019].

[3] M.D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship", Scientific data, vol. 3, article 160018, 2016. Available: https://doi.org/10.1038/sdata.2016.18 [Accessed March 14, 2019].

[4] Data Seal of Approval. Available: https://datasealofapproval.org/ [Accessed March 14, 2019].

[5] M. Grootveld and J. van Egmond, "Peer-Reviewed Open Research Data: Results of a Pilot", IJDC vol. 7-2, 2012. Available: https://doi.org/10.2218/ijdc.v7i2.231 [Accessed March 14, 2019]. GO FAIR, "FAIR principles". Available: https://www.go-fair.org/fair-principles/ [Accessed March 14, 2019].

[6] E. Thomas, "FAIR data assessment tool" [2017]. Available: https://www.surveymonkey.com/r/fairdat [Accessed March 14, 2019].

[7] P. Doorn and E. Tsoupra, "A Simple Approach to Assessing the FAIRness of Data in Trusted Digital Repositories", keynote in (meta)-data quality workshop, 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, 18-21 September 2017. Abstract Available: http://ceur-ws.org/Vol-2038/invited2.pdf [Accessed March 19, 2019].

[8] P. Doorn, M. Grootveld and E. Dijk, "FAIR metrics: starring your data sets", in Open Science Fair, Athens, Greece, 6-8 September 2017. Available: http://www.opensciencefair.eu/training/parallel-day-3-2/fair-metrics-starring-your-data-sets [Accessed March 14, 2019].

[9] P. Doorn, "One Step Forward, Two Steps Back: A Design Framework and Exemplar Metrics for Assessing FAIRness in Trustworthy Data Repositories", in 11th Plenary Meeting of the Research Data Alliance, Berlin, Germany, 21-23 March 2018. Available: https://rd-alliance.org/wg-wdsrda-assessment-data-fitness-use-rda-11th-plenary-meeting [Accessed March 14, 2019].

[10] GO FAIR Metrics Group. Available: http://fairmetrics.org/about.html [Accessed March 14, 2019].

[11] GO FAIR Metrics Group, "Our philosophy, process and the tools". Available: http://fairmetrics.org/framework.html [Accessed March 14, 2019].

[12] [GO FAIR Metrics Group,] "FAIR Maturity Indicators and Tools". Available: https://github.com/FAIRMetrics/Metrics [Accessed March 14, 2019].

[13] M.D. Wilkinson, S.-A. Sansone, E. Schultes, P. Doorn, L.O. Bonino da Silva Santos and M. Dumontier, "A design framework and exemplar metrics for FAIRness", Scientific data, vol. 5, article 180118, 2018. Available: https://doi.org/10.1038/sdata.2018.118 [Accessed March 14, 2019].

[14] M.D. Wilkinson, S.-A. Sansone, E. Schultes, P. Doorn, L.O. Bonino da Silva Santos and M. Dumontier, "A design framework and exemplar metrics for FAIRness", BioRxiv, 2017. Available: https://doi.org/10.1101/225490 [Accessed March 14, 2019].

[15] P. Doorn, FAIR Data Review Form. Available: https://docs.google.com/forms/d/e/1FAIpQLScjXupm-_RDofaVQhyiotWEgSG62ABHfQew3MIel6Tlpf6ZtA/viewform [Accessed March 19, 2019].

[16] P. Doorn, FAIR questionnaire spreadsheet. Available: https://tinyurl.com/y3lpdz8p [Accessed March 14, 2019].

[17] Research Data Journal for the Humanities and Social Sciences (RDJ), Available: https://brill.com/view/journals/rdj/rdj-overview.xml [Accessed March 14, 2019].

[18] A. Dunning, M. de Smaele and J. Böhmer, "Are the FAIR Data Principles Fair?", IJDC vol. 12-2, 2017. Available: https://doi.org/10.2218/ijdc.v12i2.567 [Accessed March 14, 2019].

[19] A. Dunning, M. de Smaele and J. Böhmer, Evaluation of data repositories based on the FAIR Principles for IDCC 2017 practice paper, TUDelft, dataset, 2017. Available: https://doi.org/10.4121/uuid:5146dd06-98e4-426c-9ae5-dc8fa65c549f [Accessed March 14, 2019].

[20] E. Tsoupra, P. Doorn, E. Fankhauser, Evaluation of the DANS EASY Repository Based on the FAIR Principles, DANS, white paper, 2018. Available: https://doi.org/10.17026/dans-2au-j2z6 [Accessed March 14, 2019]

[21] M. Mokrane, J. Recker, E. Fankhauser, P. Doorn, "CoreTrustSeal certification: enabling FAIR data

repositories", Internal DANS paper, draft version 1, 16 March 2019. To be published.

[22] CESSDA Training; "Data Management Expert Guide", Available: https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/1.-Plan/FAIR-data [Accessed March 19, 2019].

[23] Parthenos "Guidelines to FAIRify data management and make data reusable", December 2018, Available: http://www.parthenos-project.eu/portal/policies_guidelines [Accessed March 19, 2019].

[24] P. Doorn, "Does it make sense to apply the FAIR Data Principles to Software?" Presentation at Sustainable Software Sustainability Workshop, DANS & SSI, The Hague, March 7-9, 2017, Available: https://dans.knaw.nl/nl/actueel/20170307_FAIRSoftwarePrinciplesPeterDoorn.pdf [Accessed March 19, 2019].

[25] P. Doorn, P. Aerts and S. Lusher, "Software sustainability at the Heart of Discovery", working paper DANS and NLeSC, 2016. Available: http://hdl.handle.net/20.500.11755/326f699b-14f9-497c-b810-b2229d7ede48 [Accessed March 14, 2019].

[26] P.K. Doorn and P. Aerts, "A Conceptual Approach to Data Stewardship and Software Sustainability", working paper DANS, 2016. Available http://hdl.handle.net/20.500.11755/59c24848-9cf7-437c-b2d5-e943e9e4a35e [Accessed March 14, 2019].

[27] P.J.C. Aerts, "Sustainable Software Sustainability - Workshop report", Available: https://doi.org/10.17026/dans-xfe-rn2w [Accessed March 19, 2019].

[28] P. Aerts, "Software Sustainability - final report", DANS, in collaboration with the Dutch Digital Heritage Network (NDE) and the Netherlands Coalition Digital Preservation (NCDD), 2016. Available https://www.netwerkdigitaalerfgoed.nl/wp-content/uploads/2018/02/201611_DE_Houdbaar_final_report_software-sustainability_DEF.pdf [Accessed March 14, 2019].

[29] NASA Earth Science Data Systems – Software Reuse Working Group, Reuse Readiness Levels (RRLs). April 30, 2010, version 1.0. Available: https://wiki.earthdata.nasa.gov/pages/viewpage.action?pageId=49446977 [Accessed June 26, 2019].

[30] Mike Jackson, Steve Crouch and Rob Baxter, Software Evaluation: Criteria-based Assessment. SSI, November 2011. Available: www.software.ac.uk/sites/default/files/SSI-SoftwareEvaluationCriteria.pdf and www.software.ac.uk/resources/guides-everything/software-evaluation-guide [Accessed June 26, 2019].

[31] Maarten van Gompel, Jauco Noordzij, Reinier de Valk & Andrea Scharnhorst. Guidelines for Software Quality. September 30, 2016, version 1.1. Available: https://github.com/CLARIAH/software-quality-guidelines [Accessed June 26, 2019]

[32] Netherlands eScience Center, Research Software Directory, Available: https://www.research-software.nl [Accessed March 20, 2019].

[33] DANS, NARCIS, Available: https://www.narcis.nl/?Language=en [Accessed March 20, 2019].

[34] FAIRsFAIR project. Available: https://www.fairsfair.eu/ [Accessed March 14, 2019].

TABLE III

Mapping of main categories for software sustainability to the FAIR principles

CLARIAH Number	CLARIAH Criterion	SSI Criterion	Explanation	FAIR letter	No. of SSI Criteria	No. of CLARIAH Criteria	MoSCoW
5	**Usability**	**Usability**			**73**	**42**	
5.1	Understandability	Understandability	Is the software easily understood?	F	11	6	M
5.2	Documentation	Documentation	Comprehensive well-structured documentation?	F	25	12	M
5.4	Buildability	Buildability	Straightforward to build from source on a supported system?	?	11	4	W
5.5	Installability	Installability	Straightforward to install and deploy on a supported system?	?	19	10	W
5.3	Learnability	Learnability	Easy/intuitive to learn how to use its functions?	R	7	5	C
5.6	Performance	-	Does the software perform well?	R	-	5	C
6	**Sustainability & Manageability**	**Sustainability & Manageability**			**130**	**45**	
6.1	Identity	Identity	Project/software identity is clear and unique?	F	8	3	M
6.2	Copyright & Licensing	Copyright	Easy to see who owns the project/ software?	A	7	3	M
	-	Licencing	Adoption of appropriate licence?	A	5	-	(M)
6.14	Governance	Governance	Easy to understand how the project is run and the development of the software managed?	R	2	?	W
6.4	Community	Community	Evidence of current/future community?	R	11	3	?
6.3	Accessibility	Accessibility	Evidence of current/future ability to download?	A	12	7	M
6.5	Testability	Testability	Easy to test correctness of source code?	R	19	4	S
6.6	Portability	Portability	Usable on multiple platforms?	I	17*	3	C
6.7	Supportability	Supportability	Evidence of current/future developer support?	R	21	2	W
6.8	Analysability**	Analysability**	Easy to understand at the source level?	F	20	8	M**
6.9	Changeability	Changeability	Easy to modify and contribute changes to developers?	I	14	6	W
6.12	Interoperability	Evolvability	Evidence of current/future development?	R	5	1	W
6.12	-	Interoperability	Interoperable with other required/ related software?	I	6	-	S
6.13	Interoperability for community (CLARIAH)	-	Does the software comply to requirements for integration into the community (CLARIAH) infrastructure	I	-	?	C
6.10	Reusability	-	To what extent is the software reusable?	R	-	3	W***
6.11	Security & Privacy	-	Are security and privacy dealt with adequately?	R?	-	2	S

* Several PC/Mac platforms are mentioned, no platforms for mobile devices

** Combine with understandability/documentation

*** Is defined by all the other criteria

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Notes:

- The numbers in the table in the column “CLARIAH number” refer to the sections in their Guidelines. SSI did not number their criteria.
- The columns “CLARIAH criterion” and “SSI criterion” give the headings in the two original guidelines.
- The column “Explanation” gives a question to be answered for each criterion.
- The columns “No. of SSI Criteria” and “No. of CLARIAH Criteria” indicate how many (sub-)criteria both organisations specify for each (main) criterion. For example, SSI has a total of 73 (sub-) criteria for “Usability” and CLARIAH has 42, etc.
- The column “FAIR letter” gives a proposed FAIR software requirement mapped on the SSI/CLARIAH criteria. The letter is followed by a consecutive number.
- The column MoSCoW gives my proposed urgency of the criteria using the priority rules “Must have”, “Should have”, “Could have” and “Would/Won’t have”.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Preservation strategies for an internet-based artwork yesterday today and tomorrow

Claudia Roeck
University of Amsterdam
The Netherlands
c.rock@uva.nl
ORCID oooo-ooo3-2612-5672

Rafael Gieschke
University of Freiburg
Germany
rafael.gieschke@rz.uni-freiburg.de
ORCID oooo-ooo2-2778-4218

Klaus Rechert
University ofFreiburg
Germany
klaus.rechert@rz.uni-freiburg.de
ORCID oooo-ooo2-88o7-o18X

Julia Noordegraaf
University of Amsterdam
The Netherlands
j.j.noordegraaf@uva.nl
ORCID oooo-ooo3-o146-642X

Abstract This paper investigates possible preservation strategies for an internet-based artwork and assesses the strategies that best capture the authenticity of the work for future iterations. Two different preservation strategies are applied for the internet-based artwork TraceNoizer.org from 2001. A third one, a Linux Live CD, was carried out by one of the artists. They are compared and evaluated from the perspective of the long-term preservation of the work's most significant properties. Compared to software-based artworks, the characteristics of internet-based artworks shift the focus of the preservation measures from the stabilization of the software to reduction of server maintenance, protection of server and artwork from internet threats and reduction of external dependencies. This paper suggests solutions how to handle these challenges and discusses its benefits and disadvantages for long-term preservation.

Keywords internet-based art, software-based art, sustainability, long-term preservation, preservation strategies

Conference Topics Designing and Delivering Sustainable Digital Preservation

I. Introduction

Until a few years ago, Internet-based art was not widely collected in contemporary art museums and collecting institutions. The Guggenheim Museum in New York was one of the first contemporary art museums to acquire web-based art (net.flag acquired in 2002). Rhizome, an art organization in New York, has probably the widest experience in the preservation of Internet art. In contrast to most museums, they are focusing on purely digital art. Their collection consists of several hundred digital artworks. The number of museums collecting internet-based art is slowly increasing. For instance, the Stedelijk Museum Amsterdam (NL) acquired several webbased artworks jointly with the MOTI museum in Breda (NL) in 2017, LIMA in Amsterdam is hosting web-based artworks by and for Constant Dullaart since 2018 and the house of Electronic Arts in Basel (CH) acquired about 20 web-based artworks between 2016 and 2018. While the museums slowly start to acquire Internet art, the preservation, change management and hosting of these artworks is often not solved. This is the reason why an internet-based artwork was chosen as a case study for this research. Its embedding in the Internet is a specific feature relevant for its preservation that does not exist for software-based art. It makes the work vulnerable towards changes of the internet environment. While it is logical to assume that software-based art subsumes internet-based art, the term will be used more narrowly in this article: it will exclude internet-based art, in order to be able to differentiate between them.

This article shows different ways for dealing with

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

the long-term preservation challenges of an internetbased work. They are demonstrated on a case study, *TraceNoizer* (2001) by LAN, acquired by the House of Electronic Arts Basel (HeK) in 2017. After defining its significant properties, different digital preservation strategies are applied, their sustainability and their impact on the authenticity of the artwork examined and compared. One of the artists carried out a preservation strategy already in 2004, which offers an interesting opportunity to study its long-term effects. Finally, the differences between the preservation strategies for software-based and internet-based artworks will be discussed.

II. Related works and definitions

Dekker in her dissertation [1] describes the performative nature of such works. In her view, because of the great variability in their appearance and behaviour, it is not possible to conserve an actual net artwork. While we also find that it is not possible to reconstruct the actual, original appearance and functionality of *TraceNoizer*, we do propose strategies for preserving the significant properties of this work for future iterations.

Phillips et al. report the restoration of Shu Lea Cheang's Early web-based artwork, Brandon (1998-1999), at the Guggenheim Museum in [2] and [3]. They migrated the work (about 6s'000 lines of code) to a current Web server and annotated the changes they had to carry out in order to make it compatible with today's Web browsers. This huge effort was done with the help of computer science students. Future migrations will necessitate repeated efforts and introduce changes with each migration. Not many institutions have the means to do repeated migrations every few years. Hence, this paper will compare different preservation strategies in order to find more sustainable solutions.

Miksa, Mayer, and Rauber are proposing strategies for businesses whose processes depend on Web services [4]. They are suggesting to create "mock-ups" for these Web services. These mock-ups do not actually process requests but instead pull the response from a database of the mock-up. For this purpose they recorded the request and response streams of this Web service. Espenschied and Rechert suggested this "mock-up" strategy for "Apparently Infinite Objects", in particular internet-based artworks [s]. Implementation, its feasibility and eZciency, remain open. Besides, Espenschied and Rechert proposed a stub interface and the mirroring of Web services to deal with external dependencies, that will be discussed in section VII and VIII of this paper.

Web archiving can be applied to preserve certain or parts of internet-based artworks. In 2014, Mat Kelly highlighted in a presentation [6], that a Web crawler usually changes the capture context of the Web browser. As a consequence he formulates high level requirements for the creation of Web archive files (WARC). In particular, he asked for a crawling software to capture the embedded scripts of a Web page and to allow the user to execute these WARC files in an Web browser. Rhizome subsequently developed such a system with the webrecorder and webrecorder-player [7]. These approaches, however, are only capable of capturing the "surface", for a truly internet-based artwork and thus have to be extended with solutions that also preserve its logic.

For this research, we are building upon Roeck, Rechert, and Noordegraaf [8] using the same method (determine significant properties, compare different strategies and evaluate based on criteria for long-term preservation). However, the analysis was not geared towards internet-based artworks which is why they deserve further attention.

III. Significant properties and dependencies of TraceNoizer

A. *Starting from a digital ruin*

TraceNoizer is an interactive website created by the art collective LAN (Local Area Network: Annina Rüst, Fabian Thommen, Roman Abt, Silvan Zurbruegg and Marc Lee) in 2001. Anybody who had access to a computer with a screen, keyboard, mouse and Internet connection could experience it. The House of Electronic Arts Basel acquired the work in 2017 as a pile of code. *TraceNoizer* had not been online for many years. Even though the source code was available, it was not fully functional

As it is common for institutions to acquire works a few years after their creation, many websites such as *TraceNoizer* have to be reconstructed without having a functional reference. The project archive

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

delivered by the artist is in itself a blurry object, insofar, as its past significant properties are not precisely known and the code that is relevant for the work not clearly delimited. The plethora of versions and customized presentations, adaptations to free Web services and APIs, improvements of functionality and Web design, multiple backups, handwritten maintenance programs and the simultaneous lack of documentation turns the archive into a maze for a conservator or curator. Even though the artwork itself might not have been blurry, its data, which is the base of reconstruction, might be. Hence, the process of the definition of significant properties is split up in two steps:

1. The work will be described as a reconstruction of the past (what we think it worked and looked like in 2004. The 2004 version is the most recent version. It is based on the same code as the Conservix CD, s. figure 1)
2. Based on that, the significant properties of the artwork restoration will be determined (transfer to the present and future).

As this process and its results are subjective, the decisions will be supported by reasons so that others can understand the motivation to determine certain properties as significant. Hence, "significant for restoration" means "considered significant for restoration".

B. Idea of the work

In the beginning of the 2000s, the aspect of data autonomy was widely discussed among critical Internet users. The fact that it was often impossible to delete one's own traces in the Internet motivated LAN to create *TraceNoizer*. The work was inspired by the "Jam Echelon Day" on the 21 of October 1999, when the international hacker scene decided to flood surveillance agencies such as CIA with fake information. LAN applied a similar counter strategy with *TraceNoizer* in order to diffuse the traces of one's own personal data on the Internet. On the *TraceNoizer* website the user could enter his/her name. Subsequently, the *clone engine* searched for all the websites containing this name and reassembled a new personal website out of these search results by using an algorithm that follows a similar logic as the ranking of the search machines. Thus, theoretically, in a following search, the search engine should rank the generated website higher than the original ones. The generated websites (so called clones) were uploaded to free website hosting platforms that indexed these new clones so that there was a chance for the search machine to find them. The more this process was repeated, the higher up climbed the ranking of the clones until the "original" pages did not appear in the search results anymore.

Figure 1: *TraceNoizer* by LAN. Screenshot Conservix CD (Knoppix CD)

The artists highlighted in an interview, that *TraceNoizer* was a performative tool and not a static website. According to the same interview, another important aspect of the work was the automatic generation of websites without human interaction except for entering their name.

The artists called the generated websites clones, which is why this article continues to use the term clone, although they are not clones in a literal sense, but assemblies of text and images from other websites. The programs that produce the clone are subsumed under the term *clone engine*.

Significant for the restoration The cloning process was supposed to be repeated until the clones themselves appeared in the search results and were used to produce the new clones. However, there were some discussions among the artists, whether the high search engine ranking of the generated clones really worked. This doubt is confirmed by jury members of the readme festival (2002) where *TraceNoizer* was exhibited. According to their experiences their clones

would not appear in the search results. Hence, the original idea was not perfectly executed in the 2004 manifestation of the work.

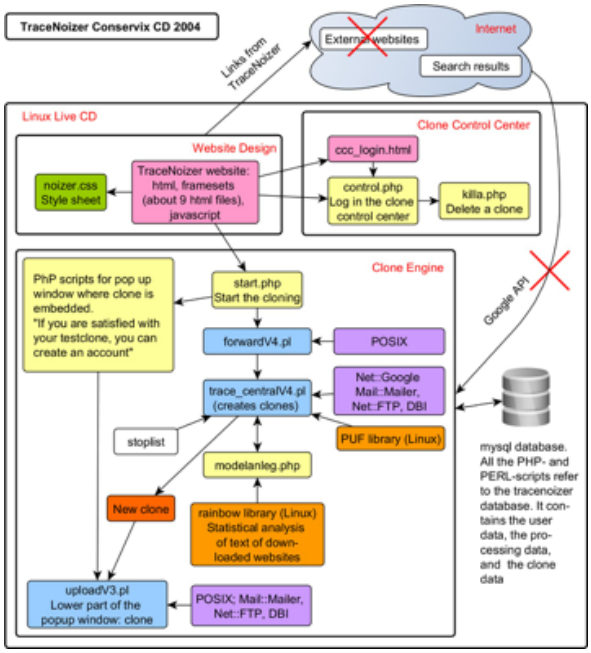


Figure 2: *TraceNoizer* in the Linux Live CD version (Knoppix CD)

C. *Processes implemented in software version 2oo4*

The *TraceNoizer* logic consists of three main parts: the graphical design of the website, the *clone engine* where the clones are generated and uploaded to the Internet and the *clone control center* where the user could access his/her clones (s. figure 2). The graphical design will be discussed in the sections D "Look and Feel" and F "External Dependencies". The trace_centralV4.pl script represents the most important script of the *clone engine* as it creates the clones. Another central element is the database that serves as a temporary data storage for the clone generation and a permanent storage for the user logins and the location of the uploaded clones. After the user entered the search term (the user's name, but other names can be entered as well), the *clone engine* started to search the Internet via a search engine interface. Up to 40 search results (websites) were saved in the database. Their text was extracted, temporarily saved in text files, then the text was split up in sentences that are saved in the database. The text was statistically analyzed by using the rainbow library[1] The ten most frequently used words were saved as keywords in the database. The texts and sentences were ranked according to the frequency of the keywords. This was the basis for the generation of the clones that mimic personal websites of the previously entered name. Each clone had a main page with ten sub-pages, each of these sub-pages corresponding to a keyword. Each keyword page consisted of an image, a text and an external link. The clone was then uploaded to a free host and the temporary database entries and text files were deleted. If the user was not satisfied with the created clone, he/she could log in the *clone control center* and delete the clone from the host. The user login could be different from the name entered in the search. However, each user could only manage one clone.

Significant for the restoration The scripts and programming languages are the material of the website. The way the scripts are programmed is typical for that time period and hence considered as significant for preservation.

D. *Look and Feel: Graphical Website Design*

The look and feel of the *TraceNoizer* website can be investigated by running the Conservix CD with Mozilla Firefox 1.3 released in 2002. There are no requirements for the Web browser such as specific plugins. The monitor resolution common at this time was 1024 x 768 pixels.

The graphical design consists of a puzzle of website pieces (frames) that are hold together by the index.htmlpage and are re-used for the different sub-pages. This feature is not supported in HTML s anymore. The background of the links change color, if the mouse hovers over it. This is achieved by JavaScript. While the website design is very clear, logical and functional, from a present-day perspective the design looks a bit outdated.

Not only the *TraceNoizer* website needs to be evaluated, but also the generated clones. The clones are built much simpler (s. figure 3). They just consist of a title, a centered image and sub-pages. The sub-pages also have a title, a centered text and an external link on the bottom. Except for the images they do not contain any graphical elements. Even for 2001

[1] 1 Information about therainbow library: https://www.cs.cmu.edu/ mccallum/bow/

standards the aesthetics of the clones are crude.

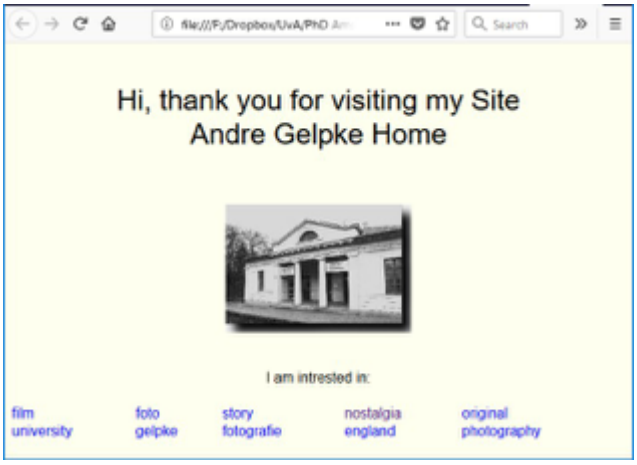


Figure 3: clone from the clone-it project 2001 in Firefox 60 (2018)

Significant for the restoration The look and feel of the *TraceNoizer* website is considered as a significant property as it points to the time of its creation. As a second priority, the framesets should be preserved, as a typical technology for that specific period.

As the clones had already looked quite crude in 2004 (They just consisted of text, images and links), there is no reason to change the clone design today. Hence, the clones generated with the *clone engine* today should look the same as they looked back then.

E. Environment

TraceNoizer contains external links that reach into the World Wide Web. Other artists such as Darko Fritz or Knowbotic used the *clone engine* to create their own artworks. The links to both artworks are broken. LAN themselves used the *clone engine* for two events in 2001 and produced a plethora of clones. These clones are not online anymore, but the artists have stored them on the Conservix CD.

Significant for the restoration The above mentioned Web pages of related artwork projects are considered significant, as *TraceNoizer* links to them directly and the artists used the *clone engine* to create these projects. Hence, these Web pages should be preserved as context of *TraceNoizer*. It could even be argued, that the indirect environment such as the artist's websites, Web pages from *TraceNoizer* exhibitions and the Echelon-Web page should be preserved as context, which is a second priority. This context demonstrates the performativity and project character of *TraceNoizer*.

F. External Dependencies

Dependencies that reach outside the artwork's Web server to Web servers that the artist and the collecting institution have no control over are typical for internetbased artworks. For *TraceNoizer*, this applies to the use of the Google search engine. The PERL interface to the Google search engine used for this work has been obsolete since many years and does not work anymore. The upload of the clones to free hosting platforms is another fragile dependency, as the free hosting platforms and their requirements to host pages for free change often. The artists experienced that too and asked friends for free server space to store the clones. For this reason the FTP clone upload was deactivated in the 2004 version of the code. Finally *TraceNoizer* depends on the HTML code of websites of the period around 2004. The *clone engine* extracts text, images and links from websites in order to reassemble them later. If this extraction does not work well, the newly assembled clones will be faulty. However, there must have been a certain percentage of faulty clones already in 2004 due to the automatic clone generation.

Significant for the restoration The different versions of the *clone engine* show that the artists switched from Google search engine in 2001 to Yahoo in 2002 and back to Google in 2004. The Google search engine itself changed considerably since 2004[1]. How the changes affect the artwork is not clear, but the artists did not seem concerned about it. In order to be consistent with the creation period a search engine from 2004 would be preferable to a current one. Knowing, that this is not possible, any search engine based on the page rank algorithm[2] is acceptable.

Due to the fact, that LAN stopped uploading the clones to free hosting platforms, the accessibility of the clones in the Internet, but not the use of free hosting services is declared as significant.

[1] "With some 1600 improvements to Google Search in 2016 alone, these are just a sample of some of the ways we have been making Search better and better over time."

[2] https://en.wikipedia.org/wiki/PageRank accessed 2019/11/02

The legibility of the clones is significant for the understanding of the work. Hence, most produced clones should be legible. In other words, they should contain legible text (not html code) and images.

G. *Conclusions*

Due to nature of internet-based artworks, dependencies on external services and external websites that are part of the ever evolving network are a real challenge. These external dependencies can be split up in technical dependencies and content dependencies. Espenschied's technical definition of a blurry object in [s] refers to technical external dependencies. Such external technical dependencies (such as the Google search engine in *TraceNoizer*) cause variations in their execution, as these dependencies change. In software-based artworks, similar variations are caused by replacement of hardware such as computer and peripherals.

The blurriness of the object extends beyond its technical external dependencies, in that internet-based artworks also have content-related dependencies with external objects/sources. As it is usually not possible to confine a context of a net-based artwork, its delineation is often subjective or dependent on circumstances. Strictly speaking one could argue that the clone-it project and the links to external websites or related art projects is not part of the work, but in order to fully understand the purpose and effect of *TraceNoizer* they become part of the work. This kind of interlocking with the environment is typical for internet-based artworks.

The performativity of *TraceNoizer* consists of the clone generation (*clone engine*) and the clone management (*clone control center*) by the user. The fact that the input is generated by search results from the Internet poses challenges for preservation, as the properties of the Internet are gradually changing.

Finally, the user management and databases are a typical property of networks. New user specific data has to be saved and becomes part of the work. Although user specific logins might also be part of certain off-line artworks, it is much less common and it is limited to the visitors of the physical artwork.

IV. Long-term preservation criteria

In order to find and evaluate preservation strategies for the above mentioned properties typical for internetbased artworks, the following criteria for sustainable preservation strategies are suggested. Criteria for assessing the long-term sustainability of preservation measures were established for software-based art in [8] and are hereby assessed for web-based artworks.

Adaptability to new hardware is relevant but not crucial for the back-end. Web servers are generic computer hardware without specific components such as super fast video cards or specific input or output devices. Almost the same can be said about client computers. In order to view the artworks, they do not use any specific hardware features except for certain generic input devices.

The *ability to deal with software obsolescence and changed network protocols* is very relevant, as the software of an internet-based work is usually updated when transferred to a newer Web server. In addition, the work needs to adapt to new network protocols that are updated periodically (for instance HTML every few years). The change of application programming interfaces (API) is another frequent cause for malfunction internet-based artworks and an example for software obsolescence.

The *stabilization of software complexity and the minimizing of the software change rate* is very relevant for webbased artworks, as they are subject to fast and frequent changes due to the Internet connection. As their environment changes fast, the risk is high that such works quickly become outdated if not dysfunctional. In order to prevent that, they need to adapt, too.

The *ease of installation* of the artwork and of connecting peripherals is not relevant for web-based artworks.

The *reduction of maintenance* is very important for internet-based artworks, as maintenance of internetbased artworks, especially of *server-side dynamic websites*, can be laborious. Tasks such as the detection and cleaning of abusive and resource-intensive processes, updating the server software,

and maintaining the database and server health can take up to one hour per week, which, in a museum context, is quit intense.

The *scalability of a preservation strategy* is relevant for web-based artworks, depending on the number of artworks an institution hosts. The more artworks, the more methodical the hosting and preservation approach needs to be in order to be able to exploit synergies and to reduce the maintenance per artwork.

V. Linux live cd (conservix cd)

In 2003, Fabian Thommen produced a CD-ROM with a bootable live operating system based on the knoppixtechnology that he named Conservix. Conservix is set up with a basic Linux operating system, an Apache Web server, a MySQL database, the programming languages PHP and PERL and the Web browser Mozilla Firefox 1.3[1]. With this live system, the user did not need to install anything or change computer configurations. The computer starts from the CD in place of its own boot system. It automatically opens *TraceNoizer* that is installed as a dynamic website (s. figure 1).

The Conservix CD fulfills the previously defined **significant properties** partly. It was possible to temporarily create clones with it, although they could not be uploaded to external Web hosting services, and the database entries could not be permanently stored as the CD is read only. For this reason, the user could not manage the clones in the *clone control center*, partly diminishing the performative nature of the work. The client computer and the Web server coincide in one machine. An important disadvantage of the CD is, that it cannot be directly accessed through an URL, but needs to be installed in an emulator in order to run. Due to the obsolescence of the search engine interface, the CD-ROM does not allow to generate clones today, while the website as a graphical interface is displayed without errors. The CD-ROM also comprised the clones of the clone-it project so that it is known today, what the output of the *clone engine* looked like. Unfortunately, the input stream (search results) of the *clone engine* was not recorded, so that it is not possible to verify today, whether a restored version would produce the same output. Other "damages" are external links on the *TraceNoizer* website that are broken in the meantime.

Regarding the **sustainability** of the Conservix CD from 2003 it can be stated, that its iso-image still exists in 2019 and that an Intel or AMD-compatible processor can still run it. The Conservix CD is *able to deal with software obsolescence and changed network protocols* such as HTML as it contains both server and client. When it comes to external dependencies such as the Google API, Conservix does not cope so well. At least, it would be convenient if the CD produced an error message saying that the Google API does not work anymore. The Conservix CD *stabilizes software complexity* as it is read only. The *Maintenance* is also low, but would include the periodical updating of external dependencies such as Google API and thus producing a new CD-ROM. The security risks are zero for the host computer, as the CD is read only and the computer does not use its own operating system. The preservation strategy is *scalable* in so far, as Linux Live CDs can be produced for other internet-based artworks.

The Conservix CD played an important role in the definition of the significant properties: It documents, how the work is installed, what libraries and program versions, and what browser were used. It recreates the look and feel of the work without having to install much, except for a generic emulator. On the other hand, a Linux Live CD is not the best solution for internet-based works with external dependencies such as a Google search API. It also gives a false sense of security, as Fabian Thommen mentioned that he did not install the work exactly the same way as on the Web server. For instance, administration programs to maintain the database are not necessary on a CD ROM.

[1] Operating system Debian 1.3-4, Web server Apache 1.3.27, database MySQL version 11.18, distribution 3.23, programming languages PHP (4.1) and PERL s.8, and Web browser Mozilla Firefox 1.3

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

VI. Preservation version "migration"

This section describes the preservation measures undertaken by Fabian Thommen in 2018. He migrated the work to a current Web server with an old PHP version s.s.9 and made the following changes:

- He replaced the Google library from 2004 with a Google Custom Search API from 2018.
- The database commands in PHP had to be replaced in order to be compatible with newer PHP versions.
- Configurations like the database connection and the Google API keys were moved to a configuration file in order to reduce maintenance and increase security.
- Security was enhanced in parts. One big security risk is the passing of variables such as user data from the client to the server. There are different methods, how a browser client can send information to the Web server. In *TraceNoizer* the user variables were passed as register_globals to the PHP script. This method is insecure, as the input can be easily manipulated. Thus, from PHP
- s.4.0. on only the GET and POST methods are possible. Fabian Thommen adapted the scripts accordingly.
- *TraceNoizer* used the rainbow library in order to analyze the text of the websites generated by the user's search. The latest rainbow version dates from 2002 and its binary was compiled for a 32bit operating system. To run the 32 bit binary on a 64 bit operating system, the library lib32z1 had to be installed.

The **significant properties** of the work were partly preserved: With this migrated version, clones can be generated and the user can login and delete his/her clone. The scripts and programming languages are only so much changed that they function. The look and feel of the website stays the same, even if the website is viewed on a current browser. This will not be the case in the future, as frames are not supported in HTMLs. However, this can be solved with a browser emulation. Almost all the significant properties are respected with one exception: the clones are faulty. They are so faulty, that they do not fulfill the purpose of pretending to be somebody's homepage (selection of keywords, missing images, s. figure 3). Most of them do not contain images and the sub-pages made of keywords are also missing. This can be caused by malfunctioning of the rainbow library, or the fact that the structure of Web pages has changed so drastically since 2004, that the extraction of sentences and images does not work properly. It is also possible that the *clone engine* has never functioned as intended and has always produced a certain amount of faulty clones.

Regarding **sustainability**, the migration strategy does not yield the best results. It is able to deal *with software obsolescence and changed network protocols* such as HTML or the Google API, but at the expense of changes in the code. In addition, these changes have to be repeated every few years to keep up with changes in Web technology. Hence migration does not stabilize software complexity but rather enhances it. *Maintenance* will be high in order to alleviate security risks but also to clean up the database periodically. The *scalability of the preservation strategy* is relevant for the House of Electronic Arts, as they host other internet-based artworks. However, as each works needs an individual Web server software environment, the migration strategy for this work does not scale.

It can be summarized, that the migration strategy met the significant properties of the *TraceNoizer* website, but not the ones of the clones. It's biggest shortcoming from the perspective of long-term preservation is the fact that it needs to be repeated regularly. Serious Internet security concerns and an expected high amount of maintenance add to the disadvantages of this strategy.

VII. Preservation version "emulation"

As an alternative, another preservation strategy was tried by emulating the Conservix CD of *TraceNoizer* using the University of Freiburg's Emulation-as-a-Service (EaaS) framework[1]. EaaS provides users with convenient access to emulators via their Web browsers. A curator can ingest a digital object, configure the right emulator and its settings, and allow any user to start the configured emulation environment. By default, each user is presented with a fresh emulation environment as configured

[1] [5]https://gitlab.com/emulation-as-a-service

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

by the curator and several users can use different sessions of the same emulation environments at the same time. Alternatively, however, emulation environments can also be "snapshotted" by users or curators, conserving their current state including any manipulations by the user.

Previously, EaaS concentrated on emulating single (unconnected) environments, e.g., a single preserved work of digital art as an archived CD-ROM image, which needs to be run on a Windows-9s environment with an installed Macromedia Flash player. As is outlined in this work, preserving single works as stand-alone entities is not suZcient for many works as their significant properties are realized through and depend on the combination of several systems. Thus, it is necessary to regard the whole ecosystem as one (connected) preservation environment.

To facilitate the emulation of a connected environment, a recent addition to EaaS allows to create a virtual network, which operates on the Ethernet layer [10]. The virtual network is represented by a URL for identification and for access control. Via this URL, the virtual network can be reached from the Internet using the WebSocket protocol over HTTPS. HTTPS/TLS encrypts the traZc and, thus, shields it from malicious access from the Internet, while the usage of the WebSocket protocol (as opposed to direct TCP/IP) shields the Internet from the emulated environments (which malicious users of the EaaS system might otherwise abuse to perform, e.g., DDoS or spam attacks on the public Internet). At the same time, it introduces a layer of emulated Ethernet traZc on top of the EaaS Web API and, thus, shields the EaaS host system from Ethernet traZc between the emulated environments. Inside the WebSocket connection, Ethernet frames from the connected emulation environments are prefixed with a simple two-octet big-endian length header (the same format as used by the VDE 2 library[1]).

Firstly, the described concept allows to connect multiple emulation environments. The emulation environments can either be specific to the preserved digital object (e.g., in a multi-machine system consisting of an application server and a database server) or generic emulation environments can be combined ad-hoc. This approach allows to easily reuse emulation environments by a curator or a user without any necessary special knowledge.

In the case of *TraceNoizer*, its Conservix-CD version was emulated by the EaaS framework using the QEMU emulator (s. figure 4, Option 1). In a second connected emulation environment, a contemporary Web browser was started. As a further step, other emulated environments containing Web browsers could be built to allow users to examine *TraceNoizer* (and any other digital objects). As *TraceNoizer* was originally built in a time in which optimizing Web sites for specific Web browsers (and build upon their non-standardized features) was prevalent, this could be essential to fully reproduce the original performance of the artwork as perceived by different users at that time.

Secondly, the approach of virtual networks allows to offer additional services in the network. E.g., it is currently already possible in EaaS to connect the virtual network to the current live Internet, and thus, allow a connected user to access current Web sites. To fully recreate the original *TraceNoizer* performance, instead of allowing access to the live Web, an archived version as, e.g., provided by the Internet Archive could be used (s. figure 4, Option 2). This would effectively operate as a transparent proxy7, operating at either the DNS and/or HTTP layer. The proxy[2] could either be preset by a curator or configured by the user to serve the Web as archived on a specific date. This would allow, e.g., to retroactively analyze the behavior of *TraceNoizer* at different points in time.

The virtual network cannot only connect emulated environments but also allows (via *eaas-proxy*) applications from outside to access the services provided in the virtual network. It, therefore, allows to map and forward an external TCP port to an internal IP address and TCP port. The termination of this connection can either occur in the public Internet (at an individually assigned network port for each user session), at the user's computer on localhost (by downloading and running a eaas-proxy binary), or, in the future, directly in the user's browser (using

[1] https://github.com/virtualsquare/vde-2

[2] Also known as interception proxy, see RFC 3040, section 2.s, https://www.rfc-editor.org/info/rfc3040

ServiceWorkers [11]). It allows the user to view *TraceNoizer* in their current Web browser, "breaking out" of the emulated environment and makes operation like interacting with the website content, copying contents, or even deep-linking to contents much easier.

The user is, though, still protected from any malicious emulation environments by their Web browser, and the operator's system is protected by the virtual network, which is directly connected to the emulated environment but separated from the actual host system.

A further problem, which can be solved by the presented approach, is the usage of ancient Google search Web APIs (utilizing SOAP) for the *TraceNoizer* system. Google has stopped supporting this API, but it can be emulated in a virtual network environment and, consequently, allows the *TraceNoizer* environment to remain unchanged. The same approach is applicable for *TraceNoizer*'s storing clones on Web hosts which have long ceased operation.

VIII. Conclusions for the preservation strategies for internet-based artworks

For internet-based artworks, that originate in the fast-changing technical and social environment of the internet, it is necessary to abstract from the concrete technical setup and formulate more high level principles. The following three principles can be used as guidelines for maintaining the functionality of internet-based artworks in a new networked environment:

a) Simulation: Part of the old environment is recreated in order to adapt to the old artwork and allow better interaction without having to change the original artwork too much.
b) Bridging: A bridge can be built between the old artwork and the new environment. This can be achieved by encapsulating the artwork and then providing an interface to translate the input/output between the old and the new environment. It could also be achieved by adapting the code of the artwork directly to communicate with the new environment, which would correspond to a migration.
c) Reinterpretation / Reconstruction: The artwork, or parts of it, could be recreated by different means, such as new platforms or new technology, in order to adapt to the new environment. For instance *TraceNoizer* could be recreated on the Facebook platform, by cloning Facebook accounts instead of websites.

Web archiving is an example for simulation. Broken external links of *TraceNoizer* could be downloaded from the Internet Archive and saved in a protected environment. This protected environment is stable and in control of the conservator.

The emulation of the Google Search Web API as described in the previous chapter is an example of the bridging strategy, as it allows to use the Google Search Engine. The exchange of the Google Search API with the current API as done in the migration of *TraceNoizer* is also seen as bridging from the "old" work to the new environment.

The above mentioned principles are carried out by applying a combination of the well-known digital preservation strategies reprogramming, migration, emulation/ virtualization and web archiving. As the examples showed, these strategies and principles can be applied to different elements of the website such as web browser, web server, or to its external dependencies such as parts of the world wide web environment and external web services.

External web services are used more and more in internet-based artworks. External web-services can be preserved or handled as described by Dragan Espenschied and Klaus Rechert in "Fencing Apparently Infinite Objects". They are suggesting the mirroring of the webservice, a stub interface that has a reduced functionality instead of the web service, and the recording of network traZc of a web service for a limited number of queries. All three proposals correspond to the principle of simulation mentioned above. If simulation is used as a strategy to preserve web services instead of bridging with an API, it is likely that a compromise with the functionality or authenticity of the work has to be accepted.

In order to link different emulated elements of the internet-based artwork, the University of Freiburg enhanced EaaS by enabling to connect these emulated environments within a virtual network. This allows a great flexibility in finding tailor-made

preservation solutions. In addition, the user can login in this network and save his/her data by taking emulator snapshots. This has the advantage that the Web server including database can be reset for maintenance purposes without losing user data.

Software-based art relies on the same digital preservation strategies as internet-based art with the exception of web-archiving. However, for software-based art, hardware can play a much more important role than for internet-based art. Hardware can be compared to an external dependency, as specific peripherals cannot be easily rebuilt and their production is dependent on the production company. Hence, the bridging principle is important when replacing old equipment with new equipment. Emulation of input/output devices is an example for this. The simulation principle would be applied when running a new piece of equipment within a case of an old device.

The networked environment in an internet-based artwork does not exist for software-based artworks. The latter rather has a physical environment that is often variable and might be determining for the installation of the software-based artwork, but its change does usually not cause malfunction.

Returning to the discussion of sustainability of preservation measures it can be said, the more external dependencies can be eliminated, the more sustainable the preservation strategy is. From that point of view, the simulation strategy is the best strategy, as it replaces the evolving Web environment with a stabilized archive. Besides the reduction of dependencies, the reduction of maintenance and security aspects are dominating in the assessment of sustainability of preservation strategies for internet-based artworks. As can be seen with *TraceNoizer*, the migrated website is unsafe and prone to web attacks. In contrast, it can be very economical to restart an emulation regularly instead of maintaining the server that can quickly take an hour or more weekly. The isolation of several internet-based artworks with different needs and dependencies from each other can be another argument for emulation / virtualization.

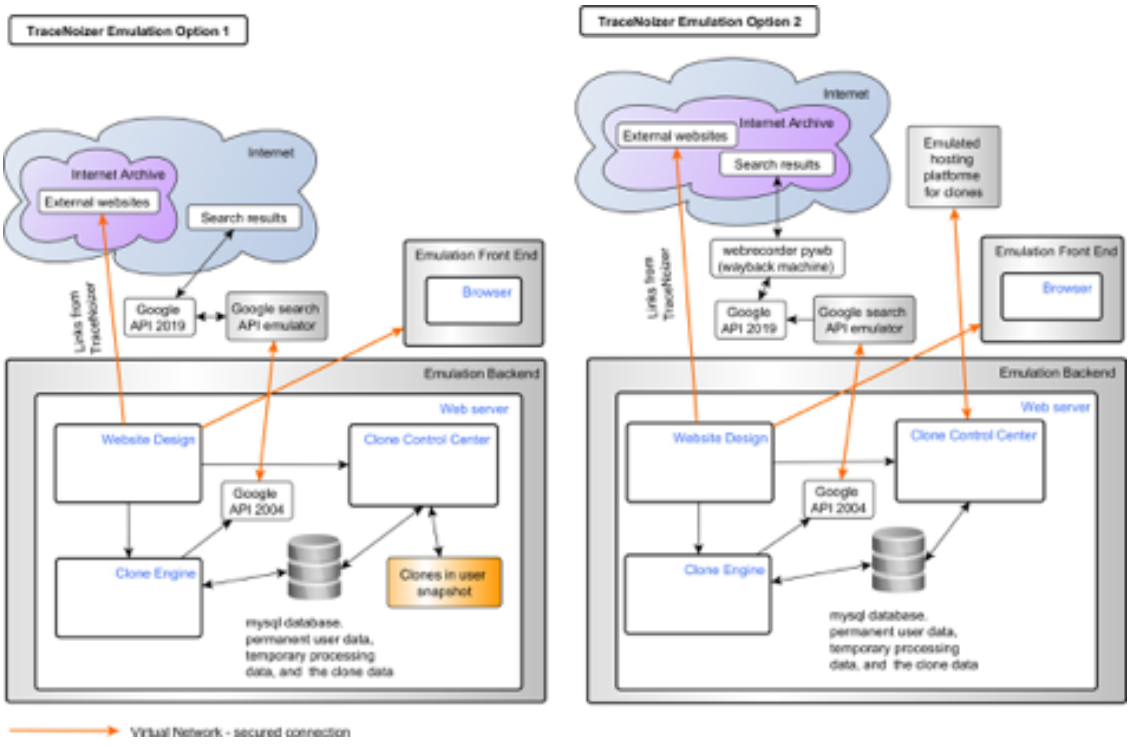


Figure 4: Emulation strategy options for *TraceNoizer*, showing different degrees of emulation. In Option 1, the emulated *TraceNoizer* interacts with the current live Web and generated clones are not exposed to the live Internet. In Option 2, Internet Archive's Wayback Machine is used to let *TraceNoizer* interact with an archived version of the Web and generated clones are exposted to the live Internet.

In the future, a good compromise between reduction of maintenance, security aspects and functionality will be crucial for the preservation of internet-based artworks.

Acknowledgment

This research was funded as part of NACCA (New Approaches in the Conservation of Contemporary Art, www.nacca.eu), a Marie Skłodowska-Curie Innovative Training Network of the European Union. The House of Electronic Arts provided the case study *TraceNoizer* (2001), LAN (the artists) provided information and an interview about the work and Fabian Thommen, member of LAN migrated *TraceNoizer*. Rafael Gieschke and Klaus Rechert, University of Freiburg carried out the emulation of *TraceNoizer*.

References

[1] A. Dekker, "Enabling the future, or how to survive forever: A study of networks, processes and ambiguity in net art and the need for an expanded practice of conservation," Doctoral Thesis, Goldsmiths, University of London, London, 2014.

[2] J. Phillips, D. Engels, E. Dickson, and J. Farbowitz, Restoring brandon. shu lea cheang's early web artwork, www.guggenheim.org, 2017. [Online]. Available: https://www.guggenheim.org/blogs/checklist/restoring brandon shu lea cheangs early web artwork (visited on 07/01/2017).

[3] J. Phillips, D. Engel, J. Farbowitz, and K. T. Rosenberg, The guggenheim restores john f. simon jr.'s early web artwork "unfolding object", 2018. [Online]. Available: https://www.guggenheim.org/blogs/checklist/ the guggenheim restores john fsimon jr early web artwork unfolding object (visited on 11/26/2018).

[4] T. Miksa, R. Mayer, and A. Rauber, "Ensuring sustainability of web services dependent processes," International Journal of Computational Science and Engineering, vol. 10, no. 1/2, p. 70, 201s, ISSN: 1742718s. DOI: 10.1504/IJCSE.2015.067058.

[5] D. Espenschied and K. Rechert, "Fencing apparently infinite objects: Defining productive object boundaries for performative digital objects," in iPres 2o18, iPres 2018, Ed., 2018.

[6] M. Kelly, Browser-based digital preservation, 2014/07/07. [Online]. Available: https://www.slideshare.net/matkelly01/browserbased digital-preservation.

[7] M. McKeehan and I. Kreymer, Symmetrical web archiving with webrecorder, a browser-based tool for digital social memory. an interview with ilya kreymer, The National Digital Stewardship Residency New York, 2016/02/23.

[8] C. Roeck, K. Rechert, and J. Noordegraaf, "Evaluation of preservation strategies for an interactive, software-based artwork with complex behavior using the case study horizons (2008) by geert mul.," in iPres 2o18, iPres 2018, Ed., 2018. [Online]. Available: https://osf.io/y3gcu/.

[9] How google search works: Useful responses take many forms, 2019. [Online]. Available: https://www.google.com search/howsearchworks/responses/#?modal_active=none.

[10] R. Davoli, "Vde: Virtual distributed ethernet," in First International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities, IEEE, 200s, pp. 213–220.

[11] M. Kruisselbrink, J. Song, J. Archibald, and A. Russell, "Service workers 1," W3C, W3C Working Draft, Nov.2017httpswwww3org/R/2017/D-serviceworkers-1-20171102/.

Towards a Universal Virtual Interactor (UVI) for Digital Objects

Euan Cochrane
Yale University Library
United States of America
euan.cochrane@yale.edu
0000-0001-9772-9743

Seth Anderson
Yale University Library
United States of America
seth.r.anderson@yale.edu
0000-0003-3304-8484

Ethan Gates
Yale University Library
United States of America
ethan.gates@yale.edu
0000-0002-9473-1394

Klaus Rechert
OpenSLX GmbH
Germany
klaus@openslx.com
0000-0002-8807-018X

Jessica Meyerson
Educopia Institute
United States of America
jessica@educopia.org
0000-0002-0180-9616

Abstract – Practitioners generally agree that providing a service to enable opening and interacting with born digital objects in their "original" software is valuable for historians, researchers and the general public so that they can experience full-fidelity experiences of the objects. Enabling this has, to date, been a difficult, time-consuming, relatively resource intensive, and tedious. In this paper, we show how we are on the verge of creating a new method and series of tools to simplify and automate the process of interacting with digital objects in their original software and greatly reduce the time and resource costs of doing so. We outline the history of the developments in the areas of emulation and software preservation that we have built on and we outline the concept of this set of tools and processes we call the "Universal Virtual Interactor". We also discuss how the UVI is being created, and finally we discuss how it may be improved upon in the future and how it may be implemented in access and discovery tools.

Keywords – Emulation, Access, Rendering, Interactivity

Conference Topics – 2. Designing and Delivering Sustainable Digital Preservation; 5. The Cutting Edge: Technical Infrastructure and Implementation.

I. Introduction and Background

From at least the 1980s, many years prior to the publication of Jeff Rothenberg's seminal article and paper "Ensuring the Longevity of Digital Documents" [1] in 1995 there have been advocates amongst digital preservation practitioners for preserving software as both information in itself but also, and importantly, as a utility for accessing other/existing digital objects over the long term [2]. As Rothenberg said:

"they [future generations] should be able to generate an emulator to run the original software that will display my document."[1 p47]

As the National Library of the Netherlands so eloquently articulated in 2003:

"There is a difference between paper and digital records. Any paper record can be perceived through the five human senses; no digital record can be perceived without going through computer hardware and software..... .

....Digital records are software dependant. They rely upon the software that was originally intended to interpret (or display) them. When that software becomes obsolete, perhaps within the space of a few years, the problem arises of how to read that record without its original software application. It is unlikely that different versions of the application will read the file in the same way, and this may well result in a change in the interpreted record (the visible or available view of the file) that affects its archival integrity. Some data may be lost altogether; in other areas, data may be gained. There may be no way to compare a new version with the original, so changes may go unnoticed. Any changes to the

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

record may affect its authenticity and integrity, which in turn may affect its archival and legal status. Depending on the nature of the record and its use, this can cause problems, not least that of losing or misrepresenting history."[3]

This was further illustrated in the "Rendering Matters" [4] research undertaken in 2011 at Archives New Zealand which demonstrated (with visual examples [5]) the necessity of ensuring we can interact with preserved digital objects using the original or representative contemporaneous software environments.

While there continue to be many valiant efforts to preserve computing hardware for future generations, particularly for pedagogical purposes[1] this approach is unfortunately neither economically scalable nor likely to be sustainable over long time frames [6]. For this reason, practitioners over the last 20+ years have instead focused on preserving the software component(s) and ensuring we can continue to maintain the ability to run legacy software as the hardware that supports it has become obsolete.

In Jeff Rothenberg's 1995 article [1], and his subsequent work with the National Library and the National Archives of the Netherlands, Rothenberg argued that emulation was likely to be the only effective general strategy for preserving the complete full-fidelity experience of digital objects over time. Rothenberg has also argued that emulation is cost-effective as a just-in-time rather than a just-in-case approach.

"few organizations can justify the cost of translating documents that they no longer use." [7 *p13]*

Emulation also couples well with other long-term digital preservation tools and strategies, such as normalization and migration, with the former facilitating cost-effective preservation of fidelity and authenticity (and "digital patina" [8]), and the latter facilitating reuse of components of digital objects that can be easily extracted (potentially on-demand) from their native contexts. Emulation can also become a tool for performing just-in-time migration when coupled with macros that interact with emulated software environments to run "open-file-then-save-as-a-new-format" operations [9].

A fruitful way of interpreting the history of emulation tools in digital preservation is to consider it as an attempt to maximise the preservation impact of our preservation tools while minimizing long-term support costs. To this end, in 2001 Raymond Lorie, while working at IBM, developed the initial design concept for what he called a "Universal Virtual Computer":

"We propose to save a program P that can extract the data from the bit stream and return it to the caller in an understandable way, so that it may be transferred to a new system. The proposal includes a way to specify such a program, based on a Universal Virtual Computer (UVC). To be understandable, the data is returned with additional information, according to the metadata (which is also archived with the data).

...we propose to describe the methods as programs written in the machine language of a ***Universal Virtual Computer (UVC)****. The UVC is a Computer in its functionality; it is Virtual because it will never have to be built physically; it is Universal because its definition is so basic that it will endure forever. The UVC program is completely independent of the architecture of the computer on which it runs. It is simply interpreted by a UVC Interpreter. A UVC Interpreter can be written for any target machine."* [10]

The UVC had quite a few pitfalls, primarily that new UVC code had to be written for every new file format. But the general approach, that of stabilizing and preserving the functionality at the "highest" possible level in the interpretation set in order to minimise the number of code revisions required to keep the overall functionality operating, was sound.

Building on this the Koninklijke Bibliotheek, National Library of the Netherlands (the KB) developed "Dioscuri" a "modular emulator" that could run anywhere the Java Virtual Machine (JVM) could run [11].

[1] Such as the work of the Living Computers Museum + Labs (https://livingcomputers.org/), Computer History Museum (https://www.computerhistory.org/), Media Archaeology Lab at University of Colorado Boulder (https://mediaarchaeologylab.com/), Maryland Institute for Technology in the Humanities (https://mith.umd.edu/0), retroTECH at GeorgiaTech University (http://retrotech.library.gatech.edu/)

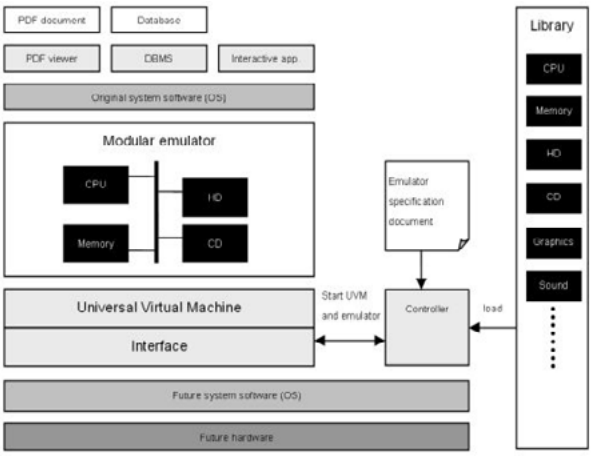


Figure 1: The Dioscuri Modular Emulator Design

Dioscuri was intended to be extensible over time with new modules being created for newer systems as needed. In the Dioscuri model the component that was to be stabilized and preserved was the Java Virtual Machine, i.e. future generations would only have to re-enable write the JVM code in the new computing system in order to maintain access to all the modular emulators and the systems and software that they support.

Following from and incorporating the Dioscuri work, the Keeping Emulation Environments Portable (KEEP) project [12] developed a desktop application that bundled a set of emulators with a configuration GUI that incorporated the concepts of Preservation Layer Models and "View Paths" from early IBM/KB work [13]:

"The PLM outlines how a file format or collection of similar objects depends on its environment. A PLM consists of one or more layers of which each layer represents a specific dependency. The most common PLM consists of three layers: application layer, operating system layer and hardware layer. However, other variations can also be created. Based on a PLM, different software and hardware combinations can be created. Each such combination is called a view path. In other words, a view path is a virtual line of action starting from the file format of a digital object and linking this information to a description of required software and hardware" [14 p148].

To implement this the KEEP developers included a method for associating file formats with configured emulated environments such that you could submit a file and it would "automatically" be attached to an environment and that environment be loaded on your desktop to be interacted with. Overall the approach of the KEEP project was less efficient than that outlined by the UVC or Dioscuri alone as it incorporated multiple emulators that would have to be supported over time. However, it did introduce the dramatic efficiency of being able to use existing file-interpreters (E.g. commercial software applications) and the ability to reuse off-the-shelf emulators developed by third parties. Concepts and approaches that are included in the contemporary work we discuss further below.

At the same time as the KEEP, the Planets Project [15] had a sub-project to build the Global Remote Access to Emulation Services (GRATE) service [14]. A method for remotely accessing emulated environments via a web browser. This approach enables resource-intensive emulation to be managed and executed remotely while the user interacts with it through a browser-based viewer.

The GRATE project was led by a team at the University of Freiburg and evolved into what became the Baden-Württemberg Functional Long-Term Archiving (bwFLA) project [16]. The bwFLA project, in turn, developed the suite of tools now commonly referred to as Emulation as a Service or "EaaS". The EaaS tools follow the basic approach pioneered with KEEP, but implemented with a browser-based interface, while adding features such as enabling the definition of derivative disk images (more on this below), the separation of objects, environments and emulators, and the addition of many reliability improvements. The browser-based approach is transformative from a user-perspective as it democratizes access. All one needs to interact with an emulated computer is a web interface.

It is upon the EaaS infrastructure that we are building the EaaSI program of work. We're expanding on the concepts of a PLM and view path to create what we are calling a Universal Virtual Interactor. The goal of the UVI project is to develop a framework into which organizations and consortia can add legacy software and metadata in order to automate presenting digital objects to users for interaction in a web browser. The objects are presented in "original" or representative interactive computing environments utilizing original or representative software

from a time period that is appropriate to the object. Or, more succinctly, the UVI automates opening old files in their "original" software in a web browser.

II. Configuring Environments in the Emulation as a Service Infrastructure (EaaSI) program of work

A. *The EaaSI Program of Work*

In the Emulation as a Service Infrastructure (EaaSI) program of work we are working with partner organizations who are hosting EaaSI nodes running instances of Emulation as a Service: the "EaaSI Network [17]". Together with a local team at Yale University Library we are configuring and documenting emulated computing environments and enabling the environments to be shared between nodes in the EaaSI network. Upon that base we are building services, workflow interfaces, and APIs to perform various digital preservation and curation functions. One of these services is the UVI that relies on this set of configured computing environments for its core functionality.

B. *Hardware Research and Configuration*

The configuration of computing environments within the EaaSI program is a fairly involved process that is time consuming and deliberately thorough. We have a team recruited from students at Yale University who are performing this important role. The workers in the environment configuration and documentation team start by selecting an application and check to see what hardware and operating system it requires. Assuming the required environment doesn't exist, they next create a virtual hard disk that is stored as an image on our servers. The workers then configure an emulated computer that has the hardware specifications required to run the dependent operating system and also to run the application itself.

We do our best to match the emulated hardware specification to representative hardware from the period during which the software was most popular, or the period we are targeting to emulate. For example, for a Windows 98 computer we can choose to emulate a contemporaneous CPU (e.g. Pentium 3), volume of RAM (e.g. 256 megabytes), and compatible sound (e.g. a SoundBlaster 16), video (e.g. a Cirrus CLGD 5446 PCI) and network cards (e.g. an AMD PCNet PCI). Sometimes this requires historical research and we have consulted various online resources from old advertisements to compiled lists of hard drive prices over time.

Figure 2. An advertisement for Cybermax Personal Computers from the late 1990s, via user @foone on twitter

Historical and performative accuracy is also weighed against long-term costs. We aim to minimize the hardware variants that we support in order to reduce the long-term cost of moving the environments to new or migrated emulators.

C. *Documentation Operating System Configuration*

The configured computer is then documented as structured metadata and defined as a configured "hardware environment". These "hardware environment" combinations can be saved as templates in our system[1], allowing future users to reuse that

[1] The hardware environments are also matched to the software applications that we later install on the hardware environments and confirm their compatibility with. These applications also have their published hardware requirements documented and associated with them. In the future we hope to use these two sets of data to automate matching newly added software applications to pre-configured "compatible" hardware environments by matching the published hardware requirements of the new software with pre-configured environments that we have confirmed are compatible with the same requirements set.

configuration when selecting their requirements (either automatically or manually using a GUI) without having to configure every sub-component (for example, they might just select the most popular pre-configured hardware environment template that supports Windows 98 SE). Next a disk (image) is connected to that emulated computer, the operating system installation media is also attached, and the computer is switched on ("booted"). The configuration user can then run through the operating system installation and configuration. Throughout this process the configuration user has to make a number of decisions about operating system configuration and settings. These decisions can affect the functions of the operating system and the applications that come bundled with it, or may be run on it, in the future. For example, setting the resolution of the desktop will affect how software displays, or choosing a language or set of locale settings can dramatically change the user experience. EaaSI configuration users select from menus and pick lists to document each of the decisions they make and add new metadata options to those pick lists where necessary. This ensures consistency and machine-readability of the captured documentation/metadata.

Having configured and documented the operating system the configuration user shuts down the emulated computer and saves the results into the disk image file. This disk image and its documented contents are defined as a new "software environment"[1]. This software environment is documented by the configuration user as structured metadata and assigned a unique identifier. Together with the hardware environment they are defined as a "computing environment" which is also documented and assigned a unique identifier. We use these concepts to organize and enable discovery of assets within the EaaSI interface.

[1] An important tangential benefit of this approach is that by preserving just one of these environments, such as a Microsoft Windows 98 computing environment running Microsoft Office 97, we have ensured that the very many digital objects created by and made accessible using the applications in the Office suite are able to be accessed for future generations. Once we have one of these environments configured we can reuse it to re-enable interaction with all of those countless digital objects at minimal incremental cost and on an on-demand basis that is useful from a financial planning perspective as it matches the burden of cost to the time of access.

D. *Installing and Documenting the Application Software*

Our next step is to install the selected application onto the existing software environment to create a new software environment. Fortunately, the EaaS software facilitates minimising the incremental cost and associated environmental impact of this by enabling the creation of "derivative" disk images that are "derived from" an existing image (either a full disk image or a derivative itself) [87]. The changes that a configuration user makes when installing and configuring the added application are all that is captured onto disk in the resultant derivative file. When the associated new software environment needs to be used in the future the full disk image (or hierarchy of image and derivatives) and the derivative file are brought together at the time of execution and integrated in real-time by the EaaS software.

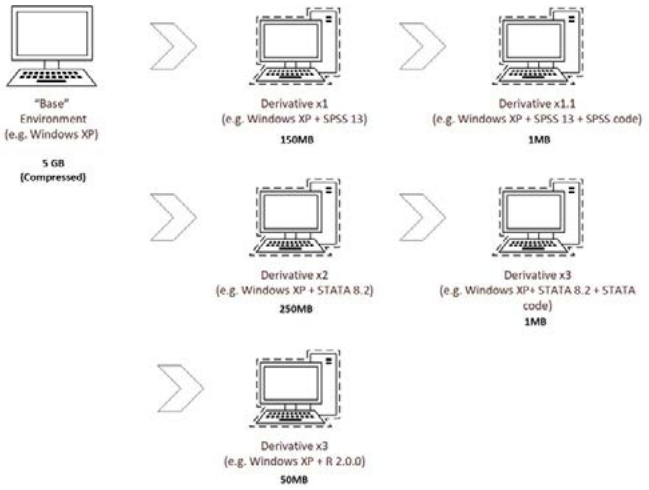


Figure 3. An illustration of the storage cost savings enabled by the use of derivative disk images

This greatly reduces the storage burden of preserving multiple software environments and frees configuration users to pre-configure software environments with only minor differences between them without raising significant concerns about storage costs. Given that it can take a number of minutes to load a computing environment and make even a small (but potentially very useful) settings change, the benefit of this becomes clear: by pre-configuring multiple environments just once each and then sharing them, this greatly reduces the time required for future users to provision a software environment appropriate to their use case, i.e. users can just pick the pre-configured software environment they want from a list.

DOI: 10.1145/nnnnnnn.nnnnnnn

While installing and configuring the application in the software environment the configuration user documents a number of facts about it in order to facilitate automated interaction with the environment in the future. For example, the configuration user will document every relevant executable program included in the application and various facts about it such as:

1. Where the executable is located within the file system
2. How the executable itself can be initiated programmatically (at system start up)
3. How the executable can be made to open a digital object programmatically during the initiation process
4. What file formats the application can open

a. This includes documenting the exact description of the format and its extension (where applicable) as displayed in the application's user interface[1]

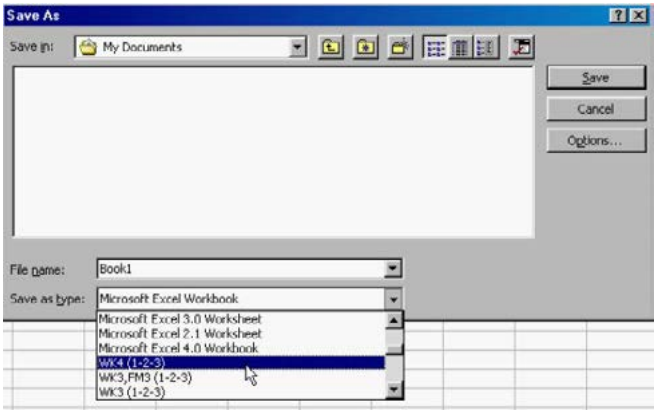


Figure 4: Save-as types as presented in Microsoft Excel 97

5. What file formats the application can import

• As above - including the specific displayed parameter.

6. What file formats the application can save to

• As above - including the specific displayed parameter.

7. What file formats the application can export

• As above - including the specific displayed parameter.

The configuration user also makes an attempt to research when the application was first released, when it was most popular, when it was first superseded (and by what software) and when it went out of general use or was no longer supported. And finally, they document the default mime-types or file extensions (depending on the operating system) that the software is configured to automatically open at the operating system level within the software environment (i.e. which types of files the application will automatically open when the file is double clicked within the operating system).

The configuration user then shuts down the emulated computer and saves a derivative environment which gets added to the pool of available software environments. Where possible the configuration user will then publish the environment so that others participating in the EaaSI network can add the environment to their local pool.

The software behind the EaaSI network is open source and available on GitLab [19], and while the current EaaSI network is using the fair-use rights available under copyright law in the United States of America to facilitate the sharing of environments outlined here [20] there is no technological reason the software and approach couldn't be extended internationally. The Software Preservation Network [21], an international cooperative of stakeholders in software preservation, is working with international partners to explore avenues for expanding the EaaSI network beyond the United States and/or to enable similar networks to be established in other legal jurisdictions.

III. The Universal Virtual Interactor (UVI)

The Universal Virtual Interactor or UVI is a concept built on the history outlined above. Its name is descriptive of its functionality: it is intended to be Universal and apply to all digital objects. It is Virtual as it uses emulation and/or virtualization ("Virtual" is also included in homage to the UVC concept described above). And it, like the objects it is used with, is Interactive.

The UVI is intended to automatically present a digital object for interaction by a user in a software environment that is either the original that the object was created and used within, or a software

[1] Such detailed application-specific metadata may be useful for distinguishing between functionally different file format variants such as "Excel 4.0 created by Microsoft Excel 97" vs. "Excel 4.0 as created by Quattro Pro 8".

environment that represents one that would have been in use at the time the object was created (and/or soon after). To enable this functionality the UVI attempts to automatically map attributes of digital objects to pre-configured software environments that could be used to interact with them. It dynamically generates view-paths based on analysis of the digital objects that are submitted to it and the metadata available about the configured environments that exist within the EaaSI network. For example, the UVI might analyse a ".doc" file and identify that it was created with WordPerfect 5.1 for MS-DOS and automatically match it to an available emulated environment containing that software. Similarly, it might identify that a ".xls" file was created in OpenOffice Calc 1.0 and match the file to a pre-configured emulated environment containing that software, as it is most likely to be the "best" option to use for interacting with that file (despite .xls being the default format for Microsoft Excel and not OpenOffice Calc 1.0).

The UVI's algorithm analyses a number of factors to identify these view-paths including:

1. It analyses the dates associated with files in the digital object (e.g. last edited and earliest created) and the available file format information and attempts to identify which environments can be used to interact with the object. The dates may come from file system metadata or embedded metadata and are evaluated for trustworthiness using a variety of tests.
2. It attempts to match the data ranges within the digital objects to first identify what software was in use and popular at the time the object was created and in use.
3. It then identifies, of that set, which of the applications could open or import the objects.
4. In additional steps, it then uses further metadata (where available) to attempt to further reduce the list of possible interaction environments, details such as which applications were popular at the time, which application created the object (information that may be inferred from metadata within the file and from information about applications that were available contemporaneously with the file), or which application created those files by default and was also generally used to interact with them.

The algorithm evolves in response to additional configuration metadata contributed by EaaSI Network users. The result is a list of environments that are available in the EaaSI network with weightings associated with how likely they are to be an appropriate representative (representing an environment that would have been used at the time the object was in use) environment to interact with the object.

Developers of discovery and access systems can choose how they want to use that list of environments. They may choose to present all options to a user, only the highest weighted option, randomly assign an environment, or use some other approach. The UVI is agnostic about this decision.

Once the files have been mapped to environments the system has a number of options for ensuring the objects are made interactable to users. In all cases the object is made available to the environment either by:

1. Including the object in a disk image that is attached to an environment
2. Editing the disk image to insert the object into a location in the file system.

We then have multiple options for enabling the content to be opened in the target software application within the emulated software environment.

1. The environment can be presented to a user with text instructions indicating how to open the object using the interface of the emulated environment
2. The disk image of the core operating system can be edited to force the object to execute on start-up (e.g. by placing the file or a link to it in the "startup" folder of a Windows environment or by inserting a script into a Linux boot process that utilizes the executable syntax metadata referenced above to open the digital object with a particular executable upon system start-up.
3. A mouse/keyboard input macro can be run after the system has loaded in order to open the object in the appropriate software.

The end result of all of these configurable options is that after a user clicks on a link to an object in a web browser they are quickly presented with the ability to interact with it in an "original" or representative

software environment from the era of the object.

A. *Progress*

The EaaS and EaaSI teams already have some of the components available to enable this automation and are rapidly building and creating more. The basic tooling to automate the steps of connecting an object to an available software environment and editing the disk image to make the object execute at system boot time are already available. A limited version of this approach is used in the German National Library to automatically present CD-ROMs to users in emulated computers running operating systems that the CDs should be compatible with [22].

The EaaSI team have acquired a large software collection and have begun configuring and documenting software environments using the approach outlined above. The initial prospects are promising and we're aiming to have 3000 software environments configured by July 2020.

IV. Future work

Programmatic Interaction with Environments

Early work with migration by emulation was completed as part of the PLANETS project by the partners at the University of Freiburg. This approach involved automating attaching a file to a computer environment, within a disk image, loading the environment, then running a macro/program that uses pre-recorded and automated mouse and keyboard inputs to open the file in an application and save the contents into a new file with a different format. This tooling still exists behind the scenes in the Emulation as a Service software that is a direct descendant from the work of the PLANETS project and a core part of the UVI. Using these features the environments created for the UVI could be re-purposed not just to serve as tools for "manual" interaction but also to be used as tools for automated interaction with at least two potential use cases:

1. The aforementioned migration by emulation, including daisy-chaining migration steps using multiple software environments.
2. Enabling "distant reading" [23] of a variety of different software environments or of sets of diverse digital objects using the same software environment.

a. For example, a researcher may be interested in comparing changes in user interfaces over time by automatically loading, automatically interacting with, and analysing the output of the environments over time. Or a researcher may be interested in automatically comparing the rendering of one digital object in a diverse variety of different software environments by automatically opening the same object in a variety of different software environments, interacting with them automatically, and analysing the outputs.

We are also following the work of the Preservation Action Registry (PAR) project [24] with great interest. As we develop persistent identifiers for computing environments there is the potential to incorporate emulation view paths into PAR with the UVI as the "tool" involved. As discussed in [25] this would enable digital preservation system developers to match digital objects to UVI compatible software environments during the ingest process and to use this information to enable access tools to automatically present the object in the appropriate environment when it is requested for access. Additionally the migration pathways enabled by the migration-by-emulation functionality should expand the PAR dataset extensively.

Our UVI is machine/algorithm driven and so the more environments that are available and the greater diversity between them, the more powerful the UVI becomes. However, we don't yet have Artificial Intelligence (AI) algorithms available to do the kind of configuration and documentation tasks described in section C. above. We are currently manually pre-configuring multiple slightly different environments and describing them with machine readable metadata. For example, we configure the same environment with multiple different pre-configured display resolutions to enable users/machines to just pick a pre-configured option rather than having to make the configuration change themselves. In the future we would like to explore using programmatic interaction with the environments to both configure and document new environments in order to further reduce the cost of populating the EaaSI network and improve the effectiveness of the UVI.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

B. *Integration into Discovery and Access Systems*

The UVI is being built as a set of APIs that enable developers to either:

1. Request an object opened in a specific environment and get back the information required to embed the environment for interaction in a browser window.
2. Submit an object and request a list of potentially appropriate environments for use in interacting with the object, with weighting data to aid in selection/presentation to a user
3. Submit metadata (dates and file format information) and request a list of potentially appropriate environments for use in interacting with the object, with weighting data to aid in selection/presentation to a users
4. Submit a file or metadata and receive back the information required to embed the likeliest appropriate environment in a web browser.

This flexibility provides developers with a number of options for how they integrate the UVI into their discovery and access workflows. They may wish to provide more or less options to end-users and may already know which environments they want to use for particular digital objects.

As discussed above, a version of this approach is already in use in the German National Library [22] and the EaaSI nodes are aiming to explore integrating the UVI into their access and discovery systems beginning in 2020.

C. *Reducing time to load environments*

The EaaS team at the University of Freiberg and their commercial offshoot OpenSLX GmbH have been working to enable computing environments to be paused at a point in time and restarted instantaneously. That functionality coupled with macro-based interaction with environments would enable reducing the time from clicking on a digital object in a finding aid or catalogue and having it presented to you in your web browser. An environment could be instantaneously loaded with an object attached in a disk image and a macro immediately run that opens the file using keyboard/mouse interactions. This could also be managed such that the user doesn't get presented with the environment in their browser until the macro has been completed ensuring no conflicts between the macro-driven inputs and the user's manual inputs.

V. Conclusion

The UVI is the conceptual legacy of more than two decades of applied research on emulation in cultural heritage contexts including the Planets Project, the UVC project and Jeff Rothenberg's early work and research with the Dutch National Library and Archives. Our current work on/within the EaaSI program further reduces barriers to using emulation and preserved software as a means of interacting with preserved digital objects. While detail-heavy and time-intensive, the collective efforts of the EaaSI Network will pay dividends in the future through economies of scale. When the environments and tooling we are developing re-enable access to potentially limitless digital objects that might otherwise be inaccessible or lose significant fidelity and content, their value will be clear. Additionally, once the UVI is standardised we will have the opportunity to open up additional services and integration points to spread the benefits throughout the digital preservation community and on to the public at large.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

References

[1] J. Rothenberg, "Ensuring the Longevity of Digital Documents", Scientific American, Vol. 272, Number 1, pp. 42-7. 1995.

[2] J. Meyerson, "Software Preservation Literature Review: The Co-Determinacy of User Needs and Advances in Preservation Methods", 2014,https://www.softwarepreservationnetwork.org/blog/software-preservation-literature-review-2014/, accessed 03/19/2019

[3] J. Slats Et al, "Digital Preservation Testbed White Paper Emulation: Context and Current Status", Digital Preservation Testbed Project Koninklijke Bibliotheek, National Library of the Netherlands, 2003, https://web.archive.org/web/20050305150902/http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf, accessed 17/03/2019.

[4] E. Cochrane, Rendering Matters, Archives New Zealand, 2012, http://archives.govt.nz/rendering-matters-report-results-research-digital-object-rendering, accessed 17/03/2018

[5] E. Cochrane, Visual Rendering Matters, Archives New Zealand, 2012, http://archives.govt.nz/resources/information-management-research/rendering-matters-report-results-research-digital-object-0, accessed 17/03/2018

[6] P. McGlone, "A guy in Minnesota is the museum world's answer to old technology", The Washington Post, 20/04/2018, https://www.washingtonpost.com/entertainment/museums/a-guy-in-minnesota-is-the-museum-worlds-answer-to-old-technology/2018/04/19/78cae5aa-3dcd-11e8-8d53-eba0ed2371cc_story.html accessed 17/03/2019

[7] J. Rothenberg, "Ensuring the Longevity of Digital Information, Council on Library and Information Resources, 1999, http://www.clir.org/wp-content/uploads/sites/6/ensuring.pdf, accessed 3/17/19

[8] E. Cochrane, The Emergence of Digital Patinas, The Digital Preservation Coalition Blog, 2017, https://dpconline.org/blog/idpd/the-emergence-of-digital-patinas, accessed 17/3/2019

[9] K. Rechert, D. von Suchodoletz, R. Welte, "Emulation based services in digital preservation", Proceedings of the 10th annual joint conference on Digital libraries, Pages 365-368, 2010

[10] R. Lorie, "Long term preservation of digital information", Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, Pages 346-352, 2001.

[11] Dioscuri team, "Dioscuri: Ideas and Key Features", 2007, http://dioscuri.sourceforge.net/dioscuri.html, accessed on 17/03/2019

[12] Keep Project Team, "Keeping Emulation Environments Portable", 2012, https://web.archive.org/web/20120121170030/ http://www.keep-project.eu/ezpub2/index.php, accessed 3/17/19

[13] E.Oltmans, R. van Diessen, H. van Wijngaarden, "Preservation functionality in a digital archive", JCDL '04 Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries, Pages 279-286, 2004.

[14] D. von Suchdoletz, J. van der Hoeven, "Emulation: From Digital Artefact to Remotely Rendered Environments", The International Journal of Digital Curation, Issue 3, Vol 4, 2009.

[15] S. Bonin, "Preservation and Long-term Access via NETworked Services: Keeping digital information alive for the future", 2009, https://planets-project.eu/docs/comms/PLANETS_BROCHURE.pdf accessed 17/03/2019

[16] K. Rechert, I. Valizada, D. von Suchodoletz, J. Latocha, bwFLA - A Functional Approach to Digital Preservation. Praxis der Informationsverarbeitung und Kommunikation 35(4): 259-267, 2012

[17] Scaling Software Preservation and Emulation as a Service Infrastructure, "About EaaSI", 2019 https://www.softwarepreservationnetwork.org/eaasi/, accessed 17/03/2019

[18] T. Liebetraut, K. Rechert. "Management and Orchestration of Distributed Data Sources to Simplify Access to Emulation-as-a-Service". iPRES 2014

[19] OpenSLX GmbH, "EaaSI Group", Gitlab.com, 2019 https://gitlab.com/eaasi, accessed 17/03/2019

[20] Association of Research Libraries et al, "Code of Best Practices in Fair Use for Software Preservation", 2019 https://www.arl.org/focus-areas/copyright-ip/fair-use/code-of-best-practices-in-fair-use-for-software-preservation, accessed on 17/03/2019

[21] Software Preservation Network, "Home" , 2019 https://www.softwarepreservationnetwork.org/, accessed on 19/03/2019

[22] K. Rechert, T. Liebetraut, O. Stobbe, N. Lubetzki, T. Steinke, "The RESTful EMiL: Integrating emulation into library reading rooms", Alexandria: The Journal of National and International Library and Information Issues, Vol 27, Issue 2, pp 120-136, 2017.

[23] F. Moretti (2013). Distant Reading. Verso, London, 2013.

[24] M. Addis, J. O'Sullivan, J. Simpson, P. Stokes, J. Tilbury, "Digital preservation interoperability through preservation actions registries: iPres 2018 – Boston" iPres 2018, Boston, 2018

[25] E. Cochrane, J. Tilbury, O. Stobbe, "Adding Emulation Functionality to Existing Digital Preservation Infrastructure", iPres Conference 2017, 2017

Who is asking? Humans and machines experience a different scholarly web

Martin Klein
Los Alamos National Laboratory USA
mklein@lanl.gov
https: // orcid. org/ 0000-0003-0130-2097

Lyudmila Balakireva
Los Alamos National Laboratory USA
ludab@lanl.gov
https: // orcid. org/ 0000-0002-3919-3634

Harihar Shankar
Los Alamos National Laboratory USA
harihar@lanl.gov
https: // orcid. org/ 0000-0003-4949-0728

Libraries and archives are motivated to capture and archive scholarly resources on the web. However, the dynamic nature of the web in addition to frequent changes at the end of scholarly publishing platforms have crawling engineers continuously update their archiving framework. In this paper we report on our comparative study to investigate how scholarly publishers respond to common HTTP requests that resemble typical behavior of both machines such as web crawlers and humans. Our findings confirm that the scholarly web responds differently to machine behavior on the one hand and human behavior on the other. This work aims to inform crawling engineers and archivists tasked to capture the scholarly web of these differences and help guide them to use appropriate tools.

Scholarly Web, Web Crawling, Scholarly Publishing Designing and Delivering Sustainable Digital Preservation; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

Web archiving has become an essential task for libraries worldwide. However, scoping this endeavor is a non-trivial issue. Fortunately, academic libraries may take guidance from a collection development policy that specifies, among other aspects, which part of the web to focus on for crawling and archiving. The collection development plans from Stanford University Libraries [1] and from Purdue University [2] are great examples that can help steer libraries' archiving efforts. National libraries, for example the National Library of Finland [3] or the Library and Archives Canada [4], on the other hand, often have a mandate to collect and archive (national) web resources. Ideally, those documents also narrow down the scope and provide direction as to which pockets of the web to focus resources on. For both types of libraries, the scholarly web typically is in scope of their archiving efforts. This can, for example, be because they are authored by representatives of the university/country or because it is understood that members of the respective communities will benefit from the long-term availability and accessibility of such resources. We refer to the scholarly web as a space where scholarly resources can be deposited (a pre-print server, for example) and where these resources have a URL and are accessible to a reader. For the remainder of this paper, we limit our definition of the scholarly web to the latter aspect, a part of the web from which scholarly resources can be consumed.

These individual web crawling and archiving efforts are organized and conducted by libraries themselves. They are more narrow in scope, smaller at scale, and run with fewer resources compared to, for example, large programs such as LOCKSS[1] or Portico[2], which are specialized in the preservation of journal publications. However, with the constantly changing nature of the web [5]–[7] and the realization that online scholarly resources are just as ephemeral as any other resource on the web [8],

[1] https://www.lockss.org/

[2] https://www.portico.org/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

[g], libraries are facing the challenge to keep up with their crawling and archiving frameworks.

When it comes to identifying scholarly resources on the web, the Digital Object Identifier (DOI)[1] has become the de facto standard. In order to make a DOI actionable in a web browser, the recommended display is in the form of a HTTP DOI e.g, https://doi.org/10.1007/ 978-3-540-87599-4_38. When a user dereferences this HTTP DOI in a web browser, the server at doi.org (operated by the Corporation for National Research Initiatives (CNRI)[2]) responds with a redirect to the appropriate URL at the publisher. From there, the browser often follows further redirects to other URLs at the publisher and eventually to the location of the DOI-identified resource. The HTTP redirection is done automatically by the browser and the user often does not even notice it. In the above example the browser redirects to the article's Springer landing page hosted at https://link.springer.com/chapter/10.1007%2F978-3-540-87599-4_38. This scenario is very typical in a way that the DOI identifies an academic journal article and, unlike the HTTP DOI, the landing page itself is controlled by the journal's publisher.

Bringing both of these considerations together, we are motivated to investigate how scholarly publishers respond to common HTTP requests that resemble typical behavior of machines such as web crawlers. We therefore send such HTTP requests against thousands of DOIs, follow the HTTP redirects, and record data the publishing platforms respond with. To put responses to machine requests in context, we compare them to responses we received from requests that more closely resemble human browsing behavior.

In this paper we report on the results of this comparative study. Our findings provide insight into publishers' behavior on the web and inform crawling engineers and archivists motivated to capture the scholarly web to use appropriate tools for the task at hand. With the insight that popular web servers do not necessarily adhere to web standards or best practices [10], we have no reason to assume that scholarly publishers are any different. To the contrary, various reports document the sometimes complex relationship between publishers and web crawlers [11], [12]. We therefore believe our work is a worthwhile contribution to the crawling and web archiving as well as to the digital preservation community at large.

We aim to address the following research questions:

RQ1: Do scholarly publishers send the same response to different kinds of HTTP requests against the same DOI? If not, what are the noticeable differences?

RQ2: What characteristics does an HTTP request issued by a machine have to have in order to obtain the same result as a human?

RQ3: Does the DOI resolution follow the same paths for different HTTP requests?

II. Related work

A study of the support of various HTTP request methods by web servers serving popular web pages was conducted by Alam et al. [10]. The authors issue OPTIONS requests to web servers and analyze the "Allow" response header used by servers to indicate which HTTP methods are supported. The study finds that a large percentage of servers either erroneously report supported HTTP methods or do not report supported methods at all. While this study is related in concept, both its scope and methodology are significantly different from our here presented work. The focus of our work is on DOI redirects from the scholarly domain and not just web servers serving popular pages. Unlike Alam et al. we are actually sending a variety of HTTP requests against resources and analyze the responses where they only sent OPTIONS requests and analyzed responses for claims of supported requests.

DOIs are the de facto standard for identifying scholarly resources on the web and therefore a common starting point for crawlers of the scholarly web. We have shown previously that authors, when referencing a scholarly resource, use the URL of the landing page rather than the DOI of the resource [13]. These findings are relevant, for example, for web crawling engineers that need to avoid duplicate crawled resources.

[1] https://www.doi.org/

[2] https://www.cnri.reston.va.us/

Similarly, the motivation behind the recent study by Thompson and Jian [14] based on two Common Crawl samples of the web was to quantify the use of HTTP DOIs versus URLs of landing pages. They found more than 5 million actionable HTTP DOIs in the 2014 dataset and roughly 10% of them as their corresponding landing page URL in the 2017 dataset

Various efforts have proposed methods to make web servers that serve (scholarly) content more friendly to machines. There is consensus in the scholarly communication community that providing accurate and machinereadable metadata is a large step in this direction [15], [16]. Aligned with this trend, sitemap-based frameworks have recently been standardized to help machines synchronize metadata and content between scholarly platforms and repositories [17].

III. Experiment setup

A. Data Gathering

Obtaining a representative sample of the scholarly web is not a trivial endeavor. Aside from the concern that the sample should be large enough, it should also reflect the publishing industry landscape since, as for example outlined by Johnson et al. [18], the Science, Technology, and Medicine (STM) market is dominated by a few large publishers.

The Internet Archive (IA)[1] conducted a crawl of the scholarly domain in June of 2018 that lasted for a month and resulted in more than 93 million dereferenced DOIs. The IA crawler followed all redirects, starting from the HTTP DOI to the URL of the DOI-identified resource and recorded relevant data along the way. We refer to the result of dereferencing a DOI as a chain of redirects consisting of one or more links, each with their own URL.

We obtained a copy of the recorded WARC files ([1g]) from this crawl and extracted the entire redirect chain for all 93, 606, 736 DOIs. To confirm that this crawl captures a representative bit of the scholarly landscape, we were motivated to investigate the distribution of publishers in this dataset. We approached this by extracting the URLs of the final link in the redirect chains and examined their hosts. For example, dereferencing the HTTP DOI shown in Section I leads, after following a number of links in the redirect chain to the final URL of the resource at https://link.springer.com/article/10. 1007/s00799-007-0012-y. The host we extracted from this URL is springer.com.

Figure 1 shows the distribution of all hosts extractedfrom the IA crawl dataset. The x-axis lists all hosts and the y-axis (log scale) shows their corresponding frequency. We expected to see a pattern as displayed in Figure 1, given the market dominance of a few publishers and a long tail of small publishers with less representation in the overall landscape. Table 1 lists the top 10 hosts by frequency extracted from the dataset6. We can observe a good level of overlap between top publishers shown by Johnson et al. [18] (cf. Table 1, p. 41) and hosts shown in Table 1. These observations lead us to believe that we have a dataset that is representative of the broader scholarly publishing landscape.

In order to scale down the dataset to a manageable size, we randomly picked 100 DOIs from each of the top 100 hosts, resulting in a dataset of 10, 000 DOIs[2].

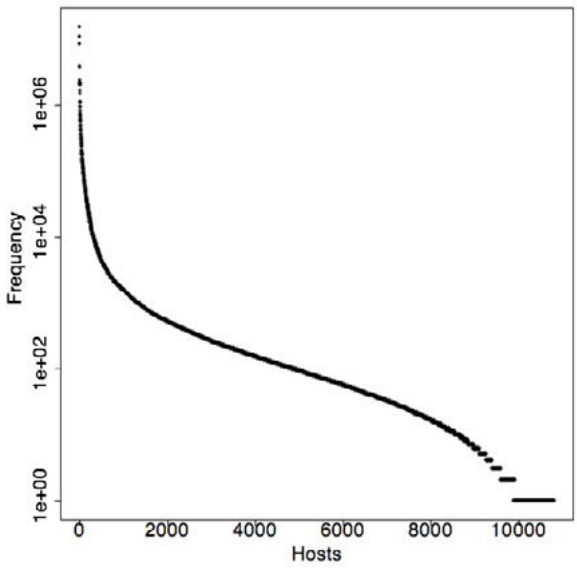


Figure 1: Dataset domain frequency

B. HTTP Requests

[1] https://archive.org//

[2] The dataset is available at: https://doi.org/10.6084/m9.figshare.7853462.v1

HTTP transactions entail request headers sent by the client (such as a web browser) and response headers sent by the web server and received by the requesting client. For a detailed description of defined request and response headers we refer to RFC 7231 [20]. RFC 7231 also specifies all HTTP request methods including the two very frequently used methods GET and HEAD. For detailed information about these methods we again refer to the RFC but note from its text that: "The GET method requests transfer of a current selected representation for the target resource." and "The HEAD method is identical to GET except that the server MUST NOT send a message body in the response...". With respect to the response headers, RFC 7231 states: "The server SHOULD send the same header fields in response to a HEAD request as it would have sent if the request had been a GET...".

Domain	Frequency
elsevier.com	15, 631, 553
springer.com	11, 011, 605
wiley.com	8, 533, 984
ieee.org	3, 941, 252
tandfonline.com	3, 780, 553
plos.org	2, 386, 247
oup.com	2, 199, 106
jst.go.jp	2, 162, 502
sagepub.com	2, 126, 063
jstor.org	2, 060, 760

Table 1: Top 10 domains of final URLs of dereferenced DOIs in our dataset

cURL[1] is a popular tool to send HTTP requests and receive HTTP responses via the command line. Listing 1 shows cURL sending an HTTP HEAD request against a HTTP DOI. The option -I causes the HEAD request method and the added L forces cURL to automatically follow all HTTP redirects. The Listing also shows the received response headers for both links in the redirect chain. The first link has the response code 302 (Found, see [20]) and the second link shows the 200 (OK) response code, which means this link represents the end of this redirect chain.

```
curl -IL http://doi.org/10.1016/j.wocn.2010.05.003

HTTP/1.1 302
Date: Sat, 16 Mar 2019 23:20:13 GMT
Content-Type: text/html; charset=utf-8
Content-Length: 20
Location:
https://linkinghub.elsevier.com/retrieve/pii/S0095...

HTTP/1.1 200
Date: Sat, 16 Mar 2019 23:20:14 GMT
Content-Type: text/html;charset=UTF-8
Content-Language: en-US
```

Listing 1: HTTP HEAD request against a DOI

```
curl -iL http://doi.org/10.1016/j.wocn.2010.05.003

HTTP/1.1 302
Date: Sat, 16 Mar 2019 23:20:13 GMT
Content-Type: text/html; charset=utf-8
Content-Length: 20
Location:
https://linkinghub.elsevier.com/retrieve/pii/S0095...

HTTP/1.1 200
Date: Sat, 16 Mar 2019 23:20:14 GMT
Content-Type: text/html;charset=UTF-8
Content-Language: en-US

<!DOCTYPE HTML PUBLIC ...>
<html>
<head>
....
</head>
<body>
....
</body>
</html>
```

Listing 2: HTTP GET request against a DOI

Listing 2 shows cURL sending an HTTP GET request against the same DOI and we can see the web server responding with the same response headers but now the last link in the redirect chain (response code 200) also includes the response body. Due to space constraints, we have removed most of the content and only show the basic HTML elements of the response body in Listing 2.

C. *Dereferencing DOIs*

For our experiment, we deployed four different methods to dereference each of the DOIs in our dataset. All four methods were run automatically by a machine since manually dereferencing 10, 000 DOIs and recording data for each link in the redirect chain is not feasible. However, since it is our intention to investigate how scholarly publishers respond to a variety of requests, we implemented two methods that resemble machines crawling the

[1] https://curl.haxx.se/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

web and two that resemble humans browsing the web. Our methods are:

1. HEAD: Use cURL to send an HTTP HEAD request against the DOI. This lightweight method resembles machine behavior on the web as humans usually do not send HEAD requests.

2. GET: Use cURL to send an HTTP GET request against the DOI. This method also resembles machine behavior as these GET requests do not include typical parameters set by common web browsers.

3. GET+: Use cURL to send an HTTP GET request against the DOI along with the typical browser parameters:
   - user agent,
   - specified connection timeout,
   - specified maximum number of HTTP redirects,
   - cookies accepted and stored, and
   - tolerance of insecure connections.

   This method, while also based on cURL, resembles a human browsing the web with a common web browser due to the setting of these typical parameters.

4. Chrome: Use the Chrome web browser controlled by the Selenium WebDriver[1] to send an HTTP GET request against the DOI. This method is virtually the same as a human browsing the web with Chrome. This method is typically used for web functionality testing [21]–[23] and is therefore commonly considered a proper surrogate for humans browsing.

Each of our four methods automatically follows all HTTP redirects and records relevant data for each link in the redirect chain. The recorded data per link includes the URL, the HTTP response code, content length, content type, etag, last modified datetime, and a link counter to assess the total length of the redirect chain. Each redirect chain ends either successfully at the final location of the resource (indicated by HTTP code 200), at an error (indicated by HTTP response codes at the 400or 500-level), or when an exit condition of the corresponding method is triggered. Examples for an exit condition are a timeout (the response took too long) and the maximum number of redirects (links in the chain) has been reached. For our methods HEAD and GET these two values are the defaults of the utilized cURL version 7.53.1 (300 seconds and 20 redirects) and both values are specifically defined for our GET+ method as 30 seconds and 20 redirects. For our Chrome method we use the default settings of 300 seconds for the timeout and a maximum of 20 redirects. The GET+ and the Chrome methods further have the user agent:

"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/65.0.3325.181 Safari/537.36"

specified for all their requests. This user agent mirrors the settings of a desktop Chrome browser to further resemble a human browsing the web. These parameters are based on inspiration from conversations with representatives from the LOCKSS platform. They are therefore based on real-world use cases and hence not subject to an individual evaluation in this work.

It is worth mentioning that we ran our experiments on a machine operated by Amazon Web Services, which means we expect the machine to not have access to paywalled content identified by a DOI. This implies that, just like for the example shown in Section I, for the most part our redirect chains, if successful, ends at a publisher's landing page for the DOI-identified resource. We do not obtain the actual resource such as the PDF version of the paper, for example. The IA crawl, on the other hand, was conducted on IA machines that may have access to some paywalled resources.

IV. Experiment results

Our four methods dereferencing each of the 10, 000 DOIs results in 40, 000 redirect chains and recorded data along the way. For comparison we also include the data of redirect chains recorded by the IA during their crawl of the DOIs in our analysis. We therefore have a total of 50, 000 redirect chains to evaluate.

[1] https://docs.seleniumhq.org/projects/webdriver/

A. HTTP Response Codes Across Methods

Our first investigation was related to our RQ1 and the HTTP response code of the last link in all redirect chains. In an ideal world, all redirect chains would end with a link that indicates "success" and returns the HTTP response code 200, regardless of the request method used. However, from experience navigating the web and educated by previous related work [10], we anticipated to observe a variety of different responses, depending on our four methods.

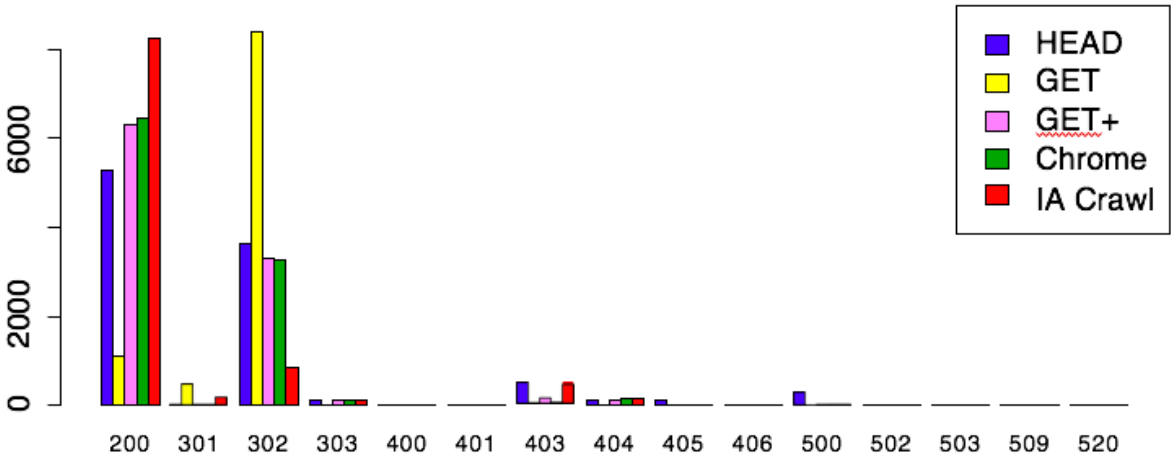


Figure 2: HTTP response codes of the last link in all redirect chains per request method

Figure 2 shows all observed response codes for last links and their frequencies distinguished by requesting method. Each set of five bars is assigned to one individual response code and each bar within a set represents a request method. Within a set, from left to right, the blue bar represents the HEAD method, the yellow bar GET, the pink bar GET+, the green bar Chrome, and the red bar the IA crawl.

We notice a spectrum of 15 different response codes from dereferencing our 10, 000 DOIs across five different methods. The distribution of our observed 50, 000 response codes is almost binary with 27, 418 being 200s and 19, 459 being 302s. Our two methods that resemble a human browser, GET+ (pink bars) and Chrome (green bars) requests result in more than 63% and 64% 200 response codes, respectively. These numbers are disappointing as we would expect more than two out of three HTTP DOIs to resolve to a successful state. The HEAD request method results in even fewer successful responses (53%). The IA crawl scores much better with 83% successful responses. We can only speculate as to the reasons why, especially since their crawls are done with the Heritrix web crawler[1] and this software is more closely aligned with our GET+ than our Chrome method, which returns the most successful results of any of our methods. It is possible though that the crawl parameters were more "forgiving" than ours, for example allowing for a longer timeout.

Our second observation from Figure 2 is that our GET request method results in a very poor success ratio (11%), rendering this method effectively useless for dereferencing DOIs. The majority of DOIs (84%) result in a 302 response code as indicated by the yellow bar in Figur 2. A redirect HTTP response code for the final link in a redirect chain intuitively does not make sense. However, after close inspection of the scenarios, we noticed that this response code is

[1] https://webarchive.jira.com/wiki/spaces/Heritrix/overview

indeed from the last link as the request most often times out. This means the web server simply takes too long to respond to such requests and our method cancels the request at some point. Since this GET method very closely resembles requests that would typically be made by machines, the suspicion arises that this web server behavior is designed to discourage crawling of scholarly publishers' resources. All other response codes do not play a significant role as they are returned in less than 5% of requests.

Figure 2 provides first strong indicators to answer RQ1: the scholarly web indeed responds differently to machines and humans.

B. *HTTP Response Codes by DOI*

Figure 3 offers a different perspective on the investigation into response codes of final links. The figure does not distinguish between individual response codes anymore but clusters them into four groups: 200-, 300-, 400-, and 500-level represented by the colors green, gray, red, and blue, respectively. Each horizontal line in Figure 3 represents one DOI in our dataset and each of them consists of five horizontal segments. Each segment represents one request method and its coloring indicates the corresponding response code. The image confirms that very few DOIs return with the same response code for all four of our methods. For example, only 880 DOIs return a 200 response code across all four request methods. If we take the IA crawl into consideration as well, the numbers drop even further, in our example to 777

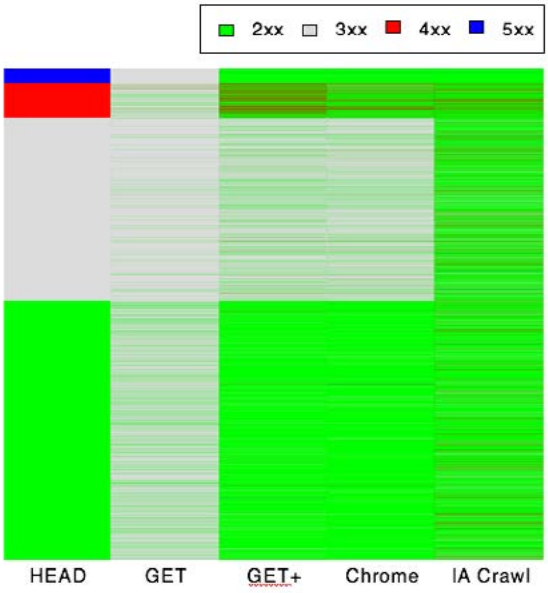


Figure 3: HTTP response codes of the last link in all redirect chains per request method by DOI

DOIs. It is interesting to note that, from visual inspection, the majority of 400 and 500-level responses for HEAD requests (690 and 286, respectively) indeed turn into 200 responses for GET+ and Chrome requests.

The impressions of Figure 3 provide further indicators that machine-based and human-based requests indeed result in different responses. They further hint at similarities between the responses for our GET+ and the Chrome method, which is relevant for crawl engineers and also part of answering our RQ2.

C. *HEAD vs GET Requests*

With our observation of the significant differences between our two machine-resembling request methods, we were motivated to investigate this matter further. In particular, we were curious to see how publishers respond to the lightweight HEAD requests compared to more complex GET requests. Figure 4 shows all DOIs that resulted in a 200 response code (indicated in green) for the HEAD method. The leftmost bar (HEAD requests) therefore is green in its entirety. The bar mirrors the 5, 275 DOIs (53% of the total) previously shown in Figure 2 (blue bar in the 200 category). The second, third, fourth, and fifth bar in Figure 4 represent the corresponding response codes of these DOIs for the respective request methods. We can observe that the vast majority of DOIs that result in a 200 for HEAD requests also result in a 200 for GET+ (93%), Chrome (96%), and the IA crawl (85%). This finding is not counterintuitive and it is encouraging in way that it would be a huge detriment to web crawling engineers if this picture was reversed, meaning we could not rely on response codes from HEAD requests being (mostly) the same for more complex GET requests. It is telling, however, that the simple GET request method does not echo the HEAD request but results in 83% 300level response codes instead.

The fraction of non-200 responses for the GET+, the Chrome, and the IA crawl are curious. As mentioned earlier, RFC 7231 states that web servers should respond with the same data for HEAD and GET requests but the shown differences indicate that the publishers' web servers do otherwise. The 5% of 400-level responses for the IA crawl (rightmost bar of Figure 4) might be explained by the different

time at which the crawl was conducted (June 2018) compared to our experiments (February/March 2019).

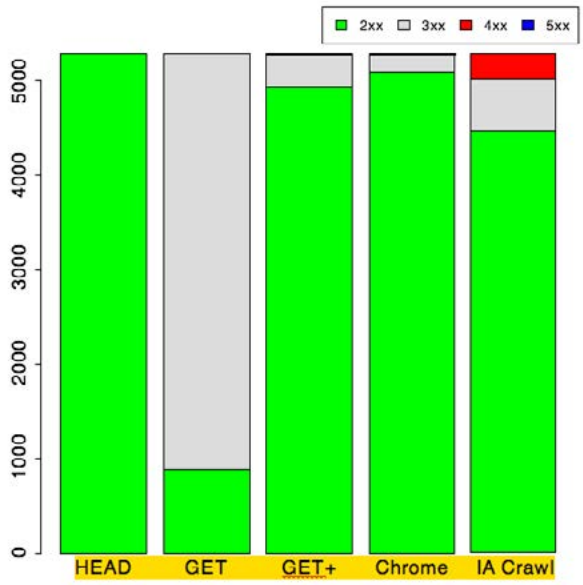


Figure 4: DOIs returning a 200 HTTP response code for HEAD requests and their corresponding response codes for other request methods

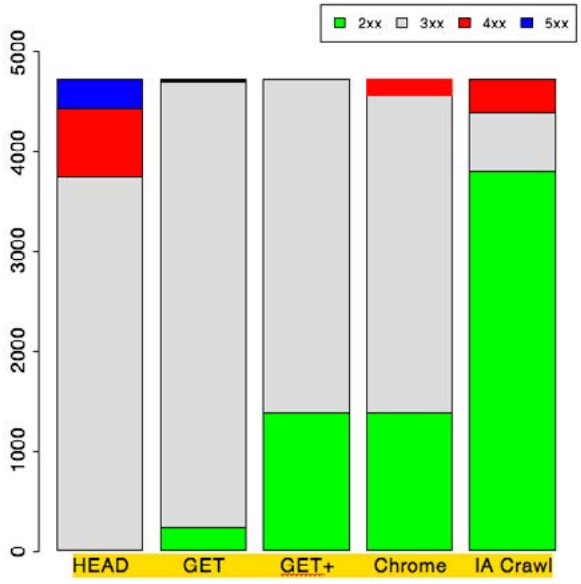


Figure 5: DOIs not returning a 200 HTTP response code for HEAD requests and their corresponding response codes for other request methods

Figure 5, in contrast, shows all DOIs (4, 725 in total) that resulted in a response code other than 200 for the HEAD request method. Consequently, the leftmost bar does not show any green portion at all. We find that 30% and 29% of these DOIs return a 200 code for the GET+ and Chrome method, respectively, and even 80% for the IA crawl. This picture does represent the above mentioned scenario where a developer can not trust the response to a HEAD request since publishers' web platforms seem to respond differently to other request methods.

Figures 4 and 5 clearly show different responses received when dereferencing the same resource with distinct request methods. They also confirm earlier findings related to answering RQ2 that the GET+ method seems to receive similar responses compared to the humanresembling Chrome method.

D. Redirects

Our next exploration was into the redirect chains and the number of links they consist of. The goal was to gain insights into whether the "path to the resource" as directed by the publisher is the same for machines if they even get that far and humans. As a first step we analyzed the total number of redirects for all chains per request method and show the results in Figure 6. We observe that the majority of chains for the HEAD, GET, GET+, and Chrome request methods are of length three or less. Given that the latter two methods result in more than 60% 200 response codes, this is relevant information for crawling engineers. The HEAD method has a noticeable representation with chains of length four (8%) and five (11%) where GET+ or Chrome methods rarely result in such long chains (around 3%). The GET method that mostly results in 300-level responses seems to fail quickly with more than 90% of chains being of length one or two. Note, however, that it may actually take a long time for a GET request to fail if it in fact waits for the timeout to expire. We can only speculate why the ratio of chains with length one is rather small for the IA crawl compared to our methods. Possible explanations are that the user agent used by the IA crawler makes a difference and that the partial access to paywalled content causes a different response and hence a different chain length. More analysis and further experiments run from different network environments are needed to more thoroughly assess this theory though. Figure 6

also shows 186 DOIs with a chain length of 21 links. 87 of them were returned from the HEAD request, two each from GET+ and Chrome, and 95 from the IA crawl. All of those DOIs are cases where the web server responds with one 302 code after another and virtually never stops. These scenarios are known as crawler traps and considered a serious detriment to crawler engineering as they can be diZcult to avoid. In our case, the maximum number of redirects was reached and hence the transaction was terminated by the client.

Figure 7 follows the same concept as Figure 6 but only shows the frequencies of chain lengths where the final link returned a 200 response. This data provides insight into how long (in terms of links, not seconds) it is worth waiting for the desired response and how many redirects to expect. We note that the majority of chains for the HEAD, GET+, and Chrome request methods are of length two, three, or four and, in addition, the HEAD method has a strong showing with chains of length four (8%) and five (10%). We also see a similar pattern with the IA crawl and a higher frequency of longer chains. It is interesting to note, however, that no chain in Figure 7 is recorded at length one. At the other end of the scale, there are indeed 15 chains of length 14 that all eventually result in a 200 response code for the HEAD request method.

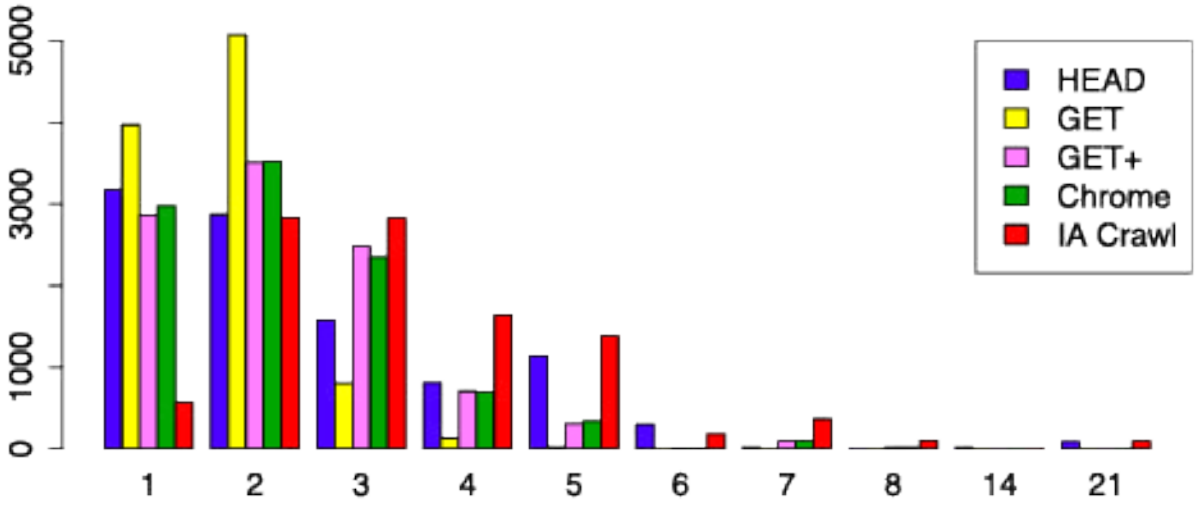


Figure 6: Frequency of number of redirects overall per request method

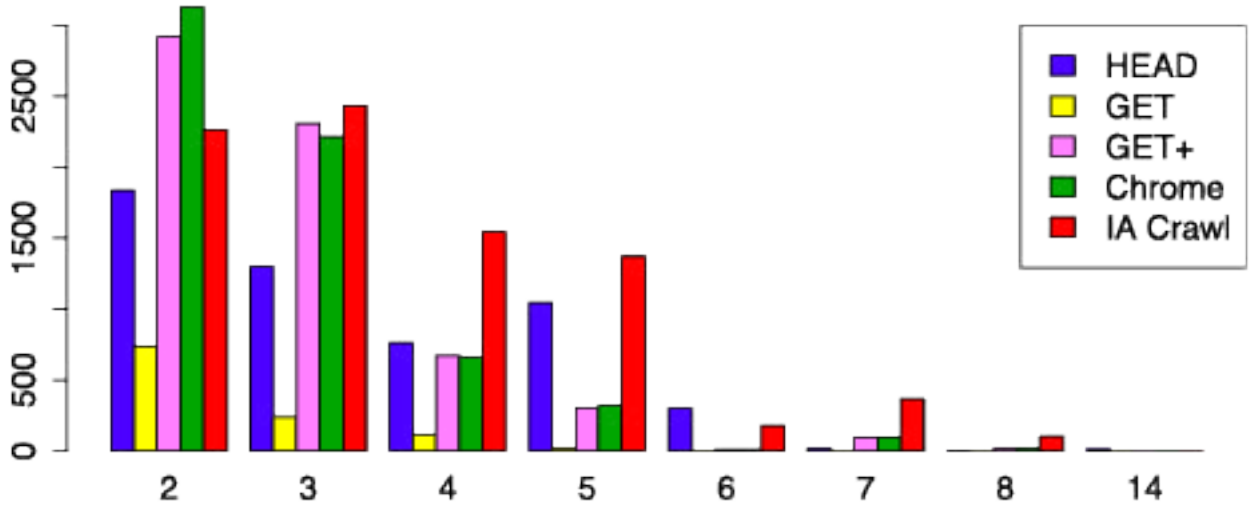


Figure 7: Frequency of number of redirects that lead to the final link with a 200 response code per request method

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Figures 6 and 7 show that responses not only differ in terms of the provided response code but also in terms of length of the redirect chain, depending on the request method. This finding confidently answers our RQ3, DOI resolutions do not follow the same path for different HTTP requests, at least not as analyzed by path length.

V. Discussion and future work

This experimentation is in its early stages and we see potential for improvements and aspects for future work. As alluded to earlier, we ran our experiments outside the institutional network. We are very interested in seeing how our results compare to those obtained when running the experiment from within an organization that has access to paywalled content. We hypothesize that differences in responses can still be observed between machine and human requests. Further, a logical extension to our series of experiments is to utilize existing infrastructure, for example, the CrossRef APIs to reliably identify individual publishers and make better assessments of their specific behavior on the web. Our dataset consists of 10, 000 DOIs obtained from a crawl by the IA. Our tests have shown that open science platforms that issue DOIs such as Figshare or DataCite are underrepresented in our sample. We are interested in repeating our analyses for these publishers as well, hoping that they might be friendlier to machines. We have utilized the established understanding that a Chrome browser that is controlled by the Selenium WebDriver is indeed virtually the same as a human browsing the web. We are motivated, however, to provide empirical evidence that this is in fact true. We are planning to pursue several approaches such as comparing screenshots taken by a human and by the Chrome method and comparing textual representations (or DOM elements) of the final link in the redirect chain. Such an extension of the experiment may also call for the inclusion of other crawling frameworks, for example, systems based on headless browsers.

VI. Conclusions

In this paper we present preliminary results of a comparative study of responses by scholarly publishers to common HTTP requests that resemble both machine and human behavior on the web. We were motivated to find confirmation that there indeed are differences. The scholarly web, or at least part of it, seems to analyze characteristics of incoming HTTP requests such as the request method and the user agent and responds accordingly. For example, we see 84% of simple GET requests resulting in the 302 response code that is not helpful to crawling and archiving endeavors. 64% of requests by our most human-like request method result in desired 200 responses. These numbers are somewhat sobering we would expect a higher percentage of successful responses but they do serve developers in managing their expectations, depending on the tool and request method used. In addition, they help to address our question raised earlier: "Who is asking?" as it nowcan clearly be answered with: "It depends!".

Acknowledgments

We would like to thank Bryan Newbold and Jefferson Bailey at the Internet Archive for their contributions to this work. We are particularly thankful for the provided dataset and input about their crawling approach. We are also grateful to Nicholas Taylor from Stanford University Libraries for his input regarding approaches implemented by the LOCKSS framework. Lastly, we appreciate Herbert Van de Sompel's contributions to this work in its early conceptual stages.

References

[1] Stanford University Libraries, Collection development, http://library.stanford.edu/projects/webarchiving/collection development.

[2] Purdue University, Web Archive Collecting Policy, https://www.lib.purdue.edu/sites/default/files/spcol/purduearchiveswebarchiving-policy.pdf.

[3] National Library of Finland, Legal Deposit OZce, https://www. kansalliskirjasto.fi/en/legal-deposit-office.

[4] Library and Archives Canada, Legal Deposit, https://www.baclac.gc.ca/eng/services/legaldeposit/Pages/legal-deposit.aspx.

[5] J. Cho and H. Garcia-Molina, "The Evolution of the Web and Implications for an Incremental Crawler," in Proceedings of VLDB '00, 2000, pp. 200–20g.

[6] J. Cho and H. Garcia-Molina, "Estimating frequency of change," ACM Transactions on Internet Technology, vol. 3, pp. 256–2g0, 3 2003, ISSN: 1533-53gg.

[7] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins, "Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay," in Proceedings of WWW '04, 2004, pp. 328–337.

[8] M. Klein, H. Van de Sompel, R. Sanderson, H. Shankar, L. Balakireva, K. Zhou, and R. Tobin, "Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot," PLoS ONE, vol. g, no. 12, 2014. DOI: 10.1371/journal.pone.0115253.

[9] S. M. Jones, H. Van de Sompel, H. Shankar, M. Klein, R. Tobin, and C. Grover, "Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content," PLoS ONE, vol. 11, no. 12, 2016. DOI: 10.1371/journal.pone.0167475.

[10] S. Alam, C. L. Cartledge, and M. L. Nelson, "Support for various HTTP methods on the web," CoRR, vol. abs/1405.2330, 2014. arXiv: 1405.2330. [Online]. Available: http://arxiv.org/abs/1405.2330.

[11] C. Hayes, Wiley using fake DOIs to trap web crawlers... and researchers, https://blogs.wayne.edu/scholar-scoop/2016/06/02/wileyusingfakedoistotrapwebcrawler-sandresearchers/,2016.

[12] L. A. Davidson and K. Douglas, "Digital object identifiers: Promise and problems for scholarly publishing," Journal of Electronic Publishing, vol. 4, no. 2, 1gg8.

[13] H. Van de Sompel, M. Klein, and S. M. Jones, "Persistent uris must be used to be persistent," in Proceedings of WWW '16, 2016, pp. 11g–120. DOI: 10.1145/ 2872518. 2889352. [Online]. Available: https://doiorg/10.1145/2872518.2889352.

[14] H. S. Thompson and J. Tong, "Can common crawl reliably track persistent identifier (PID) use over time?" CoRR, vol. abs/1802.01424, 2018. [Online]. Available: http://arxiv.org/abs/1802.01424.

[15] M. L. Nelson, J. A. Smith, and I. G. del Campo, "EZcient, automatic web resource harvesting," in Proceedings of the 8th Annual ACM International Workshop on Web Information and Data Management, ser. WIDM '06, 2006, pp. 43–50. DOI: 10.1145/1183550.1183560.

[16] O. Brandman, J. Cho, H. Garcia-Molina, and N. Shivakumar, "Crawler-friendly web servers," SIGMETRICS Perform. Eval. Rev., vol. 28, no. 2, pp. g–14, 2000. DOI: 10.1145/362883.362894.

[17] M. Klein, H. Van de Sompel, and S. Warner, ResourceSync Framework Specification (ANSI/NISO Z39.99-2017), http://www.openarchives.org/rs/1.1/resourcesync,2017.

[18] R. Johnson, A. Watkinson, and M. Mabe, The STM Report An overview of scientiic and scholarly publishing. International Association of Scientific, Technical and Medical Publishers, 2018. [Online]. Available: https://www.stm-assoc.org/2018_10_04_ STM_Report_2018.pdf.

[19] nternational Internet Preservation Consortium (IIPC), WARC Specification, https://iipc.github.iowarc-specifications/.

[20] R. T. Fielding and J. Reschke, Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content, https://tools.ietf.org/html/rfc7231, 2014.

[21] A. Bruns, A. Kornstadt, and D. Wichmann, "Web application tests with selenium," IEEE Software, vol. 26, no. 5, pp. 88–g1, 200g, ISSN: 0740-745g. DOI: 10.1109/MS.2009.144.

[22] A. Holmes and M. Kellogg, "Automating functional tests using selenium," in AGILE 2006 (AGILE'06), 2006, 6 pp.–275. DOI: 10.1109/AGILE.2006.19.

[23] C. T. Brown, G. Gheorghiu, and J. Huggins, An introduction to testing web applications with twill and selenium. O'Reilly Media, Inc., 2007.

Considerations on the Acquisition and Preservation of eBook Mobile Apps

Maureen Pennock
The British Library
United Kingdom
maureen.pennock@bl.uk
https://orcid.org/0000-0002-7521-8536

Peter May
The British Library
United Kingdom
peter.may@bl.uk
https://orcid.org/0000-0001-8625-9176

Michael Day
The British Library
United Kingdom
michael.day@bl.uk
https://orcid.org/0000-0002-1443-5334

Abstract – In 2018 and 2019, as part of the UK Legal Deposit Libraries' sponsored 'Emerging Formats' project, the British Library's digital preservation team undertook a program of research into the preservation of new forms of content. One of these content types was eBooks published as Mobile Apps. Research considered a relatively small number of apps in an attempt to better understand the preservation challenges associated with them and make recommendations for a way forwards. It found that whilst the content landscape is extremely varied, the technical challenges are similar for both Android and Apple apps. The greatest challenges appear to lie in the acquisition and access areas, particularly in enabling delivery to appropriate rendering environments and devices, though other challenges remain around the use of device sensors and content stored on remote servers rather than the user's device. Whilst these challenges may be surmountable with sufficient resource and investment, the lack of growth in this content area in recent years makes large scale investment questionable.

Keywords – mobile apps, challenges, access, acquisition, digital preservation

Conference Topics – Exploring New Horizons

I. Introduction

Mobile device ownership has increased dramatically over the past two decades. In 1996, when the Office of National Statistics first began collecting data on mobile phone ownership in UK households, only 16% of homes in the UK were known to own a mobile phone. [1] Twenty years later that figure had risen to 95%, many of which were 'smartphones', i.e. highly advanced phones with computer and internet capabilities, touchscreen, and other built-in components such as GPS or gyroscopes.

Smartphones began to appear in the early 2000s and allowed users to download, install, and run programs or 'apps' directly on their mobile phone. In the same manner, apps could also be downloaded to mobile 'tablet' computers, the popularity of which was similarly increasing. Early apps were often productivity or game based, but by the early 2010's traditional publishing houses had begun to explore how smartphone apps could be used to deliver new literature experiences. Faber and TouchPress (now known as Amphio) were early adopters, launching T. S. Eliot's 'The Waste Land' app in 2011 to widespread acclaim [2]. More apps followed, most of which had a clear link to existing analogue publications, but 2015 saw Amphio release a groundbreaking new work conceived wholly to be experienced as an app: Ian Pears' 'Arcadia'. During this time several other publishing houses began to deliver book-like content in app form, including Penguin, Random House, Nosy Crow, and Oxford University Press. Software houses such as Inkle and Simogo, alongside other smaller independent parties, also began to release book-like apps in which the textual work was the primary content type. Collectively, publishing and software houses enhanced the world of electronic books with a new way to experience titles both old and new.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

This 'new way of experiencing titles' poses a challenge for memory organizations tasked with preserving cultural heritage and electronic literature. The app genre is relatively unknown in most collecting libraries, and the overall eBook mobile app content landscape poorly mapped – how do we even begin to deconstruct the challenge and how should an organization know where to begin? If memory organizations are to collect and preserve apps, what technical issues must be addressed, and what access issues must be resolved? This paper presents work undertaken by the British Library as part of the UK Legal Deposit Libraries' 'Emerging Formats' project in an effort to begin answering those questions.

II. Exploring the Genre

Deconstructing the challenge requires a thorough understanding of the content type. During its first year, alongside exploration of other types of 'emerging formats', the project analyzed just three mobile app titles: The Solar System (Faber/TouchPress, first released in 2010); The Waste Land (Faber/TouchPress, first released in 2011), and Goldilocks and Little Bear (Nosy Crow, first released 2015). All of these were published on the iOS platform and have already been discussed within the larger context of the 'Emerging Formats' project. [3] The second year of the project increased the size of the sample in order to provide greater insight into the different types of eBooks published as mobile apps, the range of content features found within different types of eBook mobile apps, and experience of apps developed for consumption on an Android device. This increase allowed for a more in-depth evaluation of the challenges of collecting, preserving and providing access to mobile apps.

A total of twenty-two apps were selected for analysis in year two of the project. Each was chosen by content experts as an exemplar eBook mobile app, representative of different categories of content, different publishers, or notable particularly for award-winning features. Content was accessed either on a Samsung S7 phone running Android 8.0 or one of two Apple iPads running iOS9 and iOS11. Apps and their features are discussed by category, below.

A. Interactive narrative apps

Interactive narrative mobile apps are defined here as works which require choices to be made by the reader in order to establish the direction of the story, which can vary meaningfully between readers and reading experiences.

Text, particularly in narrative form, was the driving feature of all apps in this category. Some apps, such as Ian Pears' 'Arcadia' from Amphio [4], had a complex branching narrative with relatively few additional features beyond navigational aids, for example in the form of a map or index. Others such as '80 Days' from Inkle [5], 'The Kai Chronicles' from Project Aon [6], or 'The 8th Continent' from Ben Garrett [7], had significantly more game-like features built around the narrative, including functionality to accumulate and trade items, accrue funds, maintain health scores, and engage in combat. This increased functionality was often accompanied by extensive imagery, both background and foreground, and most apps in our sample also included background audio tracks. Advertisements appeared on occasion, particularly within free apps with a heavy emphasis on gameplay.

The apps in this category seemed to make limited use of built-in hardware features such as gyroscope, camera or microphone, and progress was typically made by swiping or tapping the screen.

B. Young children's apps

eBook apps for Young Children are understood here as apps aimed at young and early readers. All apps in this category made extensive use of color and imagery (as one might expect with children's literature). Most of the sample originated from Nosy Crow publishing: 'Snow White' [8] and 'Jack & the Beanstalk' [9] both provided a modern re-telling of a classic fairytale, whilst the 'Rounds' [10] series introduced young readers to the lifecycle of animals and 'Axel Scheffler's Flip Flap Pets' [11] was a modern take on traditional 'lift the flap' books. 'The Flitlits: the Funny Fair' [12] from Flitwits Ltd was the only work examined here from a publisher other than Nosy Crow, providing a bi-lingual exploration of a magical land that could be used to support teaching of the England and Wales Key Stage 1 and Key Stage 2 educational curriculum.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

All Nosy Crow apps were highly interactive. Tapping and swiping the device touchscreen were the most frequent ways to progress the story or complete in-book challenges and games, though some apps also made use of the device gyroscope, camera and microphone. All featured background music and the option for an automated 'read aloud' experience, which highlighted subtitled phrases on the screen when sounded out by the corresponding audio track. Some of the Nosy Crow apps included tracks from award-winning musicians or images from famous illustrators. In comparison, the single FlitLits app analyzed had limited potential for enhanced interaction, though the audio and textual tracks were available in English, Welsh, and American.

C. *Simple language apps*

In a similar manner to the young children's apps, the simple language apps analyzed were visually appealing, with heavy use of images and color, and relatively little text. 'Go!Gaelic'[1] was one such example from the Learning Centre of North Lanarkshire Council, which included a very limited range of English words and phrases that could be translated into Gaelic. The Interactive Welsh alphabet app 'Cyw a'r Wyddor', from Welsh-language broadcaster S4C [13], also contained a limited amount of content intended to support young children learning the language anew. Both apps made frequent use of audio, helping users to understand how translated words were pronounced. The interactive functionality of these apps was limited mainly to querying and listening to audio playbacks.

D. *Music apps*

Music apps are defined here as apps where music is the main feature of the app rather than a supporting component. Apps analyzed in this category included 'Beethoven's 9th Symphony' from Amphio [14] and 'Tido Music' from Tido [15]. Whilst music was the key feature in these apps, it was expressed in many different ways from audio-visual recordings, musical notation and descriptive textual information, to digitized content and graphical visualizations. One app, 'Steve Reich's Clapping Music' from Amphio [16], included a game as a central feature, played by tapping the touchscreen in time with the music.

The appearance and functionality of apps in this category varied, though a notable shared characteristic was that not all content was contained in the app package, with each requiring an internet connection or a fee to be paid in order to unlock access to additional content.

E. *Reference works*

Structured textual data was the predominant feature of the two reference works analyzed, though in some instances it was accessed directly from an online database rather than downloaded to the user's device. The 'Oxford Dictionaries Quick Search' app from Oxford University Press and MobiSystems [17], for example, included over 350,000 words and meanings, but whilst the full (ad-free) version cost £27.99 per year and included the option to download the entire database to your own device, the free version of the app provided limited access and required connection to a remote database. The GPC (Geiriadur Prifysgol Cymru) 'Welsh dictionary' app from the Centre for Advanced Welsh and Celtic Studies [18] offered similar install options, but in this case the full database could be downloaded directly to the app for free. Functionality was primarily limited to general searching and querying, though the Oxford Dictionary also retained a search history and allowed users to 'favourite' words.

F. *Miscellaneous apps*

A small number of apps did not fit easily into any of the categories identified. These displayed features akin to those seen across the other categories, but without some key characteristics. For example, 'American Interior' from Penguin Random House [19] and 'The 39 Steps' from Faber & TouchPress [20] were both narrative-based but absent of interaction that influenced development of the storyline, whilst the multimedia content seen in 'Shakespeare's Sonnets' [21] app was akin to that in Amphio's 'Beethoven's 9th Symphony' app [14], but with literature rather than music as the key feature.

III. Technical Dependencies

As is evident from the previous section, apps can contain an extremely wide range of content and functionality. The method by which this is packaged together and delivered to devices for consumption

[1] Go!Gaelic was withdrawn from both the Apple App store and the Google Play store in early 2019 for 'critical updates'

by readers is a container format: for iOS this is the IPA format; for Android it is the APK format. These containers hold not only the content that will populate the app, but also metadata and the software program (or application, from which 'app' is clearly derived) that will run on the access device's operating system and allow users to experience the content.

The structure of both container formats is well defined within official developer documentation. Both are based on the ZIP format, thought the APK specification is more closely aligned to the JAR format (a zip-like container format used to package together java class files, metadata and other resources prior to their distribution), and both can be unpacked on a desktop computer by simply amending the file extension to ZIP.

Both container formats are tightly linked to their relative platforms for deployment: IPA apps can only be run in an Apple environment, and APK in an Android environment. Within those environments, apps are optimised for specific versions of operating systems. Each operating system iteration typically brings about new APIs available for use by apps and together these essentially create a minimum version of the operating system environment that a given app will work in. Development best practice is to target the earliest version of the operating system that supports all the APIs required, in order to maximize the number of devices that will support the app in development. Both Android and Apple operating systems are updated relatively frequently. There have been 17 major versions of the Android operating system since its first release in 2008[1], and 12 releases of Apple's iOS since 2007[2], though early versions were limited to phones only. This rapid rate of development poses a challenge for apps with specific operating system dependencies, as these apps may not function as intended once a device's underlying operating system is updated.

Versioning dependencies should be documented in the app metadata. Specific hardware dependencies (including those upon built-in hardware features such as camera and microphone) can also be documented within the metadata file, though developers may not always utilise this feature. The handheld hardware market for Android is more varied than Apple and as a result the broader scope of hardware options means that some apps may not have full functionality (or even work) on all Android devices, which may explain the difficulty we had in getting one particular app to work properly. The dependency between apps and supplementary peripherals such as the iPad Pencil needs further research, but was not an issue for the content considered during this assessment.

The Google and Apple validation services, a pre-requisite for publishing apps via the respective app stores, should theoretically mean that all content types packaged within the app can be rendered on a subset of Android and Apple environments contemporary to the app's publication, as identified in the app metadata. Whilst all apps in our assessment were downloaded from the official app stores, some Android apps can also be downloaded from 'unknown' or 'unauthorized' sources. These may not have been through this validation check.

IV. Availability of Preservation tools

A. *Validation and identification*

Both APK and IPA files can be identified by Apache Tika. Neither currently have PRONOM entries so cannot be identified by tools reliant on the PRONOM database. APK files typically use the mime type application/vnd.android.package-archive, whilst IPA files might use variations of the MIME type application/zip, application/octet-stream, or a purposefully defined MIME type. The depth of validation applied by Google and Apple in their validation service (described above) has not been reviewed from a preservation perspective so it is unclear to what depth (i.e. individual files?) this may apply.

B. *Metadata extraction*

XML metadata is held in both IPA and APK containers. Our sampling indicated that much of it is human-readable so it could theoretically be extracted and used within preservation workflows without significant difficulties, assuming sufficient technical skills are available to write the requisite parser.

[1] As outlined at https://en.wikipedia.org/wiki/Android_version_history

[2] As outlined at https://en.wikipedia.org/wiki/IOS_version_history

C. *Migration and Emulation*

The research identified no tools that could easily unpackage and migrate IPA or APK apps for use in alternative environments. App updates that are undertaken to maintain compatibility with new versions of operating systems are typically done directly by the developer or publisher. These are often based on results from the compatibility testing function within the Software Development Kit (SDK). It is not possible for the standard user to 'update'apps in this way, though users can download new versions from the app store once the developer has released an update.

Both Android and Apple SDK's provide emulator/simulator software designed for accessing app content on a PC for development and testing purposes. The Android emulator, for example, allows developers to test apps on a variety of device configurations (e.g. screen resolutions and display sizes) and Android API levels without needing to have each physical device. The emulator provides many of the capabilities of a real Android device and can be installed via the SDK tools tab of the SDK Manager. Many other emulators are also available that would allow Android apps to be run on a PC, such as that from the Android x-86 open source project[1]. Initial experiences of using this emulator to access apps at the Koninklijke Bibliotheek in 2014 indicated that whilst it had potential, more work was required to understand its limitations. [22] Questions are also yet to be explored regarding integration of emulator software into a reading room environment and user experiences.

The iOS SDK simulator is used for development and testing purposes. This could in theory be used to deliver content on a PC but would require permission from the publisher and extensive user testing. There would also be challenges in enabling the simulator independently of the full SDK environment. The iOS operating system is closed and highly proprietary, limiting the availability of alternative emulator sources.

V. Discussion

The technical dependencies identified above, and the limited availability of preservation tools, make preserving this type of content a challenge. Furthermore, the rate at which app operating systems are updated is significantly more frequent than the rendering environments used to provide access to our other digital collection content. This has clear implications for the frequency of preservation planning activities and associated resource requirements. Any preservation solution for this content type would need emulators for all versions of the operating systems upon which a given collection of apps is dependent.

The limitations of the current generation of emulators must also be considered, particularly if they are to be deployed in a reading room environment utilising desktop computers rather than handheld devices. Emulators such as those used within Software Development Kits typically allow the developer to develop and run programmes within the emulator that generate 'mock' device functionality such as GPS or a gyroscope. Whilst this is sufficient for testing purposes, it would not work as a functional GPS or gyroscope were the emulator be deployed in a live environment. Features of an app which relied on these components would therefore not function correctly. Several of the apps we reviewed also had a dependency on built-in hardware components such as the camera and microphone, and most used the device speakers to deliver audio content. These components are more likely to be supported by a desktop machine so it could be possible to tap directly into that functionality, though it is likely this would need further development – alternatively we could take a parsimonious approach and wait for a future state emulator to emerge from the development community, though the risks of that are clear.

We note, however, that aside from this, the problem of app dependencies on particular processor architectures is an as yet unresolved issue. As discussed by Viticci [23], iOS 11 (released in Sept 2017) *only runs on modern 64-bit devices and doesn't support legacy 32-bit apps*. Viticci even notes that after the introduction of iOS 11, 'the company [Apple] began preventing apps that hadn't been updated for 64-bit devices from being launched altogether, thus rendering 32-bit apps that were still installed on users' devices useless', and that 'the biggest problem facing App Store preservation

[1] See http://www.android-x86.org/ for more details

today is the lack of any kind of emulation for old 32-bit titles that are no longer supported on modern hardware.' The proprietary nature of iOS remains at this point an unresolved challenge that limits the viability of emulation as a preservation approach for IPA apps and which will require positive engagement with Apple to reach a satisfactory conclusion.

Setting aside the challenge of access for a moment, we should remind ourselves that an access solution without content to access is no solution at all. The first line of defence against content loss is acquisition of the object, preferably in the most robust and complete form available.

Users typically acquire apps by downloading them from the app store directly to their device. This is a satisfactory approach for the immediate usage of apps, but it would require a device to be networked in order to subsequently transfer content to the repository, and even then may need third party software to perform this task. Download in this fashion also typically introduces Digital Rights Management (DRM) to an object, which is subsequently used to limit end user access to the object – this is certainly the case for IPA apps though further research is required to clarify the situation with APK apps. Acquisition-related DRM would cause problems should the content subsequently need to be accessed by different users (as likely in a reading room environment), and certainly also over the long-term when the platform's authentication mechanism is eventually changed.

An alternative approach is to consider the app primarily as software – as previously noted, apps are after all a combination of software program and content together in a single container – and to target the acquisition towards an earlier stage in the production process. Acquisition of source code direct from publishers would be one such way to avoid issues introduced by downloading content from an app store, though it would require additional deposit of a compiler in order to generate a usable app for access by readers. Whilst in theory, multiple compiler deposits supporting different platforms could allow organizations to subsequently produce apps for different platforms, in practice the source code is typically designed for either Android or iOS: the same app running on both Android and iOS is typically two separate code bases (as they are two different languages). Apps coded specifically for cross-platform development, using meta-languages to build apps that can then be cross-compiled to both Android and Apple apps, are likely to be more suitable for this process, though each app would still require production of a compilation script for each different operating system, requiring a high degree of technical skill to produce.

The problem with this is that each representation is essentially a new Intellectual Entity[1]. If an app is considered primarily as a software program, and that program is modified to work with a new version of Android (for example) then this may introduce changes to the functionality of the app: if, for example, cross-compilation from Android to iOS is performed, an organization has essentially then created something completely new. The implications of this on the perceived authenticity of the object need to be more fully considered.

Should a collecting organization pursue this approach regardless, the acquisition of technical documentation about the app would be necessary, as noted by both the National Film and Sound Archive of Australia [24] and the Library of Congress [25]. Ultimately, a 'belt and braces' approach – i.e. acquisition of the published app, acquisition of source code, compiler(s), emulators, and associated technical documentation - would keep the most options open for different access solutions at a later date, but for many organizations this may be more than is feasible.

Even then, our content analysis indicates that preservation may still not be assured. If the app is to be acquired in the most robust and complete form possible then we must find some way to

[1] 'Intellectual Entity' is an Open Archival Information System (OAIS) term defined within PREMIS as 'A set of content that is considered a single intellectual unit for purposes of management and description: for example, a particular book, map, photograph, or database. An Intellectual Entity can include other Intellectual Entities; for example, a Web site can include a Web page; a Web page can include an image. An Intellectual Entity may have one or more digital representation'. See http://www.digitizationguidelines.gov/term.php?term=intellectualentity for more information.

deal with apps which have an inherent reliance on content hosted externally to the app. These are likely to lose their integrity over time, particularly as linkage to archived web content does not yet (if at all) appear to have become standard practice in apps. Such a problem would certainly manifest significantly in apps that require a subscription (such as the 'Tido Music' app); other apps which require an active connection to an online resource would also be affected, though the degree to which this impacted on the end user and the object integrity would be variable dependent on the importance of the content affected. Loss of access to the online resource would be a high issue in, for example, the free version of the Oxford English Dictionary app. The '8th Continent', on the other hand, which used an active internet connection only to deliver advertisements, would be less badly affected. If collecting at scale, then identifying those apps with an inherent reliance on externally hosted content could be an issue of its own. If we consider this issue alongside the limitations of working directly with downloaded content that were identified earlier in this section, collectively they suggest that direct deposit of app content from the publisher is likely to result in a more 'preservable' item than if we seek to acquire content using a harvesting approach of the sort used to gather content for web archives.

Finally, we must consider the future of the mobile eBook app market. How big is the market already, and in which direction might it evolve going forwards? A report produced for the British Library in 2017 suggested that 'the total number of UK Apps which are in effect books or are book like, is probably no more than thousands and certainly less than tens of thousands'. [26] However, this figure is a cumulative one, and several are likely to have already become unavailable or unworkable. Even during the course of this project, one of our sample apps was withdrawn and another could not be made to function properly. Informal discussions with publishers has suggested that if anything, the app market is shrinking, with several withdrawing from the app market due to the high-costs of development and maintenance, especially when compared to standard eBook titles. There is also a concern that it may be difficult to ensure mobile eBook apps comply with new EU accessibility rules, and some publishers have reported an interest in ePub4[1] as an alternative. Utilizing ePub4 would effectively move us to a scenario where content is "streamed" to mobile devices (rather than held within an app), a scenario which may become more prevalent particularly with 5G mobile services. This is a different preservation challenge again. Put simply, apps may be too costly to develop and support in all but high value/niche markets. How much then should organizations invest in the development of a solution to support this type of content?

VI. Conclusion and Recommendations

Preservation of mobile apps is a relatively under-explored subject within the field of digital preservation literature. We have sought to begin to remedy that in this paper by sharing our experiences with a sub-set of the eBook mobile app genre, our exploration of the technical environment in which this type of content exists, and the preservation issues that institutions will face should they choose to begin collecting this type of content.

Though the range of content and functionality one might find in an app is not dissimilar from the range of content to be found, for example, in a web archive or a personal digital archive deposit, mobile eBook apps are certainly far removed from a typical PDF eBook. The mobile eBook apps reviewed in this research may be considered to have more similarities with eBooks in an ePub format, particularly given that the ePub format is also a container that can mask significant 'under-the-hood' complexity, but the executable combination of both content files and computer program in an app set it apart from most of our other digital collection types. Should it therefore first and foremost be understood – and preserved - as software? The many technical dependencies identified and discussed in this paper, and the challenges these subsequently pose for collecting institutions, would suggest that this would be a logical approach. Unlike mobile apps, the subject of software preservation has received significant attention in the digital preservation field over recent years, for example,

[1] ePub4 is described by EDR Lab, a W3C member involved in development and promotion of the standars as a 'Packaged Web Publication', see https://www.edrlab.org/epub/introduction-to-epub-4/ for more details.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

via the Software Heritage Foundation[1] and the Software Preservation Network[2], and it is perhaps in that direction that we must turn in order to identify a way forwards.

What then are our next steps and what recommendations can we make in order to begin preservation in the meantime? In terms of target formats for acquisition, we reach the undeniable conclusion that acquisition of the app in its packaged form (either an IPA file or an APK file) is optimal for ensuring organisations at least acquire a complete published object for preservation. Whilst this is a compressed form and would limit organisations to delivery of the app on a specific version of iOS or Android, this form should at least also include sufficient metadata about inherent technical dependencies to understand what is needed to meet them. Avoidance of embedded DRM in the apps may be best achieved by direct publisher deposit, as opposed to download, and direct deposit would further provide opportunity for ensuring that apps which are acquired are in the most complete and rigorous form possible – i.e. free from subscription or dependencies on significant content hosted remotely.

Some organizations may choose to acquire source code rather than compiled apps. In this scenario, acquisition of the uncompressed source code of the app could potentially allow more flexibility for updating and maintaining the app for use on different platforms and different versions of operating systems, though this would require access to appropriate technical skills and is likely only to be effective when apps are coded initially for cross-platform purposes. Dependent on the end access solution, apps may still need to be authenticated before they can be installed, and the limits of emulator-based access, particularly on a fixed PC, need to be carefully considered: will it ultimately provide end users with a sufficiently 'authentic' experience? If not, then should wholly alternative solutions such as recording or documentation be more seriously pursued? We recall also the observations regarding uncertain growth of this content type. It is not unfeasible to consider a scenario whereby collection of mobile eBook apps remains relatively small scale. Small scale collecting would, particularly for larger collecting organisations where this content type is only a small percentage of the overall collection, suggest that a non-standard approach to access could be envisaged, for example on dedicated and non-standard reading room terminals, reducing the need for potentially costly technical solutions and integration.

Acquisition of as many technical components as possible, including the published app, source code, compiler(s) and emulators, alongside associated technical documentation, would seem to keep the most options open for different access solutions at a later date. Engagement with software providers will be essential however in ensuring that content is acquired in its most complete form and that emulators are available for use as needed. Despite this, the functionality represented in an emulator is one step removed from actual device function, and reservations regarding the use of emulation to deliver an 'authentic' user experience still need to be more thoroughly considered. Expanding our sample and technical analysis from year one to year two of the project has been invaluable in highlighting these kinds of issues.

[1] The Software Heritage Network focuses on acquisition and preservation of publicly available source code. For more information see https://www.softwareheritage.org/#

[2] The Software PreservatioN Network is a US–based organisation established to 'advance software preservation through collective action'

Acknowledgments

Our thanks go to Ian Cooke, David Russo, Caylin Smith and Neil Wilson for their perspectives on some of the issues raised in this paper during the course of the Emerging Formats project.

References

[1] Statista, 'Percentage of households with mobile phones in the UK from 1996 to 2017', (data derived from Office for National Statistics), January 2018, https://www.statista.com/statistics/289167/mobile-phone-penetration-in-the-uk/

[2] Stuart Dredge, ' The Waste Land iPad app earns back its costs in six weeks on the App Store', article in The Guardian, 8 August 2011, https://www.theguardian.com/technology/appsblog/2011/aug/08/ipad-the-waste-land-app

[3] M. Day, M. Pennock, C. Smith, J. Jenkins, and I. Cooke, "Pres-

ervation planning for emerging formats at the British Library," in Proceedings of the 15th International Conference on Digital Preservation (iPRES 2018), Boston, MA, USA, 24-27 September 2018. https://doi.org/10.17605/OSF.IO/58PJV

[4] Ian Pears, 'Arcadia', published by Amphio and available on the Apple app store at https://itunes.apple.com/gb/app/arcadia-by-iain-pears/id1003843523?mt=8

[5] Inkle Studios, '80 days', available for multiple platforms via https://www.inklestudios.com/80days/

[6] Project Aon, 'The Kai Chronicles' (text written by Joe Dever), available from the Google Play store at https://play.google.com/store/apps/details?id=org.projectaon.kaichronicles&hl=en_GB

[7] Ben Garrett, 'The 8th Continent', available for multiple platforms from http://www.8thcontinent.co.uk/

[8] Nosy Crow's 'Snow White', available from the Apple app store at https://itunes.apple.com/us/app/snow-white-by-nosy-crow/id934867093

[9] Nosy Crow's 'Jack and the Beanstalk', available from the Apple app store at https://itunes.apple.com/us/app/jack-and-the-beanstalk-by-nosy-crow/id796383629?mt=8

[10] The Nosy Crow Rounds series, described at https://nosycrow.com/series/rounds/

[11] Nosy Crow's 'Axel Scheffler's Flip Flap Pets', available from the Apple app store at https://nosycrow.com/series/rounds/

[12] Flitwits Ltd, 'The FlitLits: The Funny Fair', available from the Apple app store at https://itunes.apple.com/gb/app/the-funny-fair/id815547838?mt=8

[13] S4C, 'Cyw a'r Wyddor', available from the Apple App store at https://itunes.apple.com/gb/app/cyw-ar-wyddor/id643054131?mt=8

[14] Amphio Limited, 'Beethoven's 9th Symphony', available from the Apple App store at https://itunes.apple.com/gb/app/beethovens-9th-symphony/id601942399?mt=8

[15] Tido (UK) Ltd, 'Tido Music: learn, play & sing', available from the Apple App store at https://itunes.apple.com/gb/app/tido-music-learn-play-sing/id1050926313?mt=8

[16] Amphio Ltd, 'Steve Reichs Clapping Music', available from the Apple App Store at https://itunes.apple.com/gb/app/steve-reichs-clapping-music/id946487211?mt=8

[17] Oxford University Press / Mobi Systems, 'Oxford Dictionaries Quick Search' app, available from the Google Play store at https://play.google.com/store/apps/details?id=com.mobisystems.msdict.embedded.wireless.oxford.dictionaryofenglish&hl=en_GB

[18] GPC 'Welsh Dictionary' app, available for multiple platforms from http://www.welsh-dictionary.ac.uk/android-ios-apps/

[19] Penguin Random House, 'American Interior', available from the Apple app store at https://itunes.apple.com/bw/app/american-interior/id857802114?mt=8

[20] Faber & TouchPress, 'The 39 Steps' app, available from the Apple Mac store at https://itunes.apple.com/tz/app/the-39-steps/id999827755?mt=12

[21] TouchPress Inc, 'Shakespeare's Sonnets' app, available from the Apple App store at https://itunes.apple.com/gb/app/shakespeares-sonnets/id528646395?mt=8

[22] Johan van der Knijf, 'Running Android apps on a PC: first impressions', published on the OPF blog, October 2014, https://openpreservation.org/blog/2014/10/23/running-archived-android-apps-pc-first-impressions/

[23] Federico Viticci, 'App Preservation: Saving the App Store's History', July 2018, https://www.macstories.net/stories/app-preservation-saving-the-app-stores-history/

[24] S. Langley, T. Carter, M. Davies, and I. Gilmour, "Managing multi-platform materials: selected case studies," in Proceedings of the 19th International Symposium on Electronic Art (ISEA2013), Sydney, Australia, 2013. https://ses.library.usyd.edu.au/handle/2123/9706

[25] Library of Congress Recommended Formats Statement 2018 – 2019, https://www.loc.gov/preservation/resources/rfs/

[26] J. Ackroyd, N. Canty, A. Watkinson, 'Research to support the British Library's work on 'Emerging Formats', CIBER internal report to the British Library, March 2017.

Eye on CoreTrustSeal

Recommendations for Criterion R0 from Digital Preservation and Research Data Management Perspectives

Michelle Lindlar
TIB-Leibniz Information Centre for Science and Technology
Germany
michelle.lindlar@tib.eu
0000-0003-3709-5608

Pia Rudnik
ZZB MED-Information Centre for Life Sciences
Germany
rudnik@zbmed.de
0000-0003-4081-9646

Abstract – The CoreTrustSeal (CTS) is an accepted trustworthy digital Repository certification process for both, research data management and digital preservation communities alike. But does it build on concepts known and understood by both of these communities? We take an in-depth look at the CTS requirement Ro-Background Information/Context, in which the applicants are asked to define their Repository Type, Designated Community and Level of Curation Performed. By extracting information from the publicly available assessment reports and cross-checking these against available supporting information, we relect on CTS from three viewpoints: the process, the institutional, and the community view. We distill concrete recommendations, which will be fed back to the CTS Board as part of the 2019 public call for review.

Keywords – trustworthy Repository certification, standardization, community watch, CoreTrustSeal

Collaboration: a Necessity, an Opportunity or a Luxury?; Exploring New Horizons

I. Introduction

Within the digital preservation (DP) and research data management (RDM) communities, institutions partake in certification for different reasons: e.g., to assure that best practice is followed; to identify gaps and improve processes; or to benchmark against other certified repositories. As certification processes like *CoreTrustSeal* (CTS) are self-assessment based, they can be understood as a form of institutional self-reflection. Publicly available assessment reports form a valuable documentation of this self-reflection process. CTS can be looked at as a collaborative product of and for both the DP and the RDM communities. But do both understand the underlying process in the same way and are the results of equal use to them? In this paper, we take a critical look at the CTS process via the available supporting information, analyzing understandability and usability. Identified gaps and ambiguities may be of use to the CTS Board to improve guidance documentation, thus potentially achieving broader community acceptance in the future as the current de-facto standard CTS is on its way to be-come an accredited European Technical Standard [1]. Our analysis is limited to CTS criterion R0-Background Information/Context and based on extracted information from 40 publicly available assessment reports and available Supporting Information. The background information requested in *R0 includes descriptions of Repository Type, Designated Community, Level of Curation Performed, Outsource Partners* and *Other Relevant Information*. R0 is undoubtedly of high importance as the context information gathered within serves as a central characterization of the *Repository*. However, unlike R1-R16, R0 does not include a self-identified compliance level, making it unclear if any objective review criteria are applied.

CTS guidance documentation describes R0 checklist options as "not considered to be comprehensive" and states that comments "may be used to refine such lists in the future" [2]. This call for recommendations is one of the main motivations for our "Eye on *CoreTrustSeal*". In taking the viewpoint of different actors of certification, this paper addresses the

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

conference topic *Exploring New Horizons*, in particular the sub-theme of community self-reflection and healthy critique regarding new or improved digital preservation standards/approaches, as well as the topic *Collaboration: a Necessity, an Opportunity or a Luxury,* in particular the reciprocal support of Research Data Management and Cultural Heritage domains.

II. Related work

CTS is considered the basic certification level in a tiered certification landscape which is followed by extended level certification via DIN 31644/nestor Seal and ISO 163637 based formal certification. CTS is the successor to the Data Seal of Approval (DSA). For a detailed introduction into CTS history, other certification processes and the European Framework for Audit and Certification of Digital Repositories we point to available literature which discusses this in detail [3],[4],[5].

As a basic or core level certification, the CTS process provides a "minimum set" of requirements, distilled from other national and international guidelines, such as DRAMBORA, the Trustworthy Repositories Audit & Certification (TRAC) Checklist and the nestor Catalogue of Criteria for Trusted Digital Repositories. Dillo and de Leuw describe CTS to be "perceived as a less 'threatening', detailed and time-consuming procedure than [...] ISO or TRAC" estimating self-assessment to take "two to four days" [1]. However, a recent study put forth that the required time for basic certification self-assessment is often underestimated [4].

CTS "emphasizes the need to conform towards the OAIS standards" [1]. In explicitly referring to ISO 14721 and reusing its terminology, CTS underlines its place in a trustworthy digital *Repository* landscape. However, this could lead to the impression that both RDM and DP communities have an equal understanding of OAIS. Using the example of the term "data", Corrado points out that this is not always the case. While CTS leverages the OAIS definition of the term-meaning any form of digital object including documents as well as images, datasets or anything else-it is, in fact, frequently misunderstood as "research data" only [6].

Regarding R0, it is notable that most literature only mentions 16 CTS requirements, therefore specifically excluding R0 [1],[7],[6]. One notable exception is a CTS Board Member's presentation from 2019, which explicitly lists R0 as a "Core TDR Requirement" [8], being in-line with CTS extended guidance, which states that "...all Requirements are mandatory and are equally weighted, standalone items" [2]. This is also underlined by Langley, who points out that there is no equivalent to R0 as a context requirement in other processes [9]. Hence, information extracted from R0 forms a unique data basis.

Only few examples could be found which exploit the data basis of certified institutions. A survey conducted by the Netherlands Coalition for Digital Preservation (NCDD) amongst DSA-certified repositories in 2016 mainly focuses on experiences made in regards to perceived clarity of instructions/compliance requirements/reviewers' comments as well as in regards to perceived benefits [10]. Along similar lines, Donaldson et al. conducted semi-structured interviews in 2017, asking representatives from different institutions about the perceived value of acquiring DSA certification [11]. Husen et al. compiled a list of DSA, ICSU-WDS, nestor, TRAC and ISO 16363/TDR certified repositories and the respective process' requirements comparing them against repositories recommended by stakeholders [12]. While all of these examples exploit available data of certified repositories, none of them looked at publicly available assessment reports as a data basis for comparison. Furthermore, none of the examples specifically dealt with the CTS, most likely due to its rather young nature dating back to the end of 2017.

III. Methodology

The Data Basis for our analysis are responses given to R0 as documented in 40 publicly available assessment reports. Within sections A and B, we briefly describe our data gathering, cleansing, and normalization. Section C gives a short overview of different explanatory materials provided by the CTS, which we refer to as Supporting Information. The different viewpoints from which we discuss the Data Basis and the Supporting Information are introduced in section D.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

A. *Data Basis-Extracted information*

Each report contains an "Assessment Information" cover page, where general information such as *Name of Institution, Country, Guideline Versions Applied, Year Achieved and Previously Acquired Seal are listed.* While we extracted that information as a general overview which is contained within the Data Basis [13], the main focus was on information from R0.

In R0, the applicants are asked to describe their *Repository* by selecting 1-n applicable *Repository Type* (list-choice incl. optional comment section), to give a *Brief Description of the Repository's Designated Community* (text box), and to select 1-n *Levels of Curation Performed* (list-choice incl. optional comment section). Furthermore, applicants can add *Other Relevant Information* (text box) as well as information about *Outsource Partners* (text box) they are cooperating with.

Our Data Basis [13] consists of data extracted from the R0 section of 40 CTS Assessment reports that were publicly available on January 15th 2019[1] The following information was gathered from each reports' R0 section:

- *Repository Type* incl. list choice answers and comment section.
- *Brief Description of the Repository's Designated Community*
- *Level of Curation Performed* incl. list choice answers and comment section.
- *Outsource Partners*
- *Other Relevant Information*
- *Reviewer Entry:* While Dillo and de Leuw [1] point out that peer review comments are part of the publicly available documentation, only few published Reviewers' Entries could be found in R0. As such entries may include helpful information about the review procedure, they are nevertheless part of our analysis.

B. *Data Cleansing/Normalization*

The extracted information was, where possible, normalized to allow for a quantitative analysis. Normalization could only be achieved for *Repository Type and Level of Curation Performed* due to underlying checklists. Since normalization was not possible for the widely differing descriptions of the repositories' Designated Communities, these were exemplary mapped against definitionsand examples provided by the CTS within the Supporting Information. The results shall highlight whether the applicants share an understanding of *Designated Community* and whether the assistance given by the CTS serves as a basis for this understanding, leading to comparable descriptions. We did not analyze extracted texts from *Outsource Partners and Other Relevant Information* further, as not every institution provided information, and as where given, it could neither be normalized, nor mapped against existing concepts. Due to this we do not consider *Outsource Partners and Other Relevant Information* to be as meaningful as the other categories.

Figure 1: Three Actors of Certification

C. *CoreTrustSeal Supporting Information*

There are three main sources of explanatory materials provided by CTS to which we collectively refer to as Supporting Information. As opposed to the CTS Board, which states that Supporting Information is "[...] primarily intended for giving reviewers guidance" [2], we primarily look at its benefit to the applicants' understanding of the requirements. The first Supporting Information consists of brief guidance on questions as well as checklists for *Repository Type and Level of Curation Performed* within the requirements themselves. The second piece of Supporting Information is the *Core Trustworthy Data Repositories Extended Guidance* [2], which includes general information on the process and more detailed information for each requirement. The third piece of Supporting Information is the *Core Trustworthy Data Repositories Requirements: Glossary* [14], which includes definitions for the terminology used in the CTS.

[1] As of March 15th, the number of certi1ed repositories has reached 49, however, assessment reports made available after January 15th 2019 could not be considered.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Regarding R0, we focused on Supporting Information about *Repository Type, Designated Community and Level of Curation Performed* to see if it can lead to an unambiguous understanding of these concepts. We examine whether the information given is easy to understand, suZcient and compliant to existing standards. In a second step, we map our Data Basis against the Supporting Information to infer if the applicants followed them and gave compliant answers, or if the assistance was misunderstood.

D. *Three Actors of Certiication*

Trustworthy *Repository* certification is shaped by three actors with their own views on the process: the Institution who undergoes the certification process, the specific certification standard itself, in our case *CoreTrustSeal*, personified by the board which signs responsible for the standard as well as the reviewers, and the Community which the institution is part of and who accepts the respective process as a (de-facto) standard (see Fig. 1).

Our discussion follows the actors shown in Fig. 1, while acknowledging that their boundaries are often transparent. From the institutional and the procedural viewpoint, we check if institutions are able to describe themselves against the given criteria within the requirement and if reviewers are able to judge these descriptions. This allows us to reflect on unclarities and different interpretations of requirements and to formulate suggestions for improving the CTS. While the CTS is an accepted de-facto standard in both, the RDM and the DP community, we specifically take the DP community viewpoint to check if CTS allows for a comparison against good practice.

IV. Findings

While general information about repositories is not the main scope of this paper, one aspect shall be mentioned as it might help to contextualize the answers: a comparison of available *Previously Acquired Seal* information put forth only 27.5% (11) were renewing certification from a previously held DSA. Many institutions therefore had no previous experience with basic level certification.

A. *Repository Type*

Repository Type is supported by a checklist which the applicant can choose all matching options from. List choices given for *Repository Type* within the Requirement, augmented here by their *Glossary* definition [14], are:

- *Domain or subject-based Repository*: Specializes in a specific research field or data type
- *Institutional Repository*: Generic, multi-subject *Repository* serving a research performing institution.
- *National Repository system, including governmental:* Multidisciplinary, national infrastructure. Has a legal mandate for certain (public or governmental) data types.
- *Publication Repository:* Generic, multidisciplinary *Repository*, focussing on data linked to publications.
- *Library/Museum/Archives:* <No definition given in *Glossary*>
- *Research project Repository*: Capture research results that require a deposit mandate by a funder or organization as a 'record of science'. Often tied to a specific (multi)disciplinary project
- *Other (Please describe)*

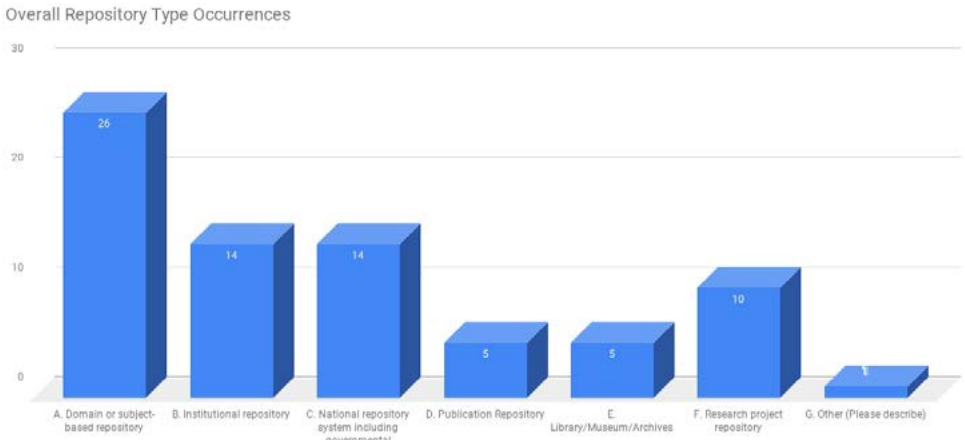


Figure 2: Overall *Repository Type* Occurrences

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The included *Repository Type*s are based on work by Armbruster and Romary [15] and were further developed by the Data Seal of Approval-World Data System partnership working group. CTS adapted this and added *Library/Museums/Archives* as well as *TOther (Please describe)* [16] to the list, the latter with regard to aforementioned possible refinements in the future.

Our analysis shows that while Armbruster and Rosary use *Publication Repository* as a super-classification of other repositories, CTS moved it to the level of a specific type of *Repository*, thereby changing the definition of *Publication Repository* from a rather content-type generic one to one "focusing on data linked to publications" [15],[14]. Surprisingly, all *Repository Type*s with the exception of *Library/Museum/Archive have a Glossary* entry-here, only the OAIS definition of *Archive* can be found: "An organization that intends to preserve information for access and use by a *Designated Community*" [14]. The *Brief Description of Repository* text field serves several purposes: for chosen option Other further elaboration seems to be expected in text field, additionally, the requirement text itself points out that the field may be optionally used if the applicant wants to provide further information to contextualize the *Repository Type* for the reviewer. In contrast, the Extended Guidance points out that a description should be given if more than one type is chosen, providing details on how these roles are fulfilled [2]-a pointer easily missed by applicants who do not carefully read through all Supporting Information.

The analysis of the Data Basis put forth 4 different types of answers for *Repository Type*. 20% (8) answered using just the list choices available, 52.5% (21) used list choices and added additional comments, 25% (10) did not explicitly state list choices first, but instead answered in prose form including keywords that allowed a mapping to list choices. 2.5% (1) did not answer the question at all. In a first step, we normalized the answers by mapping prose answers to list choices to allow for a quantitative comparison of assigned *Repository Type*s.

While *Repository Type* offers a multiple choice option, 55% (22) of the assessed institutions identified against one *Repository Type*, 15% (6) against 2, 12.5% (5) against 3, 10% (4) against 4 and 5% (2) against 5 *Repository Type*s. An interesting finding is that for institutions with multiple *Repository Type*s, all but two combinations included the option *Domain or subject-based Repository*. Overall, only 35% (14) of the certified repositories did not identify as a *Domain or subject-based Repository*. Fig. 2 shows the overall occurrences of *Repository Type*s in the Data Basis. Two reasons come to mind for the comparatively low representation of *Publication Repository* and *Library/Museum/Archives:* either, the actual number of assessed institutions do not identify against these typesorthese two options were not as comprehensible as others and therefore not chosen as often. The latter is supported by the fact that especially the option *Library/Museum/Archive* is not, like all other choices, a classification for repositories, but rather an organizational one. Further insights into the institutional view and why different *Repository Type*s were chosen shall be gained by an analysis of comments given in 31 of 40 assessment reports. Three overarching themes could be identified in those comments, in some cases more than one were mentioned. Distribution of comment types across occurrences are as follows:

- Further describing domain content (11)
- Further describing domain content & mission/history (4)
- Further describing domain content & services provided (4)
- Further describing domain content & mission history & services provided (12)

While Supporting Information asks for additional information when more than one *Repository Type* is chosen, two institutions chose several options without describing the roles further. Another 3 applicants used the comments to specifically refine *Repository Type* roles. An example for this is the following answer by the Finish Social Data Archive (FSD) [13]: "Domain or subject based *Repository*: A domain-based *Repository* with focus on research data from social sciences; *National Repository system, including governmental:* A national service resource for research and teaching; *Library/Museum/Archives:* Social science data archive".

The FSD serves as an epitome for what we see as a major point of confusion regarding *Repository Type*. It described itself against three different types,

thankfully providing further input on why these specific types were chosen. The FSD's additional information can be interpreted as follows: *Domain or subject-based Repository* describes the *Repository*'s depth (intellectual content focus), *National Repository system* describes the *Repository*'s width (of audience), and *Library/Museum/Archive* describes a function the *Repository* performs, namely archiving. While a description against such a 3-dimensional matrix (depth/width/function) would be highly beneficial and allow an adequate comparison of repositories against each other, it actually goes against the definitions of *Repository Type*s as per CTS Supporting Information. To exemplify: FSD's answer to be simultaneously a *Domain or subject-based Repository* for research data from social sciencesanda *National Repository system* is not legitimate when taking the *Glossary* into account, as per definition such a *Repository* is multidisciplinary-whereas FSD is limited to the social science discipline. FSD is not the only example, where descriptors for domain-specific and multidisciplinary were mixed-*Domain or subject-based Repository* and *National Repository system* were assigned-sometimes in combination with even further types-by a total of 11 repositories.

In a second step, we searched the extracted comments for the keywords "preserv*", "archiv*", "long-term" and "OAIS". A total of 11 assessment reports included these keywords within comments to *Repository Type*, ranging from generic statements such as providing "the opportunity for individuals and organizations to deposit collections that are considered worthwhile preserving for future generations and research projects" to specific statements such as being an "OAIS compliant web-based platform". Out of these 11 repositories, only two classified their *Repository* as being of type *Library/Museum/Archive*s, despite the fact that they are fulfilling archiving/preservation functions, according to their comments.

From a DP community point of view, the mixture of depth, width and function within *Repository Type*s creates a confusing classification schema, resulting in a diZcult to compare data basis. Additional stumbling blocks exist in non-intuitive *Glossary* definitions, as per which, e.g., a *Publication Repository* does not contain publications but rather focuses on data linked to publications. Going back to Corrado's comment regarding data as defined by OAIS, one valid reading of "data linked to publications" could then be "publications linked to publications". It remains unclear what is meant by the CTS definition of *Publication Repository*, and, furthermore, how assessed institutions understood it. Another unclarity exists on a most fundamental level: As further questions within R0 are built on OAIS concepts such as *Designated Community*, we wonder whether *Repository* is seen as a technological platform or rather understood as the *Archive* in the sense of the OAIS.

B. Designated Community

Asking the applicants to give a *Brief Description of the Repository's Designated Community* is another clear CTS reference to OAIS. OAIS defines *Designated Community* as an "identified group of potential Consumers who should be able to understand a particular set of information. The *Designated Community* may be composed of multiple user communities. A *Designated Community* is defined by the *Archive* and this definition may change over time" [17]. The *Designated Community* concept is central to OAIS, as the so-called *Representation Information* provided by an *Archive* is dependant on both, the *Designated Community*'s specific needs as well as its ability to understand information. Within OAIS, concrete preservation actions performed by an *Archive* are always connected to the *Designated Community* [17]. However, even within the DP community the concept of *Designated Community* is controversially discussed: its central importance within OAIS is felt to stand in stark contrast to its rather abstract and speculative nature, which leads to what is often perceived as a lack of practical use [18],[19],[20].

The question at hand is, whether CTS offers suZcient assistance to applicants, resulting in a clear understanding and adequate description of their *Designated Community*. What stands out first by looking at the Supporting Information is that it provides the OAIS definition of *Designated Community*, but initially contradicts this definition by referring to a *Repository's Designated Community*, whereas according to OAIS, only an *Archive* can have a *Designated Community* [17]. While CTS does not further explain this discrepancy, we assume that it happened unintentionally. Despite this impreciseness, CTS underlines the importance of *Designated*

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Community by explicitly referring to it in various requirements and their corresponding guidance (R4, R5, R6, R8, R10, R11, R14, R15, R16), as well as by requesting precise information from the applicants regarding the *Designated Community* [2]: "Please make sure that the response is specific—for example, 'quantitative social science researchers and instructors". The following description within CTS Supporting Information should further "help elucidate" the meaning of *Designated Community* [2]: "A clear *Designated Community* demonstrates that the applicant understands the scope, *Knowledge Base*, and methodologies—including preferred software/formats—they are targeting. […] A *Repository* with a very specific, narrow *Designated Community* might state the expected *Knowledge Base* (e.g., degree/level of understanding of statistics or genetics). A very broad *Designated Community* (e.g., the general public) would imply that the *Repository* has a wider range of contextual documentation to ensure its data can be understood by everyone".

In other words, the CTS wants the applicants to specifically define their *Designated Community* using terms like "broad" or "narrow" and concepts like "scope", "*Knowledge Base*" (degree of understanding; only applicable for a narrow *Designated Community*), "methodologies" (e.g., software/format preference) and necessary "contextual documentation" (the extent of which depends, according to CTS, on the broadness of the *Designated Community*). While the CTS seemingly expects the applicants to use these concepts, they are, at least in our opinion, not sufficiently and specifically enough introduced and defined within the Supporting Information. The OAIS concept *Knowledge Base* [17], for example, is neither included in the *Glossary* [14], nor elaborated on in any other piece of Supporting Information. Furthermore, it remains unclear whether it is the *Designated Community* itself or its "scope" that should be described as being either broad or narrow. Does a narrow *Designated Community* consist of few people, whereas a broad *Designated Community* means "many people"? And, what exactly is meant by "scope"? Is "scope" a further specification of a domain (e.g., "social sciences"), of a profession (e.g., "researchers and instructors"), or both? Finally, it remains unclear whether the term "contextual documentation" is used synonymous to the OAIS concept of *Representation Information* or if it means something else.

These ambiguities result in the following question for both the process as well as the institutional view: How could applicants who are not familiar with the terminology used in the Supporting Information meet the reviewers' requirements and give an adequate and "specific" definition of their *Designated Community*?

Our Data Basis shows a variety of Brief Descriptions of the *Repository*'s (sic!) *Designated Community*. We searched the descriptions for the keywords "*Designated Community*", "scope", "*Knowledge Base*", "method*" and "OAIS" to check if the applicants explicitly referred to the Supporting Information's terminology. In total, only 35% (14) use the term "*Designated Community*" in the prose text (we excluded the sub-headline *Brief Description of Repository's Designated Community* from this analysis). Instead, terminology such as "users" or "target community" are used within the answers. The term "scope" is mentioned by 10% (4), but not in the sense of "*Designated Community*'s scope". Concerning the *Designated Community*'s *Knowledge Base*, only the Finnish Social Science Data *Archive* (2.5%) explicitly refers to it. The term "method*" is used 4 (10%) times in total, but only 7,5% (3) matching responses actually use it to describe their community's methodologies. Finally, only 5% (2) explicitly refer to OAIS in their description of *Designated Community*. These results show that neither the terminology provided by the CTS via the Supporting Information, nor the OAIS terminology seem to be accepted or understood by the majority of CTS applicants.

As a quantitative analysis based on normalized data is not possible for the *Designated Community* responses, we now take a look at three representative examples for *Designated Community* descriptions [13]:

The Commonwealth Scientific and Industrial Research Organisation (CSIRO) gives an example of a specific description of a broad *Designated Community* according to CTS. What stands out first is that CSIRO does not describe its own, but the *Designated Community* of its *Repository*, the Data Access Portal. While this is in-line with CTS Supporting Information, it is not, as mentioned above, compliant to OAIS, which says that only an *Archive* can have a *Designated Community*. CSIRO defines the *Designated Community*

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

as "broad" and as "data users [that] may include: general public/industry specific groups such as agriculture/policymakers/students/researchers". Further, the institution underlines the broadness by listing 17 different research disciplines which are represented in its collections, e.g. "Environmental Sciences". Although not explicitly referring to the terms "scope", "*Knowledge Base*" or "methodologies", CSIRO's answer seems to be specific enough for the CTS reviewers considering the fact that no *Reviewer Entry* has been made. This is consistent with the argumentation in the Supporting Information, that it might not be possible to state the *Knowledge Base* of a broad *Designated Community*. Therefore it can be presumed that, from the CTS's viewpoint, CSIRO gives a specific description of a broad *Designated Community*. Despite the answer apparently fulfilling CTS expectations, we cannot fully agree with it, due to the above-mentioned inconsistency to OAIS.

Another example of a broad community is given by the Norwegian Centre for Research Data (NSD). NDS's response is accompanied by a *Reviewer Entry*. This is an important aspect for our analysis, as the entry represents the CTS's point of view and gives us an impression of what CTS considers an unspecific answer. What initially stands out is that the NSD does not refer to a "*Designated Community*", but to a "target group". This leads to the impression that the NSD is not aware of the prospective and speculative nature of the concept of *Designated Community*, and therefore alternatively presents its current target group. This group is defined as "research community" represented by "Researchers and students/ Research institutions/Finance providers/National authorities". The NSD further describes "Norwegian social science research" as its main scientific focus, but also refers to its "competence to several fields of research" and its "multiand interdisciplinary mandate". Though the NSD points out that it is not limited to Norwegian social sciences, it does not, in contrast to CSIRO, name concrete alternative scientific disciplines. This impreciseness is criticised in the *Reviewer Entry*: "It may be helpful to add to the *Designated Community* the domain specific scope ('the social sciences, humanities and medicine and the health sciences' like listed in the Strategic Plan)". Not only does this feedback clarify the meaning of "scope" as domain-specific (not profession-specific), it also demonstrates that the scope is apparently important to CTS reviewers, whereas the above-mentioned contradiction to OAIS by describing actual, not potential consumers, is not mentioned at all. This leads to the impression that in this case the CTS reviewer was not aware of this important aspect within the OAIS concept of *Designated Community*. Alternatively, it is possible that the CTS is aware of the discrepancy, but has agreed to ignore it per se without having described this approach in the Supporting Information.

We conclude the *Designated Community* analysis with an example of a speci1c description of a narrow *Designated Community* according to CTS. Due to the specialised sci-enti1c focus of the Institute for Computational Linguistics "Antonio Zampolli", its *Designated Community* can be re-garded as narrow. This is supported by the lack of a *Reviewer Entry*, leading to the assumption that the CTS re-garded this description as speci1c. Generally stating that its *Designated Community* "is constituted by the scholars of disciplines where language plays a central role", the Institute Zampolli continues more concretely by explic-itly naming "producers" and "consumers" as representa-tives of its *Designated Community*. The institute includes both, their domain-speci1c (e.g. "Humanities and the Cul-tural Social Sciences") as well as their professional scope (e.g. "Computational Linguists, Information and Commu-nication Technologies (ICT) experts and Language Engi-neers") within the description. Furthermore, examples of methodologies that are used by its Designated Com-munity are included: Producers "produce language data and digital tools to work with such data", whereas con-sumers "are interested in analyzing language data and using text processing tools available in the CLARIN infras-tructure." According to CTS Supporting Information, the narrowness of its *Designated Community* should allow the Institute "Zampolli" to state the community's *Knowledge Base*. From our viewpoint, the above-mentioned detailed descriptions can be seen as an implicit description of *Knowledge Base*, even though the term itself is not men-tioned. Apparently, CTS reviewers agree and regard the Institute's response as compliant to the CTS process. But, from a community viewpoint, does the response correspond to OAIS? At least, the Institute seems to be famil-iar with this digital preservation standard: "We ensure long term preservation [...] according to the

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

de1nition of Preservation Description Information (PDI) given in the OAIS reference model". Taking this into considera-tion, it is even more striking that, like CTS and CSIRO, the Institute Zampolli also contradicts OAIS by mistaking *Repository* for Archive: "The aim of a CLARIN *Repository* is to preserve research data sets and make them available for a *Designated Community*".

C. *Level of Curation Performed*

The second R0 question which is supported by a checklist is *Level of Curation Performed*. Like *Repository Type, Level of Curation* is also a multiple-choice option, asking the applicant to select any combination of the four offered choices. An optional comment section exists, allowing the applicant to provide further details on the *Levels of Curation* chosen. List choices provided for *Level of Curation* can be seen in Table 1. The different *Levels of Curation* were developed by the RDA-WDS Cost Recovery Interest Group and used in a survey of over twenty data centers to understand different income streams and cost recovery models in place [21].

Supporting Information clarifies that *Curation* actions are to take place on copies of the data, while original deposits should remain unchanged and should be in-line with producer-*Repository* license agreements. Additionally, CTS expects any edits made to the data to be conducted and documented by respective (domain) experts, recommending to further describe this in the comment section to allow for a better understanding by reviewers. Especially in case of Level D, "reviewers will expect a higher level of formal provenance, integrity, and version management (change logs etc.)" [2]. The same type of answer classification previously used for *Repository Type* was applied for the following data analysis. For *Level of Curation Performed*, 22.5% (9) answered using just the list choices available, 55% (22) used list choices and additional comments, 20% (8) did not explicitly state list choices, instead answering in prose form included keywords that allowed a mapping to list choices. 2.5% (1) gave an answer which could not be mapped to a *Level of Curation Performed* option. Thus, all except 1 answers were matched to the *Level of Curation Performed* choices to allow for a quantitative comparison. Despite the fact that multiple answers are possible, 55% (22) identified against only one specific Level of Curation. Here, it is interesting to note that the most frequent stand-alone choice was D. Data-level *Curation* (11), whereas no institution identified against just A. Content distributed as is.

Table 1: *Level of Curation Performed* with Answer Fre-quency

Level of Curation	Answer Frequency
A. Content distributed as deposited	9
B. Basic curation-e.g., brief checking, addition of basic metadata or documentation	22
C. Enhanced curation-e.g., conversion to new formats, enhancement of documentatio	17
D. Data-level curation-as in C above, but with additional editing of deposited data for accuracy	20
	7

It can therefore be said that every assessed institution performs at least a basic *Level of Curation* on some of its data. 42.5% (17) of the institutions chose 2-3 different *Curation* Levels, 2.5% (1) stated that all four options are applied to some (sub-)sets of data. Table 1 shows the overall occurrences of *Level of Curation Performed* in the Data Basis.

Further insights into the assessed institutions' understanding of the different *Curation Levels* were sought via the 31 comments provided. Here, two major themes could be identified: additional information on the process and additional information on applicability of different levels chosen. 14 comments included further refinements of applicability of levels chosen, e.g., pointing out that Data-level *Curation* will only take place when additional funding is available or that a lower level applies to legacy data. It seems that the commenting institutions were unsure how to answer if different *Curation Levels* apply. A full understanding of answers is therefore only possible when taking additional comments into consideration.

One example could be identified, where an institution chose one *Level of Curation (D. Data-level*

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

curation), but commented that only some data is checked against linked data for accuracy-a fact which furthermore underlines the need for an additional qualifier for *Curation* frequency. It can be assumed that the institution understood the question as asking for the highest *Level of Curation* applied, therefore omitting the level which applies to the rest of its holdings. As a consequence, the quantitative data analysis as shown in Table 1 does not necessarily provide a basis to benchmark against: it allows us to know how many institutions perform specific levels, but it does not allow us to derive if this is done as a standard procedure for the majority of data or as an exception for sub-sets under specific circumstances only.

Another question is whether the assessed institutions actually had a shared understanding of the different level choices available-we approach answering this question by analyzing the different comments given. Taking further comments for *Level D* as an example, process descriptions range from "Metadata on variable level is displayed in the online catalogue" over "include e.g. creation of new formats, enhancement of documentation and also additional editing of deposited data for accuracy" to "[...] staff review all incoming data files and apply specialized *Curation* activities such as quality assurance, file integrity checks, documentation review, metadata creation for discoverability, and file transformations into archival formats. We work closely with authors to ensure that the data is in a format and structure that best facilitates long-term access, discovery, and reuse" [13]. Based on the wide range of different process descriptions, it appears that the definition of the different levels is not clear-something that was already noted during a testbed run of the CTS process in 2016, where regarding *Level D* the point was raised that "data accuracy" may be misunderstood as pertaining to the quality of research / intellectual data accuracy, when instead a "technical quality in terms of good data" including metadata is meant [16]. Unfortunately, it seems that definitions of the different levels continue to be unclear to institutions going through self-assessment.

Of high interest from a community point of view is the fact that a definition of *Curation* is currently missing in the *Glossary*; this was already pointed out during the testbed run, where the following definition was suggested: "Activities required to make deposited data preservable or usable now and in the future. Depending on technological changes, *Curation* may be required at certain points in time throughout the data lifecycle" [16]. The implementation of this definition into the *Glossary* would be extremely helpful for several reasons. First off, the connection between *Curation and Preservation* is not always clear. As Lazorchak points out, "*Curation* is a useful concept for describing the evolving whole-life view of digital preservation, but concentrates on underpinning activities of building and managing collections of digital assets and so does not fully describe a more broad approach to digital materials management" [22]. Within the RDM community, typical *Curation* life-cycle activities include assigning persistent identifiers and curating metadata, whereas DP focuses on life-cycle activities such as preservation metadata creation and file-format based processes like file format characterization and migration or emulation. An analysis of the comments shows that only one institution mentions technical metadata creation, and while 4 comments mention OAIS terminology such as *Submission Information Package*, the majority of comments seem uniformed of DP practice.

The fact that even within the RDM community, uncertainty regarding *Curation* terminology exists, can be seen in yet another testbed comment. Initially, the example given for *Level C. Enhanced curation* was "e.g., creation of new formats", leading to the question whether this describes a (pre-)ingest/normalization process or a migration process. In order to make the answer clearer, the CTS Board decided to change this to "conversion to new formats", hoping to make the intended meaning clearer [16]. At least in our view, the meaning remains as unclear as it was before-only if read in conjunction with the (currently missing) *Glossary* definition is it understood to describe migration rather than normalization. Furthermore, the term *Levels of Curation* reminds one of *Levels of Digital Preservation* [23], a well-accepted framework within the DP community. However, neither can the concepts included in *Level of Curation Performed* be mapped to Levels of Digital Preservation, nor are they as concise and exhaustive as the latter.

V. Conclusion

Based on the available Supporting Information and our Data Basis, we have presented an in-depth analysis of the R0 criteria *Repository Type, Designated Community and Level of Curation Performed* from a procedural, institutional and community point of view. This analysis has put forth a number of weaknesses within the process, which resulted in ambiguous assessment report answers. While one of the benefits of trustworthy *Repository* certification is the ability to benchmark repositories against good practice as well as against other repositories, a data basis which is open to interpretation due to unclear results prohibits such benchmarking. Furthermore, we have taken a close look at the process from a DP community point of view, pointing out where CTS descriptions go against concepts which are standardized and accepted within the community. This final section presents concrete recommendations for the improvement of the CTS, which we derive from our findings. They are grouped by the respective R0 section and are preceded by a number of generic process recommendations. All recommendations will be passed on to the CTS Board as part of the current review of *CoreTrustSeal* requirements.

A. *Generic Recommendations for R0*

1. Heighten Relevance of R0 Assessment: Despite the importance of context information, R0 seems to be treated in a rather irrelevant manner by both reviewers and applicants alike, possibly due to the fact that there is no associated compliance level. While we acknowledge that compliance level are not necessarily applicable to R0, lacking information or incomplete answers should in our view be of consequence and not be disregarded during the review process.

2. Standardize Assessment Reports: Different templates for assessment reports exist. Some institutions have used the online form and others provided the self-assessment in a document form. These different handing-in procedures can lead to different forms of answers, where some institutions do not stick to the given list choices, instead writing prose answers which are hard to map and also hard to judge by reviewers. This should be standardized with clear routines checking for mandatory information, e.g. at least one checkbox ticked for *Level of Curation Performed*.

3. Introduce OAIS in Supporting Information: Our analysis has shown that many applicants seem neither familiar with the OAIS in general, nor with its core concepts like *Designated Community* in particular. This often leads to answers that are not compliant to the standard, not easy to compare, and diZcult to assess by reviewers. We therefore recommend an implementation of a short introduction to OAIS in the Supporting Information.

4. Include concise Definitions and Examples: To avoid unclarity concerning terminology used in R0, we recommend that clear definitions and concrete examples are added to the Supporting Information. This applies to the list choices for *Repository Type and Level of Curation Performed* as well as to terms like "scope", "methodologies", "contextual documentation", and "*Knowledge Base*" used in the *Designated Community* section. Including definitions and examples directly within the requirements might significantly improve the applicants' understanding of R0 and omit the need to query different pieces of information.

B. *Recommendations for Repository Type Section*

1. Replace mixture of depth, width and function with 3-level approach: Instead of mixing different levels within a *Repository Type* list choice, institutions should describe themselves against the three levels separately: Depth/Content-ranking from domain-specific to multidisciplinary; Width/Audience-ranking from project specific via institutional to national; Function-determining whether archiving is included or not.
2. Define *Repository*'s Boundaries: It is not clearly defined if *Repository* within CTS pertains to the technological concept or to an *Archive* in the OAIS sense. This definition should be clearly stated by the CTS, especially as further R0 questions and concepts such as *Designated Community* build on the OAIS understanding of Archive.

C. *Recommendations for Designated Community Section*

1. Change sub-headline "*Brief Description of Repository's Designated Community*": Given the fact that the expression "*Repository's Designated Community*" is not compliant to OAIS, the term *Repository* should either be

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

deleted or replaced by the term Archive. This would avoid one inconsistency to OAIS which we observed in our analysis.

2. Stimulate formalized Descriptions of *Designated Community*: A formalized way to describe *Designated Community* would be helpful. It would lead to a better understanding of the concept itself, stimulate self-reflection and result in comparable answers. The inter-dependency of the concepts "scope", "methodologies" and "*Knowledge Base*" need to be exemplified, e.g., by referring to the Digital Preservation Coalition which states: "the broader the scope of the *Designated Community*, the less specialized the knowledge associated with that community" [24]. A questionnaire, e.g., on the domain-specific and professional scope of a *Designated Community*, would be a helpful orientation.

D. Recommendations for Level of Curation Section

1. Describe Conditions for Levels Applied: Few institutions follow one *Level of Curation* for all data. Adding a tiered model, where each applicable level is described more granular, e.g. as, "applies to (a) all objects (b) sub-collections based on depositor agreement (c) subcollections based on external requirement/funding (d) sub-collections based on technical suitability" should lead to a meaningful assessment approach.
2. Include Digital Preservation-centric Model: To understand the preservation functions the *Repository* fulfills, a preservation focused model should be included in R0. This can be either the *Levels of Preservation* model, or, at the simplest level, by asking the institution which of the preservation levels bit-stream/ logical/semantic are implemented.

References

[1] I. Dillo and L. de Leeuw, "CoreTrustSeal," *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare,* vol. 71, no. 1, pp. 162–170, 2018. DOI: 10.31263/voebm.v71i1.1981.

[2] CoreTrustSeal, Core Trustworthy Data Repositories Extended Guidance. 2018, Version 1.1. [Online]. Available: https://www.*CoreTrustSeal*.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf.

[3] Digital Preservation Coaliton, *Digital Preservation Handbook.* 2015. [Online]. Available: https://www.dpconline.org/handbook.

[4] M. Lindlar and F. Schwab, "All that Work...for What? Return on Investment for Trustworthy *Archive* Certification Processes – a Case Study," in *Proceedings of the 15th International Conference of Digital Preservation*, 2018. DOI: 10.17605/OSF.IO/ 8A3SC.

[5] CESSDA, "Cessda user guide: Trusted digital repositories: Audit and certification," 2015. DOI: 10.5281/ zenodo.2595238.

[6] E. M. Corrado, "Repositories, trust, and the CoreTrustSeal," *Technical Services Quarterly,* vol. 36, no. 1, pp. 61–72, 2019. DOI: 10.1080/07317131.2018.1532055.

[7] E. Zhang and M. R. Moulton. (2018). Preparing for a Trustworthiness Assessment of the National Transportation Library's Digital *Repository* ROSA P, [Online]. Available: https :// rosap.ntl.bts.gov/view/dot/34480.

[8] R. Jenkyns, "Building trust in scientific data: Certification & the *CoreTrustSeal*," 2019. DOI: 10.5281/ zenodo.2555289.

[9] S. Langley. (2018). Digital preservation maturity model, risks, trusted digital *Repository* certification and workflow cross-walk [dataset], [Online]. Available: https://www.*Repository*.cam.ac.uk/handle/1810/287009.

[10] K.-J. Waterman and B. Sierman, *Survey of DSAcertiied digital repositories: Report on the indings in a survey of all DSA-certiied digital repositories on investments in and beneits of acquiring the Data Seal of Approval (DSA)*, English. Netherlands Coalition for Digital Preservation, Nov. 2016.

[11] D. R. Donaldson, I. Dillo, R. Downs, and S. Ramdeen, "The perceived value of acquiring data seals of approval," *International Journal of Digital Curation*, vol. 12, no. 1, pp. 130–151, 2017. DOI: 10.2218/ijdc. v12i1.481.

[12] S. Husen, Z. Wilde, A. de Waard, and H. Cousijn. (2017). Data Set-Recommended versus Certified Repositories: Mind the Gap.

[13] M. Lindlar and P. Rudnik. (2019). Eye on core trust seal-data set, [Online]. Available: https :// dx. doi.org/10.5281/zenodo.3267690.

[14] Data Seal of Approval and ICSU World Data System. (2016). *Core Trustworthy Data Repositories Requirements: Glossary*, [Online]. Available: https://drive.google.com/file/d/0B4qnUFYMgSc-REpsNVQwWDVfSkU/view.

[15] C. Armbruster and L. Romary, "Comparing *Repository Types*: Challenges and Barriers for SubjectBased Repositories, Research Repositories, National *Repository* Systems and Institutional Repositories in Serving Scholarly Communication," 2009. DOI: 10.2139/ssrn.1506905.

[16] DSA–WDS Partnership Working Group. (2016). DSA–WDS Partnership Working Group Results of Common Certification Testbed, [Online]. Available: https://www.rd-alliance.org/

system/files/DSA-WDS_Common_Certification_Testbed_Report_V0_5_Final.pdf.

[17] *CCSDS, Reference Model for an Open Archival Information System (OAIS)-Magenta Book*, CCSDS, 2012. [Online]. Available: http://public.ccsds.org/ publications/archive/650x0m2.pdf.

[18] R. S. Bettivia, "The power of imaginary users: Designated communities in the oais reference model," *Proceedings of the Association for Information Science and Technology*, vol. 53, no. 1, pp. 1–9, 2016. DOI: 10.1002/pra2.2016.14505301038.

[19] C. Keitel, "Der einzige Kompass, den wir haben: Zur Kritik der Designated Community," *Informationswissenschaft: Theorie, Methode und Praxis,* vol. 5, no. 1, 2018. DOI: 10.18755/iw.2018.4.

[20] D. Rosenthal. (2014). TRAC Audit: Lessons, [Online]. Available: https ://blog.dshr.org/2014/08/trac-audit-lessons.html.

[21] RDA-WDS Interest Group on Cost Recovery for Data Centres, "Income streams for data repositories," 2016. DOI: DOI:10.5281/zenodo.46693.

[22] B. Lazorchak. (2011). Digital Preservation, Digital Curation, Digital Stewardship: What's in (Some) Names? Library of Congress, [Online]. Available: https://blogs.loc.gov/thesignal/2011/08/digitalpreservationdigitalcurationdigital-stewardship-what % E2 % 80 % 99s-in-some-names/.

[23] M. Philips, J. Bailey, A. Goethals, and T. Owens, "The NDSA Levels of Digital Preservation: An Explanation and Uses," in *Proceedings of the Archiving* (IS&T) Conference, 2013. [Online]. Available: https://ndsa.org/documents/NDSA_Levels_Archiving_2013.pdf

[24] B. Lavoie. (2014). The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition), Digital Preservation Coalition, [Online]. Available: https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file.

Integrating Dataverse and Archivematica for Research Data Preservation

Meghan Goodchild
Scholars Portal &
Queen's University
Canada
meghan@scholarsportal.info
ORCID 0000-0001-7988-8046

Grant Hurley
Scholars Portal
Canada
grant@scholarsportal.info
ORCID 0000-0003-0172-4847

Abstract – Scholars Portal sponsored Artefactual Systems Inc. to develop the ability for the preservation processing application Archivematica to receive packages from Dataverse, a popular repository platform for uploading, curating, and accessing research data. The integration was released as part of Archivematica 1.8 in November 2018. This paper situates the integration project in the broader context of research data preservation; describes the scope and history of the project and the features and functionalities of the current release; and concludes with a discussion of the potential for future developments to meet additional use cases, service models and preservation approaches for research data.

Keywords – research data; Archivematica; workflows; Dataverse

Conference Topics – Collaboration: a Necessity, an Opportunity or a Luxury?; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

Between 2015 and 2018, Scholars Portal contracted Artefactual Systems Inc. to develop an integration between Dataverse, a popular repository platform for uploading, curating, and accessing research data, and Archivematica, an open source workflow application for creating preservation-friendly packages for long-term storage and management. Scholars Portal is the information technology service provider for members of the Ontario Council of University Libraries (OCUL), a 21-member consortium of academic libraries in the province of Ontario, Canada.1 Founded in 2002, Scholars Portal is funded by OCUL members and operated under a service agreement with the University of Toronto Libraries. Our services support both research data management via a hosted, multi-institutional instance of Dataverse2 and digital preservation services via Permafrost,3 a hosted Archivematica-based service that pairs with the OCUL Ontario Library Research Cloud (OLRC) for preservation storage.4 The Dataverse-Archivematica integration project was initially undertaken as a research initiative to explore how research data preservation aims might functionally be achieved using Dataverse and Archivematica together. The results of a proof-of-concept phase were developed into a working integration released as part of Archivematica version 1.8 in November 2018. This paper situates the integration project in the broader context of research data preservation in theory and practice; describes the scope and history of the project and the features and functionalities of the current release; and concludes with a discussion of the potential for future developments to meet additional use cases, service models and preservation approaches for research data.

II. Research Data Preservation in Context

In this paper, the term "research data" refers to a broad set of potential outputs from research activities across sectors and disciplines. The key uniting characteristic is that these materials stand as unique evidence supporting a set of research findings, whether scholarly, technical, or artistic [1].

[1] Scholars Portal: https://scholarsportal.info/.

[2] Scholars Portal's Dataverse instance: https://dataverse.scholarsportal.info/.

[3] Permafrost: https://permafrost.scholarsportal.info/.

[4] Ontario Library Research Cloud: https://cloud.scholarsportal.info/.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Furthermore, these data may constitute the research findings themselves, such as in the publication of statistical or geospatial data. The communities of stakeholders who value research findings depend on the maintenance of original data sources in a trustworthy manner that privileges ensuring their continued authenticity, availability and reliability into the future. These concepts have been codified within the sector as the FAIR Principles for research data: findable, accessible, interoperable, reusable [2]. While the FAIR Principles do not specifically cite long-term preservation as a requirement, preservation activities are crucial to the continued ability to discover, access and use research data into the future [3]. The FAIR principles therefore link to the stewardship responsibilities that repositories take on behalf of stakeholders: in order to fulfill the FAIR principles, organizations with access to sustained resources and infrastructure must commit to ensuring the long-term maintenance of the materials under their care.[1] The requirements for this maintenance are outlined in standards such as the Open Archival Information System (OAIS) reference model (ISO 14721)[2] and audit and certification frameworks including CoreTrustSeal,[3] nestor,[4] and *Audit and Certification of Trustworthy Data Repositories* (ISO 16363).[5] Repositories with stewardship responsibilities therefore seek to translate audit and certification requirements into repeatable practices to ensure that data are kept reliably into the future. A series of interrelated stages make up the lifecycle required for responsible data curation and preservation over time, including creation and receipt, appraisal and selection, preservation actions, storage, and access and discovery [4]. One tool that implements some of these stages of the lifecycle is Dataverse.[6]

Dataverse is developed and maintained as an open source project by the Institute for Quantitative Social Science (IQSS) at Harvard University since 2006 [5]. A large open Dataverse instance is hosted by IQSS and Harvard University Library.[7] Fifty individual known installations of Dataverse exist throughout the world as of the time of writing [6]. While Dataverse was developed by members of the social science community, its use is not limited to any specific disciplinary area [5]. Users can deposit and describe their data files using Data Documentation Initiative (DDI)[8] and Dublin Core-compliant standards, as well as certain discipline-specific metadata standards,[9] generate unique identifiers and data citations, and assign access permissions and license terms. Institutions can enable self-deposit or mediated workflows for depositors, and offer Dataverse to faculty members and researchers as a method of fulfilling funder requirements to deposit data in an accessible repository. Published datasets are searchable and downloadable and tabular data files can be explored using visualization tools within the platform itself.

Dataverse includes a suite of functions that contribute to the ability of a stewarding organization to reliably preserve research data. When it comes to data receipt, it enables efficient capture of materials from a depositor's individual computing systems through user-friendly upload tools, which tackles a major initial barrier of accessing data from the risky (and often inaccessible) environments of personal computers or local network storage [7]. Depositors can also describe and contextualize their submissions through a variety of metadata fields and by

[1] See also the Australian Research Data Commons' *FAIR self-assessment tool*: https://www.ands-nectar-rds.org.au/fair-tool.

[2] ISO 14721:2012 (CCSDS 650.0-M-2) *Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model*.

[3] *Core Trustworthy Data Repositories requirements*, https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf.

[4] nestor seal for trustworthy data archives: https://www.langzeitarchivierung.de/Webs/nestor/EN/Services/nestor_Siegel/nestor_siegel_node.html

[5] ISO 16363:2012 (CCSDS 652.0-R-1) *Space data and information transfer systems -- Audit and certification of trustworthy digital repositories*.

[6] Dataverse: https://dataverse.org/.

[7] Harvard Dataverse: https://dataverse.harvard.edu/.

[8] Data Documentation Initiative: https://www.ddialliance.org/.

[9] See *Dataverse user guide*, Appendix, "Metadata references," http://guides.dataverse.org/en/latest/user/appendix.html.

linking to related publications and datasets.[1] All user-submitted files receive MD5 checksums upon receipt that can enable verification of file fixity over time [8]. File format identification is also conducted as part of the Dataverse ingest workflow. Processes for file format identification include an internal service to identify tabular data files, the use of JHOVE's[2] file format validation functionality, and if these are unsuccessful, an attempt to identify based on file extension. All of these processes result in the display of the file's MIME type in the Dataverse application's database and interface [9]. The well-formedness and validity of a particular file are not recorded as an output from JHOVE.

The identification of tabular data files informs additional tabular data transformation functions. Tabular data formats (e.g., SPSS, STATA, RData, CSV, and Excel) are converted to non-proprietary tabular text data files (.tab) upon ingest, and citation-related metadata and DDI variable-level metadata are created for the tabular files [10]. Dataverse converts tabular data files as accurately as possible with the caveat that some commercial applications like SPSS have not published their specifications [11]. Tabular files also receive UNF checksums that can be used to verify the semantic content of the derivatives [12]. Users can download the data files in the original format as uploaded by the depositor, and/or in the derivative formats (tab-delimited or RData) created by Dataverse. In addition to exporting the DDI variable-level metadata as an XML file, users can also export a DDI-formatted HTML codebook for the entire dataset that also contains the variable-level metadata.

Initiatives in research data preservation, including those using Dataverse, emphasize the necessity of storing and monitoring datasets independently from the submission and discovery platforms with which users commonly interact. This approach appears to be informed by an interpretation of the OAIS reference model, which emphasizes the flow of received data as Submission Information Packages (SIPs) into stored and monitored units of content information as Archival Information Packages (AIPs) for preservation and Dissemination Information Packages (DIPs) for access. These packages may be logical rather than physical: their components may not have to be stored together so long as the total package can be retrieved and independently understood by members of the designated community [13]. Therefore, institutions could potentially use Dataverse or other repository software as an OAIS-type archive even if that software does not create and maintain physical AIPs. However, OAIS and related certification frameworks also identify in a broad sense what functions preservation systems should perform, and these features may only partially exist in a software package like Dataverse that is designed primarily for receipt, access and discovery. Creating platform-independent preservation packages means that institutions can generate and manage preservation metadata, use more than one managed method for storage, run preservation-supporting functions at ingest and over time, and audit and maintain stored packages without depending on a single system to perform all of these tasks in addition to user-facing functions.

Research on the subject of research data preservation has emphasized the desirability of storing and managing independent preservation packages. A white paper authored by members of the Canadian Association of Research Libraries (CARL)'s Portage Network Preservation Expert Group theorizes the disaggregation of OAIS-type functions among a set of potential preservation service providers who take care of particular functions such as archival storage, while communicating the results of these efforts back to a centralized administrative agency [14]. In the United Kingdom, Jisc's series of three *Filling in the Preservation Gap* reports specifically investigate the use of Archivematica in service of preserving research data.[3] A series of test implementations at the University of York and University of Hull were deemed successful and Archivematica was among the preservation providers tested with the Jisc's Research Data Shared Service pilot [15]. Therefore, Dataverse's functions primarily map to the "Producer" end of the OAIS model, where materials

[1] See also published guidance on metadata in Dataverse: *Dataverse North metadata best practices guide*, https://portagenetwork.ca/wp-content/uploads/2019/06/Metadata_V1.1_EN.pdf.

[2] JHOVE: http://jhove.openpreservation.org/.

[3] *Filling the preservation gap* project page: https://www.york.ac.uk/borthwick/projects/archivematica.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

are negotiated and accepted for ingest and some baseline preservation-supporting functions are performed. Further research is required on how platforms like Dataverse might fulfill the requirements of the Producer-Archive Interface Methodology Abstract Standard (PAIMAS)[1] and Producer-Archive Interface Specification (PAIS)[2] for structuring producer-archive interactions.

Data repositories using Dataverse are taking steps to export data and metadata from Dataverse for additional processing and/or storage, primarily as physical packages. In the Netherlands, DANS' DataverseNL service exports packages using the BagIt specification[3] to their EASY preservation repository [16]. The Qualitative Data Repository (QDR) at Syracuse University is taking a similar approach with the development of a proof-of-concept implementation of exported OAI-ORE metadata and zipped Bags from Dataverse [17]. The Odum Institute at the University of North Carolina uses scripts to push data packages to iRODS,[4] which performs preservation processing and storage replication [18]. The Dataverse software itself also includes the ability to transfer exports as Bags to DuraCloud, a hosted service for replication to cloud storage providers, as well as to the file system[5] [19].

The Dataverse-Archivematica integration takes advantage of the preservation-related actions that Dataverse performs and makes them available to an Archivematica-based workflow to create and store independent preservation packages. The scope and features of this integration are discussed in the following sections.

[1] Consultative Committee for Space Data Systems, *Producer-archive interface methodology abstract standard*. CCSDS 651.0-M-1. Magenta book, 2004. https://public.ccsds.org/Pubs/651x0m1.pdf.

[2] Consultative Committee for Space Data Systems, *Producer-archive interface specification*. CCSDS 651.1-B-1. Blue book, 2014. https://public.ccsds.org/pubs/651x1b1.pdf.

[3] *The BagIt File Packaging Format (V1.0)*, https://tools.ietf.org/html/draft-kunze-bagit-17.

[4] iRODS: https://irods.org/.

[5] Duracloud: https://duraspace.org/duracloud/.

III. History and Scope of Project

A. *Proof-of-Concept*

In response to growing community interest, Scholars Portal initiated a research project in 2015 to investigate how research datasets stored in Dataverse could be processed into AIPs using Archivematica. Initial project participants included members from Scholars Portal and the University of Toronto, Artefactual Systems, IQSS Dataverse, the University of British Columbia, the University of Alberta, Simon Fraser University, and the CARL Portage Network.

Project participants conducted an initial requirements analysis and proposed a draft workflow. Artefactual Systems developed a prototype of Archivematica that used Dataverse APIs to retrieve datasets for ingest and processing in Archivematica. The proof-of-concept integration was only available through a development branch of Archivematica and presumed an automated workflow in which all datasets in a target Dataverse would be transferred and processed by Archivematica.

The initial project provided an opportunity to explore best practices related to the preservation of research data; investigate how Dataverse handles and stores data and metadata, processes derivatives and versions files, exports data and metadata; and determine how Archivematica could accept and process Dataverse dataset packages. The project also identified the use of the DDI metadata standard within Archivematica METS files for descriptive metadata. Given DDI's capacity to comprehensively describe specific characteristics related to research data for discovery and reuse, this mapping was intended to expand the scope of descriptive metadata in Archivematica METS files and make these files more hospitable to describing research data.

B. *Production Release*

In 2018, Scholars Portal sponsored further development work with Artefactual Systems to improve the original proof-of-concept design and merge it with the public release of Archivematica in version 1.8 (developed and tested using Dataverse version 4.8.6 and above).[6] Four staff members at Scholars

[6] Archivematica 1.10 - Dataverse transfers: https://www.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Portal worked directly on the project. The authors served as project leads, including organizing meetings and managing project tasks, communicating with Artefactual, performing testing and analysis, and documenting discussions and results. Amber Leahey (Data & GIS Librarian) provided domain expertise related to research data management, and Dawas Zaidi (Systems Support Specialist) provided systems support. Alan Darnell (Director), Amaz Taufique (Assistant Director, Systems and Technical Operations), and Kate Davis (Assistant Director, Collections and Digital Preservation) provided administrative support. At Artefactual Systems, our primary contacts were Joel Simpson (Operations Manager & Solution Architect) and Ross Spencer (Software Developer). Joel led the requirements analysis process, acted as our main point of contact at Artefactual, tested iterations of the integration, and produced documentation. Ross was responsible for the majority of the software development in collaboration with colleagues at Artefactual. At Scholars Portal, the development project started in April 2018 and concluded with the release of the integration in November 2018. Our key project tasks included identifying and creating test datasets (discussed below), analyzing the outputs and identifying issues, and documenting the integration. The major result of the integration is that Archivematica can be configured to use a connected Dataverse instance as a transfer source location. Datasets are queried and retrieved using Dataverse's APIs and processed using the Dataverse transfer type, which contains specific processing micro-services (described in section IV. below).

The integration was designed with a series of assumptions in terms of its design. First, the design presumes a user has an account with a Dataverse instance and has generated an API token (a unique code for authentication). The same or a different authorized user (typically an archivist, librarian, or curator) also has access to an Archivematica instance and wishes to process certain datasets into AIPs for long-term preservation. This assumes the user has obtained the necessary rights and privileges to process and store dataset files independently from Dataverse. Secondly, the current design assumes that the user is interested in selecting specific datasets in a target Dataverse instance for preservation. This assumption conforms to specifications such as CoreTrustSeal that state that repositories must appraise and select data for preservation [20]. The current design does not include an automated function for ingest of all datasets within a Dataverse container, though we acknowledge that this functionality may meet additional use cases.

A single dataset in a Dataverse instance corresponds to a SIP. Individual files cannot be transferred from Dataverse for preservation. However, users can select individual files to be made into a final AIP by using the Appraisal function in Archivematica.[1] At present, only the current version of files and metadata can be selected for preservation, though Dataverse tracks versioning and provenance of metadata and file changes, with all versions retained by the system [21]. Finally, while users may choose to create a DIP as part of the Archivematica workflow, it is assumed that the version available to users in Dataverse will generally remain the one used for access. The scope of the integration did not include communication back with a connected Dataverse to write preservation metadata, or the replacement of user-submitted data with the DIP generated by Archivematica.[2] See section V. below for discussion of features identified for potential future development.

IV. Workflow and Functionality

Fig. 1 presents an overview of the workflow for the integration. Beforehand, an administrator of the target Archivematica installation must configure the Archivematica Storage Service to connect to a specific Dataverse instance. Then, Archivematica's transfer source locations can be set to filter based on query search terms or on a specific Dataverse container using Dataverse's Search API.

archivematica.org/en/docs/archivematica-1.10/user-manual/transfer/dataverse/#dataverse-transfers.

[1] Archivematica 1.10 - Appraisal: https://www.archivematica.org/en/docs/archivematica-1.10/user-manual/appraisal/appraisal/#appraisal.

[2] The latter is the case for the Archidora integration between Archivematica and Islandora. See T. Hutchinson, "Archidora: Integrating Archivematica and Islandora," *Code4Lib Journal* 39, https://journal.code4lib.org/articles/13150.

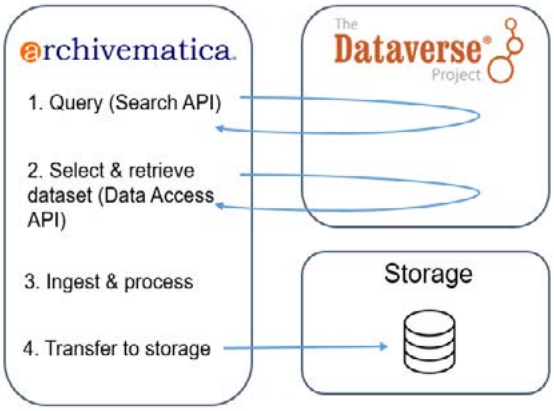


Fig. 1. Workflow for integration between Dataverse and Archivematica

To initiate a transfer, the Archivematica user sets the transfer type to "Dataverse," browses the datasets available in the Archivematica Transfer Browser, selects one dataset and starts the transfer (1). Archivematica uses Dataverse's Data Access API to retrieve a copy of the most recent version of the dataset (2). The package contains the original user-submitted data files, and if the user submitted tabular data, a set of derivatives of the original tabular files along with associated DDI variable metadata and citation metadata files describing the tabular files. Next, the Dataverse transfer type completes a set of preconfigured ingest and processing steps, including Archivematica's processing functions (3). Finally, the AIP is transferred via the Archivematica Storage Service to a connected storage system (4).[1]

Looking at the specifics of the integration, the Dataverse transfer type contains the following preconfigured ingest and processing steps:

- Creation of an initial Dataverse METS XML file describing the dataset as received from Dataverse, which includes a descriptive metadata section mapped to the DDI standard, a list of files grouped by type (original, metadata, or derivative), and a description of the structure of the files provided by Dataverse;
- Fixity checks of data files verified using the MD5 checksums that were generated by Dataverse for all user-submitted files;
- Other standard Archivematica microservices conducted as configured. These services include independent file format identification and validation, which includes mapping identified file formats against PRONOM unique identifiers. Users might also choose to apply Archivematica's preservation normalization policies.[2]

Archivematica produces a final METS and PREMIS-based XML file for the AIP (see Table 1) that copies over the descriptive metadata from the initial Dataverse METS file, outlines the relationships between original and any derivative files resulting from the tabular ingest process in Dataverse, and includes records of any actions undertaken through Archivematica's processing steps. Tabular derivatives created by Dataverse are recorded with an associated PREMIS event labeled as "derivation" in the METS file. The connected Dataverse instance's name and URI is recorded as a linked PREMIS agent in relation to the tabular derivation event.[3] Though Artefactual Systems proposed "derivation" in 2015 as part of feedback on the PREMIS controlled vocabulary for events, it has not yet been implemented in the PREMIS events controlled vocabulary.[4] Derivatives and metadata files are also identified in the METS fileGrp sections within the fileSec section.

Finally, the resulting AIP processed from Dataverse is structured in the same general format as other AIPs processed by Archivematica. As shown in Fig. 2, additional metadata files from Dataverse

[1] Storage service, *Archivematica Wiki*: https://wiki.archivematica.org/Storage_Service.

[2] Archivematica 1.10 - Preservation Planning, https://www.archivematica.org/en/docs/archivematica-1.10/user-manual/preservation/preservation-planning/#preservation-planning.

[3] This information is entered as part of the storage service setup in Archivematica and is also stored as file called agents.json in the 'metadata' folder of the AIP: https://www.archivematica.org/en/docs/storage-service-0.15/administrators/#dataverse.

[4] As Evelyn McLellan writes, "The use case is a research data publishing platform that generates tabular file format derivatives from uploaded statistical files. This is not normalization because the purpose is not preservation but rather derivation for the purpose of data manipulation and visualization." See: http://premisimplementers.pbworks.com/w/page/102413902/Preservation%20Events%20Controlled%20Vocabulary. This is as opposed to the "derivation" relationship type referred to in PREMIS s. 1.13.1.

are included, and any originally zipped folders will result in a separate directory within the AIP.

Table 1. Dataverse-Archivematica METS structure overview

METS section	Description
METS dmdSec	Descriptive metadata section in DDI
ddi:title ddi:IDno ddi:authEnty ddi:distrbtr ddi:version ddi:restrctn	DDI fields include: title, unique identifier (e.g., DOI), author(s), distributor (i.e., the Dataverse instance), dataset version, and licenses/restrictions Additional descriptive metadata from Dataverse is stored in the AIP in a file titled "dataset.json" and is referenced using an xlink in the dmdSec of the Archivematica METS file. Any DDI XML files for tabular data files are also linked in the dmdSec
METS amdSec	Administrative metadata section (for original, derivative, metadata, and normalized files)
techMD	Technical metadata (PREMIS), including file format information and extracted metadata
digiprovMD	Provenance metadata, including PREMIS events for derivation (for tabular derived files), ingestion, unpacking bundled files, virus check, fixity check, normalization, and linked PREMIS agents for associated organizations, software, and Archivematica users
METS fileSec	File section defining original files uploaded to Dataverse, derivative tabular files generated by Dataverse, metadata files generated by Dataverse, submission documentation, metadata files and normalized preservation copies generated during Archivematica processing, if applicable
METS structMap	Structural map, showing directory structure of the contents of the AIP

Fig. 2. Example Dataverse-Archivematica AIP structure [22]

V. Testing and Future Development

During the development work, nine test datasets were created in the Scholars Portal Demonstration Dataverse[1] that were representative of the types of datasets deposited in the Scholars Portal production platform. Three of these included tabular data; one included a large collection of text files and images, including zipped packages; and another consisted of geospatial data files. Three others contained Microsoft Office documents, images, and audiovisual materials, respectively. A final dataset contained only metadata (no data files). Our testing focused on ensuring the successful request and receipt of complete data packages from Dataverse to Archivematica and ensuring that an AIP with an accurate Archivematica-generated METS file was created and placed in storage. Processing and configuration details and issues specific to Archivematica, such as file format normalization, were not considered.

The main issues experienced during testing related to unsuccessful Archivematica transfers from Dataverse-originated packages. For example, transfer failures that were the result of files in Dataverse missing checksums or as a result of failed tabular data ingest processes in Dataverse.2 Testing

[1] Scholars Portal Demo Dataverse: https://demodv.scholarsportal.info/.

[2] Dataverse did not implement checksums in versions 3.6 and prior. For a list of known issues experienced during testing, see https://wiki.archivematica.org/Dataverse.

also revealed a number of issues affecting transfers and processing that were outside of the scope of the contracted development. In the following sections, we identify necessary fixes and enhancements in several areas that could be incorporated in future releases.

A. *Metadata*

Currently, only six DDI fields (title, unique identifier, author(s), distributor, dataset version, and restriction) are included in the METS descriptive metadata section (see Table 1 above). Additional DDI fields (abstract, subject, and distDate) were proposed given that the first two of these fields are mandatory in Dataverse, and the third provides additional context. However, the addition of these fields was considered by Artefactual to be outside the scope of the development contract. Rights-related metadata for individual files could also be mapped directly to PREMIS as is supported currently in other Archivematica workflows. Dataverse packages consisting of only metadata currently fail, based on the rationale that there is nothing to preserve if a package does not contain any data files.

B. *Interface*

Several improvements to the transfer browser pane were identified that would facilitate the ability to query and select appropriate datasets within Dataverse, such as showing the dataset version number and the ability to refine searches within the interface. An indication of whether a dataset has already been processed is another potential improvement. The team also outlined the need for stronger error logging and clearer notifications to users based on the issues experienced during testing noted above. Joel Simpson suggested the idea of an additional micro-service for verifying a Dataverse transfer before the transfer begins to make it easier to identify these errors and ensure compliance.

C. *Conformance with Additional Archivematica Functions*

AIP re-ingest functions present in Archivematica do not currently function for Dataverse AIPs. Development of this feature requires further discussion about use cases and approaches, such as whether re-ingest should take into account any updates made to the submitted dataset in Dataverse. The team also noted the potential benefit of relating datasets as part of a larger collection through defining an Archival Information Collection (AIC),[1] a function that needs further development to conform with the Archivematica workflow for creating AICs.

D. *Messaging to Dataverse and DIPs*

Once a dataset has been processed and stored, it would be beneficial for Archivematica to send a notification to the Dataverse platform and surface selected preservation metadata indicating to users that the dataset has been processed by Archivematica. However, this communication mechanism would require development work on both platforms. As mentioned previously in section III.B above, a larger potential development would be the automated replacement of user-submitted data with Archivematica-created DIPs, particularly when normalized access copies of files submitted by depositors might be desired for ease of access for general users. An example would be if a depositor submitted a large TIFF image: Archivematica's access normalization functions could create a smaller version in JPEG format that would be more suitable for general access purposes in Dataverse.

E. *Conformance with External Requirements*

As methods for standardization continue to develop in the field, an additional development opportunity is the ability for Archivematica-created AIPs and DIPs in Bags to be conformant with the RDA Research Data Repository Interoperability Working Group's *Final Recommendations* document. The *Recommendations* specify how repository outputs should be structured to promote data exchange, which could be used for redundant storage or access purposes [23]. Dataverse's Bag export function adheres to the RDA specification [19].

VI. Discussion and Conclusion

Currently, Scholars Portal is hosting a public Archivematica sandbox connected to its demo Dataverse installation with several test datasets.[2] Invitations to participate in testing the sandbox

[1] AIC: https://wiki.archivematica.org/AIC.

[2] Archivematica Demo Sandbox, *Spotdocs Wiki:* https://spotdocs.scholarsportal.info/display/DAT/Archivematica+Demo+Sandbox.

and to provide feedback were shared with regional, national and international groups related to research data management, digital preservation, and archives, as well as Dataverse users and Archivematica users. Community testing is crucial to provide further information about how different users might use the integration and to identify additional needs from the community. This feedback will be used to inform future platform enhancements and contribute to the ongoing discussion surrounding best practices for preserving research data. We hope that others interested in using these tools will bring additional use cases and sponsor additional developments to improve the integration. Community members who test and implement the integration on their own infrastructure will also provide new perspectives related to its capacity and limitations in different contexts.

This research and integration work contributes to ongoing research and discussions surrounding research data preservation. Several challenges exist in this area, particularly in relation to forming research data preservation policies and strategies. A recent Jisc report *What to Keep* outlined use cases for research data retention and considerations for this emerging field, noting that the practice and procedures—the what, why, how long, and where—are still evolving [24]. Another challenge in developing policies and strategies relates to the heterogeneity of research data, resulting in a large number of data types and file formats, as well as discipline-specific practices and protocols. The *Science Europe Guidance Document: Presenting a Framework for Discipline-specific Research Data Management* provides a useful guidance framework for protocols within various research domains, informed by the FAIR principles, applicable laws, regulations, and standards [25]. The significant differences across disciplines suggest inherent difficulties in developing general policies and strategies for multi-disciplinary data repositories. Increasing our shared knowledge of various curation and preservation workflows would help to ensure that the tools and policies developed in these areas assist in properly managing different types of data for the long term.

Finally, additional research and requirements-gathering needs to be conducted in the area of service models and policy development to understand how preservation approaches can flow from individual researchers to institutions and repositories that are tasked with stewarding research data, and onto potential to shared infrastructures. In addition to connecting the technical pieces of infrastructure, the stewarding institution or organization would need to develop and manage policies and costs for long-term storage and maintenance. For example, OCUL institutions that subscribe to Permafrost would have access to Archivematica instances that could be configured to their institutional containers as part of Scholars Portal Dataverse platform. In this case, datasets processed as AIPs could be stored on the OLRC and managed by the library. Other users may host Archivematica locally or take advantage of other service arrangements and still be able to connect to a target Dataverse instance of their choice. The integration also presents opportunities for centralized, collaborative services that offer Dataverse, Archivematica, and preservation storage as a service model, and therefore a consequent requirement to develop appropriate agreements and governance models for shared services.

Overall, the Dataverse-Archivematica integration project aims to connect several pieces of the research data management ecosystem, drawing on best practices and standards in the archives and digital preservation communities, and to contribute to the development and enhancement of features within these two platforms.

Acknowledgements

The authors are grateful for funding from OCUL to pursue this project and thank the Scholars Portal staff, members of Artefactual Systems, the initial project participants, and the community for their support of the project. Our thanks go to: Allan Bell (UBC), Eugene Barsky (UBC), Peter Binkley (University of Alberta), Eleni Castro (Boston University, formerly IQSS), Alan Darnell (Scholars Portal), Kate Davis (Scholars Portal), Philip Durbin (IQSS), Alex Garnett (SFU), Geoff Harder (University of Alberta), Chuck Humphrey (Retired, formerly CARL Portage), Larry Laliberte (University of Alberta), Amber Leahey (Scholars Portal), Victoria Lubitch (Scholars Portal), Steve Marks (University of Toronto), Evelyn McLellan (Artefactual Systems), Umar Qasim (University of Alberta), Joel Simpson (Artefactual Systems),

Justin Simpson (Artefactual Systems), Ross Spencer (Artefactual Systems), Amaz Taufique (Scholars Portal), Leanne Trimble (University of Toronto), and Dawas Zaidi (Scholars Portal). Finally, the authors would like to thank the three anonymous reviewers of the manuscript for their valuable comments.

References

[1] CASRAI, Research data, Aug. 12, 2015. Accessed on: Nov. 1, 2019. [Online]. Available: https://dictionary.casrai.org/Research_data

[2] GO Fair, FAIR Principles. Accessed on: Nov. 1, 2019. [Online]. Available: https://www.go-fair.org/fair-principles/

[3] M.D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data 3, Mar. 15, 2016. doi:10.1038/sdata.2016.18

[4] Digital Curation Centre, DCC curation lifecycle model. Accessed on: Nov. 1, 2019. [Online]. Available: http://www.dcc.ac.uk/resources/curation-lifecycle-model

[5] M. Crosas, "The Dataverse Network: An open-source application for sharing, discovering and preserving data," D-Lib Magazine, vol. 17, no. 1-2, January-February 2011. doi:10.1045/january2011-crosas

[6] Dataverse, Dataverse repositories: A world view. Accessed on: Nov. 1, 2019. [Online]. Available: https://services.dataverse.harvard.edu/miniverse/map/

[7] T. H. Vines et al., "The availability of research data declines rapidly with article age," Current Biology vol. 24, no. 1. doi:10.1016/j.cub.2013.11.014

[8] Dataverse, Dataverse installation guide: Configuration – Database settings - :FileFixityChecksumAlgorithm, Oct. 23, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: http://guides.dataverse.org/en/latest/installation/config.html?highlight=md5#filefixitychecksumalgorithm

[9] Dataverse, Dataverse installation guide: Installation, Oct. 23, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: http://guides.dataverse.org/en/latest/installation/installation-main.html

[10] Dataverse, Dataverse user guide: Tabular data ingest, Oct. 23, 2019. Accessed on: Nov. 1, 2019 [Online]. Available: http://guides.dataverse.org/en/latest/user/tabulardataingest/index.html

[11] Dataverse, Dataverse user guide: Tabular data ingest - SPSS, Oct. 23, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: http://guides.dataverse.org/en/latest/user/tabulardataingest/spss.html

[12] Dataverse, Dataverse developer guide: Universal Numerical Fingerprint (UNF), Oct. 23, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: http://guides.dataverse.org/en/latest/developers/unf/index.html

[13] B. Lavoie, The Open Archival Information System (OAIS) reference model: Introductory guide (2nd edition). Great Britain: Digital Preservation Coalition, 2014. doi: 10.7207/twr14-02

[14] U. Qasim, C. Davis, A. Garnett, S. Marks, and M. Moosberger, Research data preservation in Canada: A white paper. Portage Network, 2018. doi:10.14288/1.0371946

[15] J. Mitcham, C. Awre, J. Allinson, R. Green, and S. Wilson, Filling the digital preservation gap: A Jisc research data spring project. Phase Three report. Jisc, 2016. doi: 10.6084/m9.figshare.4040787

[16] V. Tykhonov, P. Doorn, and M. Wittenberg, The development of DataverseNL data repository as structured data hub, 2017. Accessed on: Nov. 1, 2019. [Online]. Available: https://www.slideshare.net/vty/dataversenl-as-structured-data-hub

[17] James Myers, Data and metadata packaging for archiving, Jan. 10, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: https://github.com/QualitativeDataRepository/dataverse/wiki/Data-and-Metadata-Packaging-for-Archiving

[18] J. Crabtree, R. Moore, D. Sizemore, Odum Institute iRODS policies to support preservation, 2016. Accessed on: Nov. 1, 2019. [Online]. Available: https://slideplayer.com/slide/14487733/

[19] Dataverse, Dataverse installation guide: Configuration - DuraCloud/Chronopolis integration, Oct. 23, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: http://guides.dataverse.org/en/4.17/installation/config.html#bagit-export

[20] Core Trustworthy data repositories requirements, Nov. 2011. Accessed on: Nov. 1, 2019. [Online]. Available: https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf

[21] Dataverse, Dataverse user guide: Dataset + File Management – Dataset versions, Oct. 23, 2019. Accessed on: Nov. 1, 2019. [Online]. Available: http://guides.dataverse.org/en/latest/user/dataset-management.html#dataset-versions

[22] E. McLellan, "Dataverse: AIP Structure," Archivematica [wiki], Nov. 17, 2015. Accessed on: Nov. 1, 2019. [Online]. Available: https://wiki.archivematica.org/Dataverse#AIP_structure

[23] RDA Research Data Repository Interoperability Working Group (WG), Final Recommendations. Accessed on: Nov. 1, 2019. [Online]. Available: https://www.rd-alliance.org/system/files/Research%20Data%20Repository%20Interoperability%20WG%20-%20Final%20Recommendations_reviewed_0.pdf

[24] Neil Beagrie, What to keep: A Jisc research data study, Feb. 2019. Accessed on: Nov. 1, 2019. [Online]. Available: https://repository.jisc.ac.uk/7262/1/JR0100_WHAT_RESEARCH_DATA_TO_KEEP_FEB2019_v5_WEB.pdf

[25] Science Europe, Science Europe guidance document: Presenting a framework for discipline-specific research data management, Jan. 2018. Accessed on: Nov. 1, 2019. [Online]. Available: https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf

Sustainability through Community

ffmprovisr and the Case for Collaborative Knowledge Transfer

Andrew Weaver
University of Washington, USA
weaveraj@uw.edu
https://orcid.org/0000-0002-9792-4036

Ashley Blewer
Artefactual Systems, MediaArea
USA
ashley.blewer@gmail.com

Abstract – This paper will present the online resource ffmprovisr as a case study in using open, online documentation as a shared educational resource. ffmprovisr is a website that supports the digital audiovisual community through technical education. This paper will cover how the website was developed, its role in fostering a collaborative educational environment, and its impact on the field and beyond.

Keywords – Collaboration,Open-source, Education, Audiovisual

Conference Topics – Building Capacity, Capability and Community; Collaboration: a Necessity, an Opportunity or a Luxury?

I. Introduction

Facilitating knowledge sharing between practitioners is one of the most foundational functions amongst the myriad missions of professional organizations. While this is accomplished via a variety of mediums, such as journals, conference presentations and organizational publications, the level of openness in both content creation and dissemination can vary widely across fields and formats. This paper will present 'ffmprovisr', an online resource maintained under the auspices of the Association for Moving Image Archivists' Open Source Committee, as a case study in using open source methodology for specialist knowledge transfer. In covering the development of ffmprovsr, its role in fostering a collaborative educational environment, and its impacts both within and outside of the field of digital preservation, this paper will demonstrate how open and collaborative approaches both complement and contrast with more traditional resources generated by professional organizations.

II. Problem Space

ffmprovisr's inception was brought about by the need to address the steep learning curve of the open source multimedia manipulation framework, FFmpeg. FFmpeg's website defines FFmpeg as "A complete, cross-platform solution to record, convert and stream audio and video."[1] In the field of audiovisual preservation this software suite, consisting of FFmpeg, FFprobe and FFplay, is one of the core tools for practitioners for performing tasks such as digitization, file characterizations, bulk file operations, data conversion and fixity checks on material targeted for preservation and access.

As is noted by archivist Dave Rice in [1], audiovisual preservationists have traditionally been reliant on software tools that were designed for the broadcast industry; a reliance that is increasingly problematic as needs diverge and support for legacy formats is dropped from commercial solutions. FFmpeg can play a significant role in empowering practitioners to perform preservation actions with highly targeted control, independent of ill-fitting consumer or broadcast oriented products, and as such, skill in its use is increasingly essential.

Despite its growing importance within the preservation community, a significant obstacle to its use is that FFmpeg is a program that runs only on the command line. For people who have not previously used a command line interface, interacting with their computer solely via text can be a new and daunting endeavor. Even when a user is comfortable with using the command line, FFmpeg is such a large and complex program that understanding the correct

[1] https://ffmpeg.org/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

commands and syntax can also be intimidating and confusing. Users may also receive unexpected results and not understand the nuances of how files are being created or modified.

An issue with FFmpeg that is especially pertinent to people working in preservation is the implicit versus explicit actions that can be performed using certain parameters or "flags." FFmpeg is ideal for preservation as it allows very granular control of actions via specific flags. However, when those flags are not present, FFmpeg will make assumptions about what the user implicitly wants to do with a file, leading to unexpected results. This often occurs when a user specifies the container for a video file but does not specify the type of video codec to use for the video datastream inside of the container. Another example is failing to specify audio bit depths, leading to possible undesired loss of detail. Additionally, depending on interactions between flags (or lack thereof) FFmpeg may fail to create a file, exiting instead with an error.

The creation of FFmpeg scripts must be given careful thought so as to avoid unintentional repercussions, particularly when working on scripts that change many different kinds of video files into one specific kind.

III. History

ffmprovisr began as a small, personal proof-of-concept project in 2014 by archivist/developer Ashley Blewer, following her introduction to FFmpeg at the 2013 AMIA/DLF (Association of Moving Image Archivists and Digital Library Federation) Hack Day, where she worked with a group on a project to make a wiki-style resource for people interested in FFmpeg called "Fast Forward." In 2015, Blewer brought the ffmprovisr project to the AMIA/DLF Hack Day via the following proposal:

"I think it'd be fun to combine and continue to build up these two projects [my ffmprovisr and the Fast Forward wiki] into something better because ffmpeg continues to live on as a mysterious but necessary component of a/v archival practice. This project would be mostly R&D with some basic front-end web development skills (building forms). I feel this is a little out of the scope of hack day (and those greedy for rewards may seek refuge elsewhere) in that it's more of a REMIX project and a mostly- hack-the- docs- with- some- coding project, but if there is interest (there was last year, for ffmprovisr) – we will build the hell outta this!"[1]

During this event, the project had a team of around a dozen contributors, including contributions that came in remotely from Ireland and New Zealand. From that point onward, ffmprovisr was no longer a single person's proof-of-concept project but a community effort, and it has continued to move forward in that way. During this event, Blewer moved the resource site from her personal Github repository to being hosted at the AMIA open source committee's Github repository and web domain, where it can be accessed today[2].

For a few years after this, Blewer was the de-facto maintainer of the project and continued to update the resource with new scripts when discovered, or facilitate the addition of new scripts by other people.

In April 2017, archivists Reto Kromer[3], Katherine Nagels, and Kieran O'Leary joined Blewer in the role of Maintainer[4]. In November 2017, librarian Andrew Weaver was asked to join the group as a Maintainer due to his active code contributions and dedication to the project[5]. The purpose of a Maintainer is defined as a role supporting other people's issues and contributions, and striving to generally keep ffmprovisr as a happy and well-functioning space to learn about FFmpeg. The maintainers also encourage learning and support for related skills, such as building web pages and using git and Github. ffmprovisr's maintainers are distributed across the globe and are readily available to answer questions and provide support. In addition to the officially listed maintainers, due to the project's open nature,

[1] https://wiki.curatecamp.org/index.php/Association_of_Moving_Image_Archivists_&_Digital_Library_Federation_Hack_Day_2015#ffmprovsr

[2] https://amiaopensource.github.io/ffmprovisr/

[3] Reto Kromer has since stepped down as maintainer, but remains an active contributor.

[4] https://github.com/amiaopensource/ffmprovisr/commit/89039f55b3012f75c5b908c80cd2ebdc77b2f6a6

[5] https://github.com/amiaopensource/ffmprovisr/commit/5a3e437d76570f8f6ab78820626b12861709a922

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

anyone can help improve it by such means as direct code contributions, FFmpeg advice, questions and requests, reviews of commands and testing.

Since its inception, ffmprovisr has extended its reach to include not just helpful recipes for using the FFmpeg framework, although that remains its primary purpose, but also now includes support, education and recipes for related multimedia tools used in preservation such as ImageMagick, and audio extraction tools.

IV. Dual model for education

ffmprovisr functions as a model for open documentation and collaboration through both its usage as a shared internationally-referenced resource and through its educational function to help preservationists learn new skills through practicing them in a welcoming environment. ffmprovisr is an educational resource not solely as a webpage for looking up recipes or a space for skill-honing, but something more powerful by being both of these things working to benefit each other. The purpose of ffmprovisr thusly is two-fold, both for reading and gaining of knowledge and through the active support and practice of tool-building and contributing to the field. We will discuss first the benefits in ffmprovisr as a traditional educational resource and then go deeper into how ffmprovisr exists as a new kind of educational resource, expanding the ways in which archivists and librarians disseminate field-specific knowledge amongst each other to overall benefit the field at large, as a collaborative effort.

A. Traditional Educational Resource

ffmprovisr is a resource that works like a classic cookbook. It provides users with tasks they may want to do, such as changing a video format or creating a short clip from an existing video. Excluding an introductory section with higher-level concepts, each task contains a script that a user may copy and paste into their computer's terminal window. With some small modifications like adjusting filenames or output paths, the script can easily be adapted to a user's local settings or configurations. Below each script is a breakdown of each component of the script, explaining how it was built, why the specific component exists within the context of the larger task, how the user could modify the values to produce different results, and links to more robust documentation when necessary. This provides the user with a quick answer to their solution, an ability to know where and how to modify the script, general education about how FFmpeg works and information about any caveats that might cause inconsistent results from modifying the script. Additionally, ffmprovisr supports an offline mode and is installable via the Homebrew package manager.[1]

ffmprovisr, while a simple site and simple concept in nature, has a wide and lasting impact on practitioners in the field, many of whom are lone technicians working with limited time and budgets. By serving as a clearinghouse for frequently used commands in A/V preservation, as well as an annotated source of pertinent knowledge it facilitates both daily tasks as well as continuing professional skill development.

B. New Model of Educational Resource

In addition to being an example of a traditional online resource available to all, ffmprovisr exists as a new, unique model for knowledge transfer within and by professional communities. This resource is a living, active document featuring the latest information gathered and reviewed by dozens of experts in the subject. ffmprovisr goes beyond open publishing, by allowing for not just open access, but open collaboration at all points of creation.

This is in contrast to the traditional academic process, where the collection and dissemination of knowledge tends to follow a hierarchical pattern, wherein a researcher will perform their research, possibly keeping the findings closed until a final publication. This can then take months to years for information to be disseminated to a wider audience. Another common model is for a professional organization to assemble a group tasked with the generation of a technical document. These traditional processes can and do result in the creation of useful resources, however, their relatively closed nature can make them inflexible as well as exclusionary to a range of voices.

ffmprovisr is vetted by at least several maintainers and also the general open source community and public audience, so each script has been

[1] https://brew.sh/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

reviewed by multiple experts in the field before being published on the site. The resource can also be reviewed and updated at any point in time. This is, in essence, a more robust version of the practice of peer review in traditional academic institutions, but with faster turnaround, wider breadth, more encouragement of broader participation. This allows for more diverse and interdisciplinary contributions, increasing the overall scholastic rigor of the document. The contributors to ffmprovisr range from being internationally-recognized consultants in audiovisual preservation, archivist-developers at film and broadcast archives, digital preservationists at academic libraries, FFmpeg developers, professional media engineers, and many others. At the time of writing, ffmprovisr has had over thirty total contributors, representing over thirty institutions, and that number continues to increase.

ffmprovisr purposefully does not track users or gather statistics about usage, but its influence can be seen through conference proceedings, references to the tool on twitter, remixing, use in other applications, references in job postings[1] and anecdotal praise given to maintainers of the project from people working both within and outside of the archival field. ffmprovisr has been cited as an invaluable resource in peer-reviewed journal articles introducing FFmpeg to a wider Digital Humanities community[2]. A significant benefit of ffmprovisr is that it does not exist in a niche bubble, only to be seen and contributed to by a small pedagogical coterie within only one discipline. Rather, ffmprovisr can be seen and reviewed by anyone, allowing it to tap into a broad pool of expertise.

Beyond providing support for an increased understanding of FFmpeg, the maintainers of the ffmprovisr project aim to use the project to support education in other technical skills such as using the version control software git, the popular web platform Github, and writing HTML and CSS. This is done through active solicitation of contributions and coaching/encouragement to individuals seeking to add to the project. This has in turn fostered greater participation within the audiovisual open-source community and helps fill a core need within the libraries/archives professional field for technical training opportunities.

The codebase and website declares itself as having open Creative Commons license, CC-BY (free to use with a request for attribution)[3]. Because of this open licensing model, ffmprovisr has been "remixed" at least three times to produce similar resources based on the existing source code and project ethos: a guide for analog audiovisual cables, The Cable Bible[4], by Ethan Gates, and two sites to facilitate simple scripting in libraries and archives: Script Ahoy![5] by Dianne Dietrich and Jarrett Drake and The Sourcecaster[6] by James Baker and Thomas Padilla. ffmprovisr is a successful project and it is made more successful as a result of its permission to be shared and widely distributed, and have smaller sites get built off of the same model of open, collaborative resource sharing.

V. Conclusion

The ffmprovisr model of open, interdisciplinary and shared collaboration between colleagues is one that can and should be modeled by more professional organizations with a focus and mission around the development of professional practice. Its open nature, both to information creation and publishing, allows for a wide pool of contributors as well as vetters, and presents a strong alternative to more hierarchical processes. Additionally, it is flexible and responsive enough to remain current amongst changing technologies - a necessary trait for a digital preservation resource.

Acknowledgement

The authors of this paper would like to acknowledge the other maintainers of ffmprovisr, Katherine Francis Nagels and Kieran O'Leary, former maintainer Reto Kromer and all contributors to the ffmprovisr project. Thank you to all the FFmpeg

[1] http://web.library.emory.edu/documents/pa_staff_Audiovisual%20Conservator_Nov2018.pdf

[2] https://programminghistorian.org/en/lessons/introduction-to-ffmpeg

[3] https://amiaopensource.github.io/ffmprovisr/

[4] https://amiaopensource.github.io/cable-bible/

[5] http://dd388.github.io/crals/

[6] https://datapraxis.github.io/sourcecaster/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

developers. Thanks to Dave Rice for his mentorship and influence.

References

[1] Rice, D. (2015). Digitization Software Obsolescence, Too?. International Association of Sound and Audiovisual Archives (IASA) Journal, (45), 11–19. Retrieved from http://journal.iasa-web.org/pubs/article/view/79

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Fedora and the Oxford Common File Layout

Enhancing Support for Digital Preservation

David Wilcox
DuraSpace,Canada
dwilcox@duraspace.org
0000-0001-5411-9208

Abstract – Fedora is an open source repository platform for managing and preserving digital objects. While Fedora has always been associated with digital preservation, recent releases have focused on exposing linked data and aligning with modern web standards. The Oxford Common File Layout (OCFL), which defines a shared approach to file hierarchy for long-term preservation, provides an opportunity to bring the focus back to digital preservation in Fedora. The OCFL supports application-independent, transparent file persistence that can be used to rebuild a repository in case of disaster. These features address the current needs of the Fedora community, so a group of Fedora committers met in person to design a version of Fedora that implements the OCFL. This will be the focus of the next major release, Fedora 6. 0. This paper introduces the OCFL and describes the proposed design for Fedora 6. 0, including the next steps for development and implementation.

Keywords – Fedora, repository, OCFL, preservation, standards

Conference Topics – Designing and Delivering Sustainable Digital Preservation; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

Fedora [1] is a flexible, extensible, open source repository platform for managing, preserving, and providing access to digital content. For the past several years the Fedora community has prioritized alignment with linked data best practices and modern web standards. However, the focus is now shifting back to Fedora's digital preservation roots with the help of the Oxford Common File Layout (OCFL) [2]. The OCFL, which began as a discussion at the Fedora and Samvera Camp at Oxford, UK in 2017, is an effort to define a shared approach to file hierarchy for long-term preservation. This approach includes both transparency and the ability to rebuild a repository from the contents on disk, both of which address key needs in the Fedora community. With the support of the Fedora governance group, a team of Fedora committers met in person in February of 2019 to design the next major release of Fedora, which will implement the OCFL at the persistence layer. This paper introduces the OCFL and describes the proposed design for Fedora 6. 0, including the next steps for development and implementation.

II. The Oxford Common File Layout

The Oxford Common File Layout (OCFL) is "an application-independent approach to the storage of digital objects in a structured, transparent, and predictable manner. It is designed to promote long-term access and management of digital objects within digital repositories. " [3]. The OCFL arose from the need to insulate digital objects, which tend not to change much after being accessioned, from the churn of software applications, which tend to change much more often. These application updates often involve data migrations, which put the data at risk. They also store data in application-dependent ways, making it difficult or impossible to understand the data without the software. The OCFL mitigates these issues by specifying a file and directory layout that applications must conform to.

The most basic element of the specification is the OCFL Object, which is "a group of one or more content files and administrative information, that are together identified by a URI. The object may contain a sequence of versions of the files that represent the evolution of the object's contents. " [3]. OCFL Objects contain administrative information that identifies them as OCFL Objects and allows changes to be tracked over time. The content files of an OCFL

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Object can be anything at all; however, in order to support rebuilding the repository from the files on disk, OCFL Objects should contain "all the descriptive, administrative, structural, representation and preservation metadata relevant to the object. " [3].

An OCFL Object contains a file declaring its conformance with a particular version of the specification, along with a sequentially numbered folder for each version of the content files. A version folder (e. g. v1) contains a content folder (if it has contents), an inventory file, and an inventory digest file. The inventory file is a JSON document with a manifest of all the files in the version's content folder. Each file in an OCFL Object has an associated digest, which both provides a fixity value that guards against degradation over time and allows for a content-addressable reference to the file.

III. Motivations for Implementing OCFL In Fedora

Fedora is a digital repository for the long-term storage and management of digital objects. Fedora has gone through several upgrades over the years as the needs of its user community changed and technologies improved. Some of these upgrades have required data migrations, which makes them very challenging for institutions to absorb, especially if they have large amounts of data. It also puts the data at risk of corruption during the move. By making Fedora OCFL-compliant, future upgrades should not require data migrations. This is because the application will be made to conform with the files on disk, rather than the other way around.

Another motivation is transparency. Currently, Fedora objects are stored in a database and file structure that is application-dependent, meaning the contents of the repository cannot be inspected and understood without going through the Fedora application. This presents a risk to future access; if a hard drive with the contents of the repository were recovered without the Fedora application layer, the contents would be difficult to interpret.

Finally, the ability to rebuild the repository from the files on disk is an important motivator. Currently, backup and restore tools must be used to recover the repository in the case of a problem. This new functionality would allow the repository to be rebuilt by reading the contents of the files on disk. For all of these reasons,, the Fedora community has decided to implement the OCFL in the next major version of Fedora.

IV. Designing Fedora 6. 0

A. Implementing OCFL

A group of Fedora committers met in-person in February, 2019 to design Fedora 6. 0, the next major version of the software. The team went into the meeting with several design goals:

1. Implement the OCFL in Fedora
2. Improve performance and scale
3. Support complete repository rebuilds from the contents on disk
4. Don't make major changes to the API

With these goals in mind, the team discussed how best to implement the OCFL in Fedora in a way that would be scalable and performant without causing undue problems for users who might have written applications against the API. The first problem to address was the mapping between Fedora objects and OCFL objects, which are not exactly the same. Specifically, Fedora objects are based on the Linked Data Platform (LDP) specification [4], and contain Fedora-specific information. In the interests of backward compatibility, scale, and the ability to rebuild the repository from the file system, the resulting objects will contain fedora-specific metadata and may be required to follow specific naming conventions, and have other structural requirements placed on them.

In the interests of archival transparency, an opt-in extension to the Fedora API will allow grouping of resources that are persisted and versioned together as a single archival unit. All resources underneath an "Archive group" container will be persisted within a single archive (i. e. within an OCFL object). This will allow for the creation and maintenance of compound OCFL objects containing several files within them.

It will also be possible to drop an instance of Fedora on top of an existing OCFL storage root; in this case, Fedora will be able to read and make sense of the contents of the file system. The existing OCFL data are not required to contain any fedora-specific

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

metadata, or follow any specification, convention, or be otherwise related to Fedora in any way. This approach has many advantages, one of which is to create a plausible migration path from Fedora 3. x by converting the contents on disk to be OCFL-compliant before dropping Fedora 6. x on top.

B. *Other Features*

In addition to implementing the OCFL, Fedora 6. 0 will also include a number of other features and improvements. One of these will be a built-in query endpoint for simple, common repository queries. Since version 4. 0, Fedora has not supported an internal query service, instead delegating such functionality to external tools like Apache Solr. However, the community has expressed a need for a synchronous, internal query service, so this will be added in Fedora 6. 0. The supported queries will include:

1. List all resources
2. List resources by mimetype
3. List resources by parent
4. List resources by mimetype, parent, and modified date (<>=)
5. List resources where modified <> x date.

Users will still need to use an application like Solr for more complex queries, but there is already an out-of-the-box integration based on Apache Camel that can be set up and used with a standard Fedora installation.

While Fedora currently supports fixity checking, the community has expressed a need for a more robust, proactive fixity service. This new service will automatically check the fixity of all items in the repository at a frequency and schedule specified by the administrator. It will log the results and report errors, and it will also maintain a full report of the health of the repository that can be requested on-demand.

C. *Architecture*

Fedora 6. 0 will be architected to support greater performance and scale while complying with the OCFL. This will be achieved by replacing the current ModeShape backend with an OCFL-compliant file system while optimizing reads and lookups with an internal database. This database will act as a kind of fast cache on top of the relatively slow filesystem.

V. Next Steps

Following the design meeting in February, the Fedora committers put together a high level summary of the design to share, first with the Fedora Leadership Group, and then with the broader Fedora community. The goal of this effort is to get buy-in, both from the Fedora governance group and the community as a whole.

Once we have buy-in from the Fedora governance group and the community, we will proceed to schedule code sprints to complete the work. While the Fedora project has full-time staff through its relationship with DuraSpace, these staff members do not write the majority of the code for the software, instead playing roles as community and technical coordinators. The bulk of the code and documentation will be written by members of the community, which is why achieving buy-in is important. We will also get commitments from institutions to adopt Fedora 6. 0 when it is ready, and we will work with these institutions as we develop the software to ensure we are building the application that the community wants.

VI. Conclusion

Over the years, the Fedora community has prioritized and focused on different aspects of the software. The 4. x line of Fedora releases put the emphasis on support for linked data and alignment with modern web standards. This culminated with the release of Fedora 5. 0, which implements the recently completed Fedora API specification [5]. Having reached this milestone, the community has returned to a focus on digital preservation, which coincides with the development of the OCFL specification. The OCFL represents a return to the digital preservations sensibilities of Fedora 3. x, but as a more standardized, community-focused effort. With the completion of the initial design of Fedora 6. 0, the community will proceed to put together a development plan, including a combination of code sprints and funded development effort. We are targeting late 2019 for the 6. 0 release, which will bring together the linked data and web standard features with the strong digital preservation sensibilities of the OCFL.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Acknowledgement

This paper would not have been possible without the support of the Fedora community, particularly those institutions that are members of DuraSpace in support of Fedora.

References

[1] Fedora - The Flexible, Modular, Open Source Repository Platform. [Online]. Available: https://duraspace. org/fedora/. [Accessed March 15, 2019].

[2] Oxford Common File Layout. [Online]. Available: https://ocfl. io. [Accessed March 15, 2019].

[3] A. Hankinson, et al, "Oxford Common File Layout Specification 0. 2," March 12, 2019. [Online]. Available: https://ocfl. io/0. 2/spec/. [Accessed March 15, 2019].

[4] S. Speicher, J. Arwe, A. Malhotra, "Linked Data Platform 1.0," February 26, 2015. [Online]. Available: https://www.w3.org/TR/ldp/. [Accessed March 19, 2019].

[5] B. Armintor, E. Cowles, D. Lamb, S. Warner, A. Woods, "Fedora API Specification 1.0," November 22, 2018. [Online]. Available: https://fedora.info/2018/11/22/spec/. [Accessed March 19, 2019].

OAIS VERSION 3 DRAFTUPDATES

David Giaretta
PTAB Ltd UK
david@giaretta.org
0000-0001-8414-7509

John Garrett
Garrett Software USA
garrett@his.com
0000-0003-4460-9902

Mark Conrad
NARA USA
mark.conrad@nara.gov

Eld Zierau
Royal Danish Library
Denmark elzi@kb.dk
0000-0003-3406-3555

Terry Longstreth
Washington Acad. Sciences USA
longstreth@acm.org
0000-0001-8228-6185

J Steven Hughes
Jet Propulsion Laboratory USA
john.s.hughes@jpl.nasa.gov
0000-0003-4851-293X

Matthias Hemmje
FernUniversität Hagen Germany
matthias.hemmje@fernunihagen.de
0000-0001-8293-2802

Felix Engel
FernUniversität Hagen Germany
felix.engel@fernunihagen.de
0000-0002-3060-7052

Abstract – This paper provides a high-level view of the changes proposed in the OAIS Reference Model.

Keywords – OAIS, digital preservation

Conference Topics – Designing and Delivering Sustainable Digital Preservation; Exploring New Horizons.

I. INTRODUCTION

The aim of this paper is to describe and explain the most significant updates which have been made to version 2 OAIS [1], which was published in 2012, from the point of view of the authors, who have all been deeply involved with the revision. These updates resulted in a draft which, at the time of writing, is the text to be submitted for the Consultative Committee for Space Data Systems (CCSDS) and ISO review. Further changes may be made, either before submission or during the review process, after which version 3 of OAIS will be published.

In the body of this paper the text in italics is taken from the current draft.

ISO 16363, which is the basis of auditing trustworthy digital repositories, is being updated to be consistent with the changes to OAIS.

II. OAIS REVIEW PROCESS

The previous update of OAIS was primarily debated at physical CCSDS meetings augmented by email exchanges via a mailing list, together with weekly telecons. This did allow wide participation but the CCSDS MOIMS-DAI [2] working group, which oversees the latest revision, wanted to improve and widen access. With this aim in mind, the http://review.oais.info website was set up, which allowed anyone to register and contribute to the discussion. It allowed everyone, whether registered or not, to view all the contributions and discussions.

The page for each suggested change showed the various contributions and the final decision on whether to reject, accept or modify that proposal. Changes to the evolving document were reviewed and dispositioned at weekly telecons and by in person and remote attendees of MOIMS/DAI sessions at the

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

semi-annual CCSDS plenary meetings. The disposition reflected the consensus reached in these telecons and meetings, as reflected by agreements recorded in http://review.oais.info.

A marked-up Word document was maintained, with comments linking each change to the discussion on the website.

When all 200+ suggested changes to OAIS had been resolved, a second round of comments were collected on this marked up document to check for inconsistencies and small errors. These were resolved via the telecons and appropriate changes made, with comments to record the justification for the change.

The remainder of this document describes the major changes made in the draft which will go for formal CCSDS and ISO review.

To some the changes described here may seem unimportant or irrelevant but they have been made in anticipation of new challenges to the preservation community which may arise over the 5 or more years before version 4 will be produced

III. Updates to OAIS concepts

A. *Representation Information*

One of the key OAIS concepts is Representation Information, which, when combined with a Data Object, produces an Information Object. The question as to how much Representation Information is needed is determined by the definition of the Designated Community and its Knowledge Base.

The amount of Representation Information will change over time as the Knowledge Base of the Designated Community changes. The OAIS needs to ensure that it has Long Term access to all the relevant Representation Information. A choice must be made whether the OAIS collects all the relevant Representation Information itself or references the existence of the Representation Information in another trusted OAIS Archive. That choice is an implementation and organization decision.

The updates make it clear that in **special cases** the initial amount of Representation Information required may be very minimal. For example, for a specific Data Object and a specific Designated Community, the Knowledge Base of the Designated Community is adequate for its members to understand or use the Data Object. In such cases the Representation Information could be the statement that no additional Representation Information is needed for that specific Designated Community at that particular time, but further Representation Information may need to be collected in future. The revised text goes on to say, …"any Representation Information that can be gathered at ingest should be included since it will likely be costlier to rediscover and add it at a later time."

B. *Preservation Description Information (PDI)*

In the versions of OAIS up to now the components of PDI, namely Provenance Information, Reference Information, Fixity Information, Access Rights Information and Context Information, referred to the Content Information, i.e. the Content Data Object plus its Representation Information. Although these are a consistent and useful set of definitions, it does cause some problems in terms of potential implementations. Consider the case where one deals with a distributed network of Representation Information, which changes with the Designated Community's Knowledge Base. A change in some part of the Representation Information network would mean that all the elements of the PDI would change.

The update concerning PDI is that all the components of PDI would now refer to the Content Data Object rather than Content Information.

There are several reasons for this change. The consensus was that for most, perhaps all, repositories, the PDI components do refer to the Content Data Objects. For example, the Fixity Information is often essentially a digital digest of the Content Data Object. This focus on Data Objects would also make audits of repositories more practical since the auditor can perform checks on specific Content Data Objects. Of course, even the Content Data Object may be complex, for example consisting of many files, but at least changes in the Knowledge Base of the Designated Community does not cause it to change.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

A related point considered by the group was that, for example, the Representation Information should have Fixity also. To clarify this point the following note was added to emphasize the fact that, from the very first version of OAIS, the Information Model applies to every one of the things which are called "Information", including, for example, Representation Information and Provenance Information.

Defining PDI (as well as its components - Provenance Information, Context Information, Reference Information, Fixity Information, and Access Rights Information) as relevant to the Content Data Object does not mean that those concerns are any less important for other data objects or at other levels, for example, it is important to apply reference, fixity, provenance, context and access rights to Representation Information, or to any other information the Archive is preserving. Definition of these terms as relevant to the Content Data Object is simply to ease discussion of these concepts at the Content Data Object level.

In other words when one is talking about, for example, Representation Information as the target of preservation, then one can regard it as Content Information in its own right, as well as being part of another instance of Content Information. To some readers this may seem a strange way to describe things, but it is similar to what should be the familiar arrow in the OAIS Information Model which "loops back" from Representation Information back to itself.

C. *Preservation Objectives*

Usability has played a central role in defining preservation. However, there was a feeling that the meaning of usability needed to be clearer, and more testable. To this end the concept of a "Preservation Objective" has been introduced and defined as a specific achievable aim which can be carried out using the Information Object.

Preservation Objectives can then be used in the definition of other terms including:

- Representation Information: The information that maps a Data Object into more meaningful concepts. so that the Data Object may be understood in ways exemplified by Preservation Objectives.
- Independently Understandable: A characteristic of information that is sufficiently complete to allow it to be understood by the Designated Community, as exemplified by the associated Preservation Objectives, without having to resort to special resources not widely available, including named individuals

Preservation Objectives are intended to allow the repository to make it possible to test and demonstrate whether the information actually is Independently Understandable by members of the Designated Community now and into the future.

- Examples of Preservation Objectives are provided in the updated OAIS:
- The ability to render documents, images, videos or sounds in a way which is sufficiently similar to the original. This could be checked by verifying that, for example, the document is readable or the image is viewable. An analysis of the colours could also be compared. A spectral analysis could be performed on the sounds and compared with that of the original.
- The ability to process a dataset and generate the data products expected. This could be checked by comparing with something generated earlier, for example on Ingest.
- The ability to understand a dataset and use it in analysis tools to generate results, for example the density of electrons in the upper atmosphere or the structure of a molecule, given certain measurements. These could be compared with results generated earlier.
- The ability to re-perform an artistic performance. This could be compared with a recording of a previous performance.

IV. Update to the OAIS Functional Model

There have been many small clarifications made to the various text and diagrams which make up the Functional Model, introducing unambiguous shapes for diagram entities; MOIMS-DAI hopes that CCSDS/ISO will allow the publication of the new version to include the colors which give visual clues as to the grouping of the boxes.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

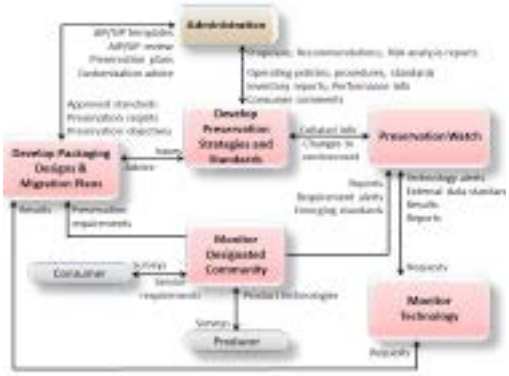


Figure IV-1 Updated Preservation Planning Functional Entity

The one area where an extra function has been added is to the Preservation Planning Functional Entity.

The additional function is the already widely used "Preservation Watch". This is described in the update as follows:

The Preservation Watch function is the role of collating preservation related information from a variety of internal and external entities. The Preservation Watch function also brings in reports, requirement alerts and emerging standards from the Monitor Designated Community function and technology alerts, external data standards, results and reports from the Monitor Technology function. Changes in the environment of the Archive (financial, political, and environmental) can be part of the Preservation Watch function.

Previously, Preservation Watch functionality was primarily located within the Develop Preservation Strategies and Standards.

V. Updates to the OAIS information model

The major updates to the Information Model carry forward the changes which have been described in section III. These are summarized in the following diagram where the PDI connects to the Data Object rather than the Content Information:

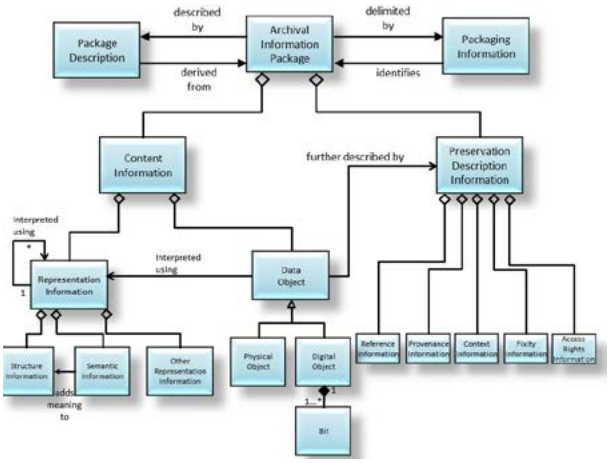


Figure V-1 Updated AIP diagram

VI. Updates to preservation perspectives

Major changes have been made to the section of OAIS which describes practices that have been, or might be, used to preserve digital information and to preserve access services to digital information.

Up to now, essentially the only preservation practice which has been explicitly described has been Migration and Preservation of Access, e.g. Emulation. However clearly the OAIS mandatory responsibilities require that there be adequate Representation Information, and that the OAIS should preserve information against all reasonable contingencies, including the demise of the Archive.

The changes in the new draft now include explicitly that the Content Data Object being preserved may be

(1) kept by the Archive but may be changed or
(2) kept by the Archive unchanged or
(3) not kept by the Archive, but instead be handed on to another Archive

Each of these three imply the following:
In case (1) the Archive may Transform the Content Data Object
In case (2) the Archive may add Representation Information to ensure the Content Information is Independently Understandable
In case (3) the Archive may hand over the AIP which contains the Content Data Object
This change makes the text as a whole more consistent and clearer.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

VII. Updates to archive interoperability

A major change to the discussion of various possible types of archive interactions is the way in which the distribution of OAIS functionality may be described. Such a distribution of functionality could be motivated, for example, by cost reduction or the availability of a comprehensive functionality offer. These descriptions should allow archives to be described more accurately and make it even clearer that an OAIS has never been required to be a monolithic organisation.

The text describes some possible categories (not an exhaustive or mutually exclusive list) of Archive associations. The first set of three categories has successively higher degrees of organizational interaction:

- Independent: Archives motivated by local concerns with no management or technical interaction among them.
- Cooperating: Archives with potential common Producers, common submission standards, and common dissemination standards, but no common Finding Aids.
- Federated: Archives with both a Local Community (i.e., the original Designated Community served by the Archive) and a Global Community (i.e., an extended Designated Community) which has interests in the holdings of several OAIS Archives and has influenced those Archives to provide access to their holdings via one or more common Finding Aids.

Another set of categories, somewhat orthogonal to the previous set, differentiates according to how internal Archive functions and functional areas are addressed and by styles of resource sharing.

- *All In-house: Archives that perform all archival functions in-house.*
- *ïShared resources: Archives that have entered into agreements with other organizations to share resources, perhaps to reduce cost. This requires various standards internal to the Archive (such as ingest-storage and access-storage interface standards) but does not alter the user community's view of the Archive.*
- *ïDistributed: Archives that have distributed the OAIS functionality either geographically or organizationally. Different levels, forms and organization of the distribution are possible. In every case, the Archive is required to oversee and manage the Archive's use of the distributed functions, but does not alter the user community's view of the Archive*

An important classification of distribution is where the supporting organizations, which supply the required functionality, are themselves each an OAIS. One can describe the arrangement as a primary OAIS using one or more supporting OAIS for specific services. In such a case, each supporting OAIS, as well as the primary OAIS must fulfill all requirements for OAIS conformance, namely the Mandatory Responsibilities and support for the Information Model. Therefore, service level agreements are required to guarantee proper implementation of the functionality distribution. Particularly, the primary OAIS must monitor that the supporting OAIS is meeting its service agreement. The conformance of each supporting OAIS may be used as a piece of evidence.

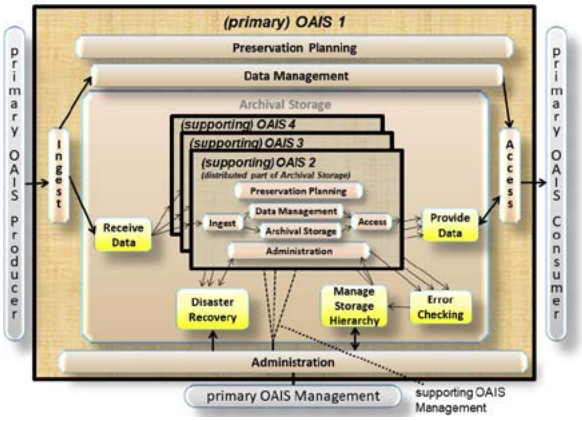


Figure VII-1 Primary/Supporting OAIS distributed functionality

The term 'Outer OAIS-Inner OAIS' has been used in the literature [3]. This usage is consistent with the "Outer OAIS" being the primary OAIS and the "Inner OAIS" being the supporting OAIS in cases where the "Outer OAIS" and "Inner OAIS" are each totally conformant to OAIS requirements. To exemplify the use of distributed functionality with supporting (inner) OAISes the Figure VII-1

shows how a set of supporting OAISes complete the functionality of the primary OAIS Archival Storage.

VIII. Conclusions

The updates made to the current version of OAIS, to be submitted for CCSDS and ISO review, provide significant clarifications and, when integrated into ISO 16363, improve the auditability of repositories, for example by giving auditors specific tests to verify understandability by using the Preservation Objectives, where they are available.

The changes add further clarity to OAIS and bring in a number of useful concepts developed by others since version 2 of OAIS was published. They will allow repositories to be described more clearly, despite increasing complexity. The consensus was that the updates will not require archives which are currently conformant to OAIS to make major changes but will instead allow such archives to provide evidence about their conformance more clearly. In addition, the changes should keep OAIS fit for purpose as archives are faced with new challenges in the coming years.

Acknowledgements

Thanks are due to all the participants in the CCSDS MOIMS-DAI working group and others, in particular those who contributed over 200 suggested changes to OAIS and helped in their resolution.

References

[1] Reference Model for an Open Archival Information System (OAIS). Magenta Book. Issue 2. June 2012, available from https://public.ccsds.org/Pubs/650x0m2.pdf also known as ISO 14721:2012.

[2] Data Archive Interoperability Working Group https://cwe.ccsds.org/moims/default.aspx#_MOIMS-DAI

[3] "Supporting Analysis and Audit of Collaborative OAIS's by use of an Outer OAIS – Inner OAIS (OO-IO) Model" by Eld Zierau and Nancy McGovern. In Proceedings of the 11th International Conference on Preservation of Digital Objects (iPres) 2014, pp. 209-218, available at http://www.ipres-conference.org/ipres14/sites/default/files/upload/iPres-Proceedings-final.pdf

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Data Stewards and Digital Preservation in Everyday Research Practice

Esther Plomp
TU Delft, Netherlands
E.Plomp@tudelft.nl
https://orcid.org/0000-0003-3625-1357

Yan Wang
TU Delft, Netherlands
y.wang-16@tudelft.nl
https://orcid.org/0000-0002-6317-7546

J.S. Love
TU Delft, Netherlands
j.s.love@tudelft.nl

Abstract – Data Stewards at TU Delft promote digital preservation by incorporating actions in everyday research practice. Typical activities include requirement scoping, tool selection and policy drafting, all of which are tailored to a specific group, project or faculty. Here we discuss the situation of the Data Stewards within the university and examples of preservation work, including the creation of data repositories and a trial of Webrecorder.

Keywords – data stewardship, architecture, design, Webrecorder, policy

Conference Topics – Collaboration: a Necessity, an Opportunity or a Luxury?

I. The TU Delft Data Stewards

Since the latter part of 2017, Delft University of Technology in the Netherlands has been running a pilot programme aimed at placing professional research data management (RDM) specialists into daily practice within university faculties [1]. These specialists hold the novel title of Data Steward, and they are charged with engaging research communities with research data issues and aiding them in defining what constitutes good data management within a specific discipline. Stewards are granted considerable latitude in identifying issues relevant to their communities as well as coming up with solutions on how to meet them and add value to daily research practices. As the entire domain of research data management falls within the Data Steward scope of interest, they partake of a broad range of data-related subjects: storage solutions, sharing technologies, legal & ethical concerns, etc. Digital preservation represents one key component of the Data Steward agenda with the stewards acting as mediators or facilitators, suggesting improvements to work practices and gaining insight on implementation which can be translated into feedback for further development.

A credo of the Data Stewards is 'incremental cultural change', so large-scale shifts in work patterns are neither common nor necessarily viewed as desirable [2]. Instead, as members of a disciplinary research community, stewards join projects at various stages and act as analysts and consultants to make recommendations for improved durability of their research materials. The recommendations vary to suit person and project and may include actions such as suggesting software and tools, providing assistance with data sharing or helping draft documentation. Much of the work is simple advocacy, however, and often takes the form of raising awareness of why open formats are preferable or why and how to secure backup copies of data.

At this conference we aim to present a selection of cases to demonstrate how the TU Delft Data Stewards weave digital preservation into ongoing research practices. These are by no means an exhaustive list, and we are constantly innovating and iterating to improve operations. As jacks- and jills-of-all-trades, the Data Stewards are also by design not experts in digital preservation, and we would welcome discussions on how to better educate

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

ourselves in the current tool sets available for digital preservation of research materials and how to translate this knowledge to our constituents.

II. Research Data Management Policies at TU Delft

Since June 2018, there has been a central policy for research data management at TU Delft [3]. Part of this policy is focused on making digital research data and outputs more accessible and sustainable. For example, the TU Delft library is expected to provide robust, high-quality infrastructure to facilitate data management and preservation. The TU Delft library hosts the DSA (Data Seal of Approval)-certified data repository, 4TU.ResearchData, that guarantees the availability of research data for at least fifteen years, in addition to advice and support for researchers in the preservation of their research outputs. The data repository has a dedicated team with the Data Stewards acting as satellites in their respective faculties.

In preparation for this central policy, a series of interviews and consultations were conducted with researchers from 2016 onward in order to gain a better understanding of prevailing perspectives on various issues relating to RDM [4]. It became clear from these consultations that practical guidance for diverse disciplinary practices was needed. In consultation with the Data Steward, members of each faculty at TU Delft are currently constructing faculty-specific RDM policies, in which the roles of the various faculty positions are outlined (principal investigators, PhD supervisors, PhDs, Department Heads and the Data Stewards). Researchers stand at the centre of the RDM policies and are consulted for their construction, an approach that is favoured at TU Delft [5]. In such a bottom-up approach, researchers are more inclined to invest their time and effort in long term preservation of their digital research materials [6]. Furthermore, the policy will reflect disciplinary practices and allow it to be a relevant document rather than being perceived as an administrative burden. The policies stimulate researchers to appropriately document and preserve their data by sharing their research outputs in a data repository (in accordance with the FAIR principles, [7]). For example, PhD students starting in 2019 will be required to deposit their underlying research, together with their PhD theses, before they are allowed to defend their thesis [3].

III. Reaching out to Researchers

Being a large technical university, TU Delft hosts eight faculties covering a wide variety of science and engineering disciplines, as well as numerous interdisciplinary research groups. This composition poses challenges in providing preservation support at the university level, since one-size-fits-all solutions would never cover all needs per faculty or even per department. By being embedded in each faculty, Data Stewards have the chance to work closely with researchers, discover the challenges in their RDM practices and subsequently understand the unique preservation requirements from each field via both top-down and bottom-up approaches. Top-down approaches include contacting researchers through the RDM policy engagement described above, faculty research councils, or graduate school and departmental executive secretaries. Bottom-up contact results from engagement with individual researchers on data management plans, providing researchers with information on RDM practices, and training sessions on tooling and long-term data archiving. Furthermore, researchers who advocate and implement good RDM practices are encouraged to become local Data Champions [8].

By reaching out to researchers, some insights into the varieties of research data they possess and challenges that they face are revealed, and it becomes clear that data diversity is often found within the same discipline. Figures 1 and 2 are word clouds representing primary research data types according to researchers in the fields of architecture and housing study respectively, both situated within the Faculty of Architecture and the Built Environment. They highlight the variation in research data even at the local/departmental level. Such overviews constitute essential input for building preservation infrastructures.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Figure 1 Research data types in architecture

Figure 2 Research data types in housing study

Another practical challenge faced by researchers is the lack of capacity in preparing datasets for publishing and archiving. Wide application of technology has led to increasing amounts of collaborative, interdisciplinary research. For instance, scanning technology developed in an aerospace engineering group could contribute to heritage studies, and satellite imagery processed by geoscience researchers can be used in research on disaster management. However, these research opportunities gives rise to new challenges concerning data structuring and preservation. Sometimes, perhaps even often, new projects start with insufficient technological knowledge and skills from both researchers and research support to perform the desired tasks. Given the array of technologies and research instruments available, some researchers are struggling with structuring data sources, cleaning datasets or preparing documentation. As part of ongoing efforts to address these issues, the Data Stewards organize training workshops for the stewards themselves and for researchers on basic software skills, such as Software Carpentry [12]. At the same time, an increasing number of projects also demand a more customized preservation environment to ensure the reuse and reproducibility of the research. The case on the digital humanities at TU Delft will explain more about this preservation challenge.

IV. Cases

A. *Webrecorder*

Since early December 2018, the Data Stewards have been conducting a small pilot aimed at preserving research websites using Webrecorder (http://webrecorder.io/). The primary goal of this pilot is to perform preservation actions on a set of content-rich digital-born research outputs (websites) which currently have a high risk for loss. Pilots of this variety are initiated ad hoc by instigation of the Data Steward.

For this pilot we compiled a quick survey of existing websites produced by research groups involving members of the Faculty of Industrial Design Engineering (IDE) at TU Delft. The list was not intended to be exhaustive but rather to identify representative work and can be expanded as desired. Inclusion criteria were broad, and consideration for the pilot required only that a website be somehow affiliated with or produced by a member of IDE and that it be hosted outside of the TU Delft network (pages within the TU Delft domain currently have an institutional preservation solution through Archiefweb: http://tudelft.archiefweb.eu/). The cursory inventory yielded almost twenty sites of varying content and complexity. Of these, four were selected for a test with Webrecorder. Webrecorder was selected for this purpose following a recommendation obtained at the 'Memory Makers' DPC event in Amsterdam in November 2018. WARC files for portions of all four sites were produced using Webrecorder. These provided a short-term preservation solution as well as tangible, usable products with which content producers could engage. These sites currently being tested are:

- CRISP, a design sector framework programme: http://selemca.camera-vu.nl/index.html
- Design & Emotion (D&E), a design society: http://www.designandemotion.org/
- DINED, an ongoing anthropometry production: https://dined.io.tudelft.nl/en
- 'Involving the CROWD …', a site for a conference held in 2016: https://museumsandcrowds.wordpress.com/

Of these four sites, it was determined that Webrecorder was a suitable tool for preserving two, namely the CRISP and Design & Emotion sites, both

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

of which are no longer currently being maintained but were deemed to possess continuing value. The 2016 conference site was judged to be temporary and not worth retention, while the DINED site continues to be developed and proved to be too complex in structure for Webrecorder to capture even superficially.

The most valuable insights from discussions with contributors to all four sites included the reasoning behind why these sites should be preserved (or not) and in what format. Within the design industry, CRISP was viewed as a flagship project, and its outputs continue to be consulted by practitioners in the field. Similarly the D&E Society provides historical context to the current Delft Institute of Positive Design. Webrecorder also sufficed to meet the demand that these sites remain integral and easily accessible online rather than as a complex set of HTML pages and associated media. At present the WARC files are accessible through the IDE Data Steward's personal Webrecorder account, and plans are being made to have a more centralized location as well as a backup copy of these sites in faculty archives.

B. Digital Humanities at TU Delft and GIS Platform

Although the primary research emphases at TU Delft lie in the sciences, a number of disparate research groups around the university also conduct research in the humanities. These include the history of architecture and urban planning, the Museum Futures Lab, preservation technology for imagery and heritage and a number of others. Unlike many other universities where the humanities have their own, separate research centers, these groups are embedded within major science and engineering disciplines. The demand for creating a digital humanities community across multiple engineering disciplines at TU Delft was recognized and picked up by the Data Stewards in November 2018. Since then, several community-driven events have been organized. This initiative is led by researchers from the heritage groups in the Faculty of Architecture and the Built Environment, and Data Stewards from a few faculties followed it closely with the purpose to steer discussions and capture requirements on digital preservation.

Within this new community, some group members have elected to generate specifications for a GIS platform. This platform is intended to supplement TU Delft's 4TU.Researchdata repository by allowing an expanded range of search mechanisms and improved visualizations of GIS data. In a requirement gathering meeting in early March 2019, several existing open source geoplatform solutions (such as QGIS https://www.qgis.org/en/site/about/index.html , GeoBlacklight https://geoblacklight.org/ , GeoNode http://geonode.org/gallery/), and current research projects involving spatial data were showcased and discussed. One insight derived from this activity is that maps are perceived as both research data and research outcomes by project members. Therefore both functions require consideration when designing preservation solutions. For an open repository of geographic data and related research outcomes, textual descriptions or simple metadata attached to the datasets were deemed insufficient for making datasets reusable. Instead, it was decided that potential users need a visualization of the datasets in order to understand their contents before downloading and re-using them. During the discussions of the platform, we were also able to have an open discussion about the merits of preparing datasets in open formats and that this was a more fundamental need beyond the software tools employed by any individual or group. With this understanding in place, or at least introduced, the envisioned geographic platform will be more accessible to any interested parties through the use of interoperable formats. Following a series of fruitful discussions, a list of features for the platform was compiled, which, together with the insights obtained, could serve as baseline requirements and initial input for other service providers at the university to proceed with selection and implementation.

V. Discussion and Future Work

The cases presented here represent only two examples of the types of digital preservation activities which are needed for research work at TU Delft. Data Stewards are in a fortunate position to gather and interpret further requirements as part of daily research, and they stand in a prime position to discover practical challenges to digital preservation within the university research environment as well as negotiate between researcher needs and the design of digital preservation solutions. Ideally this will all be done incrementally and iteratively in order

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

to discover and reinforce good practice wherever possible. As part of this process, the Data Stewards would like to seek out new avenues by which they might gain relevant knowledge concerning good practices recommended by preservation experts.

References

[1] Teperek, M. "Vision for Data Stewardship at TU Delft", 2018. Retrieved from http://doi.org/10.5281/zenodo.1251989

[2] Teperek, M., & Dunning, A. "The main obstacles to better research data management and sharing are cultural. But change is in our hands", November 14 2018. Retrieved from https://openworking.tudl.tudelft.nl/2018/11/14/the-main-obstacles-to-better-research-data-management-and-sharing-are-cultural-but-change-is-in-our-hands/

[3] Dunning, A. "TU Delft Research Data Framework Policy", 2018. Retrieved from https://doi.org/10.5281/zenodo.2573160

[4] Teperek, M., Cruz, M. J., Verbakel, E., Böhmer, J., & Dunning, A. "Data Stewardship addressing disciplinary data management needs." *International Journal of Digital Curation*, vol. 13 no. 1, pp. 141–149, 2018. https://doi.org/10.2218/ijdc.v13i1.604

[5] "TU Delft Strategic Framework 2018-2024", 2018. Retrieved from https://d1rkab7tlqy5f1.cloudfront.net/TUDelft/Over_TU_Delft/Strategie/Towards%20a%20new%20strategy/TU%20Delft%20Strategic%20Framework%202018-2024%20%28EN%29.pdf

[6] Teperek, M., Higman, R., & Kingsley, D. "Is Democracy the Right System? Collaborative Approaches to Building an Engaged RDM Community". *International Journal of Digital Curation*, vol. 12 no. 2, pp. 86–95, 2017. https://doi.org/10.2218/ijdc.v12i2.561

[7] Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* vol. 3, art. no. 160018, 2016. https://doi.org/10.1038/sdata.2016.18

[8] Plomp, E. "Data Champion kick off meeting", 14 January 2019. Retrieved from https://openworking.tudl.tudelft.nl/2019/01/14/data-champion-kick-off-meeting/

[9] M. King, B. Zhu, and S. Tang, "Optimal path planning," *Mobile Robots*, vol. 8, no. 2, pp. 520-531, March 2001.

[10] H. Simpson, *Dumb Robots*, 3rd ed., Springfield: UOS Press, 2004, pp.6-9.

[11] M. King and B. Zhu, "Gaming strategies," in Path Planning to the West, vol. II, S. Tang and M. King, Eds. Xian: Jiaoda Press, 1998, pp. 158-176.

[12] Kurapati, S., & Teperek, M. "4TU.Centre for Research Data partners with The Carpentries: Impressions from the first workshop at TU Delft". 14 March 2019. Retrieved from https://openworking.tudl.tudelft.nl/2018/12/16/4tu-centre-for-research-data-partners-with-the-carpentries-impressions-from-the-first-workshop-at-tu-delft/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

What is the Standard Format for Digitized Audio?

Approaches for Storing Complex Audio Objects

Nick Krabbenhoeft
New York Public Library
United States of America
nickkrabbenhoeft@nypl.org

Abstract – The best practices for representing analog audio with digital bitstreams are relatively clear. Sample the signal with 24 bits of resolution at 96KHz. The standards for storing the data are less clear, especially for media with complex configurations of faces, regions, and streams. Whether accomplished through metadata and/or file format, the strategy chosen to represent the complexity of the original media has long-term preservation implications. Best practice guides rarely document these edge cases and informal discussions with practitioners have revealed a wide range of practices. This paper aims to outline the specific challenges of representing complex audio objects after digitization and approaches that have been implemented but not widely adopted.

Keywords – Audio, Digitization, Object Modeling
Conference Topics – Collaboration: a Necessity, an Opportunity or a Luxury?; Building Capacity, Capability and Community

I. Introduction

In response to the deteriorating sustainability of magnetic media, many organizations are pursuing digitization as their preservation strategy for audio and video collections. The New York Public Library has opted for this approach and has digitized over 200,000 objects in order to maintain the accessibility of their contents past the deterioration of the original media and/or playback equipment.

According to the OAIS Framework, organizations are responsible for defining the specifications for SIPs and AIPs, including the Content Objects those packages contain. Community guidelines such as IASA TC-04 simplify the process of defining Content Objects. For example, in order to represent the original audio signal at an equal or higher fidelity to what human ears can distinguish (20kHz), guidelines recommend sampling audio signals at a minimum of 44.1 kHz and at even higher rates to capture qualities of the recording medium outside the auditory range.

Best practice documents are less exact on how to store the bitstreams. Recommendations to keep audio signals as uncompressed PCM streams wrapped in a Wave or Broadcast Wave format leave room for interpretation. Some workflows store left and right stereo tracks in separate files while others may interleave them into a single file. Some workflows limit audio file sizes to 4 GB[1] while others use different file formats for long audio streams.

Reviews of the audio digitization literature have shown relatively little guidance on questions like this, and informal conversations have revealed a range of approaches. IASA TC-04 devotes three paragraphs in total to target formats. [1] In the Sound Directions project, Indiana University and Harvard University documented their approaches in greater detail, but they did differ. [2] As the scale of digitization increases, the number of situations not addressed within guidelines increase as well.

[1] The Wave file format based on the Resource Interchange File Format (RIFF), which allocates bytes 4-7 to specifying the file size. This limits the size to 2^32 bytes (about 4.295 GB). RF64 extension defined in EBU 3306 [7], allows for daisy chaining of additional audio data in 18 EB chunks.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

This paper presents edge cases in digitized audio file specifications as encountered by the New York Public Library and documents potential options in hopes of spurring more public discussion.

II. The Challenge of Complex Audio Objects

Magnetic media is composed of metallic particles attached to a flexible tape by a binder. This composition does not inherently limit how information is recorded to the media. Audio may be stored as a stream of information in any location or orientation along this magnetic tape. The dependence on equipment for recording and playback equipment restricts the possibilities, but there is still great variety possible in the usage of a given format.

For example, the Compact Cassette format initially debuted in the 1960s as a format to record dictation. Early machines record a sequence of audio linearly within the upper or lower half of the tape. At the end of the tape, the cassette is flipped and audio is recorded to the other half. (Table 1.A)

Stereo content such as commercial music has a very similar layout, except the area used to record a single stream of mono audio is divided into narrower areas for left and right channels with a gap in between. (Table 1.B)

Other machines allow Compact Cassettes to be used as relatively low-cost studio recorders. Up to 4 inputs can be recorded simultaneously onto a tape, each perhaps representing an instrument like vocals, guitar, bass, and drums. The areas are arranged much like a 2-sided stereo cassette, except the tape is recorded in only one direction. (Table 1.C)

Finally, layouts can be a mixture of the above examples. Any machine that supports Compact Cassettes can record to them regardless of their prior use. For example, a tape first used in a dictation machine and then used to record music from the radio would have a mixture of mono and stereo arrangements. (Table 1.D)

Discovering and responding appropriately to the layout of audio is an important skill of audio engineers engaged in preservation. Each portion of the layout must be extracted with machinery appropriate to the layout of the recorded signal. Colloquially, a number of terms are used, such as streams, tracks, and channels. Frustratingly, these terms are imprecise in usage. What some may consider 2 mono tracks other may call 1 stereo channel. This paper uses the following terminology as defined in AES-57. [3]

1. Stream - a single linear sequence of audio signals
2. Region - a group of streams to be played back synchronously
3. Face - a group of regions to be played back sequentially

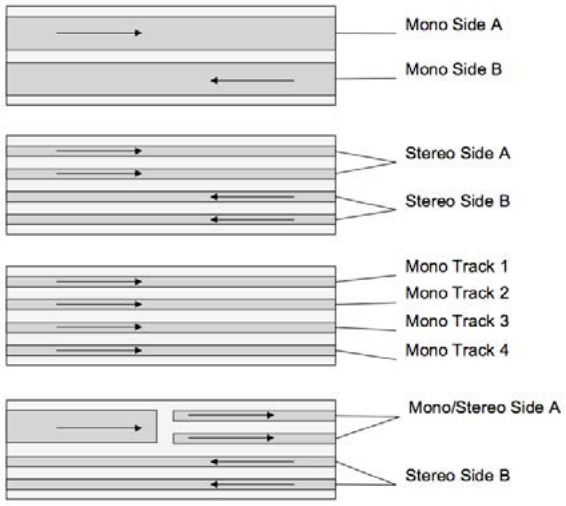


Table 1. Increasingly complex example layouts of audio on a Compact Cassette.
A dictation tape
A commercially released

Using those terms, the layouts in Table 1 would be described as follows:

A. 2 faces (Side A and B) each with 1 region and 1 stream
B. 2 faces each with 1 region and 2 streams
C. 1 face with 1 region and 4 streams
D. 2 faces. The first face has 2 regions. The first region has 1 stream and the second has 2 streams. The second face has 1 region and 2 streams
E. Strategies for Representing Complex Audio Objects

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The complexity of Table 1.D demonstrates how a few choices made during recording can create intricate branching relationships. This is matched by the ability of collecting organizations to make choices on how to transform it into Content Information. Reviewed guidelines do not prescribe specific strategies. This section introduces potential strategies that use a combination of documentation and file formats.

To simplify this discussion, options will be illustrated through example objects with the following layouts:

- 1 face, 1 region, 1 stream
- 1 face, 1 region, multiple streams
- 1 face, multiple regions
- Multiple faces

1 face, 1 region, 1 stream

Example Item: oration on an open-reel audio tape recorded as mono

Even in this base example, the Wave and AIFF formats generally recommended for use may not be appropriate due to technical limitations.

The base specification for Wave and AIFF files stores the total file size as a 4-byte, unsigned integer in bytes 5-8 of the file. [4] [5] As a result, these formats are limited to a valid file size of 2^32 bytes (4 GiB or 4.295 GB) or 4 hours and 9 minutes of audio digitized at the typical digitization specifications of 96 kHz/24-bit. This length of mono is rare, but NYPL holds examples such as recordings of long-form speeches.

Assuming that a digitization program will produce valid files and retain a given audio, there are two potential strategies.

1. split the audio across multiple files
2. use a different file format

Creating multiple files per audio stream requires two additional considerations. First, the choices of how to divide the data such as appropriate point in the stream and whether or not to include overlap. These are not trivial choices and may require discussions with engineers on a case-by-case basis. Second, there are multiple methods to store the relationship between the files.

Splitting audio data across multiple files means the relationship between the object and the files is no longer obvious. To address this, a file naming convention can be modified to include the information, for example "part1", "part2", etc, but this should not be the sole form of metadata. As advised in Sound Directions, "filenames are not a reliable means of storing information." Filenames are directly editable from the file manager level as opposed to embedded metadata or metadata stored in sidecar files, and so they are more vulnerable to keying errors and accidental edits.

Major audiovisual metadata standards such as AES-57, PBCore, EBUCore, and AudioMD do not directly address situation, but generic structures within them that define one-to-many object-to-sub-object relationships could be applied. For example, the <pbcoreInstantiation> element can be used to describe any "unit that typically (though not always) comprises a whole representation of the asset." [6] Similarly, the <file> element in METS could be used to document the relationship between these files. [7] But, these methods are generally hypothetical and do not appear in literature. With digitization dating back to 2005, NYPL's metadata strategy is roughly based on AES-57. [8]

The second option for storing long durations of audio is to use another format. The Interchange File Format (IFF) that underlies AIFF and WAV was originally created in 1985. The technical constraints and assumptions of that era were fossilized into the specifications, file size being the most obvious of these.[1]

Extension specifications have been created for both AIFF and Wave that expand the total possible file size to 16EiB (roughly 2 million years of 96/24 audio). Sony published the Wave64 extension in 2003. Apple published the CAF extension for AIFF in 2005. and EBU published the RF64 extension in 2008. [9] [10] However, none of these extensions have received universal uptake in software used for audio digitization as an export option and many still consider them to be entirely different formats.

As an example, trying to export 4GB of audio

[1] New formats based on IFF continue to be developed and supported. For example, the WebP image format published by Google in 2011 has the same hard-coded 4GB size limit.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

from Audacity to Wave or AIFF prompts an error message. To create an RF64 or CAF, a user must find a separate export menu that lists those formats as options. Other programs default to ignoring the IFF file size metadata limitation and create invalid files in the base format specification.

Another possibility is to use another format entirely. More recent formats such as FLAC, MXF, and Matroska address the size issue in their base specification. [11] [12] 13] Again, there is uneven support for these formats within software used for audio digitization. However if they are supported, the file size issue is unambiguous audio unlike RF64, and they enable additional preservation friendly features such as embedded checksums and lossless compression. [14]

1 face, 1 region, multiple streams

Example Item: studio recording on an open-reel audio tape with 24 simultaneous tracks for different instruments

In audio file formats, streams are typically stored as channels corresponding to the expected speaker output. Wave and AIFF both natively support the most common multi-stream arrangement, stereo. But, they are poorly suited for storing more streams. Because the file size limit applies to the entire file, total possible duration decreases inversely to the number of streams. A WAV can hold roughly 4 hours of a single stream of 96/24 audio but only 1 hour of 4 simultaneous streams of 96/24 audio. Additionally, because larger number of streams are associated with surround sound speaker setups, the default file interpretation may not match the context of the original audio.

For example, in music studio production use, a stream or group of streams would capture a single instrument or voice during a recording session. This allows the instruments to be edited individually before being mixed down into a single stereo song. These streams are not intended to be played simultaneously without further mixing

Other formats support additional more complex layouts, including the MWBF extension to Wave, MXF, and Matroska.

The formats allow for further abstraction of audio arrangement through a concept called tracks. Multiple streams can be grouped as channels within a track separate from other streams while maintaining a synched timing. For example in a studio recording, instruments may be captured as a mixture of mono and stereo. Tracks can be used to keep organize this data within the file. As with any format choice, the biggest hurdle is ensuring export support from authoring software.

In production workflows, a common strategy is to save each stream to its own file. Used in a preservation workflow, this avoids format support issues, in exchange for requiring a metadata schema that records the relationships between files. It also requires specifying when to employ this strategy. Stereo audio is a multi-stream format. Interpreted stringently, a 1-file-per-stream strategy would save left and right audio streams were saved to separate files, instead of interleaving them.

1 face, multiple regions

Example Item: open-reel audio tape used to record sessions of dictation (mono) at different speeds

For media with regions, engineers must adjust the setup of playback equipment in accordance with the changing characteristics of the layout. Each of the changes, such as swapping a mono head for a 4-track head or adjusting the playback speed from 7.5 inches per second to 15 inches per second, requires stopping the playback process. Many workflows also require capturing the following audio as a new digital object.

Both Wave and AIFF support only a single audio chunk per file. If it is important to maintain a distinction between audio data from different regions, Wave and AIFF require creating a file per region. As with other multi-file strategies this also requires support in the metadata schema for maintaining the relationship between files. An interesting feature in Broadcast Wave is the TimeReference field that can be used to record the temporal relationship between two files on a shared timeline. [15

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Container formats such as MXF and Matroska define an abstraction to demarcate playback often called a chapter. This provides the ability to sequence playback of the tracks within the container using chapter metadata. However, chapters assume sequential playback. During digitization, engineers will overlap the beginning and ends of neighboring regions to ensure total information capture. Experiments with container formats have not yielded a strategy for creating an unambiguous shared timeline within the container.

Multiple faces

Example Item: open-reel audio tape recorded in mono across four tracks

Faces have a sequential relationship, and the recommended strategy has been to store each face as a separate file. Although storing faces as chapters in a container file is a potential strategy, the difficulties in using chapters for regions would greatly complicate the representation of any audio with both faces and regions.

III. Discussion

There is a garden of forking paths when it comes to storing digitized audio. It would be helpful for digitization guidelines to go past 96/24 BWF recommendation and present options for file structures, but examples are difficult to find in the literature. Greater discussion and documentation of the approaches above would be particularly useful for two communities, digitization labs and repository developers.

In the first instance, the support for custom metadata formats, embedded metadata, Wave extensions, and container formats varies across digitization software and vendors. If every collecting institution chooses its own combination of strategies, labs are forced to support a wide range of strategies, increasing expense and likelihood of confusion or errors. After digitizing materials through in-house and vendor workflows, complex audio configurations is still a difficult class of media to design QC processes for. Documentation of even a few shared strategies would greatly simplify target selection for collecting organizations and support for labs.

In the second instance, representing the semantic relationship between files is one of the most challenging aspects of repository development. Documenting edge cases and migrating from previous strategies occupy outsized portions of time. Again, complex audio has presented a particular challenge for the development of ingest workflows at NYPL and, based on conversations, at other institutions as well.

While all of the summarized strategies are viable, it is from this perspective that the author finds container formats to be most worth investigation. NYPL has experimented with using the Matroska format to store 24 tracks of mono audio from a studio recording in a single file with an image of the track-listing. Doing so proved to be far simpler for object modeling than storing the relational metadata in a sidecar and developing a parser. However, as an experiment, it bears examination if such strategies impede access in the future.

This paper has discussed only strategies of how to reflect the structure of the physical object in a digital form. It does not discuss how intellectual content interacts with this organization. The layout of intellectual content might be entirely defined by the physical layout, such as the two sides of a tape being used to record different meetings. It may cut across the layout, such as a speech captured across two regions when it became the recording speed had to be lowered before the tape ran out. It may exist within the layout, like songs on a compact cassette. And it is most often a combination of the two. For the preservation of the original media, this paper advocates the primacy of the physical layout in creating digital objects while leaving the intellectual layout to presentation frameworks such as IIIF.

IV. Conclusion

This paper is a provocation to discuss and document how digitization projects encode and package outputs. It does not believe there is a single optimal strategy but hopes that as the scale audio digitization continues increasing and classes of edge start numbering in the thousands that common strategies may be developed.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

References

[1] TC-04 Guidelines on the Production and Preservation of Digital Audio Objects (web edition). IASA, https://www.iasa-web.org/tc04/key-digital-principles.

[2] Sound Directions. Indiana University and Harvard University. http://www.dlib.indiana.edu/projects/sounddirections/.

[3] AES 57 AES standard for audio metadata - Audio object structures for preservation and restoration. Audio Engineering Society. http://www.aes.org/publications/standards/search.cfm?docID=84

[4] Wave 1.0 Specification. http://www-mmsp.ece.mcgill.ca/Documents/AudioFormats/WAVE/Docs/riffmci.pdf

[5] AIFF 1.3 Specification. http://www-mmsp.ece.mcgill.ca/Documents/AudioFormats/AIFF/Docs/AIFF-1.3.pdf

[6] PBCore Part Type, PBCore. https://pbcore.org/elements/pbcorepart

[7] METS. http://www.loc.gov/standards/mets/

[8] AMI Metadata Analog Reel Sample, NYPL. https://github.com/NYPL/ami-metadata/blob/master/versions/2.0/sample/sample_digitized_audioreelanalog.json#L48

[9] Multiple Channel Audio and Wave Files, Microsoft. http://www.microsoft.com/whdc/device/audio/multichaud.mspx

[10] EBU 3306 MBWF/RF64, European Broadcasting Union. https://tech.ebu.ch/publications/tech3306

[11] FLAC Specification. https://xiph.org/flac/format.html

[12] SMPTE MXF RDD 48. http://www.digitizationguidelines.gov/guidelines/rdd48-2018_published.pdf

[13] Matroska Specification. https://www.matroska.org/technical/specs/index.html

[14] Rice, Dave. Reconsidering the Checksum for Audiovisual Preservation. https://dericed.com/papers/reconsidering-the-checksum-for-audiovisual-preservation/

[15] A Primer on the Use of TimeReference, AVP. https://www.avpreserve.com/wp-content/uploads/2017/07/AVPS_TimeReference_Primer.pdf

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The Matterhorn RDF Data Model

Formalizing Archival Metadata With SHACL

Tobias Wildi
docuteam GmbH
Baden, Switzerland
t.wildi@docuteam.ch

Alain Dubois
Archives de l'Etat du Valais
Sion, Switzerland
alain.dubois@admin.vs.ch

Matterhorn RDF is a linked data-based model for archival metadata with the goal of improving the contextualization of archival records. It covers the three standards ISAD(G), ISAAR(CPF) and ISDF, as well as the areas "Preservation Description Information" and "Representation Information" of the OAIS information model. For the implementation of Matterhorn RDF, classes and properties of existing ontologies are used. The formalization of the model is realized with the help of SHACL shapes. [1]

Keywords: Archival metadata model, linked data, ontology, SHACL, RiC, contextualization

Conference Topics: Exploring New Horizons.

I. Introduction

This paper describes a model for archival metadata based on semantic technologies. The model represents both descriptive and technical metadata, specifically the standards ISAD(G), ISAAR (CPF) and ISDF of the International Council on Archives (ICA), as well as "Representation Information" and "Preservation Description Information" from the OAIS information model. The model also takes into account the current work of the ICA's Expert Group on Archival Description (EGAD), but chooses a different design approach than their conceptual model Records in Context (RiC).

The first part of this document defines the goal and scope of Matterhorn RDF. The second part substantiates why semantic technologies are used for the model and how they eliminate the disadvantages of today's XML-based data models. The third part outlines the design principles of Matterhorn RDF. This includes the decision not to develop an new ontology but rather exclusively use classes and properties of existing ontologies. The Shapes Constraint Language (SHACL) is used to formalize and validate Matterhorn RDF. The fourth and fifth parts explain the concept model and the class model of Matterhorn RDF. The most important and at the same time unspectacular finding of both these parts is the realisation that the innovation of Matterhorn RDF lies in the adaptation of existing models and ontologies for use in archives. The last part provides an outlook on the potential of Matterhorn RDF in terms of its technical implementation.

II. Improved contextualization as a goal

Archival metadata have the function of keeping the context in which documents were created comprehensible over a long period of time. Archival material has to be placed in a context to have any value. Thus, documents are contextualised through the description of their content (What?), the actors involved (Who?) and the process of creation (How?). The triangle of what, who and how has been covered to date by the three standards ISAD(G), ISAAR (CPF) and ISDF. While EAD and EAC can be coded in XML, the same is not true for ISDF. The three standards were developed by ICA over several years, with the result that they partly overlap and it is now unclear as to how relationships between them are to be mapped. The aim of Matterhorn RDF is firstly to ensure the encoding of the three standards and secondly to show how relationships between them can be modelled.

The need to revise, standardize and improve the relationship between the existing standards also manifested itself within the ICA. The Expert Group on Archival Description (EGAD) was founded in 2012 with the task of developing a new model under the title "Records in Context". Matterhorn RDF is not to be seen as an alternative to RiC, but rather seeks to

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

elaborate the RiC concept model in a future version, taking into account, however, different design considerations to those which EGAD currently implements.

The perimeter of Matterhorn RDF goes beyond descriptive metadata: the model also includes technical metadata necessary for the long-term preservation of digital objects. These are "Preservation Description Information" and "Representation Information" from the OAIS information model. Matterhorn RDF thus lays the foundation for a model that contains both the content and the technical contextualization of a record.

III. Semantic technologies instead of XML

Matterhorn METS, the predecessor of Matterhorn RDF, was registered with the Library of Congress in 2012 in the form of a METS profile. [2] Today, Matterhorn METS is used by around 25 institutions in Switzerland, Germany and France. This XML-based model is based on the standards METS, PREMIS, EAD and EAC. [3]

The modelling of archival metadata in XML leads to problems in the technical implementation for several reasons. Firstly, the typical hierarchies for archives (tectonics) generate deeply nested structures in XML. Secondly, the two standards EAD and PREMIS require elaborate XML constructs compared with the information actually transported. Thirdly, the use of persistent identifiers in XML is by no means self-evident and must be explicitly specified.

For a successor model, semantic technologies were the obvious choice in order to simplify structures and better model relationships between individual resources. There were three reasons for using Linked Data. Firstly, each resource can be uniquely identified using a URI. This is an advantage over the original XML-based approach, where identifiers were unique only within a single METS file. Secondly, the relationships between resources can be qualified. For example, not only is a relationship between two people propagated, the relationship is additionally qualified with the help of so called predicates like "child of" or "married to". The third and most important reason is that the use of external resources and knowledge sources for cataloguing is greatly simplified. Archival cataloguing today largely consists of filling in free text fields in database applications. In contrast to library cataloguing, this procedure is less systematic. With Linked Data, the full text description is at least partially replaced by linking to already existing knowledge sources. These can be entries in Wikidata, GND or VIAF, for example, each of which can be uniquely referenced via a URI. The reference to long-term stable external resources promotes the efficiency and accuracy of archive cataloguing. And vice versa, resources in one's own archive can be used much more easily by third parties.

IV. Design Principles of Matterhorn

The central design principle of Matterhorn RDF is that, as a linked data-based model, it does not have its own ontology. The model is based exclusively on classes and properties of existing ontologies. It regroups and correlates them with each other using a conceptual model. This design principle is derived from the Best Practices for Publishing Linked Data of the W3C, which state: "Standardized vocabularies should be reused as much as possible". [4] State actors, including many archives, are especially called to account: "Government publishers are encouraged to use standardized vocabularies rather than reinventing the wheel, wherever possible."

The decision not to create a domain-specific ontology for archival metadata allowed for the development of a data model in a relatively short period of time and resource-saving manner. The fact that no data dictionary had to be written in order to precisely execute the semantic meaning of each property, was especially time-saving. It was sufficient to refer to the descriptions of the respective ontologies.

V. Overview and most important elements

The Matterhorn RDF model is conceptually based on the three standards ISAD(G), ISAAR (CPF) and ISDF,[1] as well as the specifications for Preservation Description Information and Representation Information from the OAIS information model. The model is very similar to the PREMIS3 ontology and works with the following three core classes:

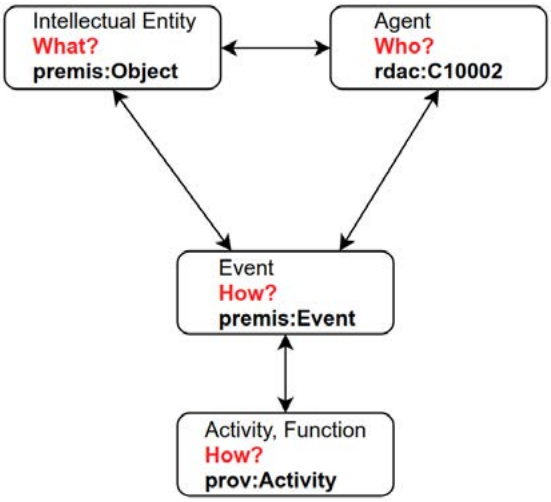
[1] As soon as RiC is consolidated, the RiC concept model will be implemented in the next version of Matterhorn RDF.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

- **Intellectual Entities (Records):** premis:object from PREMIS3 ontology
- **Agents:** rdac:C10002 from RDA ontology
- **Functions and Events:** prov:Activity from PROV ontology of the W3C

These classes are structured hierarchically into subclasses.

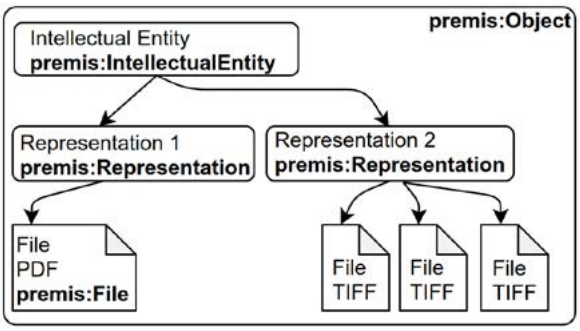
The classes are related as follows:



Only some of the used properties in the Matterhorn RDF model come from the ontologies of the corresponding classes. In addition, properties from Dublin Core, Ebucore or the standardized preservation vocabularies of the Library of Congress are used. The PREMIS standard does not include properties for descriptive metadata, therefore, attributes from other ontologies must be used. Dublin Core, Dublin Core Terms and RDA (Resource Description and Access) contain attributes that semantically correspond to the respective ISAD(G) fields.

The description of archival content takes place in the premis:IntellectualEntity class, a subclass of premis:Object. Intellectual entities are brought into a hierarchical relationship to each other via "has part" relationships, thus modelling the ISAD(G) tectonic. Horizontal or associative relationships between intellectual entities can also be modelled. An important feature is that a record or a single intellectual entity can be displayed by several representations at the same time. For example, a text document (= Intellectual Entity) can be represented by a PDF file as well as several TIFF files. To model this, the two following premis:Object subclasses, premis: Representation and premis:File, are used. These subclasses do not contain any descriptive metadata, they do, however, contain technical metadata from the PREMIS ontology. Thus, descriptive and technical metadata are combined in a single data model.

The graphical representation is as follows:



The actors defined by ISAAR(CPF) are represented in the class rdac:C10002. RDA is a set of library rules for cataloguing and publishing. [5] With FRBR, RDA has its own data model, which we are not concerned with in this context, because it is based on the concept of the "work", which is relevant only to libraries and has no bearing on archives. The part of the RDA-Ontology concerningthe so-called "Agent Properties" is, however, semantically largely congruent with the ISAAR (CPF)-Standard. Therefore, the already existing and widespread RDA-Ontology can be used to encode ISAAR (CPF). The class rdac:C10002 includes "A person, family, or corporate body", i.e. exactly the same concepts as ISAAR (CPF). Corresponding properties to the ISAAR(CPF) fields can be found in RDA and for auxiliary fields (versioning, language etc.) Matterhorn RDF uses the Dublin Core Terms ontology.

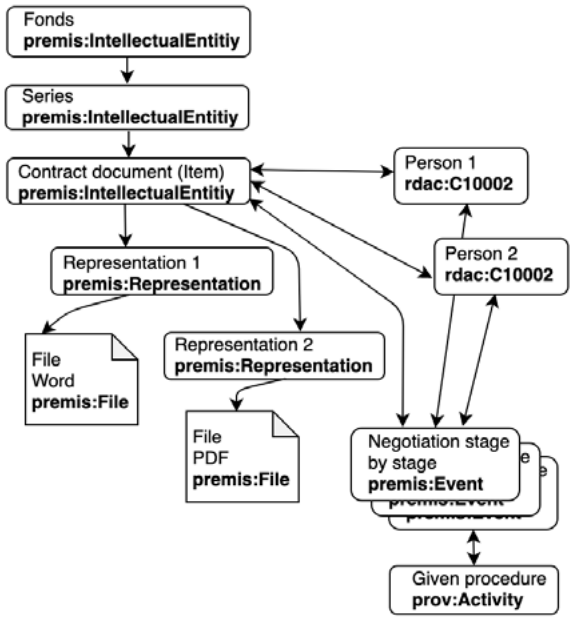
Functions, i.e. administrative tasks, processes and activities, are described with the help of the ISDF standard. These metadata form the basis for documenting the creation (and use) of records. The PROV data model and the PROV ontology of the W3C can be used to implement this. PROV is widely used and recommended by W3C for the modelling of "entities, activities and people". Matterhorn RDF, however, exclusively uses PROV's area of activities.

Two prov:Activity-subclasses model the process description on the one hand and the process documentation on the other hand. In prov:Activity the generic description of a business process or administrative procedure can be found in the form of a

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

sequence of various related activities. An activity is a generic concept for the work that a person or organization performs. It can stand alone or be composed of sub-activities. In the premis:Events class, a subclass of prov:Activity, the actual course of a business process is documented by means of individual events.

The negotiation of a fictitious contract between two persons shall give an exemplary illustration of the entire model. The content of our contract document is described using the premis:IntellectualEntitiy class. There are two representations of the contract document (premis:Representation), a first premis:File in the form of a word file and a second premis:File in the form of a PDF. The premis:File class also stores technical metadata such as checksums and file format information. The contract was signed by two persons who are described using the rdac:C10002 class. The negotiation of the contract followed a given procedure, which is stored in prov:Activity. Each step in this process, including several rounds of negotiation, is documented in premis:Events. This provides us with metadata for our contract on all three questions What, Who and How, as well as technical metadata that form the basis for Preservation Planning. Thus, the contract is put into context and its creation is documented in a comprehensible way.



VI. Formalization and validation

Matterhorn RDF does not formulate its own ontology. The development and ongoing maintenance of a new ontology requires much time and effort. Nevertheless, it is possible to formalize the model. This should entail a description of the classes the model consists of as well as the definition of the necessary properties and their purpose. For each property, restrictions regarding value ranges, minimum or maximum occurence and data types are to be formulated. For XML-based data models the proven schema language is available for this purpose. For semantic models the equivalent Shapes Constraint Language (SHACL) has been available since 2017. [6] [7] SHACL is used to formulate so-called shapes, against which the statements made in the RDF triples are validated. The formulation of shapes is therefore an elegant way to describe an RDF-based data model built on existing classes.

The shapes are published online. [8] The development of the shapes for all elements of Matterhorn RDF should be completed by the end of 2019. The following example of the ISAD(G)-field “Title” of will show how such a shape looks like.

```
sh:property [
        sh:path dc:title ;
        rdfs:label “Title”@en ;
        rdfs:label “Titel”@de ;
        rdfs:label “Titre”@fr ;
        rdfs:comment “ISAD 1.2” ;
        owl:sameAs rico:title ;
        sh:datatype xsd:string ;
        sh:minCount 1 ;
        sh:maxCount 1 ;
        sh:nodeKind sh:Literal ;
] .
```

In this shape the property “dc:title” is specified in greater detail. The labels of the title field are defined in different languages, an important functionality for a multilingual country like Switzerland. A comment line refers to the ISAD(G) standard field 1.2. A further reference to the corresponding field in the RiC ontology is made with the help of owl:sameAs. The title field has to contain data of the type “string” and must appear exactly once. The entry of a value (“Literal”) is expected and not a reference to another node (“IRI” or “IRIOrLiteral”).

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

VII. Conclusion

The Expert Group on Archival Description (EGAD) is currently formulating its own ontology with RiC-O for the "Records in Context" concept model. With Matterhorn RDF we propose an alternative way to contextualize records. Our model is based on already existing and widely used ontologies, which brings an increase in efficiency not only in the development but especially in the maintenance of the model. The model can be formalized even without an ontology of one's own. SHACL is a suitable tool for this purpose. Matterhorn RDF and RiC-O should not be competing models. By using the SHACL-shapes to store the semantic equivalents of RiC-O, the matterhorn RDF-model ensures the necessary crosswalk between the two models.

The transition from encoding archival metadata in XML or relational databases to linked data-based solutions will fundamentally change the way archives are described. Today, the primary access to archival material takes place through a single hierarchy structured according to ISAD(G). In the future, access and entry points will also be possible via actors or business processes. The origin context of records is therefore no longer documented in rigid, non-adaptable XML schemas but in a flexibly extendable model.

The activity of archival description is shifting away from a barely systematized textual description in free text database fields towards linking archival content to already existing and clearly referenceable knowledge resources. The search and access to the archive will also change. Today's full text search for terms and character patterns is being replaced by structured access to clearly identifiable resources.

Matterhorn RDF is thus a new approach to encoding and modeling archival metadata. The innovation lies in the new combination of existing ontologies for the contextualization of records in archives and in the fact that both descriptive and technical metadata are mapped with the model.

References

[1] All information about Matterhorn RDF is available at http://matterhorn.tools

[2] Matterhorn METS profile: http://www.loc.gov/standards/mets/profiles/00000041.xml

[3] Wildi, Tobias. Spezifikation Matterhorn METS. Baden, 2017. http://matterhorn.tools/matterhorn-mets.pdf

[4] Best Practices for Publishing Linked Data. W3C Working Group Note 09 January 2014: https://www.w3.org/TR/ld-bp/#VOCABULARIES

[5] RDA Ressource Description & Access: http://www.rdaregistry.info/

[6] Shapes Constraint Language (SHACL). W3C Recommendation 20 July 2017: https://www.w3.org/TR/shacl/

[7] For validation of RDF see: Ying Ding, Paul Groth, "Validating RDF Data". Morgan & Claypool Publishers, 2017.

[8] https://bitbucket.org/docuteam/matterhorn

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

An Overview of the Digital Preservation Storage Criteria and Usage Guide

Eld Zierau
Royal Danish Library, Denmark
elzi@kb.dk
0000-0003-3406-3555

Nancy Y McGovern
Massachusetts Institute of Technology, USA
nancymcg@mit.edu
0000-0002-7733-1516

Sibyl Schaefer
University of California, San Diego, USA
sschaefer@ucsd.edu
0000-0002-7292-9287

Andrea Goethals
National Library of New Zealand, New Zealand
Andrea.Goethals@dia.govt.nz
0000-0002-5254-9818

Abstract – The Digital Preservation Storage Criteria (or "Criteria") resulted from a community discussion at iPres 2015 on providing guidance to organizations that either use or provide digital preservation storage. First developed in 2016, they have been refined in iterative versions over the last three years based on feedback gathered at conference sessions and through a survey. The Criteria are intended to help with developing requirements for, or evaluations of, preservation storage solutions; to seed discussions about preservation storage; or to use within digital preservation instructional material. The latest version of the Criteria contains sixty-one criteria grouped into eight categories: content integrity, cost considerations, flexibility, information security, resilience, scalability & performance, support, and transparency.

The key new development since the Criteria was presented at the iPRES 2018 workshop is a usage guide, developed to accompany the Criteria. It includes sections on key topics to consider for preservation storage in addition to the Criteria: risk management, independence, elements in establishing bit safety, and cost considerations. The usage guide will be released publicly for review as one of the next steps in the project, along with developing version 4 of the Criteria and taking steps to further build the community around the Criteria.

Keywords – digital preservation storage, archival storage, criteria, risk management

Conference Topics – Designing and delivering sustainable digital preservation; The cutting edge: technical infrastructure and implementation; Collaboration: a Necessity, an Opportunity or a Luxury?

I. Introduction and Background

The Digital Preservation Storage Criteria (or "Criteria") are a result of a collaborative process based within the digital preservation community. This paper provides some context that traces the development and implementation of the Criteria and looks ahead to current and possible future developments. The development of the Criteria has involved iterative cycles of definition and elaboration by a working group, followed by opportunities for community review and feedback, and then finally the integration of community feedback into a series of versions that are publicly available on a project website [1]. Since the advent of computers, storage and processing capacity have framed the development and evolution of preservation strategies; the Criteria are meant to address evolving organizational requirements as digital preservation programs mature, as technological options emerge and evolve, and as opportunities and challenges become clearer.

A. Definition of Digital Preservation Storage

One of the prerequisites for identifying and elaborating the Criteria was developing a working definition of Preservation Storage, absent a shared and authoritative definition within the digital preservation

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

community. Defining "digital preservation storage" requires first defining "digital preservation." The definition adopted as a starting point is from the Digital Preservation Coalition: "the series of managed activities necessary to ensure continued access to digital materials for as long as necessary" [2].

Building on this base definition, the working definition of digital preservation storage for the purposes of the Criteria is: a fundamental component of digital preservation that supports and enables ongoing digital preservation activities. The term digital preservation storage encompasses the functions of the OAIS [3] functional entity *Archival Storage* as well as related OAIS functional entities that are needed to store, maintain in storage, and retrieve Archival Information Packages (AIPs) from storage [4].

For example, preservation storage includes parts of the following:

- ***Preservation Planning*** responsible for monitoring technology for storage options, relevant standards and practices, and media migrations.
- ***Data Management*** that ensures the relationship between preserved content and its associated metadata.
- ***Administration*** concerned with policies and standards pertaining to preservation storage management.
- ***Ingest*** concerned with the coordination of input and updates to different data replicas in storage.

The Criteria are intended to continually enable the digital preservation community to weigh the potential opportunities and risks of modern storage services and options while addressing the expectations of modern digital preservation practices.

B. Background on the Criteria Creation

The roots of the Criteria trace back to an initial digital preservation community discussion of digital preservation storage that was convened by the iPres 2015 conference organizers, which in part highlighted the lack of a guiding document related to preservation storage. Several of the participants then put forward a call for volunteers to establish a working group to design a set of preservation storage requirements. It quickly became clear that "requirements" would vary from organization to organization, and thus were unrealistic and unhelpful to outline. What was helpful was a list of criteria from which to select and further develop into specific requirements. Thus, the Criteria were born.

The working group culled requirements from several Requests for Proposals that they had used in various organizational settings, and then abstracted specific requirements into more general criteria. In preparation for the 2016 iPres workshop on the Criteria, the working group listed these starter criteria in a survey that was delivered to workshop participants prior to the conference. The survey asked participants to rank each criterion according to their value. This activity was successful in getting the participants to engage deeply with the Criteria and the result was a productive conversation during the workshop. The feedback generated in this iPres workshop, as well as during an earlier workshop held at the annual Library of Congress Designing Storage Architectures meeting, was then incorporated into the second version of the Criteria.

The Criteria working group then used this same pattern -- revision of the Criteria, presentation and workshopping of them at iPres and the Library of Congress Designing Storage Architectures meetings, followed by incorporating feedback to create a new version -- during 2017 and 2018. The working group also created a Google email group[1] for interested community members to discuss and comment on the work and new versions.

The working group is currently drafting version 4 following a series of presentations at 2018 conferences and a workshop at iPres 2018[2].

C. Potential Uses and Audiences

The Criteria have been developed as a set of design attributes, and considerations for digital preservation storage services. Some of the uses for the Criteria include:

- Guiding evaluations of preservation storage

[1] See groups.google.com/forum/#!forum/dpstorage

[2] The forums where the Criteria has been presented for community feedback are listed on the project website wiki (osf.io/sjc6u/wiki).

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

services and options
- Identifying gaps in existing digital preservation storage implementations
- Assisting with Request for Proposals (RFPs) and related documents
- Contributing to instructional materials on digital preservation
- Informing infrastructure design and planning with Information Technology (IT) and other domains
- Framing discussions within the digital preservation community.

The possible audience(s) for the Criteria include digital preservation managers who need to implement and manage digital preservation storage, providers of digital preservation storage services, auditors of digital preservation programs, digital preservation instructors and students, and practitioners in affiliated domains who rely upon digital preservation storage.

A guiding principle for the versions of the Criteria has been ensuring that the Criteria remain generally applicable to digital preservation storage in any context by avoiding the inclusion of local practices. The Criteria provide a bridge to implementation by including a usage guide and accumulating examples to demonstrate the local use of the Criteria.

II. Structure of the Criteria

A. *Presentation*

The Criteria are organized into a table with five columns and one row per criterion shown in Table 1.

TABLE I

Structure of the Preservation Storage Criteria

No.	Criteria	Category	Description	Related Criteria & Refe-rences
1	Integrity checking	Content Integrity	Performs verifiable and/or auditable checks to detect changes or loss in or across copies ...	
2	...	...	...	
...				
61	...	...	...	

The columns are for the 'Number' (sequential ID for the criterion), 'Criteria' (short descriptive name for the criterion), 'Category' (one of eight topical areas used to group the Criteria), 'Description' (short definition for the criterion), and 'Related Criteria and References' (a placeholder to map relevant standards or related criteria to the criterion). For example, in Table 1, the first listed criterion is "Integrity Checking" in the category of "Content Integrity." The Integrity Checking criterion indicates that the preservation storage "Performs verifiable and/or auditable checks to detect changes or loss in or across copies." There currently are no related criteria or references listed for this criterion.

B. *Categories*

Starting with the second version of the Criteria, the initially unwieldy list of criteria was organized into categories to group similar criteria together and to provide an overall organization. Currently, the eight categories are:

1. **Content Integrity** refers to practices ensuring the state of stored data has not changed.
2. **Cost Considerations** reflect the financial impact of storage decision making.
3. **Flexibility** refers to the adaptability, interoperability, and overall ability to customize preservation storage solutions to an organization's needs.
4. **Information Security** refers to data protection methods to ensure that the data cannot easily be tampered with or accessed without proper authorization.
5. **Resilience** refers to the durability and availability of the storage system.
6. **Scalability & Performance** refers to computational performance and ability to be scaled up or down according to organizational needs.
7. **Support** refers to support contracts as well as services like training and additional preservation services such as migration.
8. **Transparency** refers to the visibility into the storage system's functions, e.g. auditing, reporting, error notification, and documentation.

C. *Revisions*

As mentioned previously, the Criteria have been revised several times because of feedback from workshops, presentations at conferences, and a survey. The introductory narrative of the current version of the Criteria (version 3) has been enhanced to add:

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

- more clarity on the definition and scope of "preservation storage"
- clarification that the audience for the Criteria includes both consumers and providers of preservation storage
- additional key considerations to consider in addition to the Criteria

Changes were also made to the Criteria table to include categories (see Table 2) and to normalize the Criteria names (bolded) and definitions. Finally, a reference list and an accompanying usage guide were developed.

TABLE II

Evolution of the Criteria Categories

	2016 - Version 1	2017 - Version 2	2018 - Version 3
No. of Criteria	48	58	61
Categories	None	Content Integrity (3)	Content Integrity (2)
		Cost Considerations (3)	Cost Considerations (3)
		Flexibility & Resiliency (12)	**Flexibility** (7)
		Information Security (11)	Information Security (15)
		Scalability & Performance (11)	Scalability & Performance (10)
		Support (3)	Support (4)
		Transparency (11)	Transparency (14)
		Storage Location (4)	**Resilience** (6)

IX. Usage Guide

The Criteria cannot stand alone; they need to be set in context of basic preservation principles. Therefore, the Criteria are supplied with a usage guide focusing on preservation storage principles.

Preservation is about preventing the loss of data, therefore managing the risks that could cause data loss is an essential practice for all types of preservation. The usage guide therefore includes the following key concepts that should be considered in relation to the Criteria: risk management, **independence between copies, elements in establishing bit safety and cost analysis.**

A. *Risk Management*

The usage guide includes a short description of the general concepts and processes of the practice of risk management to help organizations using the Criteria. Digital preservation requires storage solutions that can be sustained over the long-term. Risks to digital preservation storage operations may come from one or many events, incidents or situations. The usage guide includes a list of examples of these.

An organization can use risk management practices to identify and isolate risks that are specific to digital preservation over the long-term to reduce and mitigate impacts on digital preservation operations. Similarly, an organization can use a risk assessment to compare the risks of storage solutions that address different sets of criteria. Because digital preservation storage solutions must be sustained over time, it is useful to have a consistent methodology for risk management that can be used by the organization over time as solutions change, and as organizations use the Criteria to propose solution changes over time.

B. *Independence Between Copies*

For Preservation Storage, risk management must consider the goal that no or only an acceptable amount of data may become lost. There are risks that one event, agent, or technology can harm several copies of data in a way which imply loss of all data or an unacceptable possibility of data loss. The best way to mitigate such risks is to ensure independence between copies in a way that prevents the same event or incident for doing such harm. Independence means that any one event, agent, or technology cannot affect a majority of copies. The independence must be considered on any level where risk of loss can exist, e.g. organizational level, technical level, environment level etc. The total risk assessment must take all three key elements (number of copies, independence between copies, and integrity checks of and among copies) into account for each type of risk.

It is important to note that independence between copies may include the use of checksums. This is especially the case when there is a minimum number of copies (two full copies and one checksum), since loss of both checksum and one copy will make it impossible to verify whether the surviving copy is correct.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

C. *Elements in Establishing Bit Safety*

A full risk assessment of Preservation Storage needs to include more than independence between copies; it needs to include all the three essential elements which are needed for evaluating whether a Preservation Storage solution provides the required level of bit safety. These are:

- **Number of copies** - There should be enough copies available to survive the loss of some number of the copies.
- **Independence between copies** - The copies should exist independently of one another
- **Integrity checks (of copies and among copies)** - The copies must undergo periodic integrity checks to assure their fidelity.

The decisions on how many copies are needed can be determined with a complete risk assessment with focus on risk of losing all copies or losing the ability to verify correctness of surviving copies based on consideration of all three elements. Risk assessment may vary due to which risks each organization is willing to take. The absolute minimum number of copies is two, since an error in one copy requires having a healthy copy to be repaired from. The risk of keeping only two copies is that unless information like a checksum is also kept, there may be no way to tell which copy is valid if one becomes erroneous. When using such a minimum setup it is very important to consider the risks of loss.

Another important part of Preservation Storage is to consider how requirements for confidentiality and availability and costs of the preserved data are ensured, e.g. it may be hard to ensure confidentiality for data that has 100 copies spread all over the world, and it may be difficult to provide fast access to data that is only placed on off-line media. Such issues need to be considered as part of the risk analysis along with the other bit preservation elements.

D. *Cost Analysis*

The usage guide includes a short description of the general concepts and processes of the practice of cost analysis, to help organizations using the Criteria. An organization can use cost analysis to identify and isolate storage solution costs that are specific to digital preservation, and/or to compare the costs of different storage solutions that address different sets of criteria.

Cost analysis is a systematic approach to estimating resource expenditures, either to compare potential or existing situations, or to establish an approach for valuing resources for a specific decision or course of action. For example, a cost analysis can help identify and compare the resources required to implement and sustain two different storage solutions which are based on different sets of digital storage criteria. The usage guide includes an introduction to cost assessment and how it is used as well as tools and additional resources.

X. Future Work

While much of the content of the usage guide was presented in recent iPRES and PASIG conference sessions, at the time of this writing the usage guide has not been released. The next step for the working group is to complete the first version of the usage guide and to release it publicly for feedback by members of the dpstorage Google group and the broader digital preservation community.

There is also work planned for the Criteria document itself. Version 4 of the Criteria will map the Criteria to applicable standards and will incorporate feedback from recent conferences.

Additionally, effort will go into building the community around the Criteria project. The project website will be improved to expose more of the project outputs and roadmap. Examples of organizations using the Criteria will be documented and shared through the project website. Lastly, an organizational host for the Criteria project will be sought to provide a stable home for the Criteria and to help engage the community to use and improve it.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Acknowledgment

The latest version of the Criteria has been shaped by the feedback of many individuals in the digital preservation community over the last three years. In addition to the authors of this paper, core members of the Digital Preservation Storage Working group include Jane Mandelbaum, Gail Truman, and Cynthia Wu. Steve Knight and Kate Zwaard made important contributions to early versions. People who joined the dpstorage Google Group have provided valuable feedback. Special thanks to the attendees of the iPRES 2017 Digital Preservation Storage workshop in Kyoto who gave us in-depth feedback on the Criteria and the newly introduced categories.

References

[1] A. Goethals, N. McGovern, S. Schaefer, G. Truman, and E. Zierau. 2018. Digital Preservation Storage Criteria. Retrieved June 19, 2019 from https://osf.io/sjc6u/ DOI: 10.17605/OSF.IO/SJC6U

[2] Digital Preservation Coalition. 2015. Digital Preservation Handbook (2nd. ed.). Retrieved March 13, 2019 from https://dpconline.org/handbook

[3] ISO 14721:2012. "Space data and information transfer systems - Open archival information system (OAIS) – Reference model", 2012.

[4] N. McGovern and E. Zierau. "Supporting Analysis and Audit of Collaborative OAIS's by use of an Outer OAIS - Inner OAIS (OO-IO) Model," Proceedings of the 11th International Conference on Preservation of Digital Objects (iPres 2014), pp. 209-218, 2014.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Giving Access to Multimedia Objects in Libraries with Emulation

Implementing a Solution

Tobias Steinke
Deutsche Nationalbibliothek, Germany
t.steinke@dnb.de
https://orcid.org/0000-0002-3999-1687

Abstract – The German National Library makes collected multimedia objects available onsite within emulated environments. The framework is based on Emulation as a Service and optimized for automatically handling big collections. The paper describes the development in a German research project and the practical integration in the access environment of the library. It also gives an overview of legal aspects related to the emulation framework.

Keywords – Emulation, Access, Multimedia, Libraries, Licenses

Conference Topics – Designing and Delivering Sustainable Digital Preservation; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

The German National Library has a huge collection of multimedia publications on data carriers. The legal deposit includes digital publications like encyclopedias, educational applications and interactive literature. These multimedia objects were published on floppy discs, CD-ROMs, DVD-ROMs and other data carriers. As these are mostly computer programs, a suitable environment of hardware and software is needed. Although mostly made for different versions of Windows, there are also programs for older computer systems like Apple II, C64 and Atari ST. Access to the collection is only available at Windows based terminals in the reading rooms of the German National Library. In order to provide access to all these kinds of objects a flexible and scalable framework with support for emulations of many systems is needed.

II. The German Project EMiL

In the project Emulation of Multimedia objects in Libraries (EMiL) [1] an emulation-based flexible and automatable access framework for multimedia objects in libraries and museums was developed. The project was funded by the German Research Foundation (DFG) and ran from 2014 to 2016. Project partners were German National Library, Bavarian State Library, Karlsruhe University of Art and Design and University of Freiburg.

In a first step the project partners defined their requirements. The libraries have to deal with huge collections of very different objects. No manual effort for preservation actions on each object is possible on this scale. In the context of museums on the other hand the quality of the presentation of each object (mainly art) is most important. In any case the access is restricted to onsite access due to legal reasons. A flexible framework was needed, that supports emulations for many different systems and that could work automatically for many objects as well as with pre-installed highly optimized emulation environments for special objects. There was no need to work over the internet, but the access should scale for a certain amount of local users.

In a second step existing emulation frameworks were validated based on the requirements and with a set of test objects by the project partners. These were basically the framework of the European project KEEP [2] and the Emulation-as-a-Service system of the German project bwFLA [3]. The result of the evaluation was that the system of bwFLA provided the best starting point for the development. It was also developed by the project partner University of Freiburg.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

One of the key developments needed for EMiL was a module to automatically assign a fitting emulation environment for a given object. There are catalogue entries for the huge amount of multimedia objects in the collections of the libraries, but these lack accurate technical metadata. There is no existing information of the needed software and hardware for accessing the object, which could be used to start a best fitting emulation. Manual examination of each object is no option for collections of thousands of objects. The university of Freiburg developed a module that tries to identify all executable files of an object (usually CD-ROM images) and match one of the pre-installed emulation environments. If no executable file is found, an environment is chosen with an installed player for the mostly found file format in the disk image.

The basic access use case for the framework was defined as: A user at a Windows based terminal uses the library catalogue in the reading rooms and finds the entry of a multimedia publication (floppy disk, CD-ROM, DVD-ROM). After clicking on a link in the catalogue entry page in the browser, a new browser window will open and automatically start a suitable emulation. The selected object will be mapped as drive within the emulation so it can be used in the same way as if the disk would be inserted in a disk drive at the original machine. By closing the browser window the emulation will be closed and everything done within the emulation will be wiped out.

The developed framework by University of Freiburg supported several existing emulators for different machines and allowed adding new environments based on existing ones. E. g. a Windows 95 environment could be enhanced with an installation of MS Office and then be saved as a new additional environment.

At the end of the project in 2016 there were test installations of the developed frameworks at the partner institutions. These test installations were not integrated in the local infrastructures and featured not all the functionalities wished by the partners.

III. EMiL Integration

The German National Library decided to start an internal project called EMiL Integration in 2017 to further develop the EMiL framework according to the institution's needs and to integrate it in the existing infrastructure.

The team of University Freiburg that had developed the EMiL framework established a company called OpenSLX[1]. The actual software is available as open source[2], but OpenSLX continues developing the system and supports it.

The main new functionalities that were identified as requirements for the usage in the reading rooms of the German National Library were session management and printing. Session management means that users should be able to save the state of the used emulated environments for later usage. Many Windows CD-ROMs need to be installed before using them and allow individual user configurations. Without session management a library user wanting to use the same object on several days would have to re-install and configure the object every day. A certain percentage of a publication is allowed to be printed out by users. This should be possible for publications in emulated environments as well. OpenSLX was able to implement this depending on the emulation. For some emulators it is possible to re-direct the emulated printer output in a PDF file that can then be used by the local Windows machine for printing.

The thousands of multimedia publications on data carriers (mainly disks) in the collections of the German National Library are not yet migrated to the digital preservation system. Until now disks with multimedia content needed to be ordered by users and are then converted to disk images for access via the commercial software Virtual CD.

At the end of 2018 the EMiL system became available for the users in the reading rooms of the German National Library, but just with a handful of objects. In a next step the existing created images for access with Virtual CD will be stored in the preservation system and this will make them available for access with the EMiL framework. That way at least the most requested multimedia publications will be available for emulation, but it is of course only a small percentage of all objects in the collection. A

[1] http://openslx.com/

[2] https://gitlab.com/emulation-as-a-service

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

systematic migration of the content of all disks will be needed in the future.

The usage of available objects is implemented like the basic use case of the original project. Access links were added to the catalogue entries of the multimedia objects. A click triggers the EMiL system installed within the infrastructure of the German National Library. The software is called with the parameters object ID and user ID (if a user is logged in). In a new browser window the interface of EMiL is displayed together with a screen of the emulated system (see figure 1).

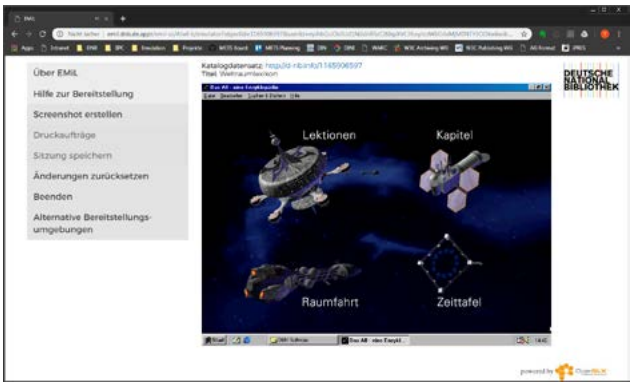


Figure 1 Emulation of a Windows 98 system with a learning program about space travel

Within the emulated environment the selected disk is mapped to a drive. The user has to to open the drive and start the program or the installation process if needed. A help button shows general information on how objects are normally started in the currently running environment. If the same user used the same object before and saved the session, the emulation will start with the saved environment. It is possible to reset the current emulated environment to its default state. It is also possible to get a list of other available environments that would (probably) work with the selected object and start one of these alternative emulated environments. E. g. although Windows 98 was automatically selected as best fitting environment for this object, the user could decide to use it in a Windows XP environment. If a specific information is available for the selected object like a registration code for installation this will be displayed. Users are able to save a screenshot of the emulated system and they can use existing printing functionalities within the emulation (for supported systems). The print output will be stored as PDF in a queue for using with the local printers in the reading rooms. Session saving is only available after shutting down of the emulated system. This is necessary to avoid corrupted environments.

An administration interface gives an overview of all existing environments, saved sessions and objects with assigned environments. Within the browser based interface it is possible to edit or add environments. New environments based on existing ones can be created with the same interface as the normal access. Default software that needs to be added will be mapped within the emulation, can be installed and then saved like a session. The resulting new environment might be manually assigned to existing objects or might be used by the automatic assigning process. Completely new emulators can only be added by OpenSLX, because these need to be integrated in the framework.

IV. Legal Aspects

One task of the project EMiL Integration was to examine legal aspects related to the emulated environments and acquire licenses if necessary. The EMiL framework itself and all integrated emulators are free software. But operating systems (OS) are needed for the emulation of hardware systems. Most of the OS - even very old ones – are (still) copyrighted and therefore usage requires getting a license by the right holders. For some of the older systems it was not so easy to find and contact the right holders. Emulation can be treated as normal usage unless license agreements say otherwise. For some OS there are free alternatives.

These are the results of the project for the emulated systems in the EMiL framework:

- MS-DOS 6.22: Use FreeDOS as a free alternative.
- Windows 3.11 to 7: Licensed as part of an existing framework contract with Microsoft.
- Apple II DOS, System 7: License agreements prohibit copying of ROMs.
- Apple MacOS X: License agreements prohibit usage on non-Apple machines.
- Linux: Free software.
- C64, Amiga: Bought license from right holders Cloanto Corp.
- Atari ST: Use EmuTOS as free alternative.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Solutions were found for all systems except for the systems by Apple. Several requests to Apple to acquire licenses to emulate Apple II and Macs remain unanswered. There was no answer at all. So for now all emulations of Apple systems are not available in the EMiL installation of the German National Library.

Of course all software installed within emulated environments (viewers, players, plugins) need to be licensed as well. For every creation of an additional environment a check of the newly installed software in this regard will be necessary.

V. Summary and Outlook

For interactive digital objects in the collections of libraries emulation is the only suitable preservation strategy. The emulation framework developed in the German project EMiL and integrated in the access environment of the German National Library implements a practical solution for giving access to those kinds of objects. It is flexible enough to integrate a variety of emulated environments.

Next steps will be migrating more multimedia content of disks in the collection of the German National Library to the repository to make these objects available with the EMiL framework. The emulation framework could also be used for other object types like eBooks or archived web sites.

References

[1] T. Steinke, F. Padberg, A. Schoger, and K. Rechert, "Project EMiL – Emulation of Multimedia Objects," iPRES 2016 proceedings, 2016, pp. 167-170.

[2] B. Lohman, B. Kiers, D. Michel, and J. v. d. Hoeven, "Emulation as a Business Solution: the Emulation Framework," iPRES 2011 proceedings, 2011.

[3] K. Rechert, I. Valizada, D. v. Suchodoletz, and J. Latocha, "bwFLA – A Functional Approach to Digital Preservation," PIK – Praxis der Informationsverarbeitung und Kommunikation, 35, 4, 2012, pp. 259–267.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

A Perspective from Digital Preservation on Sustainability of E-Infrastructures

Simon Lambert
UKRI-STFC, UK
simon.lambert@stfc.ac.uk
ORCID 0000-0001-9570-8121

Abstract – Much effort is expended on assuring sustainability of e-infrastructures and the services within them. Some of these services have a dimension of long-term preservation, in particular the infrastructure around persistent identifiers (PIDs). It is therefore worth asking whether established approaches to long-term digital preservation can offer insights into planning for sustainability. This paper examines some aspects of the OAIS Reference Model and their applicability to sustainability planning for e-infrastructures.

Keywords – E-infrastructure, Sustainability, OAIS
Conference Topics – Exploring New Horizons

I. Introduction

Sustainability is a fundamental attribute of e-infrastructures. The European Open Science Cloud Declaration uses the word repeatedly as a *sine qua non* of the EOSC [1]. Sustainability means the capacity of the operations, services and activities of the e-infrastructure to continue dependably into the future. It means that users can rely on the e-infrastructure, and that the initial investment in creating it is not at risk of being wasted. Sustainability is not the same as financial viability, though often considered as such. There are other ways that sustainability can fail apart from problems with profitability or cash flow. Many smaller social network websites lost out to the irresistible march of Facebook, not because of financial unviability but because of loss of members to the more attractive alternative. The same is true for digital preservation: having the money to keep a repository or archive running, though necessary, is not sufficient—it might be that the archive's contents eventually become unusable, for reasons which are now very well known.

Some e-infrastructure components that need to be sustained also have long-term preservation aspects. Obviously, services that specifically offer long-term data storage such as EUDAT's B2SAFE [2] and Preservica's products [3] are of this type. The infrastructure for persistent identifiers (PIDs) is another example: a whole complex of interconnected access would be in jeopardy if the e-infrastructure failed, putting the record of science at risk.

Not all e-infrastructure services have this long-term dimension. A service for uploading and sharing datasets among collaborating researchers or a cloud computing offering should be sustained, but do not have such ramifications—their value is in their use at one particular time.

Given that some e-infrastructure services have a long-term dimension, it is natural to ask whether anything can be learnt from the principles and practices of digital preservation when thinking about and planning for sustainability of such services. The contention is that some concepts and models from the OAIS (Open Archival Information System) standard [4] can be related to sustainability and provide a fresh perspective on it.

OAIS is a conceptual framework, proposing a variety of concepts and models for describing the responsibilities and functions of an archive that aims to preserve information for the long term. This paper examines some of those concepts and models and relate them to sustainability in which there is a dimension of long-term preservation (as for PIDs). Clearly some of the concepts and models are highly specific to the functions and organization of an archive rather than an e-infrastructure, and so the correspondence cannot be taken too far, but nonetheless there are clear and valuable parallels.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

II. Basic Concepts of OAIS

The fundamental concept of OAIS is the Archive:

"An OAIS is an Archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community."

The Archive corresponds to the whole e-infrastructure, and the information that the Archive preserves corresponds to services within that e-infrastructure. Those services are made available to some community and the intention is that the services will continue to be available into the future. The e-infrastructure may not be an organization in quite the same was an Archive (though OAIS recognizes the possibility of distributed archives); the "organization" is whatever structures and mechanisms have been put in place to manage, govern and develop the e-infrastructure. The European Open Science Cloud (EOSC) has its governance structures; the EUDAT Collaborative Data Infrastructure has a network of service providers; PID service providers such as DataCite and Crossref have their business models based on membership; the FREYA project envisages a "PID Commons" for governance of the PID infrastructure [5].

Other fundamental concepts of OAIS are the Producer and Consumer. Producers "provide the information to be preserved" while Consumers "interact with OAIS services to find and acquire preserved information of interest". In e-infrastructure terms, Producers correspond to service providers and Consumers correspond to service users—whether end-users or third-party providers of other services. Table 1 summarizes the correspondence of these fundamental concepts.

TABLE 1

Correspondence of OAIS and E-Infrastructure Concepts

OAIS	E-infrastructures
Archive	E-infrastructure as a whole
Content Information (the original "target of preservation")	Services
Producer	Service provider
Consumer	Service user

III. Specialized OAIS Concepts and Sustainability

Having established this basic mapping, it is now possible to examine some more specialized OAIS ideas to assess whether and how they can be applied to thinking about sustainability of e-infrastructures. It should be stressed again that the sustainability in question relates specifically to services with a long-term dimension; and that this is only a preliminary view, so that more thorough analysis might well throw up further analogs.

A key idea of OAIS is the Designated Community, which is defined as:

"An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time."

Thereby the Designated Community is contrasted with the wider group of potential Consumers, and the Archive has assumed an explicit responsibility to ensure that the information it holds should be understandable to the Designated Community over the long term. The relevance to sustainability is in the selection of a set of services that a certain community positively requires over the long term, and consequently the assumption of responsibility for maintaining the usability of those services. For example, the "Designated Community" (always with the long-term perspective) in the case of a PID infrastructure might be the body of researchers at large who require long-term access to the scientific knowledge base—meaning that resolution of PIDs appearing in articles, whether to other articles, datasets, software or whatever other entity must be maintained. This defines the requirement for long-term sustainability, even if the ability to mint new PIDs (of a certain type) ceases.

As an incidental note, thinking in terms of the Designated Community in OAIS terms might help to avoid the pitfalls of referring to "the community", a term which according to context (it is seldom clearly defined) might refer to distinct fields of research ("the social science community", "the particle physics

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

community") or merely to the set of individuals who take an interest in the functioning of the e-infrastructure (as in "engaging with the community").

A further basic idea of OAIS is that of Representation Information: "The information that maps a Data Object into more meaningful concepts," i.e. what must be provided to supplement the knowledge base of the Designated Community so that they can understand and use the information in the Archive. The idea is broad enough to encompass not only explanatory material, data dictionaries and the like but even emulation software. The analog in e-infrastructures is what must be provided to ensure continued usability of services as the environment changes. This might include new APIs or user interfaces, training offerings, best practice guides or "ambassadors" for particular domains.

The OAIS model defines a number of Functional Entities playing particular roles in the preservation endeavor. Some of these such as Ingest (accepting material from Producers into the Archive) are very specific to preservation of materials—though it might be that the activity of accepting a new service into an e-infrastructure throws up some parallels worth exploring. The two Functional Entities that do have clear analogs are labelled Administration and Preservation Planning: the former "provides the services and functions for the overall operation of the Archive system" while the latter "provides the services and functions for monitoring the environment of the OAIS, providing recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, the Designated Community ...." This makes clear a valuable distinction between day-to-day operation of an e-infrastructure and the higher-level tasks of tracking trends in the user base and in technology. From a sustainability perspective, it should be possible for an e-infrastructure to point out how these two distinct functions are carried out.

IV. Preservation Strategies and Sustainability

The three basic strategies of preservation implied by OAIS can be encapsulated as "hand over", "transform" and "add Representation Information" (the last including emulation as a special case).

Handing over refers to the transfer of the Archive's holdings to a successor Archive, which may be necessitated if the original Archive ceases to operate or otherwise becomes incapable of fulfilling its preservation responsibilities. It is tempting to relate this to individual services—ensuring continuity if a service provider fails—but it must be remembered that the hand-over refers to the whole Archive, equivalent to the e-infrastructure, not to an individual object within it. It would presumably correspond to a situation in which the e-infrastructure as a whole undergoes significant change, perhaps through a major shift in funding or top-level governance, and raises the question of how the "planning" and "administration" functions previously noted can be sustained in the new environment—in other words, what would the "successor" e-infrastructure look like?

Transformation in OAIS is a type of operation on the holdings that changes the content information—a typical example being a transformation from one data format to another. For the services of an e-infrastructure, this would correspond to reimplementation of the same functionality on a different platform. An interesting consideration is OAIS's idea of Transformational Information Property, sometimes called "significant property", which may be used to give assurance that information content has been preserved after a transformation (typically relating to accurate rendering of a document or image, though not limited to that). Trying to identify such properties for services of an e-infrastructure could be an avenue for focusing sustainability planning on what is essential in the long term, or what will give assurance of "authenticity" as things change over time.

The equivalent of adding Representation Information would be the case in which existing services are maintained but something has changed in the environment or user base that necessitates or at least makes desirable the additional "information". Tracking what might change and determining how to react are important functions that must be present in sustainability planning, whether the changes are in the technological base or community behavior. In fact there are two requirements: identifying what kind of thing might change in future, and monitoring to see if it actually does change. Both these are reflected in the OAIS description of the

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Preservation Planning Functional Entity, which refers to "risk analysis reports, and monitoring changes in the technology environment and in the Designated Community's service requirements and Knowledge Base."

V. Conclusion

Examination of the concepts and models of OAIS reveals that some of them at least have correspondences in the sustainability of e-infrastructures, in some cases suggesting fresh ways of looking at and planning for sustainability, such as transferring the idea of the Designated Community or the three basic strategies for preservation. The FREYA project intends to make use of this perspective as one contribution to its work on sustainability of the global and European PID infrastructure.

Acknowledgement of funding

The FREYA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777523.

References

[1] "EOSC Declaration", Brussels, October 2017. https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf

[2] EUDAT Service Catalogue: B2SAFE. https://www.eudat.eu/catalogue/B2SAFE

[3] https://preservica.com/digital-archive-software

[4] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information Systems", June 2012, https://public.ccsds.org/pubs/650x0m2.pdf

[5] https://www.project-freya.eu/en/about/mission

Together Forever,

or How We Created a Common and Collaborative Digital Preservation Service

Johan Kylander
CSC – IT Center for Science
Finland
johan.kylander@csc.fi
https://orcid.org/0000-0002-8084-8233

Heikki Helin
CSC – IT Center for Science
Finland
heikki.helin@csc.fi
https://orcid.org/0000-0003-4002-8203

Kimmo Koivunen
CSC – IT Center for Science
Finland
kimmo.koivunen@csc.fi
https://orcid.org/0000-0002-9614-8370

Juha Lehtonen
CSC – IT Center for Science
Finland
juha.lehtonen@csc.fi
https://orcid.org/0000-0002-9916-5731

Abstract – Collaboration is a key to success in digital preservation. In Finland we have, within the national digital preservation services, established a successful model of collaboration with archives, libraries and museums to preserve our digital cultural heritage. Collaboration is essential through the whole life cycle of digital assets as digital preservation needs to be a dynamic activity in a constantly changing environment. In association with archives, libraries and museums we are able to build capabilities that are needed for digital preservation. We stipulate that we have been able to create a unified and centralized preservation service with a set of common specifications only through a close collaboration with the data curators.

Keywords – collaboration, community building, digital preservation services, common specifications

Conference Topics – Collaboration: a Necessity, an Opportunity or a Luxury?; Building Capacity, Capability and Community

I. Introduction

"Knowledge begets agony!" The Finnish proverb for ignorance is bliss, but with the process reversed, was an often heard expression cried out in meetings during the early days of planning and sketching the Finnish national digital preservation service. The meetings were actively attended by both the producer of the preservation service as well as various cultural heritage organizations. The expression conveyed both a sense of despair at the staggering task at preserving Finnish digital cultural heritage but also acted as a signal that we are in this together and that all problems are shared problems. This communal spirit that was prevalent during the planning turned out to be vital in successfully creating a functional service.

The national digital preservation services are a result of a long term and profound cooperation between the services' producer CSC – IT Center for Science (CSC) and the organizations that curate the data. The Digital Preservation Service for Cultural Heritage preserves digital assets from the cultural heritage sectors, represented by archives, libraries and museums, whom are referred to as partner organizations. The national services are currently expanding to include preservation of research data, covering disciplines such as geophysics, astronomy and political sciences. The digital preservation system, the technical solution behind the service, is common for both cultural heritage data and research data.

Together we have managed to create a centralized

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

digital preservation service that employs a shared technological solution for processing different types of data, as well as developing common specifications that define the format and structure of the data ingested to the service. The collaboration is an ongoing effort where we build preservation capabilities for cross-disciplinary and cross-institutional needs.

In this paper we briefly present our services and the forms of our collaborative effort. We present how the services were designed, what are their visible products from the user's point of view and how the services are managed in cooperation with our partner organizations.

II. National Preservation Services

A. *Building the Service*

Back in 2008, after years of unofficial discussions, the National Digital Library of Finland (NDL) initiative was formed within the remit of the Ministry of Education and Culture of Finland. The project aimed at creating a nationally unified structure for contents and services, ensuring the effective and high-quality management, dissemination, and digital preservation of cultural digital information resources. Libraries, archives, and museums from the cultural heritage sector formed the basis for the NDL.

It was decided in the early days of the NDL that a common and shared digital preservation service should be created. This was based on a profound cost-benefit analysis. It was estimated that common infrastructure and services reduces costs, increases system integration, strengthens cooperation, and brings the practices of partner organizations closer together. Besides technical solutions, the collaboration between partner organizations was an essential goal of the NDL.

During the next years, the Ministry of Education and Culture, CSC, and partner organizations, in tight co-operation, designed a national digital preservation service, suiting the needs of all cultural heritage sectors. A lot of planning was done in the preservation support group consisting of the aforementioned actors. The preservation support group evaluated new features, devised the common specifications, and gave indispensable input on the release plan for the services. The support group also provided a valuable forum for partner organizations for sharing their experiences on digitization and data management processes, which in turn helped individual organizations in improving their processes.

Several targeted working groups were established to solve particular technical and social challenges. These working groups, consisting of invited experts on certain topics, were tasked with solving specific issues, such as recommending and evaluating file formats for audio-visual data or suggesting technical and/or structural metadata for different types of content. The working groups reported to the preservation support group, thus providing valuable input in creating the service.

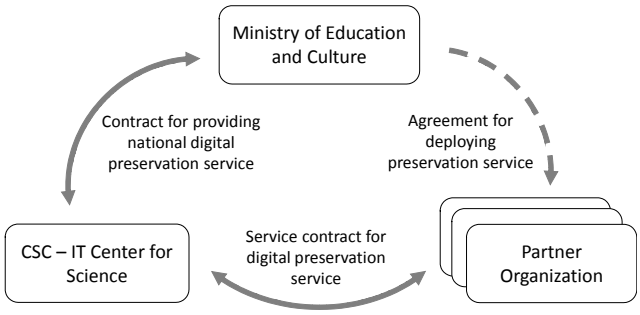


Figure 1. The contract model of the national preservation services.

The framework for planning the service came from the national NDL project. As one key area of focus was increasing interoperability, foster common practices and integrate systems, a national enterprise architecture for the cultural heritage sector was created. The enterprise architecture describes common high-level principles for managing digital assets, common processes for achieving the goals of the NDL project, and describes centralized services and technological solutions that are to be used. This architecture includes describing processes for improving the long-term digital preservation of cultural heritage by assigning the preservation to the national preservation service. The enterprise architecture also includes the Standard Portfolio. The portfolio is a specification giving an overview of all approved metadata standards and common interfaces that are allowed within the NDL project.

The successful collaboration which was conducted on many levels resulted in a set of common national

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

digital preservation specifications and eventually in the national digital preservation service. The first AIP was created on November 2015, roughly seven years after the NDL project started. The volume of content to be preserved in the service is growing steadily. In 2018 the volume nearly doubled with approximately 107 terabytes of new content received for preservation. By the end of 2018, there was a total of over 222 terabytes of content to be preserved in our services

B. A Formalized Service

The NDL initiative ended in 2017 but the collaboration continues as partner organizations have recognized the importance of cooperation and knowledge sharing. Our goal is to develop and maintain the digital preservation service on a sustainable basis. Currently we have preserved more than a million AIPs and the number of partner organizations preserving their digital content in the national preservation service grows constantly.

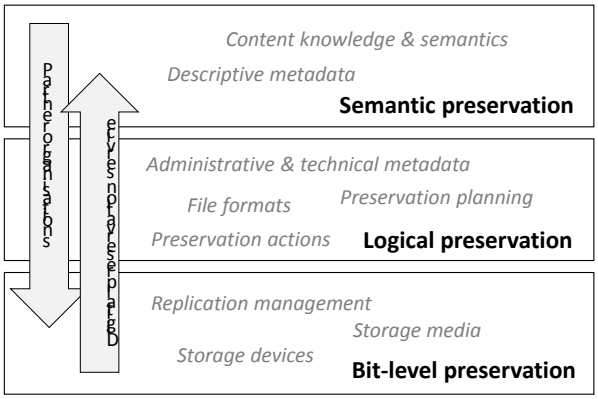


Figure 2. Levels of preservation

The national digital preservation services are owned by the Ministry of Education and Culture and are provided by CSC. The ministry and CSC have signed an agreement in which for example governance of the services is defined. Furthermore, many of our partner organizations have statutory obligations in preserving certain digital assets which increases the long-term scope of our services. Direct preservation costs are funded by the ministry, meaning that for partner organizations the utilization of the service is free of charge

Partner organizations, under the remit of the Ministry of Education and Culture, may utilize the national digital preservation services after they have agreed with the ministry about borderlines. These include recognition of the collections to be preserved and quota allocated for these collections. After that, the partner organization makes a formal contract with CSC in order to start preservation, as depicted in Fig. 1.

III. Digital Preservation is Cross-Disciplinary

The foundation for effective collaboration begins with the simple acknowledgment that the data producers and curators are experts in their respective fields and know their own data and data models best. Therefore, they know what data to preserve, what it contains, and what it requires. The service provider, CSC, acts as a facilitator to discuss digital preservation issues together and between organizations. CSC also provides deep know-how about digital preservation and produce preservation tools for partner organizations.

Active collaboration is a necessity in digital preservation for several reasons. In addition to the cost-effectiveness of sharing knowledge and common tools we recognize the notion that digital preservation requires knowledge about a wide area of topics. These include knowledge about the contents of the digital assets, knowledge about how and why the assets have been created and included in a repository, as well as technical expertise on digital preservation methods.

The theory behind the collaboration rests on defining different levels of digital preservation that require different actions and know-how. As depicted in Fig. 2, digital preservation can be divided into three levels [1]. These are, starting from the bottom, bit-level preservation, logical preservation, and semantic preservation. The arrows to the left indicate from which angle the producer of the preservation service, from the bottom going up, and the partner organization, from the top going down, approach digital preservation. It implicates that certain areas, like bit-level preservation, requires more technical know-how and its implementation is (mostly) the responsibility of the service producer. The semantic preservation requires deep found knowledge about the data and why it is to be preserved. This level is naturally (mostly) the responsibility of the curators.

This division is something that we have

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

implemented in our service as the technological solution for the bit-level preservation is mostly managed by the service's provider. Reports on the bit-level preservation, such as hard disk failures or other statistical information, is regularly shared with our partner organizations. It is however in the areas of logical and semantic preservation that our collaboration with partner organizations really bears fruit. At the logical level, various tasks are shared between the service provider and partner organization, depending on the needs of the organization. The organization does not, for example, have to care about details of preservation actions if it does not have enough competence or resources to do so. On the other hand, the semantic level concentrates on semantics of assets to be preserved. These may vary a lot between disciplines. The service provider cannot have enough competence about various semantic details and therefore partner organizations are mainly responsible for the issues in this layer, possibly with help from the service provider.

IV. Collaboration equals Common Requirements

A. *Common Specifications*

The most visible product of the collaboration, uniting organizations from the diverse cultural heritage sector, is a set of published specifications common to all digital assets ingested into the services. The specifications were created to enable us to build a service that can receive, process, and preserve digital assets from the whole cultural heritage sector. They also define the interfaces of the national preservation services in order to aid partner organizations to integrate their production systems with the preservation services. The specifications have been developed and approved in close collaboration with all partner organizations. They are updated annually undergoing a yearly review process. Up-to-date specifications are available at our website.

Because the quantity of data ingested into the services is considerable, a fully automated processing of data during the ingest phase is essential. This, in turn, requires that all metadata in the submission information packages (SIP) ingested to the preservation services are machine readable, thoroughly defined, has to follow a standard and can be automatically validated. The automated processing requires both administrative, technical, structural, as well as descriptive metadata to be included as a part of the SIP in a controlled way.

As an answer to this, we have created the specification *Metadata Requirements and Preparing Content for Digital Preservation* that describes the required metadata and the mandatory structure of the information package. Work on the specification started during the planning phase of the preservation service as a part of the NDL project. Input and approval from partner organizations has been constant already from the beginning. The allowed metadata standards are defined in the Standard Portfolio. Together with the national METS profile, the Standard Portfolio acts as the framework for our specifications, as shown in Fig. 3.

The metadata specifications are not only functional requirements for our preservation services. They also have a pedagogical value in creating a common understanding of digital preservation metadata and of its quality by steering partner organizations toward implementing the required metadata and using recommended standards. This approach has been very successful, as some smaller organizations did not take these kinds of issues into account before introduction to our specifications. Further, some national organizations, not specifically performing any digital preservation activities, have utilized our specifications in order to improve their daily processes.

It is a common practice in digital preservation to carefully select file formats for preservation. These formats should be technically viable for a long time in the future. In our case the data to be preserved is diverse and includes text, image, audio, moving image, and database files. That is why we and our partner organizations have surveyed what file formats are commonly in use in the cultural heritage sector. We have created a set of approved file formats for preservation and a second set of file formats approved for ingest. The sets are based on an evaluation criteria that for example evaluate whether a file format is open, standardized, technically stable, and preferably backwards compatible with earlier versions. File formats that meet these

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

criteria are approved for preservation. Common formats widely in use but that do not fully meet our evaluation criteria can be approved for ingest, implicitly with the notion that they will be migrated to another file format during the preservation.

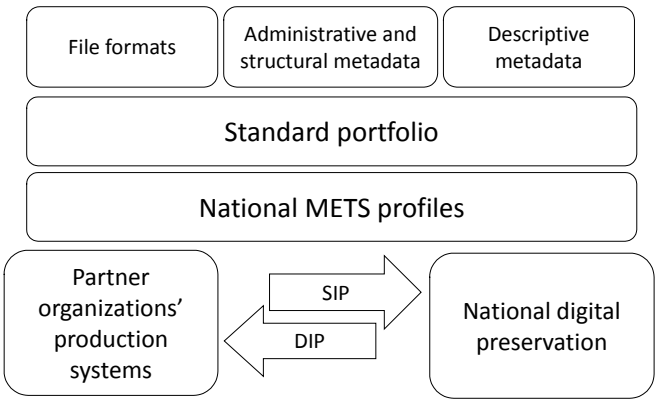


Figure 3. The specifications and standard portfolio govern the information packages

The file format sets form the backbone of our specification *File Formats*. Every year the partner organizations can propose new formats if there is a need for it. These are reviewed according to the evaluation criteria and added to the specifications if they pass the criteria and are approved by the preservation collaboration group.

B. *Regular Forms of Cooperation*

The digital preservation services are administered by a steering group and the development is overseen by a collaboration group. The collaboration group is an open group consisting of partner organizations and others who are interested in the development of the services. The collaboration group discusses the long-term development of the service, approves the specification updates, and gives valuable input to the producers of the preservation services. At the same time it is a forum for discussing general digital preservation and presenting topics ranging from risk management, validation tools, and interfaces. The collaboration group meets at least four times a year.

Annually, in the fall, new demands and suggestions for updating the specifications are collected in the form of a survey sent to interested parties throughout Finland, including all partner organizations. The survey results are processed, resulting in specification updates that are reviewed and approved by the collaboration group. The new features are typically implemented in the beginning of the following year. The new specification versions are published at the same time as the features are implemented.

C. *Increasing Capabilities and Knowledge*

One cornerstone of our fruitful collaboration is a shared understanding of digital preservation and the goals of the service. Regular meetings and other support activities helps us maintain this shared vision. We arrange workshops and training sessions on different preservation topics several times a year in which participants are sharing knowledge with each other.

An important aspect of collaboration is raising the understanding of digital preservation issues among all participating actors. Its importance is twofold. Firstly, the quality of the digital preservation services improves as the services' producer gains more insight on preservation actions available as the input from the partner organizations increase. Secondly, curating digital assets and processing digital assets and metadata for ingest requires know-how. Therefore, an increased knowledge about digital preservation across the whole field is a necessity for increasing the volume of data ingested into our services.

Sharing knowledge and providing support for partner organizations as well as other institutions in Finland is an important part of the digital preservation services. The services maintain a support email address through which we annually answer about one hundred issues.

Extended support is given when a partner organization joins the service and starts preparing data for ingest. We also provide support in the form of tools which are published on GitHub[1]. These are tools for assisting partner organizations in creating SIPs for ingest and validating their data, for example the pre-ingest tool [2]. We strongly believe that sharing our technological knowledge in the form of a published set of tools is an important aspect of both collaboration and in the end of the quality of our service.

As we have stressed above that successful digital

[1] https://github.com/Digital-Preservation-Finland

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

preservation requires collaboration, we must also look beyond the national boundaries. Examples of somewhat similar approaches to our can be found for example in Germany and the Netherlands. The nestor Network [3] in Germany is doing somewhat similar forms of collaboration in Germany to what we do in Finland. Perhaps the biggest difference however is that they do not provide a concrete preservation service. The same is true for the Dutch Digital Heritage Network [4]. Both of these are successful national examples. Nonetheless, organizations fostering wider international cooperation are needed. Examples, such as the Digital Preservation Coalition (DPC) or Open Preservation Foundation (OPF), are needed. CSC joined the OPF in early 2019. International cooperation is essential in a wider perspective but international cooperation still cannot solve all national challenges.

V. Conclusions

While developing and maintaining a centralized preservation service for cultural heritage we have encountered several obstacles while trying to solve things in a cross-disciplinary and cross sectoral environment. We have learned that collaboration is the key to success and we consider it the only way to enable organizations from different fields and people with very different backgrounds to build a shared vision for digital preservation.

Even within the scope of cultural heritage, sectors can have different requirements and needs. Even sharing a common language can be challenging. However, with thorough collaboration, mainly in the form of regular meetings with commonly set targets, a consensus can be achieved. We have realized that the self-organizing nature of the early collaboration, tasking the organizations with leading expert groups and contributing heavily to the requirements, was a key factor in achieving a common understanding. From the common understanding grew the foundation for common requirements and ultimately a common service.

Maintaining this collaborative spirit is important throughout the whole life-cycle of preserving digital assets - it is a continuing effort. Addressing issues on the level of logical preservation, such as evaluating file formats and devising material specific preservation planning, requires input from both the data curators and the service provider. These are issues that we continue to process collectively through the preservation collaboration group.

Collaboration in the form of sharing knowledge results in better quality of data and its management. Sharing knowledge can be in the form of seminars, workshops and hands-on support. It also comes in the form of meetings, common specifications, and tools that automate certain tasks, easing the requirements for technological know-how and allowing organizations to spend their resources on curating their data instead.

As our services mature and expand to include more organizations with more limited technological and personnel resources, the importance of tools and even services that aid and automate certain processes for preparing content for ingestion into our services will increase. Ultimately, an increased knowledge on preservation issues in the cultural heritage sector as well as tools that lighten the workload of our partner organizations, enables the amount of data sent to our services to increase and the services role as an important centralized repository for Finnish digital cultural heritage and research data to be realized.

An important insight and lesson has been that submitting data to the national digital preservation services has forced the organization to think about data quality, the quality aspects in the creation of digital data, and proper management of digital resources. In some cases even changes in organizational culture has been a side effect of becoming a partner organization and submitting data to our services. None of this would have been possible without the collaborative process of common meetings, seminars and workshops.

Given all this, it is clear, at least to us, that collaboration for successful digital preservation is essential, and it should be broadened even more. There might be slightly different approaches to collaboration, but still it is very important, no matter in what shape. We are in this together, hopefully forever.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Acknowledgments

The authors would like to thank all members of the Digital Preservation team at CSC, as well as the national preservation collaboration group for their valuable comments and input during the preparation of national digital preservation services. Especially, we thank the Ministry of Education and Culture of Finland for making this all possible.

References

[1] K. Thibodeau, Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years, CLIR 2002.

[2] K. Lehtonen, P. Somerkoski, J. Törnroos, M. Vatanen and K. Koivunen, "Modular Pre-Ingest Tool for Diverse Needs of Producers", in 14th International Conference on Digital Preservation (iPRES 2017), Kyoto, Japan, September, 2017.

[3] S. Schrimpf and Y. Tunnat, "15 years of nestor: German Network of Expertise in Digital Preservation", in 15th International Conference on Digital Preservation (iPRES 2018), Boston, MA, USA, September, 2018.

[4] Dutch Digital Heritage Network, https://www.netwerkdigitaalerfgoed.nl/en/

Provenance for Internet Art

Using the W3C PROV data model

Lozana Rossenova
London South Bank University / Rhizome
UK / USA
mehandzl@lsbu.ac.uk

Dr Karin de Wild
University of Leicester
UK
Kdw15@leicester.ac.uk

Dragan Espenschied
Rhizome
USA
dragan.espenschied@rhizome.org

Abstract – When a scholar, curator or archivist is researching an artwork, they need provenance, an essential piece of information that can help them evaluate as to whether a source can be trusted. This paper will investigate how to express the provenance of Internet art as linked data. One of the standards that has emerged to describe the provenance of digital data is the W3C PROV. It provides a model which facilitates description of the entities, agents and processes involved in producing data. This generic model has proven to be applicable in various contexts, including the cultural heritage domain [1, 2]. However, its potential to describe the provenance of Internet art is not yet fully explored [3]. This paper demonstrates how the PROV model can be used to describe the provenance of Internet art by applying it to a case study from Rhizome's ArtBase, an online archive dedicated to preserving works of Internet art. This paper is aimed at digital art conservators, digital curators, Web archivists and art historians.

Keywords – Provenance, Internet art, Rhizome, W3C PROV, linked data

Conference Topics – Collaboration: a Necessity, an Opportunity or a Luxury; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

During the nineties, shortly after the wider introduction of the World Wide Web, artists started to create artworks within this online environment. These digital artworks were (and some of them still are) embedded within the Web. At the time of writing, there is not a definite name for these artworks. Various terms can be found in the literature, including "Internet art", "Net(work)-based art", "Web art" and "net art". Throughout this paper we will use the term "Internet art". Characteristic for Internet art is that the work evolves over time, often into various instantiations (or versions). In Rhizome's ArtBase, instantiations are referred to as "variants", a term developed by Dragan Espenschied, which we will also adopt in this paper.

Over the last 20 years, Internet artworks have increasingly entered museum collections and archives, and the study of these artworks is becoming part of contemporary art historical research, which brings new complexities. For instance, scholars are advised to proceed with caution when studying an Internet artwork as there is limited provenance information available (or made accessible), which is essential for critically evaluating the reliability of the source as evidence. Normally, researching the provenance of artworks includes an object study, going through resources about the artist (e.g. catalogues raisonnés) and other owners (e.g. auction results and exhibition catalogues). Not all of this is available, and sometimes not even applicable for Internet artworks.

In this paper we will demonstrate an approach to describing provenance for Internet art by testing the application of the PROV Data Model (PROV-DM). Instead of a history of ownership, PROV-DM describes "the people, institutions, entities, and activities, involved in producing, influencing, or delivering a piece of data or

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

a thing" [4]. It was developed as part of a family of documents published by the Provenance Working Group at W3C, which also include an OWL2 ontology (PROV-O), developed for mapping PROV-DM to RDF[1].

PROV-DM has several key characteristics which are applicable to the case of Internet Art. First, this model makes it possible to give an overview of the lifecycle of an artwork. The PROV model not only captures the creation of the artwork, but also how various actors contribute to or influence the work over time. For instance, these may include individuals or institutions, who commission, acquire, transfer or modify the work. Furthermore, PROV-DM can capture the different variants of a single artwork, even when these are preserved across various institutions. A single Internet artwork can be included in multiple (museum) collections, (Web) archives, whilst being part of the live Web. There is not yet a standard way of describing provenance adopted by all. PROV-DM is useful in this regard, as it can function as a provenance interchange model between heterogeneous systems, e.g. across (data) collections held by different memory institutions. Finally, PROV can also be used in conjunction with other ontologies in linked data repositories.[2]

The main reason we chose PROV-DM for the case study presented in this paper is that it was developed specifically for expressing provenance data on the Web. While other metadata schemas have been developed to model all data about a cultural heritage object (e.g. CIDOC-CRM), or focus on the preservation of highly abstracted digital objects (e.g. PREMIS)[3], no schema has been developed specifically to address the challenges of provenance description for Internet art. Comparing PROV to other existing schemas or ontologies is not the purpose of this paper. Rather, our objective is demonstrating how PROV-DM can be applied in a practical way towards modelling data and conceptualizing provenance for Internet art.

We continue with a description of our methods in section II. Next, we investigate how PROV-DM can be applied to an Internet artwork, including how it can be implemented in a linked data knowledge management system. Section IV discusses key issues that came up in the modelling and implementation processes, some of which may also require future research. Finally, section V concludes the paper.

II. Methods and Methodology

A. *Application of the PROV model*

For this paper, we applied the PROV model in three steps: First, we traced the lifecycle of the artwork, based on archival research. Next, we translated this information in a PROV-DM application that illustrated the key components of provenance we considered necessary (section III.B). In the final step, we used PROV-O to map PROV-DM to RDF in Rhizome's Wikibase (section III.C).

PROV-DM consists of six components, of which we tested three for this paper (the PROV core structures) [4]. Component 1 (C1) describes the "entities" and "activities". An entity can be a variant of the artwork, and/or physical, digital or conceptual elements of the artwork. An activity is something that affects an entity within a certain time period. Component 2 (C2) describes "derivations". In our case this means how one variant of an artwork derives from or relates to another. Component 3 (C3) refers to agents and their responsibilities. An agent can be a person, as well as an organization or a piece of software. Using these components - C1, 2 and 3 - it becomes possible to pose and answer questions such as: "Who [agent] did something [activity] to this variant of this artwork [entity]?"; or: "How does this variant relate to other variant(s) [derivation]?"; etc.

B. *Interdisciplinary collaboration*

All findings presented in this paper are a result of interdisciplinary collaboration. At each stage, we made prototypes, which we reflected on and discussed, bringing our own areas of expertise. Through iterations, the outcomes were further refined. We adopted a practice-based research method, in which collaborative prototyping is a mode of enquiry [5].

[1] https://www.w3.org/TR/prov-overview/

[2] PROV-O properties and concepts are already integrated in the latest OWL specification of PREMIS.

[3] Jefferies, et al, cite these reasons to explain why they chose PROV-DM, over other schemas, as a more practical approach to describe provenance in the Oxford Research Archive

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

C. *Case study approach*

Although this paper focuses on a single case study, we prepared models for various Internet artworks. The results were compared to further understand PROV-DM and, in particular, how to model the relationships between entities, agents and activities. Based on these initial tests, one artwork was selected and studied in more detail. All artworks tested are part of the ArtBase, one of the largest archives of Internet art that is accessible online since 1999. Maintained by Rhizome (USA), the ArtBase is also one of the few online art archives to support a functional linked data infrastructure, which can facilitate the implementation of a linked data provenance model.

The case study that is discussed in this paper is "untitled[scrollbars]" by artist Jan Robert Leegte. It was selected, because the ArtBase features several variants of the artwork, which offered the opportunity to model how the artwork evolved over time. Additionally, the record for this particular artwork features more detailed information about the accession and preservation of the work, compared to other records in the archive. Furthermore, unlike many other examples of Internet art, this artwork does not contain external media or data sources dependent on third-party services. Such media and services would require additional provenance research outside the scope of this initial study.

III. Case Study

A. *"untitled[scrollbars]"*

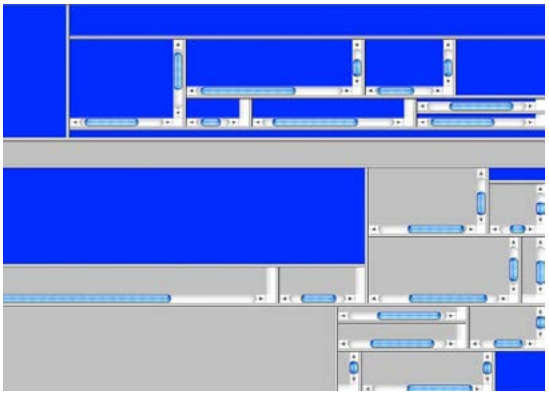


Figure 1. Jan Robert Leegte, "untitled[scrollbars]" (2000), web project, source: Rhizome ArtBase (Q2508).

The artist Jan Robert Leegte lives and works in Amsterdam, The Netherlands. He is part of a generation of artists, also known as the net art movement, who have been making art on the Web since the nineties. His early works rebuild basic, interactive elements of the (classic) Windows interface, such as buttons, window frames or scrollbars. The artwork "untitled[scrollbars]" was published online in 2000. Over time, the title of the artwork has changed from "untitled" to "untitled[scrollbars]" and "scrollbarcomposition", additionally the artwork can be encountered at various URLs. It can also be understood as part of a larger corpus of works, including a physical installation with the same title, consisting of a wooden structure and a projection of a Windows 98 or a Mac Aqua version of the scrollbar (2005, 2011). In 2001, "untitled[scrollbars]" was accessioned by the ArtBase, where an archived variant (in HTML), as well as a WARC (Web ARChived) file are preserved.

B. *Application of PROV-DM*

We developed a data model for the provenance of "untitled[scrollbars]" [Appendix A]. In this model, we applied three views on provenance that PROV supports: "data flow view", "responsibility view" and "process flow view", including the associated classes and properties from PROV-O [4].

The "data flow view" shows how one variant of an artwork derives from another. First, we identified all variants of the artwork that can be found in the ArtBase. These include two archived variants – in the custody of Rhizome, and two further URL's that pointed outside the ArtBase. In the provenance information we had available, it was not possible to find the derivation relationships for all of the variants. Building upon the data structure already in use in the ArtBase, we added one additional entity in our data model to represent the "artwork" as a general concept. PROV-DM provides support for modelling relationships between a general concept of an entity (the artwork) and its specific instantiations (the variants) with the properties *prov:alternateOf* and *prov:specializationOf* [4]. An entity that is a "specialization" of another shares all aspects of the latter. On the other hand, two "alternate" entities share some aspects of the same thing, but may also differ and may or may not overlap in time [4]. Since the way an Internet artwork evolves over time is unpredictable and variants can differ substantially

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

from one another, we found the latter, broader term more suitable in our case study.

The "responsibility view" involves assigning the agents responsible for different events in the artwork's lifecycle. We distinguished between the concepts of "attribution" and "association". In our model, we use attribution specifically in relation to the artist/creator of the work, whereas we use association to refer to any other contributors or the archivist/archival institution. In our case study, all variants of "untitled[scrollbars]" can be attributed to the same artist – Jan Robert Leegte. Therefore, we assigned the attribution property to the general concept of the artwork. In contrast, we assigned association agents to each specific variant. We found this to be a useful way of documenting custodial care, i.e. whether the variant is in the care of an artist (*:variant* **>** *prov:wasAssociatedWith* > *:artist*) or an archive (*:variant* > *prov:wasAssociatedWith* > *:archivist* > *prov:actedOnBehalfOf* > *:archivalOrganisation*). Association can be related to entities, or activities. In the latter case, it becomes part of the "process flow view".

In the "process flow view", we refined our data model by including "generation" activities (i.e. creation), and appended dates and locations.[1] The generation activities in our case study include "Cloning" and "Webrecorder capture". "Cloning" refers to creating a file directory copy of the artwork from the artist's server to Rhizome's server. "Webrecorder capture" refers to creating a WARC (Web ARChive) of the artwork, using Rhizome's tool Webrecorder. This activity is subject to the decisions of an archivist performing the capture. In order to document this agent's influence, we assign the association directly to the activity, rather than the variant. Furthermore, this association can be qualified (modelled as a *prov:qualifiedAssociation* in PROV-O) by additional properties, e.g. adding a "plan" to the activity, for example the archival instructions used during capture. A generation activity can also be qualified (modelled as a *prov:qualifiedGeneration)* by dates and times with the *prov:AtTime* property. When no particular activity of generation is assigned to a variant, the variant can still be dated using the *prov:generatedAtTime* property. Lastly, in addition to multiple times and activities of generation, the variants in our case study had different URL addresses. We used *prov:atLocation* to assign URL locations to each variant.

C. *Implementation in Rhizome's Wikibase*

1. *Wikibase and Wikidata*

Rhizome is one of the first cultural heritage organisations to use Wikibase as a collection management system for its archive, the ArtBase [6]. Wikibase is the open source software environment built to run Wikidata – a knowledge base of public domain structured data maintained by the Wikimedia foundation (WMF). Originally, the software infrastructure was not designed as a linked data system. Linked data capabilities were added later to serve the community needs for interoperability with existing linked data sets [7].

Wikidata's knowledge base follows RDF principles, and is organised in subject-predicate-object triples. These translate to item-property-value statements in terms of Wikidata syntax (e.g. artwork [item] > attributed to [property] > artist name [value]) [7]. Statements can have bibliographic references, too. This is how PROV-O is currently used in RDF data that can be exported from Wikidata. The *prov:wasDerivedFrom* property is used to link a bibliographic source to a particular statement. This application of PROV is insufficient in the case of Internet artworks and additional concepts and properties are needed.

Crucially for our use case, Wikibase can be deployed as a stand-alone instance, independent from Wikidata. While it still follows the RDF data modelling conventions of Wikidata, a Wikibase installation requires a custom configuration of concepts and properties. [6]. This is how Rhizome's Wikibase can adopt some PROV-DM concepts and PROV-O properties. Derivation and attribution, for instance, can easily be modelled as item-property-value triples. Owing to its legacy ties to Wikimedia

[1] Please note that the level of abstraction in this model is different compared to other digital preservation standards, such as PREMIS. This level of abstraction is concerned with artistic and historic integrity, not just technical integrity. We are modelling activities and actors involved in the creation, acquisition, or modification of an artwork variant, rather than tracking file system activities related to individual files such as checksum creation, etc.

software, Wikibase is very good at representing data related to things (e.g. Wikipedia pages), which become "items" in its RDF statements. The concept of the item can be mapped to the PROV-DM concept of an entity. The concept of the qualified activity (or process) from PROV-O, however, is more challenging to represent in Wikibase.

2. Mapping concepts and properties

The possibility to map concepts across different concept schemas is an integral part of the design of linked data. The SKOS (Simple Knowledge Organisation System) data model was developed specifically to facilitate such linking across knowledge organization systems on the Web.[1] The SKOS mapping property (i.e. *skos:exactMatch*) is already being used in Wikidata as a way to match a concept from one standard schema to another[2]. Using *skos:exactMatch*, it is also possible to link properties and concepts from Rhizome's Wikibase to corresponding PROV-O properties and concepts [Table I].

The "data flow" and "responsibility views" can be fully represented in Wikibase via such mapping. The "process flow" view presents challenges with regards to the qualified relations, which in order to be represented in Wikibase may have to be broken down and simplified [Appendix B][3]. In some cases, it is possible to model PROV qualified processes with Wikibase "qualifiers" – these are sub-properties which can be added to statements, providing additional detail such as time periods, locations, etc. Qualifiers add flexibility to data modelling in Wikibase, however, this flexibility can make querying more difficult, because a user would need to know the exact structure of the data model in order to make a meaningful query [7].

While this may be considered a limitation of the system, it also provides an opportunity to model provenance data in statements that are both easier to present to end-users accessing the data via a graphical user interface (GUI), as well as easier to query by users who want to find the provenance of artworks without being experts in the particular data model used in the ArtBase.

[1] https://www.w3.org/TR/skos-reference/

[2] https://www.wikidata.org/wiki/Property_talk:P2888

[3] See the current record for "untitled[scrollbars]" in the ArtBase with partial PROV implementation: https://staging.catalog.rhizome.org/wiki/Item:Q2508

TABLE I

Mapping PROV-O properties to properties in the ArtBase

Art Base Property	Property ID	skos:exact Match mapping	PROV view
variantOf	P56	prov:alternateOf	Data flow
derivedFrom	P102	prov:wasDerivedFrom	Data flow
artist	P29	prov:wasAttributedTo	Responsibility
collaborator	P120	prov:wasAssociatedWith	Responsibility
associated With	P118	prov:wasAssociatedWith	Responsibility
onBehalfOf	P119	prov:actedOnBehalfOf	Responsibility
generatedBy	P117	prov:wasGeneratedBy	Process flow
inception	P26	prov:generatedAtTime	Process flow
accessURL	P46	prov:atLocation	Process flow
startTime	P11	prov:startedAtTime	Process flow
endTime	P13	prov:endedAtTime	Process flow
archivalPlan	P121	prov:hadPlan	Process flow

IV. Discussion

A. *Linking variants to a general concept*

To gain insights into the provenance of an artwork, it is important that a query can retrieve all variants of the work. While a general concept is not required in PROV-DM, we used *prov:alternateOf* to connect all variants of the artwork to a general concept. This strategy is compatible with other cultural heritage and bibliographic semantic models, such as CIDOC-CRM and FRBRoo, where our concept of the "artwork" is equivalent to E28 Conceptual object (CIDOC) or F1Work (FRBR), and "variant" is equivalent to E73 Information Object (CIDOC) or F2 Expression (FRBR).[4] [5] This compatibility offers the potential for information exchange between different cultural heritage collections.

[4] http://www.cidoc-crm.org/

[5] https://www.ifla.org/publications/node/11240

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

B. *Modelling historical gaps*

Although closing knowledge gaps within the provenance of an artwork remains the goal, a 'complete' story can rarely be established. Considering the complexity of historical data, it is important to find ways to model gaps, inconsistencies and/or errors. PROV-DM offers the opportunity to state partial or incomplete provenance about an entity. For example, in the provenance for "untitled[scrollbars]" we added a date towards the general concept of the artwork (2000), but not to the variant that entered the ArtBase. This indicates that it is known that the artwork was created in 2000, but it is not precisely known when the particular variant accessioned by the ArtBase was created. Not only is it possible to leave out unknown information, PROV-DM also allows modelling in increasing levels of detail. For example, the provenance for "untitled[scrollbars]" indicates that the agent who accessioned the artwork is an archivist (their role). When provenance research reveals additional information, it is possible to further refine the provenance by adding, for example, a person's name (taking into consideration that their identity can be revealed) or more details about their actions (e.g. archival plans, etc).

C. *Accessing provenance data*

The considerations for implementing PROV-DM in a linked data art archive extend to how it will be accessed by end-users. Despite its limitations with regards to expressing qualified processes, the Wikibase system does provide a GUI, where the complexity of a graph database is made intelligible to end users [6]. While full integration of the PROV-O into Wikidata RDF expressions is not yet possible, if equivalent properties and concepts are accurately mapped, users will have the ability to query the ArtBase for PROV statements using a query service, such as the Wikidata Query Service. The formal terms for running federated queries across knowledge bases remain a matter of debate within the Wikidata community. A universal adoption of a standard mapping notation such as *skos:exactMatch* would improve the usability of the query service, particularly for users who would like to use it via a GUI. Until this adoption is implemented, users will need to first express the desired mapping in SPARQL and then formulate the particular query of interest.

V. Conclusion

In this paper we have presented a practical approach to expressing provenance for Internet art using PROV-DM. This approach facilitates the description of the lifecycle of the artwork, including any changes that were made over time and who was responsible for them. Additionally, it offers the opportunity to provide an overview of all the variants of an artwork, even when they are included in different collections and archives.

In future work, we will test PROV-DM against further case studies to include other types of provenance entities, agents and relationships. Further research is also needed to test the application of our proposal on a larger scale – e.g. entire collections. Another area for future research is how to fully integrate PROV-DM in Wikidata RDF expressions. For the time being, we have demonstrated alternative ways of working with PROV-O and PROV-DM in the ArtBase. We see future collaborations between digital preservation professionals, historians and the Wikimedia community as a key route to the wider adoption of PROV as a standard practice for preserving and presenting provenance of Internet art as linked data on the Web.

Acknowledgment

The authors would like to thank the following colleagues for their support and contributions to this research: Dr Annet Dekker, Anisa Hawes, Dr Katherine Thornton, Rhizome's Digital Preservation team, and Wikidata's development team at Wikimedia Deutschland. Lozana Rossenova would like to acknowledge that her PhD research is funded by AHRC (UK) with additional support from Rhizome.

References

[1] C. Li and S. Sugimoto, "Metadata-driven approach for keeping interpretability of digital objects through formal provenance description," in *Proceedings of the 14th International Conference on Digital Preservation iPRES 2017, Japan*, 2017.

[2] N. Jefferies, et al., "Contextual and provenance metadata in the Oxford Research Archive (ORA)', in *Proceedings of 9th Research Conference, MTSR 2015,* Manchester: Springer International Publishing, 2015.

[3] K. de Wild, *Internet art and agency*, PhD diss. Dundee: University of Dundee, 2018, pp. 183-184.

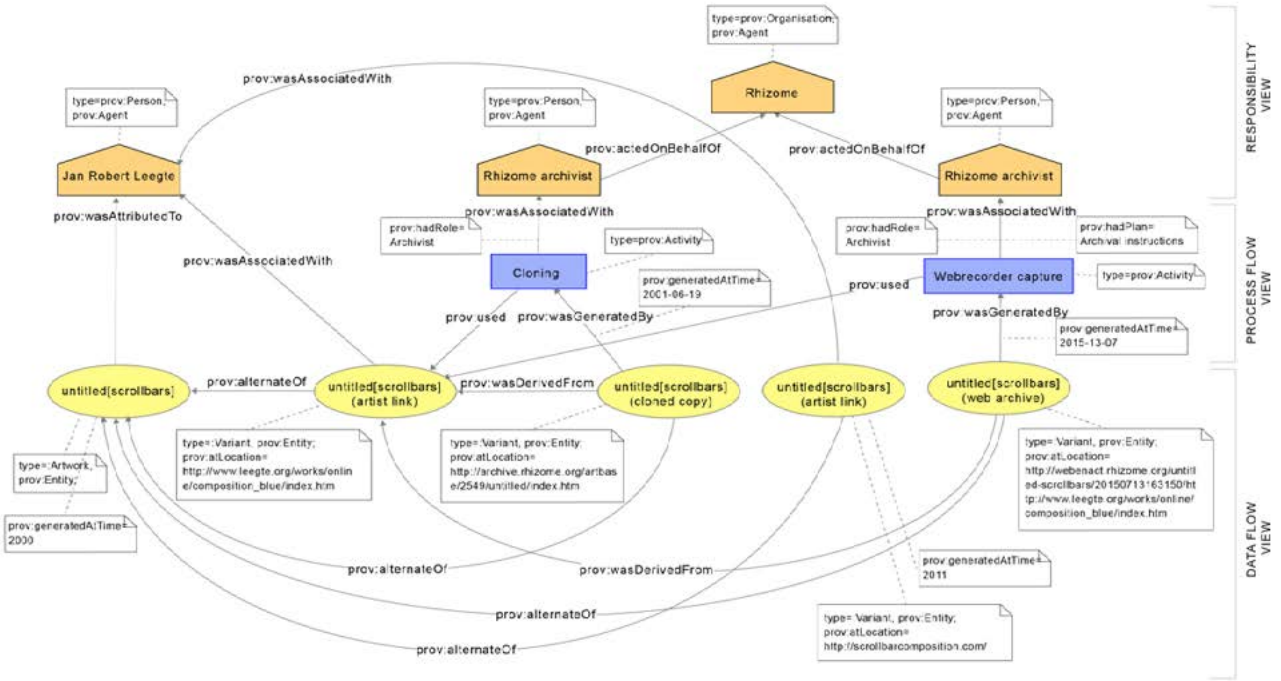
[4] L. Moreau and P. Groth, *Provenance: An introduction to PROV.* CA: Morgan & Claypool Publishers, 2013, pp. 3, 22, 33.

[5] F. Kräutli and S. Boyd Davis, "Digital humanities research through design," in *Proceedings of DH Early Career Conference,* King's College London, 2016, pp. 4.

[6] S. Fauconnier, "Many Faces of Wikibase: Rhizome's Archive of Born-Digital Art and Digital Preservation: Interview with Dragan Espenschied, Lyndsey Moulds and Lozana Rossenova", Wikimedia blog, 2018. https://wikimediafoundation.org/2018/09/06/rhizome-wikibase/

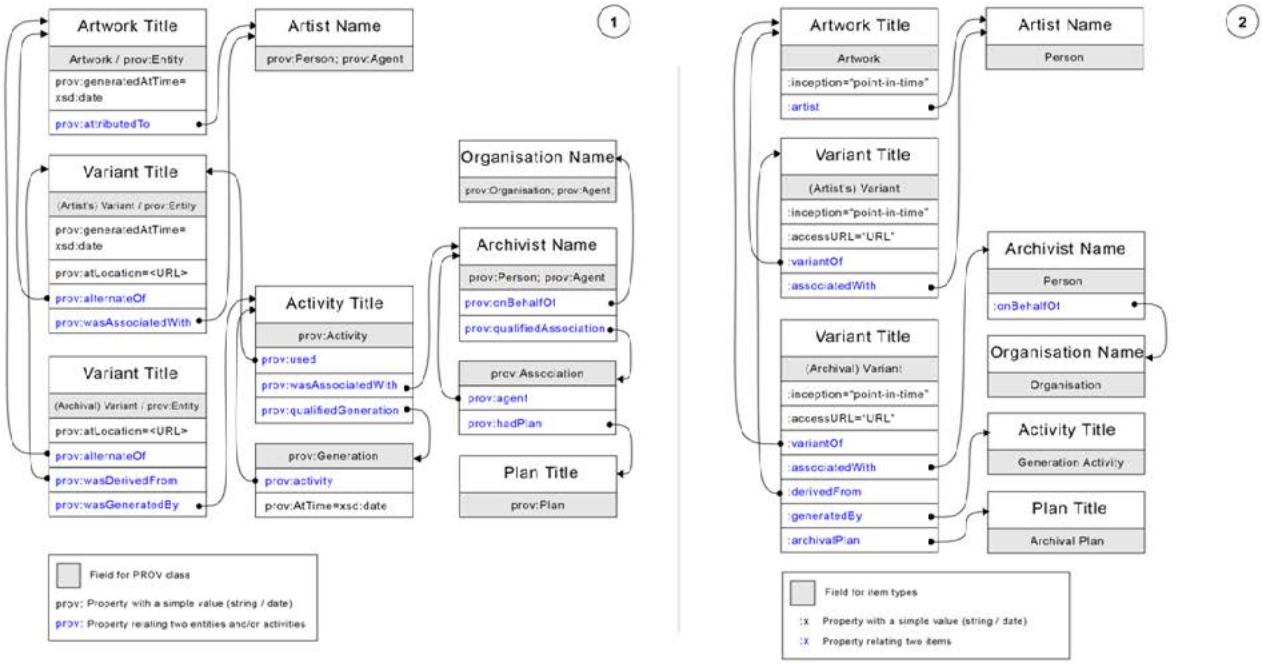
[7] K. Thornton, et al., "Modeling the Domain of Digital Preservation in Wikidata", in *Proceedings of the 14th International Conference on Digital Preservation iPRES 2017*, Kyoto, Japan, 2017, pp. 2, 5-6.

APPENDIX A



PROV-DM application for "untitled[scrollbars]". Visualisation follows the PROV Graph Layout Conventions, specified by the W3C recommendation: https://www.w3.org/2011/prov/wiki/Diagrams. Image can also be accessed in the Open Science Framework repository.

APPENDIX B



A comparison between 1) the generalised PROV-DM application for the ArtBase developed in the case study, vs. 2) the proposed practical implementation of PROV concepts in Rhizome's Wikibase.
Image can also be accessed in the Open Science Framework repository.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Saving Data Journalism

Using ReproZip-Web to Capture Dynamic Websites for Future Reuse

Katherine Boss
New York University, USA
katherine.boss@nyu.edu
0000-0003-2148-8386

Vicky Steeves
New York University, USA
vicky.steeves@nyu.edu
0000-0003-4298-168X

Rémi Rampin
New York University, USA
remi.rampin@nyu.edu
0000-0002-0524-2282

Fernando Chirigati
New York University, USA
fchirigati@nyu.edu
0000-0002-9566-5835

Brian Hoffman
Independent, USA
brianjhoffman@gmail.com
0000-0003-2333-0603

Abstract – While dynamic and interactive Web applications are becoming increasingly common to convey news and stories to people all around the world, their technological complexity makes it hard to archive and preserve such applications, and as such, they are being lost. We present ReproZip-Web, an open-source prototype aimed at saving these news applications from extinction. ReproZip-Web leverages ReproZip, a computational reproducibility tool, and Webrecorder, a tool for recording Web resources, to automatically and transparently capture and replay dynamic Websites. The prototype creates a bundle that contains all the information needed to reproduce a news application, and its lightweight nature makes it ideal for distribution and preservation. We will present our ongoing work on the prototype, and also discuss some use cases and avenues for future development.

Keywords – Web archiving, Emulation, Data Journalism, Emulation-based web archiving, ReproZip

I. Introduction

Data journalism stories are among the most complex, innovative, and original stories being produced by newsrooms today. These projects, created by news organizations in dozens of countries, are custom-built websites that display content dynamically in the browser. On the back end, many of these works also allow readers to explore, query, and inspect data related to a news story. Iconic examples of data journalism projects include "Dollars for Docs" by ProPublica, "Gun Deaths in Your District" by The Guardian, and the stories produced by The Upshot team at The New York Times [1]–[3]. These works are often called interactives or "news applications," and have increased dramatically in production and popularity as societies have become more data-driven [4, p. 154]. Yet because of their technological complexity, these sites cannot be fully or systematically captured by current web archiving tools. Current web archiving technologies, which have been successful in capturing snapshots of static news content, fail to capture the look, feel, and functionality of a significant amount of dynamic content, including social media feeds, interactive maps, visualizations, and database-reliant websites. While technologies like WebRecorder and Perma.cc (which is built on WebRecorder) have alleviated some aspects of this problem, there are several limitations to these tools [4], [5]. WebRecorder records resources as they are loaded by the browser and stores them as Web ARChive (WARC) files, but

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

at this time the capture and record process is not automated; users must click on each link of a site to initiate recording. For database-driven websites this is an untenable solution to capture content at scale, as it would require an archivist to click thousands to hundreds of thousands of links to fully capture a single project. Scalable web archiving of dynamic content requires an "emulation-based" approach that could capture assets located on the web server, many of which are protected, rightly, by firewalls and other security measures [6].

Beyond the technical challenges of capturing and archiving dynamic websites, there are organizational barriers to advancing a solution. In the majority of newsrooms in the United States, newsroom libraries have long been shuttered [7], [8]. This trend, which accelerated in the 2000s as newspapers became less profitable and budget cuts became common, has left newsrooms bereft of any archiving or preservation expertise within the organization. Few, if any, newsroom staff wake up each day thinking about how to save their digital content. Consequently, web archiving has always been an afterthought; in 2002 only 7% of newsrooms with libraries (already a minority) were conducting any sort of web archiving [7, p. 44]. Though outside organizations such as the Internet Archive have stepped in to save millions of pages of articles [9], much more content has been lost.

These losses have a detrimental impact on the collective cultural record and the future of research based on journalism [10, p. 1208]. News stories are known as the "first draft of history," and this makes them an important and frequent object of study for scholars across the academy. News websites are an important research artifact [11], and demand for them will likely only grow, given that the way the public finds, reads, and shares news is increasingly online [12], [13].

To address this problem and save interactive news websites, our research team has built an emulation-based web archiving tool, which, to our knowledge, is the first of its kind.

II. About the Project

Emulation as an archiving and preservation strategy was introduced as a concept decades ago by Rothenberg [14], though the infrastructure, skill, and knowledge to create emulators has only recently made it a feasible, economical, or practical option [15, p. 2]. Advances in cheaper and more abundant digital storage in the last decade have paved the way for emulation projects, and coincided with the belief in the digital archiving community that to save digital objects for the long term, we must emulate them in their original computational environment [15]–[19]. Preservation of encapsulated projects, rather than websites (both static and interactive), is currently underway at several institutions, including Rhizome, the Internet Archive, Carnegie Mellon, New York University, Yale and the Software Preservation Network, Deutsche Nationalbibliothek, and the British Library [15], [20], [21],. These pioneering projects have advanced the capture and preservation of system images and the frameworks that allow users to replay them on modern machines [10]. However, none of these initiatives have yet addressed a scalable, full-stack, emulation-based web archiving tool that could systematically capture the large volume of interactive news projects being published daily. Our project addresses this need.

III. A Prototype to Preserve News Apps

To this end, we extended an existing open-source project, ReproZip, originally designed for computational reproducibility [22]. ReproZip is a tool that automatically captures all the dependencies of a software application originally run in a Linux environment, and creates a single, distributable bundle that can be used to reproduce the entire experiment in another environment (e.g., on Linux, Windows, or Mac). ReproZip works in two steps:

Packing. In the *packing* step, the tool traces all system calls related to the execution of the application, capturing all of the dependencies at the OS level, including software, data files, databases, libraries, environment variables, and OS and hardware information. Using this information (which can optionally be customized by the user), ReproZip creates a bundle for it: an .rpz file containing all of the dependencies.

Unpacking. In the *unpacking* step, given an .rpz file, other users can use ReproZip to automatically

and transparently set up the packaged application in their environment, even if their OS is different than the one used for the creation of the application. This is possible thanks to emulation- and container-based tools leveraged by ReproZip.

ReproZip successfully captures and reproduces the software environment, including involved scenarios such as the client-server ones that are common to news apps. One of the current limitations of ReproZip, however, is that it cannot capture front-end remote dependencies. As we discovered in the course of this research, news apps often depend on remote front-end files, e.g., JavaScript, cascading style sheets (CSS), fonts, and other resources. Consequently, the bundle created during the packing step is incomplete. When unpacking, these front-end files will only work assuming they are still accessible and live: if they become inaccessible, the look, feel, and interactivity of these news apps is entirely lost.

Our extension to ReproZip, called ReproZip-Web, aims to address this limitation, and therefore to fully capture and preserve news apps. To capture these remote resources and add them to the .rpz bundle, we leverage the Core Python Web Archiving Toolkit (pywb) software library from Webrecorder [4], [23].

Packing and Recording the News App. The prototype assumes that the back-end of the news app (e.g., databases, web framework) has already been packed by ReproZip, thus creating a .rpz bundle. Given this .rpz file as input, our prototype simultaneously launches the emulated news app, a Webrecorder server, and an instance of the Chromium browser, which is controlled via the Chrome DevTools Protocol (CDP). With the application unpacked and being reproduced, the browser makes requests to Webrecorder, which acts as a proxy while it builds a WARC archive containing all the resources that were requested. Once the browser has finished loading the news app, our tool consolidates the .rpz and WARC data into a single package: a new .rpz file. Figure 1 depicts the full workflow for packing and recording a news app.

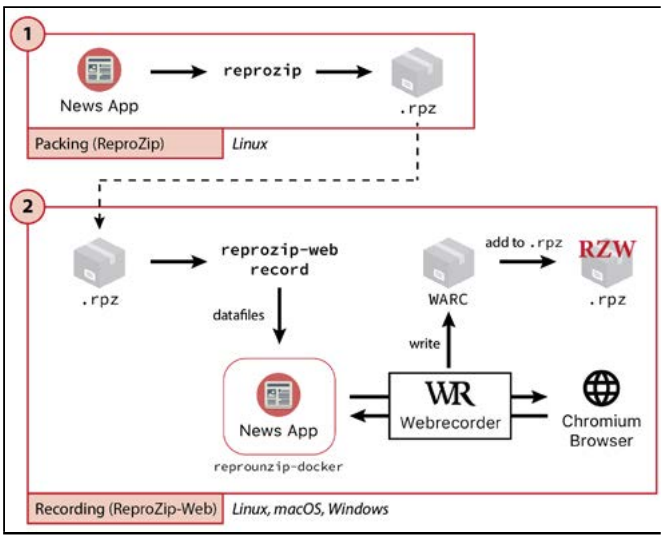


Figure 1: Packing and recording a news app with ReproZip-Web.

Replaying the News App. When a user replays the .rpz package with the prototype, the two servers are again launched simultaneously: the emulated news app in a Docker container (via reprounzip-docker), and a Webrecorder server providing the web resources archived in the package's embedded WARC data. A proxy server (Nginx) receives all network requests from the browser and routes them to the appropriate server, using the domain of the request URI to determine which server can fulfill the request. This configuration requires a browser with customized proxy settings pointing at the dockerized nginx server; our tool also has a mode in which the requests are handled directly by the Wayback server, allowing archival packages to be played back over the internet and without modification to the researcher's browser. Figure 2 depicts the full workflow for replaying a news app.

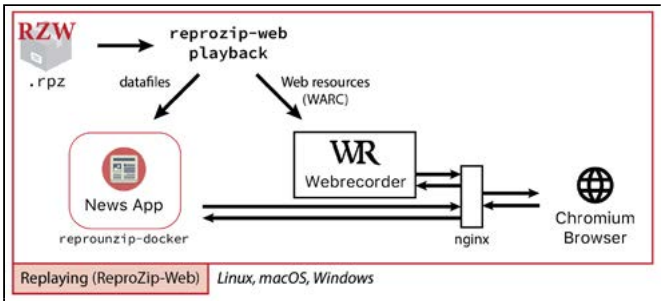


Figure 2: Replaying a news app with ReproZip-Web.

Our final product is a tool that effectively harnesses ReproZip and Webrecorder simultaneously to create a single archival package emulating both the state of the web server on which the news app runs, and the state of the relevant parts of the

world wide web at the time the app was published. Our prototype is available at https://github.com/reprozip-news-apps/reprozip-web.

IV. Use Cases

Our primary test case for the tool was a data-driven news app called Dollar for Docs, from ProPublica [2]. This news app was built using Ruby on Rails, MySQL, and Elasticsearch on a Linux server. Our first step, prior to development, was to capture the app and its environment using ReproZip, and to review the emulated site. Our first observation was that the look and feel of the app broke entirely when the browser was restricted from accessing resources on the world wide web (as expected).

When we implemented ReproZip-Web, we found that we could indeed view the news app with the correct look and feel, and began to drill down to specific limitations. One unexpected issue was the inoperability of the paginated search feature, which led us to the realization that certain Ruby on Rails files had not been captured by ReproZip due to their "lazy loading": one of the search results pages raised a server error because the HTML template had never been captured in the original package. This led to our decision to modify ReproZip with additional rules aimed at capturing a complete Ruby on Rails application, and the recognition that such rules may be required for each major web publishing platform

We also tested our prototype on The Guardian's Elections Poll Projection, an open-source news app from The Guardian that was built using Node.js [24], [25]. We were able to successfully reproduce and replay this news app as well, even when restricting the browser from accessing external resources. A demonstration video of our prototype using this news app is available at http://bit.ly/2O3Q4Ee.

V. Future directions

Much work is needed in testing, developing, and generalizing the prototype. ReproZip was developed to only capture what is executed during an application, but many interactive websites require that all parts of the environment (e.g., gems for Ruby) be included, and front-end content to be recorded and packaged. To account for this, we implemented an extra rule in ReproZip-Web to detect and automatically capture Ruby gems even if they were not executed while the application was running. In the next phase of development, we plan to implement more rules for other languages, optimizing for languages commonly used in data journalism projects. This will generalize the tool for a wider range of dynamic websites.

In the course of this project, while assessing different news apps, we also found that some news apps require access to external APIs and data (e.g.: sites that dynamically upload and download data from Amazon S3 stores). Reproducing this scenario is challenging and we plan to investigate different solutions in the future.

We also plan to gather more information about the needs of data journalists, system administrators, and other stakeholders that would use this tool. We will work closely with potential users in testing the prototype and gathering their feedback on its usability, the time commitment it would require, and the types of software that users would be willing and able to install on the production servers where ReproZip-Web would be deployed. This will also include an investigation of packing multiple apps at one time. All the feedback will be instrumental in building a finished tool that fits the needs and workflows of newsrooms.

Finally, a graphic user interface (GUI) is necessary to make the tool user-friendly and simple to deploy. While secondary users of ReproZip-Web have access to a GUI to replay the news applications from the .rpz file, there is currently no GUI for those originally packing the work, though there are in-depth command line utilities. Acknowledging that the command line is a high learning curve, adding a packing GUI would allow newsrooms to utilize ReproZip-Web to capture and archive their interactive news content at scale, potentially packing dozens of projects each month.

VI. Discussion

Cultural heritage institutions can leverage ReproZip-Web to create distributable and preservation-quality bundles of complex web applications that can be replayed in-browser or on desktop

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

machines. ReproZip-Web not only captures the applications and all their dependencies, but also automatically re-configures these in any other environment. This has not only simplified the process of capturing these important cultural artifacts for posterity, but allows anyone, on any type of computer, to access and replay them in their original computational environment — seeing the applications as they were experienced from the onset.

The .rpz is ideal for preserving complex applications such as these because it is extensible, lightweight, and open. The ReproZip ecosystem is such that after a journalist or archivist captures a news application with ReproZip-Web, they can then use any current or future ReproZip unpacker to replay it. Currently, to replay a news app from an .rpz file, users can choose four unpackers: reprounzip-directory, reprounzip-chroot, reprounzip-vagrant, and reprounzip-docker. However, a fifth is on the way (reprounzip-singularity, to unpack via Singularity containers), and a sixth has been contributed to the project from an outside colleague (reprounzip-benchexec), only possible because ReproZip and it's file format are open source. ReproZip bundles can be unpacked and replayed with any virtual machine or container software; so as these software wax and wane out of popularity and use, so can the ReproZip ecosystem be adjusted to create new unpackers, or depreciate old ones, without compromising the ability to use and replay old .rpz files [26].

Additionally, unless the size of the input data for a news application is on the Terabyte scale, the .rpz files are quite small and easily distributable. To date in our testing, we have yet to create an .rpz file over 800MB. These archival bundles are easily shared and distributed, as well as stored at a much lower cost, without compromising on the ability to reuse, replay, and preserve the contents of the news applications.

Likewise, the capturing process is scalable in that it captures an entire manifold of search results (or other database-driven content) without requiring the manual web recording of every possible search. However, our application isn't without a need for some case-by-case attention; it includes an extensible library of language and platform specific heuristics to address capturing issues arising from the idiosyncrasies of web publishing tools. ReproZip-Web allows newsrooms and archives to package complex news applications with all their dependencies into a single distributable and preservable .rpz bundle, from which users can replay the news application and archivists can ensure long-term preservation. Wide-scale adoption of this software and archiving practice would be a giant leap forward in saving data journalism projects for history, posterity, and the cultural record.

Acknowledgements

We wish to acknowledge the Institute for Museum and Library Services for their support of this project (LG-87-18-0062-18), as well as the PI of that grant Dr. Michael Stoller, and our project partner ProPublica for allowing us to test our prototype on their news apps. Thanks to Dr. Juliana Freire, the PI of the ReproZip project, for her continued support, and to Bofei Zhang, our research assistant, for testing the prototype and writing protocols. Thanks to Rhizome and especially Ilya Kreymer for the work on Webrecorder, without which this project would have had even greater barriers to overcome. Lastly, we would like to acknowledge the support from the Gordon and Betty Moore Foundation as well as the Alfred P. Sloan Foundation via the Moore-Sloan Data Science Environment for supporting continuing development of ReproZip.

References

[1] K. Davis, R. Harris, N. Popovich, and K. Powell, "Gun deaths in your district: what have your elected representatives done?," The Guardian, 2015. [Online]. Available: http://www.theguardian.com/us-news/ng-interactive/2015/dec/14/gun-control-laws-congress-shooting-deaths-nra-lobby-campaign-donations. [Accessed: 12-Mar-2019].

[2] L. Groeger, C. Ornstein, M. Tigas, and R. G. Jones, "Dollars for Docs," ProPublica, 2010. [Online]. Available: https://projects.propublica.org/docdollars/. [Accessed: 12-Dec-2015].

[3] The New York Times Company, "The Upshot," 2016. [Online]. Available: http://www.nytimes.com/section/upshot. [Accessed: 05-Apr-2016].

[4] I. Kreymer and D. Espenschied, "Webrecorder: A project by Rhizome." [Online]. Available: https://webrecorder.io/. [Accessed: 10-Jan-2018].

[5] Harvard University Library Innovation Lab, "Websites change. Perma Links don't.," Perma. [Online]. Available: perma.cc. [Accessed: 28-Jun-2019].

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

[6] K. Boss and M. Broussard, "Challenges of archiving and preserving born-digital news applications," IFLA Journal, vol. 43, no. 2, pp. 150–157, Jun. 2017.
[7] N. Paul and K. A. Hansen, "Reclaiming News Libraries," Library Journal, vol. 127, no. 6, p. 44, 4/1/2002 2002.
[8] K. A. Hansen and N. Paul, Future-Proofing the News: Preserving the First Draft of History. Lanham: Rowman & Littlefield Publishers, 2017.
[9] The Internet Archive, "Wayback Machine Hits 400,000,000,000!," Internet Archive Blogs, 09-May-2014. .
[10] M. Broussard and K. Boss, "Saving Data Journalism," Digital Journalism, vol. 6, no. 9, pp. 1206–1221, Oct. 2018.
[11] N. Brügger, "Website history and the website as an object of study," New Media & Society, vol. 11, no. 1–2, pp. 115–132, Feb. 2009.
[12] Pew Research Center, "Where People Get Their News," Pew Research Center | Global Attitudes & Trends, 04-Oct-2007. [Online]. Available: https://www.pewglobal.org/2007/10/04/chapter-7-where-people-get-their-news/. [Accessed: 20-Mar-2019].
[13] Pew Research Center Journalism Project, "The Growth of Digital Reporting," Pew Research Center's Journalism Project, 26-Mar-2014. [Online]. Available: http://www.journalism.org/2014/03/26/the-growth-in-digital-reporting/. [Accessed: 26-Mar-2014].
[14] J. Rothenberg, "Ensuring the Longevity of Digital Documents," Scientific American, vol. 272, no. 1, pp. 42–47, Jan. 1995.
[15] D. S. Rosenthal, "Emulation & Virtualization as Preservation Strategies," 2015.
[16] S. Granger, "Emulation as a Digital Preservation Strategy," D-Lib Magazine, vol. 6, no. 10, Oct-2000.
[17] L. Johnston, "Preserving News Apps | The Signal," 11-Mar-2014. [Online]. Available: //blogs.loc.gov/thesignal/2014/03/preserving-news-apps/. [Accessed: 16-Jul-2017].
[18] K. Rechert, I. Valizada, S. D. von, and J. Latocha, "bwFLA – A Functional Approach to Digital Preservation," PIK - Praxis der Informationsverarbeitung und Kommunikation, vol. 35, no. 4, pp. 259–267, 2012.
[19] D. von Suchodoletz and J. van der Hoeven, "Emulation: From Digital Artefact to Remotely Rendered Environments," International Journal of Digital Curation, vol. 4, no. 3, pp. 146–155, Jul. 2009.
[20] S. Anderson, E. Cochrane, E. Gates, and J. Meyerson, "About EaaSI – Saving Software Together," About EaaSI – Saving Software Together, 2018. [Online]. Available: https://www.softwarepreservationnetwork.org/eaasi/. [Accessed: 13-Mar-2019].
[21] J. Kim, "Access and Discovery of Born-Digital Archives," New York University, Jun. 2015.
[22] F. Chirigati, R. Rampin, D. Shasha, and J. Freire, "ReproZip: Computational Reproducibility With Ease," presented at the 2016 ACM SIGMOD International Conference on Management of Data (SIGMOD), San Francisco, USA, 2016, pp. 2085–2088.
[23] I. Kreymer, Core Python Web Archiving Toolkit for replay and recording of web archives: webrecorder/pywb. Webrecorder, 2013.
[24] C. Zapponi et al., "Election 2015: The Guardian poll projection," The Guardian, 07-May-2015. [Online]. Available: http://www.theguardian.com/politics/ng-interactive/2015/feb/27/guardian-poll-projection. [Accessed: 16-Mar-2019].
[25] C. Zapponi, C.-J. (Apple) C.Fardel, and S. Clarke, Tracking the UK election polls for 2015. The Guardian, 2016.
[26] V. Steeves, R. Rampin, and F. Chirigati, "Using ReproZip for Reproducibility and Library Services," 1, vol. 42, no. 1, pp. 14–14, 2018.

Jisc Open Research Hub

Integrate, Deposit, Discover and Preserve

Paul Stokes
Jisc, UK
Paul.stokes@jisc.ac.uk
0000-0002-7333-4998

John Kaye
Jisc, UK
John.kaye@jisc.ac.uk
0000-0002-4400-4252

Dom Fripp
Jisc, UK
dom.fripp@jisc.ac.uk
0000-0001-5352-4666

Tamsin Burland
Jisc, UK
Tamsin.burland@jisc.ac.uk
0000-0002-5129-979X

Tom Davey
Jisc, UK
tom.davey@jisc.ac.uk
0000-0002-0319-5574

Matt McGowen
Jisc, UK
matt.mcgowan@jisc.ac.uk

Abstract – Jisc's Open Research Hub (JORH) integrates a number of repository, preservation, reporting and storage platforms as a one stop shop for researchers and research managers. The service offers both open source and proprietary systems and allows data and metadata to be shared openly if required. The platform has been developed through years-long consultation with the UK HE research sector and sector bodies, along with contributions from both in-house Jisc and third-party experts.

The need for such a solution has arisen from the sector's desires to achieve several, shared aims, including: greater collaboration; tackling the reproducibility crisis; enabling better research; and meeting funder requirements.

Jisc's custom-built repository—the Jisc Research Repository—is part of the Jisc Open Research Hub. It's built upon an extensive data model and rich messaging layer, providing users with a clean, simple, and easy-to-learn interface for the deposit, approval, and discovery of a range of outputs. In particular it allows for a seamless end to end experience for the user; from deposit straight through to preservation.

Jisc's position in the UK higher education / research sector, as well as the scale of the service provides us with many domain-specific insights to share with iPRES delegates, ranging from the broad methods mentioned above, down to individual design decisions informed by our research and domain expertise.

Keywords – Research Data, Integration, Preservation, Shared Services, Repository to Preservation

Conference Topics – The Cutting Edge: Technical Infrastructure and Implementation;

Designing and Delivering Sustainable Digital Preservation.

I. Introduction

Jisc's Open Research Hub, integrates a number of repository, preservation, reporting and storage platforms as a one stop shop for all users of repository and preservation systems. The service offers a range of systems—both open source and proprietary—and a range of potential integrations, both in terms of data (and metadata) sources and endpoints. It allows data and metadata to be managed, preserved and shared as openly as possible and as securely as needed.

This paper explores the design and development philosophy of JORH and presents the features that make it well placed to fulfil many different types of repository and preservation use cases.

This paper will be of interest to content generators, developers, integrators, vendors, repository managers, curators, research data managers, support staff, and data end-users.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

II. Origins and Drivers

Jisc has been involved in development of research data management projects for over a decade. Between 2009-11 Jisc funded work that developed a number of important tools, technologies and services within universities. Programmes that run between 2011-13 built on earlier work and also supported the development of skills and expertise for specific disciplines and support staff. After this, the next step was to broaden the implementation of previous work

This led to the Research at Risk programme (2014-2016) led by Jisc in partnership with RLUK, RUGIT, SCONUL and UCISA, and informed by numerous stakeholder consultation events. The stated aim of the programme was *"To provide infrastructure, advice and tools to support universities in establishing good data management practice as a core part of their research function."* An additional aim was to take the lead in developing a sector-owned direction of travel for Research Data Management (RDM) and develop a community governed infrastructure.

All this work led to the Jisc Research Data Shared Service project—the output of which transitioned to service last year.

A. *Sector sanctioned*

In 2015 a report [1] by Jisc and a number of other sector bodies including SCONUL, RLUK and UCISA looking at RDM and universities identified 5 key areas for action in research data management

1. Policy development and implementation
2. Skills and capabilities
3. Infrastructure and interoperability
4. Incentives for researchers and support stakeholders
5. Business case and sustainability

6. Identified within the Infrastructure and interoperability actions strand the following themes were identified *"...there is clear demand for national shared services for research data management... ...the potential economies of scale are attractive to the community... ...there is also demand for a national approach to data perseveration..."*

B. *Drivers*

Research active HEIs now operate in an environment which requires them to address issues around research data management and 'open research'. There are also mandates from publishers and research funders regarding making data findable and accessible to others, and preserving the data for a given length of time.

In addition, the recent changes in the laws around storing personal data means that universities need to take more interest in how researchers manage datasets containing personal information.

There are also increasing concerns around research integrity and reproducibility, calls for greater access to the original research data.

Finally, there is an increasing recognition of the academic and societal benefits of open research, and initiatives to openly share both the findings of the research and the data these were based on.

Given all these issues, it's increasingly important that universities, carefully manage, store, share (if appropriate) and preserve their digital research outputs.

Jisc was tasked with building a service primarily for Institutions to address the emerging requirements and challenges of RDM

It's a big challenge for institutions to take on individually, especially small institutions without much expertise in the area of research data and digital preservation. Hence the concept of a shared service.

Given their position in the sector, and previous work we had been involved in, Jisc was perfectly placed to lead such a project.

III. Development process

A. *Shared service and co-design*

The platform has been developed through years-long consultation with the UK HE research sector (over 70 universities took part in the consultation) and sector bodies, along with contributions from both in-house Jisc and third-party experts. The functional requirements fell out broadly into 4 categories:

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Capture & reuse; Preservation; Reporting; Advice and best practice.

The requirements collected posed a challenge for building the shared service. To meet these requirements ultimately we needed:

- A multi-tenant system—which enabled it to be an affordable and scalable cost effective solution for the sector.
- Multi content types—institutions wanted to be able to store all the outputs of research—data, articles, theses, software and method—in one place.
- A flexible system—one where institutions used only the parts of the system they needed and which allowed for interoperability with other systems.
- A good user experience—one that eased the pain of preservation and reduced the need for expertise in digital archiving.
- Reporting
- Most importantly community governance, a national service build for the UK sector, by the UK sector.

To build a service which could meet these challenges, we mapped out the work flow shown in

To achieve this workflow and interoperability, we developed a publish-subscribe messaging layer, based on an open, canonical data model and open APIs.

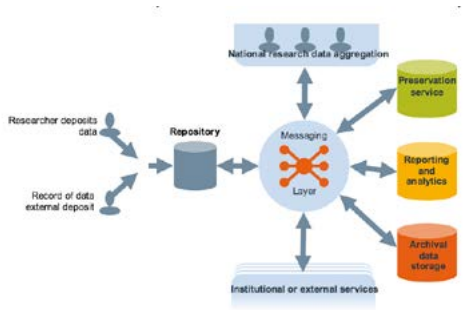


Figure 1 - Service workflow summary

Our alpha Minimum Viable Product (MVP) demonstrated automatic ingest of data files into the preservation system from two different repositories (one open source and one commercial), into two different preservation systems (again one open source and one commercial).

B. *Agile development and the current Open Research Hub*

The project has been developed using agile techniques throughout. It has also been extensively tested with a group of Pilot institutions through a co-design process. Inevitably this has meant an evolution from the original proposed architecture devised to fulfil the initial set of requirements. In the early stages of the project many more than the current two repositories and two preservation components were incorporated into the overall system. However, in order to achieve a working MPV is as fast a time frame as possible, these were winnowed down to the current two of each. On the other hand, other systems can still be used with JORH through the use of connectors and the open APIs.

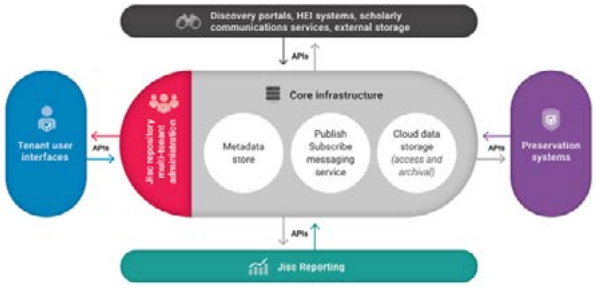


Figure 2 - System architecture

The core infrastructure is where the messaging layer sits, along with the metadata store and data storage.

Initially we were developing an existing open source repository to sit alongside this, but we couldn't find a product which match all our requirements. It became clear that we could get the functionality required from the Core infrastructure, and just needed to build a light weight front end for our tenants and our administration purposes. These front ends (the admin interface and the tenant interface) make up the Jisc Research Repository, a multi-tenant, multi-content, interoperable repository.

The infrastructure interacts with the two other components of the system—Preservation and Reporting—via open APIs

Interoperability with other systems, such as other repositories, Current research Information Systems (CRIS), scholarly comms services, external storage,

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

and discovery portals is also facilitated by adaptors based on the open API specification.

The service as a whole forms the Jisc Open Research Hub. It allows users to deposit data quickly and easily (with some of the required metadata being pulled in from other institutional systems), and have that data automatically preserved, reported upon and exposed to discovery portals.

C. *What are the next challenges in research and are we prepared for them?*

Jisc built JORH around a national need for affordable, accessible and compliant research data management infrastructure in higher education. Now the platform exists, how can it further address the needs of the wider UK research sector and the international open science agenda

JORH has the capacity to support and apply metadata standards for publishing various stages of research in a standardized and robust fashion. This could include anything from publishing a hypothesis, to methods, to null and significant results, bringing visibility to the vast array of findings that are never published. JORH can bring interconnectivity to these component pieces of research within by integrating with tools from across specific research domains and catalyse new habits in research that supports openness, and verifiable science.

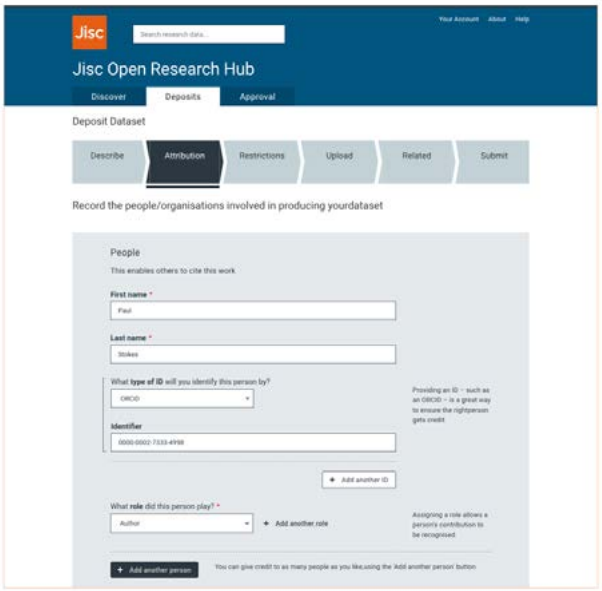
While the Hub was built to address the needs of research data curation, its adoption of open, best practice standards means it has the potential to allow the service to handle a much wider range of digital research objects, including Open Access articles, theses and software. The data model, rich messaging layer and an open API facilitate interoperability with other institutional and scholarly communications systems. This provides the potential for the Hub to underpin infrastructure capable of meeting the requirements of an ever-evolving open research agenda.

Artificial intelligence enables great opportunities for the automation of discovery, computation and analytics for data driven research at scale. Yet training effective algorithms takes effective training data. JORH can help by providing a place for well described and verified datasets to be made openly available. This is a first but important step to a more transparent and interpretable version of AI.

IV. Key design features

A. *Open and extensible*

JORH is a cloud-based, community governed, multi-tenant solution for universities and other research institutions to manage, store, preserve and share their published research data. Based on existing open standards, the service's open and extensive data model incorporates best practice from across the sector, including DataCite, CrossRef, CERIF, Dublin Core and PREMIS.



B. *User Experience*

Jisc recognised the importance of delivering this service with a compelling user experience, and invested greatly into achieving this aim.

A key challenge for the viability and effectiveness any digital solution is achieving engagement from users. This is particularly true in the research arena, where people are required to engage with a number of complex systems in order to achieve their goals, as well as the aims of their institution, funders, and other stakeholders.

To this end we engaged a user-centered-design approach. Beginning by researching to develop a

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

rich understanding of the people who will use the system (their motivations, their pain points, their skills and knowledge) we were able to set about designing a solution that is meaningful, compelling, and engaging. This informed a range of design decisions, some fundamental (for example, how to break complex tasks into smaller but meaningful ones); and some much smaller ones (for example,. which of the myriad words we could use for this field will people most likely understand).

The ability to deliver a good quality and compelling user experience requires an iterative approach to problem solving: a willingness to generate, test, and improve upon solutions; and to do so through ongoing engagement with real users. Several rounds of user acceptance testing (UAT) at key points in the design process have enabled Jisc to have confidence in the fundamentals of the service, and (through benchmarking) allowed the team to demonstrate continual improvement in usability, as well as identifying areas that require improvement.

C. *Data model*

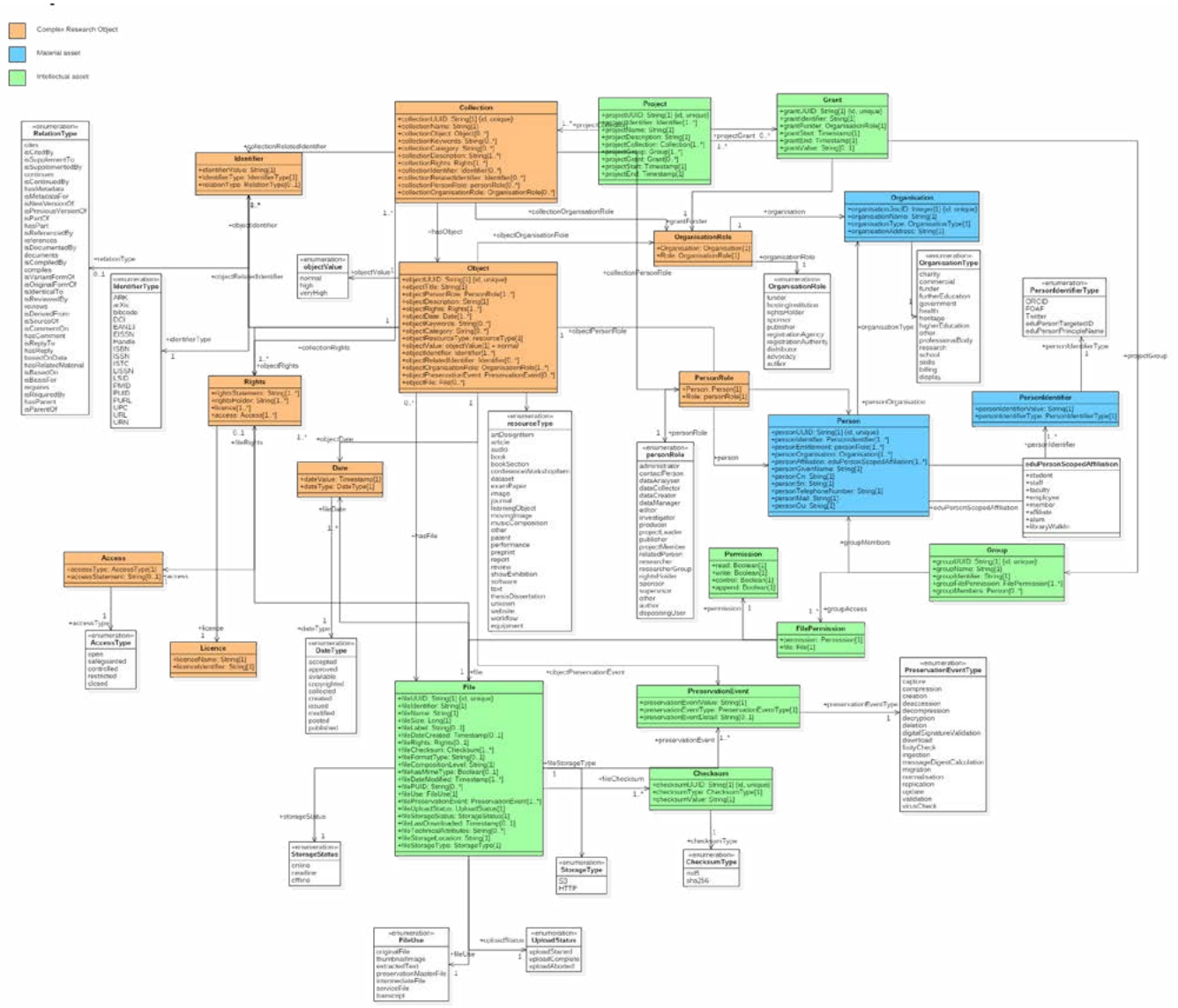


Figure 4 - JORH canonical data model

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The canonical data model shown above in Figure 4 underpins the message specification used throughout (and in particular in the messaging layer and APIs). Built through consultation with the UK HEI sector and utilises best practice in popular metadata standards, schemas and ontologies, such as Datacite, Premis and Cerif. The data model is a living document, iterates based on the requirements of the service and the third-party integrations it supports. The documentation and the model (including every version) are openly available in a Github repository [2] and can be used as the single source of truth for metadata mapping between the service and any other domain. The interoperability framework the data model provides is critical for JORH to integrate with other services and systems, and why it is essential to make this document public and transparent.

V. Conclusions

In developing this service Jisc have taken an innovative approach to a problem—that of preserving research data—which is often approached purely from the technical angle. It's not a trivial task, even for specialists in preservation, but it is a task that is relatively well understood.

But in the real world preservation is in the hands of non-specialists, for whom this isn't their day job. They have little or no digital curation experience, would rather someone else did it and would rather be doing their research. They need to just be able to do it without re-training to be a digital archivist and with very little effort on their part.

So JORH has been deigned with this goal in mind; to be as cutting edge as possible, to be as open as possible in order to be as integrated as possible, to have the potential to be metadata rich, but without the need for manual entry and curation expertise.

In a nutshell; Integrated, innovative, extensible and user friendly.

References

[7] S. Brown, R. Bruce, D. Kernohan "Directions for RDM in UK universities" https://www.fosteropenscience.eu/sites/default/files/pdf/1240.pdf

[8] https://github.com/JiscRDSS/Canonical-data-model/tree/master/Data-Model

The Portico Straight-to-Ingest Project

Sheila Morrissey
Portico
USA
sheila.morrissey@ithaka.org
https://orcid.org/0000-0003-2069-121

Amy Kirchhoff
Portico
USA
amy.kirchhoff@ithaka.org
https://orcid.org/0000-0002-3136-142

Abstract – This paper describes a soon-to-be-concluded six-month project at Portico. The goal of this project is to attack preservation cost drivers with automated, data-driven analysis of problematic content, revamped policies and manual procedures, and targeted development of automated management and remediation tools. The desired outcome, in addition to more quickly ingesting content into the archive, is the most efficient, economical application possible of technical staff to the automation of problematic content handling, as well as a streamlining, rather than expansion, of manual production processes for content continually growing in quantity and complexity.

Keywords – Sustainability, economies of scale, automation, managed ingest, minimal ingest

Conference Topics – Designing and Delivering Sustainable Digital Preservation; The Cutting Edge: Technical Infrastructure and Implementation.

I. Introduction

The activities associated with selection, acquisition, and other pre-ingest processing (including quality assurance) of content are well-understood to be critical cost drivers for digital preservation (See, for example, [1], [2], [3]).

Sustainable preservation services must realize economies of scale and of the strategic application of automation, so that the growing size of an archive -- the volume of content processed, the ever-increasing amount of storage required -- doesn't mean comparably increasing costs. Additionally, these services must continually monitor what are new or, often, continually changing input streams of content, to ensure that variable content does not result in equally variable, unpredictable, and (in the worst case) runaway costs. As the Digital Preservation Coalition (DPC) noted in its review of digital preservation of non-print legal deposit materials at the British Library,

"Increasing volume, complexity and unpredictability of content place considerable strain on digital preservation workflows in a variety of ways. Greater volumes (both in numbers of items and sizes of component files) place strains on the workflows that must process them, requiring more resilient software processes and greater workflow automation to enable issues to be resolved without backlogs arising.... Unpredictability requires greater flexibility to react to changes in content and its supply. Deposited data that doesn't conform to previously encountered norms must be detected and workflows adapted to process it. The accuracy and completeness of digital preservation activities will be impacted with adaptation to meet these challenges." [4]

Certainly this has been a challenge for Portico. Portico is a community-supported digital preservation service for electronic journals, books, and other content. Portico is a service of ITHAKA, a not-for-profit organization dedicated to helping the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. Portico serves as a permanent archive for the content of, at present, 606 publishers (from 60 countries, and on behalf of over 2000 learned societies and associations), with 32,004 committed electronic journal titles, 1,379,448 committed e-book titles, and 220 committed digitized historical collections. The archive currently contains over 103 million archival units (journal articles, e-books, etc.), comprising over 1.67 billion preserved files. Portico is sustained by the support of over 1000 libraries in 23 countries.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

II. Process and Procedures: Content Workflows

How does content make its way into the Portico archive?

First, Portico works out an agreement with a publisher. Besides ensuring the legal right to preserve content, the discussion in this pre-processing phase includes sharing sample content, so that Portico can analyze publisher metadata, as well as the conventions used in packaging together and delivering a collection of files that comprise, for example, the articles of a single issue of a journal.

Portico reverse-engineers those conventions into declarative XML "profiles," which direct the aggregation of individual files in a delivery into complex multi-file digital objects, such as journal articles. These profiles also enable Portico automatically to assign a functional descriptor to each file making up the complex digital object (what we term a "functional unit type"); to determine if an expected component is not in the package; and, once all files are grouped, to detect "left-over" files that are not attached to an archival unit, suggesting investigation is required to determine if the file a new component now being provided, or simply "noise" that can be ignored in subsequent deliveries (for example, "thumbs.db" files in Windows folders). We also develop an XSL transform to normalize publisher descriptive metadata into the Journal Article Tag Suite (JATS) or Book Interchange Tag Set (BITS). Both the original and normalized metadata are preserved in the archival unit.

Those discussions also elicit details necessary for automating publisher submission of content, typically via FTP. These details are again translated into declarative XML rules in Portico "Fetcher" and "Loader" profiles, which drive the automated fetching of content, and assembling of files into batches to be loaded into the content processing (ConPrep) automated workflow system.

When these steps comprising a "publisher setup" are complete, automated processing can begin. The ConPrep workflow, using rules in the profile, assembles the various files in a batch into archival units. Publisher metadata is validated against publisher-provided XML schema, and transformed to JATS or BITS. Technical metadata is assembled about each component file in an archival unit, and preservation metadata (including descriptive, technical, rights, and event metadata) is created and packaged with each archival unit. Accepted archival units are then ingested into the archive, where they are replicated, and where periodic fixity checking is performed.

Portico's content and processing model can accommodate updates to the content that publishers might provide. The updated version of the content is packaged with the original, and both versions are maintained, with accompanying metadata, as a single archival unit. They also accept, for example, technically "deficient" component files such as supplementary images that do not pass JHOVE validation.

What the processing model would not permit to be ingested into the archive are archival units that fail the profile rules described above: those missing mandatory components such as the PDF of an article if the publisher does not provide XML full text, or those with ill-formed or invalid XML descriptive metadata, or those missing XML metadata files entirely. Further, since the workflow is a batch processing system, all archival units in the same batch as a defective one are retained in the ConPrep system, until all problems for all archival units are resolved.

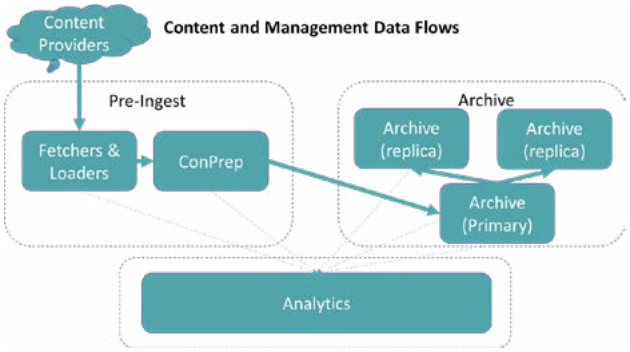


Figure 1

III. Processes and Procedures: Adapting to Growth and Change

By long-standing policy, Portico does not make editorial decisions: Portico does not "correct" the scholarly record. If publisher-provided metadata, for example, states an article belongs in issue 42, when it actually appeared in issue 24, Portico preserves

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

that information as it was received. We preserve the content as it was published.

Nevertheless, there is a quality control function in the handling of publisher-supplied content. Portico's original ingest policies with respect to the cohesion and completeness of content, and the content workflows built in conformance to those policies, were predicated on the detection of certain classes of content defects (for example, missing article components such images, or invalid XML metadata), and on obtaining repaired content and metadata before ingest into the archive. These policies in turn were based on other assumptions:

- that publishers are both willing and able to provide corrections to content and metadata
- that all such defects are equally significant barriers toward making content available, accessible, and useful over the long term
- that there would be a minimal amount of content in problem state, since the content is actively in production at the publisher

The consequence of this policy of "perfect-only" content in the archive was that a significant body of content, across all publishers and content types, was stalled in the content ingest workflow holding queue – even if that content was published with those flaws. This content is expensive. It requires staff people to frequently touch it and manually manage the storage space in ConPrep. In addition, unlike content in the archive, content in the ConPrep queue is not replicated and is not subject to the archive regimen of regular checksum computations to detect "bit rot."

Because bibliographic metadata about archival content is an outcome from processing by ConPrep, the considerable amount of content in the holding queue also was not visible in Portico's holdings data, effectively making them "invisible" to us and to our participants.

And, perhaps most crucially from a cost management perspective, because the ConPrep system is, by design, a transactional system, focused on processing of specific content streams, Portico did not have capabilities for looking across content in those streams, or the processing information detailing defects detected, to make informed decisions on aggregating and prioritizing problem-resolution of content.

Additionally, Portico's very success in attracting more and more publishers to entrust content to our care meant that the staff burden of managing problem resolution, and publisher interactions, was continually increasing.

As we have described elsewhere [5], Portico's recently re-architected, horizontally scalable technical infrastructure is easily able to accommodate geometric growth in content over the past fifteen years. What has proved more problematic has been the relative increase in the number of small and medium publishers, as the total number of publishers and content streams has grown year by year. Currently, over 55 per cent of publishers providing content to Portico are classfied as small or medium.

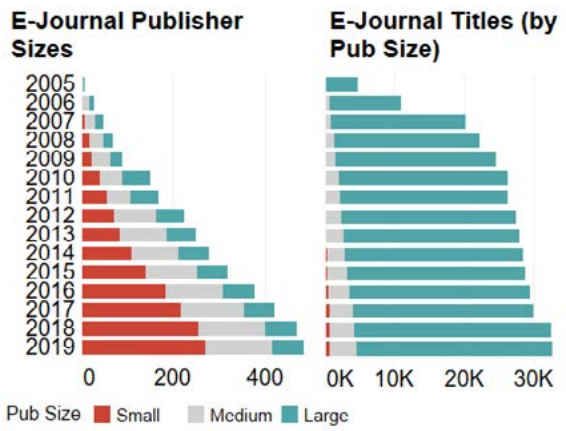


Figure 2

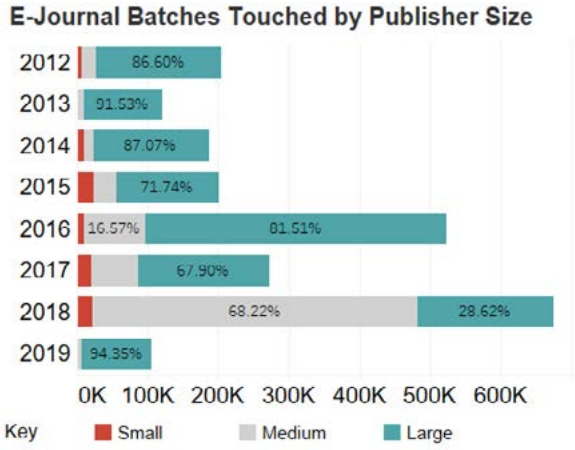


Figure 3

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Given that many of these new content streams are in the "long tail" of scholarly publishing, comprising publishers with fewer technical resources, or simpler, sometimes manual production processes, our experience has been that the number of problems per provider has begun and will to continue increase as well.

Another way of saying this is that an increasingly small subset of the content coming into Portico necessitates an increasingly greater amount of staff time in manual interventions to deal with problems in the automated processing of that content. In 2018, a third of all batches, across all content types, required some sort of manual handling.

So, while we now had the capability to scale our automated systems horizontally in a reasonably effortless and economic fashion, the same was not true of our manual processes. A fundamental change to ingest policy was required to meet this challenge.

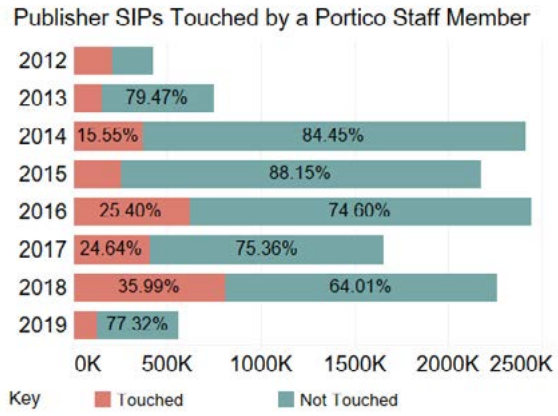


Figure 4

IV. The Straight-to-Ingest Project

A. *Goals*

The goal of the Straight-to-Ingest (S2I) project was simple: to make changes to our production processes (both automated and manual) that would enable us to move all content, regardless of its "cleanliness" or completeness, into the archive. This would enable us to eliminate the backlog of defective content in the ingest queue, to scale our manual production processes to meet the demands of increasing growth and variation in quality of the content entrusted to Portico, and to make informed decisions on how to prioritize the handling of the most seriously defective content.

This does not mean that we intend to ignore defects in content preserved in Portico. Rather, we have moved the point in our automated and manual workflows at which we deal with bad content, from its original location (correcting all defects in ConPrep before ingest) to new, post-ingest processing. There, we can employ the new analytic tools of the archive, as well as a comprehensive view of all content deposited, to inform our interventions to make repairs to defective content, if possible.

Specifically, S2I was designed to:

- scale our manual production processes to leverage the horizontal scalability of the new technical architecture to deal with increasing volumes, types, and sources of content, of widely varying quality, without necessitating increasing staff
- move all content into secure, managed, long-term preservation and out of the transactional content processing system where defective content is often stalled
- provide greater transparency to both publishers and libraries about the current state of content committed and submitted to Portico
- bring information about all content submitted to Portico into our holdings metadata, and provide a more complete and correct picture of content in our care
- leverage new analytics capabilities to provide us with a deep understanding of what problems exist in the content, and to enable us to make informed decisions about how to allocate staff resources to address them
- flexibly and rapidly address major and urgent content defects
- eliminate redundant interactions with publishers over already-corrected content, or content for which publishers are unable to provide corrections
- minimize or eliminate manual interventions to correct defective content

B. *Process and Procedural Changes*

We now "grade" archival units as they move into the archive. These "grades" will enable us to prioritize resolution of the biggest problems of "broken"

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

content. Some of those problems are with the content itself, and are of varying severity for the "community of use" of preserved content. For example, if we have descriptive metadata in an XML file that references a missing figure graphic in the article abstract, but we also have a PDF file with all the content of the article, including that figure, the article is completely usable. That is, a reader has enough information, through the PDF, to understand the full intellectual content of the article. That article would be ingested into the archive with a grade of B, along with information about the file referenced, the fact that it was missing, and the rationale for giving the article less than an A grade.

Another benefit of "grading" is that it surfaces an accumulation of errors in a particular content stream which seems to indicate a dramatic shift in the regular practices of the publisher, that has not been communicated to Portico, but which perhaps mean we need to make changes to our profiles, transforms, or tools.

We updated publisher profiles to indicate, on a per-stream basis, the minimum grade allowable for content from that stream to be ingested into the archive. Also, in addition to ConPrep workflow changes to categorize errors and grade content, we updated our preservation metadata schema to include grade information about each archival unit. This grade information is displayed when the archival unit is accessed.

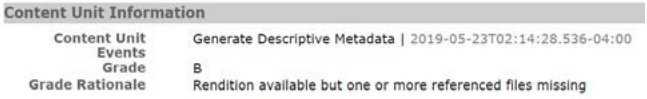


Figure 5

In addition, for any archival unit with a grade less than 'A,' a new functional component has been added to the package containing the preserved object – a JSON file containing detailed error tracking information about the nature of the detected defects. A long standing requirement of the Portico archive is that it is "bootstrap-able" – everything one needs to reconstitute the archive is contained in the archival units themselves. This means we must capture all these errors in a machine (and human) readable way within the archival units. All error information is also cached in our analytics system.

We have developed new reports, both for production staff and for communicating to publishers about problematic content. These reports enable us to manage and report problems by publisher and by defect type. Additionally, we have developed new workflows to capture the "feedback loop" of publisher responses (or of updates to problematic content without accompanying feedback from the publisher). This enables us to eliminate duplicate or out-of-date reporting, as well as to detect publisher remediation, or, should it be the case, to record a publisher response indicating they are unable to repair defective content.

C. *Early Outcomes*

The first use case of problematic content implemented in the project is "Grade B" content that is missing referenced ancillary files, but has a component that provides the full "intellectual content" of the archival unit. In the first few weeks of processing, we are finding, as we hoped, that batches containing such problems are spending less time in ConPrep. Additionally, previously blocked content without problems, but in the same batch as problem content, is now going into archive, where it is being replicated and is available for access

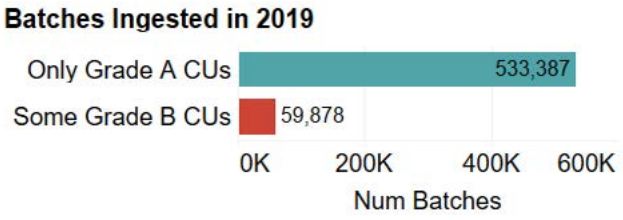


Figure 6

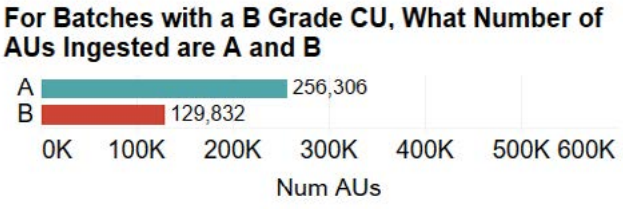


Figure 7

We will be looking over the next several months to see if our new reports and automated feedback loops, along with streamlined processes for communication with publishers, is significantly reducing the management burden on our production staff.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

D. *Next Steps*

Over the coming months, we will expand the grading scheme to include other categories of errors. A next likely use case is using alternative methods (regular expression parsing, natural language processing) for extracting at least minimal descriptive metadata for archival units where publisher-supplied XML is not well-formed and valid.

Additionally, we will be building improved "dashboard" reporting and alerts that, based on accumulated process analytics and grade information, indicate some uncommunicated change in a publisher's content stream – or, perhaps, a previously undetected error in Portico's workflow configuration, tools, or transforms.

V. Implications for Best Practice

As noted by Jurik et al in their description of minimal effort ingest at the State and University Library, Denmark [7], moving content that, according the policies of a given archive or the "submission agreement" worked out with a contract provider, is somehow incomplete or deficient into the archive implies a change to the function model of OAIS [9]. It moves a function from the "Ingest Functional Entity" to the "Archive Function Entity". OAIS assumes only wholly perfect or complete content is allowed into the archive; it also assumes on-going, as well as tractable, communication with content providers. Nor does OAIS provide a very rich model for what Caron et al referred to as enrichment and enhancement of digital content after ingest [6]. Realistically, we feel that not only must we find a way to accommodate what we would consider to be less than "perfect" data, but also that we must make pragmatic decisions to ensure the overall sustainability of the archive, and of digital preservation as a whole

Ideally, Portico would like all content coming into the archive to be "born preservable" – complete, correct, and supported by ample contextual information. We feel however, as Stephen Abrams has suggested [8], that while OAIS provides a useful model for measuring the trustworthiness of archival processes and procedures, this must be balanced against, and completed by, a measure of the effectiveness of communication with a future user – a robust topic to be addressed by the next iteration of preservation standards.

References

[1] H. Uffen and T. Kinkel, "Controlling the Cost of Digital Preservation," Dutch Digital Heritage Network, 2018 https://www.netwerkdigitaalerfgoed.nl/wp-content/uploads/2019/02/Article_DDHN_Controlling-the-cost-of-digital-preservation.pdf.

[2] Charles Beagrie LTD, "Costs Factsheet," CESSDA Cost-Benefit Advocacy Toolkit, DOI: 10.18448/16.0003.

[3] Mosely, S., Moran, J., McKinney, P., Gattuso, J., 2016. Conceptualising Optimal Digital Preservation and Effort, iPRES2016 *13th International Conference on Digital Preservation,* Bern, Switzerland.

[4] Digital Preservation Coalition, *Non-Print Legal Deposit Digital Preservation Review: Final Report*, Dec 2017, https://www.bl.uk/britishlibrary/~/media/bl/global/digital%20preservation/non-print-lega-deposit-digital-preservation-review.pdf.

[5] K. Wittenberg, S. Morrissey, "The Evolution of Infrastructure: Making a Renewed Investment in Preservation at Portico," *The Scholarly Kitchen,* Oct 2018, https://scholarlykitchen.sspnet.org/2018/10/17/guest-post-the-evolution-of-infrastructure-making-a-renewed-investment-in-preservation-at-portico/

[6] B. Caron, J. De La Houssaye, T. Ledoux, S. Reecht. "Life and Death of an Information Package: Implementing the Lifecycle in a Multi- Purpose Preservation System," iPRES2017 *14th International Conference on Digital Preservation,* Sep 2017, Kyoto, Japan.

[7] B. Jurik, A. Blekinge, K. Christiansen, "Minimal Effort Ingest," iPRES2015 *12th International Conference on Digital Preservation*, Nov 2015, Chapel Hill, NC.

[8] S. Abrams, "Nothing succeeds like success: An approach for evaluating digital preservation efficacy," iPRES2018 *15th International Conference on Digital Preservation*, Sep 2018, Boston, MA. https://escholarship.org/uc/item/8cp160c9

[9] Space Data and Information Transfer Systems. ISO 14721:2012 Open Archival Information System (OAIS) - Reference Model. The International Organization of Standardization, 2012.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Getting digital preservation data out Wikidata

Katherine Thornton
Yale University Library USA
katherine.thornton@yale.edu
https://orcid.org/0000-0002-44990451

Kenneth Seals-Nutt
Yale University Library USA
kenneth.seals-nutt@yale.edu
https://orcid.org/0000-0002-5926-9245

The Wikidata knowledge base provides a public infrastructure for creating and syndicating machine-readable data about computing resources. We have prepared a set of queries that can be used to gather data sets relevant to digital preservation from Wikidata. We present these data sets in the context of the Wikidata for Digital Preservation portal (Wikidp). Wikidp is a free software portal that allows people to explore data related to digital preservation from the Wikidata knowledge base. Structured data about file formats, the many versions of software titles, and computing environments, are already available in Wikidata. The content of Wikidata is licensed under the Creative Commons Zero license, meaning that anyone can reuse the data for any purpose. The content in Wikidata is available in more than 300 human languages. The data in Wikidata is FAIR data, and it is linked open data. Our portal provides an interface designed for the needs of the digital preservation community.

Wikidata, digital preservation, linked open data Designing and Delivering Sustainable Digital Preservation; Building Capacity, Capability and Community

I. Introduction

Wikidata is the knowledge base that anyone can edit. Twenty thousand people edit Wikidata each month [1]. These editors add structured data in the form of statements of facts, and references for those statements, to the knowledge base. The Wikidata community has created more than five thousand properties for the knowledge base, and editors compose statements using these properties. Many people in the Wikidata community are personally interested in the domain of computing. Collectively, these editors have curated hundreds of thousands of statements related to software and hardware.

II. What is wikidata?

Wikidata is a cross-domain knowledge base of structured data. Simply put, it is a database of facts that both humans and machines can edit and consume. Wikidata went live in late 2012 [2]. The infrastructure of Wikidata is collaboratively built via commons-based peer production [3]–[5]. Commons-based peer production is the name given to open collaboration systems where users are creating content under the agreement that all content will remain in the public domain. This means that all of the work products of the community are free to be reused by others. The peer-production aspect refers to how users coordinate work themselves. Wikidata is edited by volunteers from all over the world in more than 350 languages [6].

In addition to a free software infrastructure, the Wikidata community also publishes all content in the knowledge base under a Creative Commons Zero License. The Wikidata community makes dumps of previous versions of the content of the knowledge base available. The infrastructure of the Wikidata knowledge base is maintained by an international community of people. For cultural heritage institutions who find structured data in Wikidata relevant for their work flows, this means that there will be much less staff time necessary to design, build and maintain infrastructure for this data. For cultural heritage institutions with limited digital preservation budgets, this means that they can now access descriptive and technical metadata for tens of thousands software titles and more than three thousand file formats without having to create manage or maintain that data locally.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

III. What data can I reuse from Wikidata?

The Wikidata community maintains a public SPARQL endpoint. As of October, 2017 the endpoint was consistently handling 8.5 million SPARQL queries per day [7]. Writing SPARQL queries for the Wikidata endpoint allows users to search for data about resources in a flexible way. SPARQL queries enable us to search for file formats by media type, or by file extension, etc. They allow users to search for software titles by their readable file formats, or to search for software titles published within specific windows of time. We can use SPARQL queries to search for software titles by genre, to search for technical specifications that describe a particular file format, or to search for a digitized copy of a user guide for a piece of legacy software. Because Wikidata is a cross-domain knowledge base, the range of data combinations allow users to query data that span technical metadata as well as descriptive metadata aspects of these resources

A. Getting Data from Wikidata

Humans can view data in Wikidata via any of the wiki pages. To access data in bulk, users can access the MediaWiki API[1] or the Wikidata Query Service[2]. Users of the Wikidata Query Service SPARQL endpoint can request subsets of the data contained in Wikidata that match specific patterns. Users can design queries that take advantage of different Wikidata properties such as the examples in the figures below.

1. Return all software titles known to read .dxf files, see Figure 1.

```
SELECT DISTINCT ?app ?appLabel WHERE {
  ?app (wdt:P31/(wdt:P279*)) wd:Q7397;
    wdt:P1072 wd:Q691652.
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language
      "[AUTO_LANGUAGE],en".
  }
}
```

Figure 1: A screenshot of a SPARQL query to request software titles that can read .dxf files.

Try this query! Code for this query.

2. File formats used for 3D graphics, see Figure 2.

```
SELECT ?item ?itemLabel WHERE {
  ?item wdt:P31 wd:Q235557;
    wdt:P366 wd:Q189177.
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language
      "[AUTO_LANGUAGE], en".
  }
}
```

Figure 2: A screenshot of a SPARQL query to request file formats used for 3D data.

Try this query! Code for this query.

3. Sequence alignment software with date of publication, programming language and license, see Figure 3.

```
SELECT DISTINCT ?software
  ?softwareLabel ?licenseLabel
  ?langLabel WHERE {
  ?software (wdt:P31/(wdt:P279*))
    wd:Q7397. {
    ?software wdt:P366 wd:Q827246.
  } UNION {
    ?software wdt:P366 wd:Q1377767.
  }
  OPTIONAL {
    ?software wdt:P275 ?license.
  }
  OPTIONAL {
    ?software wdt:P277 ?lang.
  }
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language
      "[AUTO_LANGUAGE], en".
  }
}
```

Figure 3: A screenshot of a SPARQL query to request sequence alignment software with date of publication, programming language and license.

Try this query! Code for this query.

[1] https://www.mediawiki.org/wiki/API:Main_page

[2] https://query.wikidata.org/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

4. File formats to which the defining ISO standard been linked, see Figure 4.

```
SELECT ?format ?formatLabel
  ?standardLabel WHERE {
  ?format (wdt:P31/(wdt:P279*))
    wd:Q235557;
    wdt:P503 ?standard.
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language
        "[AUTO_LANGUAGE],en".
  }
}
```

Figure 4: A screenshot of a SPARQL query to request the a list of file formats described by ISO standards.

Try this query! Code for this query.

5. What is the signature of the SteroLithography file format, see Figure 5.

```
SELECT ?signature ?codingLabel WHERE {
  wd:Q1238229 p:P4152 ?signatureStmt.
  ?signatureStmt ps:P4152 ?signature;
    pq:P3294 ?coding.
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language
        "[AUTO_LANGUAGE], en".
  }
}
```

Figure 5: A screenshot of a SPARQL query to request the file format signature of the .stl file format.

Try this query! Code for this query.

IV. WikiDP Reports

The Wikidata for Digital Preservation portal (WikiDP) is a specialized interface that provides a view of the data in Wikidata tailored to the domain of digital preservation[1]. The Wikidata for Digital Preservation portal can be used to access and download multiple datasets derived from Wikidata. Users will find reports on a dedicated portal page[2]. Each time a person runs a query, the result set is computed live, thus results for many queries will change over time. The purpose of this page is to gather queries that return datasets of potential interest to the digital preservation community.

Reusing these queries over time allows us to gain a deeper understanding of how the data changes over time.

The reports featured on the portal website are a subset of the total reports we track[3]. As we write additional queries this inventory of useful datasets will also grow.

V. Fair data

Long-term preservation and governance of metadata for the domain of computing is an important issue for the digital preservation community [8], [9]. Centralization of technical metadata for the domain of computing benefits all creators and users of this metadata. Wikidata is a multilingual knowledge base, leveraging the mappings created through years of conceptual alignment among the different language versions of Wikipedia and Wikidata items [10]. This means that more users will have access to metadata related to the domain of computing in their language, an important step in reducing the dominance of the English language which disadvantages other linguistic communities.

The data contributed to Wikidata is compliant with the FAIR data principles [11]. By creating data that aligns with the FAIR data principles, we ensure that this metadata is easy to find and easy to reuse. This technicalital preservation professionals must be able to identify and refer to, will be more complete if we distribute our effort. Redundant, fragmented descriptions in siloed repositories are frustratingly incomplete. Many governmental bodies and international consortia have endorsed the FAIR data principles as a key aspect of their open science or open data initiatives [12]. The data contributed to Wikidata is linked open data[4]. Experts from libraries, archives, museums and technologists of the World Wide Web Consortium (W3C) recommend linked data for library metadata published on the web[13].

FAIR is an acronym for findable, accessible,

[1] A description of the WikiDP system.

[2] www.wikidp.org/reports

[3] For a more complete list see https://github.com/emulatingkat/SPARQL

[4] https://www.w3.org/DesignIssues/LinkedData.html

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

interoperable and reusable. Metadata for the domain of computing that we contribute to Wikidata are **findable** in that Wikidata items are indexed by all large search engines. The Qids assigned to Wikidata items are their unique, persistent identifiers.

These metadata are **accessible** because the entity data associated with their unique ids (all statements and references asserted about an item) are dereferencable via the HTTP protocol. They are **interoperable** in that they link to many other databases and systems through the collection of external ids as seen in Figure 7.

These metadata are **reusable** due to the use of the CCO license for the content of Wikidata. Anyone can reuse Wikidata data for any purpose. Publishing data in the Wikidata knowledge base fulfills the most complete degree of FAIRness, level F, "FAIR data, Open Access, Functionally Linked", as described in [12].

The Wikidata for Digital Preservation Portal provides direct links to items in Wikidata. If a user would like to consult Wikidata to view additional information related to vocabularies that have been stored, they may consult the item of interest by following the links provided in the Portal.

VI. Discussion: A centralized repository of FAIR, linked open data

Wikidata is growing. We have been participating in Wikidata by structuring data in the domain of computing since August, 2016. In the years of our participation we have seen growth in Wikidata as a whole, and improved data coverage for computing topics.

Wikidata is a project of Wikimedia Deutschland[1] and has been supported by the chapter budget, grant awards and donations. In 2016 the Wikimedia Foundation announced that it would begin funding the software engineering activity for Wikidata[2]. This is a strong signal that the infrastructure of Wikidata will continue to be supported in the future.

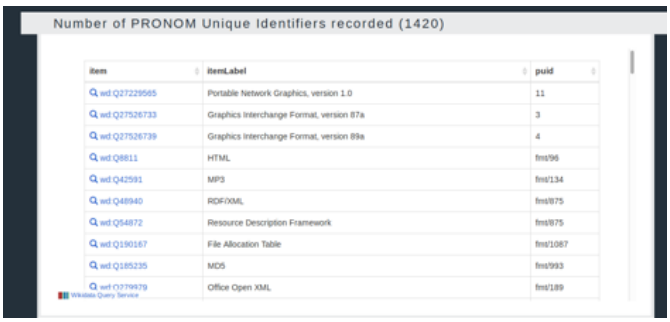


Figure 6: This is a screenshot of the Wikidp Reports page. When a user selects a report, a query is performed on WQS and the most recent result set is returned.

Figure 7: A screenshot of the Wikidata Item for the Sterolithography file format showing the collection of links to external resources that also describe the format on the right-hand side of the interface.

A. *Infrastructure and Maintenance*

The Wikidata for Digital Preservation portal can support the work of many users, and only one local developer [14] with knowledge of Wikidata is required. The local developer inspects the data models and the infrastructure of Wikidata in order to make recommendations about the user interface and the interaction design of the portal, effectively articulating work for many users.

We conclude that reusing the infrastructure of the Wikidata knowledge base, which has been assigned to the public domain, is a compelling model for cultural heritage institutions looking for a centralized repository for metadata. Commons-based peer production of infrastructure allows for distributed stakeholders to collaboratively maintain the infrastructure [4]. The wiki software is developed by engineers who have agreed to make their work available to all by releasing it under free software licenses

[1] https://wikimedia.de/wiki/Hauptseite

[2] https://blog.wikimedia.org/2016/10/04/supporting-the-future-of-wikidata/

The Wikidata community maintains a public SPARQL endpoint for the knowledge base. This SPARQL endpoint allows users to write flexible, powerful queries to retrieve subsets of the data in the knowledge base. In contrast, maintaining a public SPARQL endpoint is not often feasible for a software or format registry developed within an institution or project context. The Wikidata community has enhanced the SPARQL endpoint by providing multiple visualization options for the data returned in queries, from bubble charts to graphs of many varieties. The developers who work on the SPARQL endpoint have also created an interface that supports users who do not yet know SPARQL in writing or modifying SPARQL queries[1]

B. Collaboration

The structure of Wikidata allows the crowd to collaborate. A boundary object is a tool for thinking that allows people from different communities of practice to use a shared form to bridge the differences in their experiences and effectively collaborate [15]. When multiple boundary objects are used in conjunction they can become parts of systems of boundary objects [16]. Star and Bowker introduced the concept of "boundary infrastructure" to theorize about systems of boundary objects. Boundary infrastructure allows for collaboration without consensus [16]. Wikidata, the knowledge base of structured data that anyone can edit, is an example of boundary infrastructure that allows people from many communities of practice, from many walks of life, specialists and non-specialists across many domains, to effectively collaborate to structure data, and make it available for reuse.

C. Sustainability

Digital Preservation is a expensive activity for many institutions. Institutions with limited budgets for digital preservation can reuse this data at no cost. The boundary infrastructure of Wikidata provides a means for digital preservation professionals from different parts of the world, working in different languages, to collaborate by creating structured data in the knowledge base. This reduces the risk of redundant effort to describe the same file format in numerous local format registries. The boundary infrastructure of the knowledge base also supports contributions from the crowd, people who have interest in, and information about, the domain of computing. This allows for collaborations that otherwise might not happen without the boundary infrastructure that facilitates communication in a community of practice.

Members of the general public will also have access to this information. Having this information in an accessible, structured repository will allow more people to consult it, which could lead to people making different computing choices in their lives, for example choosing an open format, which could impact the work of future generations of digital preservation professionals. Wikidata's CC0 license ensures that this data will have an equalizing force, as it will not be controlled by any single institution, or even any consortium of institutions. Anyone with access to the internet will be able to inspect and reuse this data for their own projects or systems. Institutions that do not yet have budgets for digital preservation will have access to this metadata and will not have to recreate it in their local systems.

VII. Conclusion

The Wikidata for Digital Preservation Portal facilitates increased communication between members of the Wikidata community and the international digital preservation community.

Centralizing the metadata for the domain of computing, eliminates redundant labor of individual institutions creating structured data within their local systems. When we collaboratively create metadata and publish it in Wikidata, anyone can reuse it. This allows metadata professionals to focus on the administrative, preservation, and use metadata pertinent to their local settings.

Making use of infrastructure supported by the Wikimedia foundation, built and maintained by an active community of tens of thousands of contributors is a new option for cultural heritage institutions. The fact that this infrastructure is built in conformance to open standards and is comprised of free software means that we can audit this system and to

[1] https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

see if we can continue to trust it to store and access our data.

Acknowledgment

This portal was inspired by the Su Lab and their Wikigenomes portal. We thank Wikimedia Deutschland[1] for supporting the Wikidata project. We thank the Wikidata community for sharing thoughts, feedback, data curation and data modeling work. Specifically, would like to thank all participants of WikiProject Informatics[2] for engaging with us as we work to describe the domain of digital preservation in Wikidata. To review the source code for this project please visit our Github repository[3]. We appreciate the informative conversations we had with participants of Wikicite 2016, Wikicite 2017. This work is supported by the Council on Library and Information Resources, Andrew W. Mellon Foundation, Alfred P. Sloan Foundation, and the Open Preservation Foundation.

References

[1] Wikidata, Statistics, 2019. [Online]. Available: https://www.wikidata.org/wiki/Special : Statistics.

[2] D. Vrandečić, "Wikidata: A new platform for collaborative data collection," in Proceedings of the 21st International Conference Companion on World Wide Web, ACM, 2012, pp. 1063–1064.

[3] Y. Benkler, "Coase's penguin, or, linux and the nature of the firm," Yale Law Journal, pp. 369–446, 2002.

[4] Y. Benkler, A. Shaw, and B. M. Hill, "Peer production: A modality of collective intelligence," Collective Intelligence, 2013.

[5] C. Müller-Birn, B. Karran, J. Lehmann, and M. Luczak-Rösch, "Peer-production system or collaborative ontology engineering effort: What is wikidata?" In Proceedings of the 11th International Symposium on Open Collaboration, ACM, 2015, p. 20.

[6] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, and D. Vrandečić, "Introducing wikidata to the linked data web," in The Semantic Web–ISWC 2014, Springer, 2014, pp. 50–65.

[7] L. Pintscher, "State of the project," WikidataCon 2017, https://upload.wikimedia.org/wikipedia/commons/b/b0/WikidataCon_2017-_State_of_the_Project.pdf, year=2017.

[8] P. McKinney, S. Knight, J. Gattuso, D. Pearson, L. Coufal, D. Anderson, J. Delve, K. De Vorsey, R. Spencer, and J. Hutař, "Reimagining the format model: Introducing the work of the nsla digital preservation technical registry," New Review of Information Networking, vol. 19, no. 2, pp. 96–123, 2014.

[9] P. McKinney, D. Pearson, D. Anderson, J. Hutař, S. Knight, L. Coufal, J. Delve, J. Gattuso, K. DeVorsey, and R. Spencer, "A next generation technical registry: Moving practice forward," iPRES 2014: 11th International Conference on Digital Preservation, 2014.

[10] S. Burgstaller-Muehlbacher, A. Waagmeester, E. Mitraka, J. Turner, T. Putman, J. Leong, C. Naik, P. Pavlidis, L. Schriml, B. M. Good, et al., "Wikidata as a semantic framework for the gene wiki initiative," Database, vol. 2016, baw015, 2016. https://www.wikidata.org/wiki/Wikidata:WikiProject_Informatics12 https://github.com/WikiDP

[11] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg,G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., "The fair guiding principles for scientific data management and stewardship," Scientiic data, vol. 3, p. 160 018, 2016.

[12] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, and M. D. Wilkinson, "Cloudy, increasingly fair; revisiting the fair data guiding principles for the european open science cloud," Information Services & Use, no. Preprint, pp. 1–8, 2017.

[13] T. Baker, E. Bermès, K. Coyle, G. Dunsire, A. Isaac, P. Murray, and M Zeng, "Library linked data incubator group final report," report, W3C Incubator Group, October, vol. 25, 2011.

[14] M. Gantt and B. A. Nardi, "Gardeners and gurus: Patterns of cooperation among cad users," in Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, 1992, pp. 107–117.

[15] S. L. Star and J. R. Griesemer, "Institutional ecology,translations' and boundary objects: Amateurs and professionals in berkeley's museum of vertebrate zoology, 1907-39," Social studies of science, vol. 19, no. 3, pp. 387–420, 1989.

[16] S. L. Star, "This is not a boundary object: Reflections on the origin of a concept," Science, Technology and Human Values, vol. 35, no. 5, 601–617, 2010.

[1] https://www.wikimedia.de/wiki/Hauptseite

[2] https://www.wikidata.org/wiki/Wikidata:WikiProject_Informatics

[3] https://github.com/WikiDP

Reformat Required:

Addressing "Paper Minds" in Archives through a New Collaborative Model for Digital Knowledge Transfer

Angela Beking
Library and Archives Canada
Canada
angela.beking@canada.ca

For many memory institutions, policies, procedures, and practices are built on the realities of analog records. The belief that digital content can be appraised, acquired, described, and made accessible using the same methods as paper records can inhibit the development of end-to-end digital preservation programs. To start to address this challenge, I developed a collaborative model for digital knowledge transfer based on adult education theory. The model has shown great promise for building digital capacity, capability, and community amongst my colleagues at Library and Archives Canada. This paper outlines the concepts that drive the model, as well as the three steps that are required for its implementation. As analog preconceptions often influence the thinking of those who are responsible for relationships with records creators, as well as the acquisition and processing of digital content, such collaboration and capacity building is a necessity for the success of end-to-end digital preservation programs.

Collaboration, capacity building, learning, knowledge transfer, digital archives

Collaboration: a Necessity, an Opportunity or a Luxury?; Building Capacity, Capability and Community

I. Introduction

Many of the policies, procedures, and practices concerning the acquisition of documentary heritage in memory institutions are built on analog foundations. In 2007, Canadian archivist Terry Cook observed that despite the "fundamental changes" necessitated by digital records, "despite the consequent need to reorient or reinvent or reconceive our work, almost all the concepts, practices, procedures, and even accepted terminology of the [archival] profession reflect our legacy of paper records. We have paper minds trying to cope with electronic realities" [1]. The intellectual divide Cook observed twelve years ago still exists, as "paper-minded" approaches continue to prevent meaningful engagement with digital archival and digital preservation approaches and programs [2]. Digital content is often acquired without a full examination of the feasibility of such acquisition in terms of long-term preservation and access. How do we, as digital preservation professionals, develop collaborative relationships that will overcome "paper-minded" approaches and thinking to develop our digital archival and digital preservation programs?

In 2018, I developed a collaborative model for digital knowledge transfer based on andragogy, "the art and science of helping adults learn" [3]. The model has four unique elements that make it a good basis for successful collaboration. First, it is focused on the self-concept of the learner; second, it is conducted in small group learning environments; third, it is problem, not subject, oriented; and fourth, it responds dynamically to the learner's shifting needs. The model's first participants have developed the ability to engage with digital archival and digital preservation approaches and issues, representing the beginnings of a true shift in "paper mind" thinking. Collaboration, therefore, has had the greatest impact when we share knowledge based on best practices for adult learning. As digital preservation needs and concerns become tied to the professional self-concepts of our non-digital preservation colleagues, we will start to see true growth in digital capacity. In this manner, we pave the way for a reorienting or "reformatting" of the "paper mind."

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

A. *Institutional Context*

As the center of expertise for digital archival records at Library and Archives Canada (LAC), the Digital Integration unit has tried a variety of initiatives to build digital capacity. Such initiatives have included large-scale training sessions and presentations on LAC's digital procedures. While such initiatives have provided short-term motivation for acquiring staff to engage with digital archival and digital preservation issues, they have not sparked a deeper self-reflection on what it means to apply a "paper mind" to digital issues. Such initiatives are not best suited to engage the learning needs of adults, which results in a lack of effectiveness.

The collaborative model for digital knowledge transfer was developed in the context of Digital Integration's efforts to address our institutional backlog of digital content. Responsibility for this content is assigned to archivists according to subject-matter portfolio, and digital archivists are assigned to provide strategic direction, guidance, and support. In the digital archivist role, I saw an opportunity to achieve three goals: first, to help process the backlog content and make it accessible to our clients; second, to build the skill sets of portfolio archivists, so that they could process subsequent content independently of us; and third, to start to raise institutional awareness and build digital capacity.

II. Developing a Collaborative Model for Digital Knowledge Transfer

A. *Step 1: Understand and Incorporate the Self-Concept of Your Collaborator*

Andragogy posits that adult learners see themselves as self-directing, deriving self-fulfillment from their performance in certain roles, such as worker, spouse, or parent. Adults no longer see themselves as full-time learners, but rather as "producers" or "doers." This self-concept must be understood and engaged by those who wish to transfer knowledge to adults [4]. In many ways, the adult defines the self by experience, and those experiences should feed into any educational activity they undertake [5].

1. *Diagnosis of Needs*

The learner's self-concept has important implications for the development of successful digital knowledge transfer. In the model, the development of the learning program itself must be a collaborative effort.

In curriculum development, the "diagnosis of needs" occurs when learning experiences are selected and organized on the basis of learner requirements, which can include interests, abilities, background, motivational pattern, social needs, or values [6]. Andragogy places emphasis on self-diagnosis, under the assumption that an adult is more deeply motivated to learn what he or she identifies as something he or she needs to learn, in order to enhance an aspect of the self-concept. The teacher serves as a facilitator, guide, or resource; the learner diagnoses their own needs, and collaborates with the teacher to translate those needs into specific educational objectives and learning experiences [7].

How can this theory be applied to the transmission of digital archival and digital preservation knowledge? How can our non-digital preservation collaborators be in a position to diagnose their own learning needs, when "paper minds" underpin much of their thinking and approaches?

The collaborative model for digital knowledge transfer addresses this by starting small in scope, with a tangible goal (or goals) linked to one or more basic professional objectives of the collaborator. Often, the immediate identified learning need is to develop the skills necessary to process backlog digital content. This goal has a concrete link to the self-concept of many of our collaborators, who are archivists or other curatorial professionals who find fulfillment in meeting professional expectations.

B. *Step 2: Plan Your Initial Program*

1. *Problem-Centered Focus*

A crucial difference between childhood education and adult education is that the former is often *subject-centered*, while the latter should be *problem-centered.* Andragogy suggests that children gather information that does not necessarily apply to their everyday life challenges; adults, however, are motivated to address the problems they currently face. Thus, andragogy places emphasis on immediacy. The goal of a learning experience should be to help adults develop approaches and solutions to

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

current problems, rather than learning a particular subject matter [8]. Expertise is not taught directly, but will emerge gradually over time.

Processing backlog content is a good problem-centered objective, well suited to the needs of many archivists or other curators. This objective can and should, however, be adapted to meet the needs of collaborators with different professional goals. The essential task in planning a collaboration for digital knowledge transfer is to ensure that all learning experiences are tied to real, measurable goals, such as clearing assigned backlog, acquiring a new digital collection, or testing a new technique or software tool, that is relevant to the collaborator's self-concept.

This is an area where the small-scale collaborative model differs significantly from large-scale presentations of content. In smaller, hands-on groups, it is possible to develop individualized activities that meet the needs of specific collaborators. The facilitator may also receive knowledge from their collaborator, including a better understanding of the way in which the "paper mind" is influencing their approaches to digital work. Through such sharing, messages can be customized to explain differences and similarities between analog and digital records in ways that will be meaningful to the learner. Such sharing is not possible in a large-group setting. In such a setting, an archivist interested in processing backlog and an archivist interested in furthering their specific subject-matter expertise would receive the same information, such as a general tutorial. Neither would feel the same sense of professional ownership over their learning and development, which is why I believe much of the motivation falls away shortly after such large-scale sessions.

Thus, the goal of the collaborative model is to facilitate the efforts of "paper-minded" colleagues who are struggling with very specific digital challenges. How do I ensure that the content on this hard drive I acquired is preserved? A creator wants to transfer a database, what should I do? How do I set access restrictions on individual email messages? These are the types of questions the collaborative model for digital knowledge transfer is built to address. In so doing, over time, expertise emerges through experience.

2. Learning Environment

After needs are identified, planning specific learning experiences should also be, as much as possible, a collaborative effort. In my role as facilitator, guide, and resource, I often suggest that the program begin with facilitated processing. This often takes the form of addressing digital object content categories, starting with textual files and proceeding to more complex content such as graphic or audiovisual files in non-standard formats.

One such collaboration aimed to process the digital-born records of the National Roundtable on the Environment and the Economy (NRTEE), which was rendered defunct in 2013. As LAC is mandated to assume the "care and control of all records of a government institution whose functions have ceased", NRTEE's digital assets were transferred to LAC on an external hard drive [9]. The content included an export from the NRTEE's Records, Document, and Information Management System (RDIMS), an electronic document and records management system used by the Canadian government, the contents of a shared drive, and the NRTEE's email accounts.

By 2018, this content was considered "backlog." The immediate learning need, self-identified by the portfolio archivist, was problem-centered: to develop the skills necessary to select, arrange, and describe the material so that it could be preserved and made accessible. My task was to develop a learning program to meet this goal.

The NRTEE data posed several challenges, including file formats that did not conform to LAC's Guidelines on File Formats for Transferring Information Resources of Enduring Value. The data also represented a fundamental disruption to the method by which government records archivists appraise and select archival records at LAC. "Macroappraisal" places emphasis on the context of records creation over the content of records. The records of a unit of government whose business functions are deemed to create records of archival value will be declared to have archival value, even if the archivist never sees those records. The strategy relies on file classification plans that detail the intellectual and physical arrangement of (usually analog) records.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The NRTEE data was exported from RDIMS by user name. The archivist could not identify relevant program functions with such a data structure, as it was impossible to link individual employees to business functions. The files had also exported with system identifiers as file titles, which did not make sense to the archivist (for example, a Word document was titled "6mq01!"). The user-assigned file titles were stored in a poorly structured Microsoft Access database that had accompanied the transfer. As NRTEE was defunct, it was not possible to work with the department to export the data in a structure more amenable to archival processing. How could I explain that selection work based on macroappraisal strategy would not be possible for these records? How could I help the archivist develop a workable path forward?

Andragogy places emphasis on experiential teaching techniques and practical application of new concepts. The theory suggests that the transmittal techniques prevalent in youth education, such as readings, lectures, and audiovisual presentations, are not well suited to adult learners, who seek self-direction and thrive when they have a sense of ownership over their learning [10]. As such, small group, hands-on workshops, with ratios of 1:1, 1:2, or a maximum of 1:3, seem to be most beneficial for digital knowledge transfer. The learner "drives" the work (in most cases, this equates to conducting all mouse navigation and clicks), while the facilitator provides strategic guidance and direction. Weekly meetings are beneficial, with collaborators agreeing on deliverables for the next week at the end of each session, such as files to be processed or research to be completed.

For the NRTEE content, the portfolio archivist and I met for one hour, twice a week, for six months, as part of our regular operational work. In 1:1 sessions, I explained how to mobilize software tools like TreeSize Professional and Quick View Plus to perform archival selection, arrangement, and description on the shared drive content. The portfolio archivist conducted all navigation and clicks, building confidence in completing this kind of work. Our second task was to work through the export from RDIMS. I provided explanations of the challenges presented by this content, the most significant of which being that it required item-level review, rather than macro-level review, due to its structure. Between sessions, the portfolio archivist reviewed user-generated file titles from an Excel file generated by the Digital Preservation team. By manually comparing this list to the NRTEE files, the archivist gradually identified records of archival interest. As time passed, realizations emerged from this work that would radically change the nature of our knowledge transfer sessions.

C. Step 3: Respond Dynamically to Your Collaborator's Shifting Needs

The success of the collaborative model is also predicated on the facilitator's ability to respond to their collaborator's shifting needs. A collaboration may begin with an identified learning need of "develop the skills necessary to process digital backlog." As work progresses, however, a collaborator may become interested in other areas of digital archival work, such as approaches to acquisition. The learning program should expand in response. In this example, the focus should grow to include deeper engagement with the theoretical underpinnings of digital archival practice as they apply to acquisition. Sessions might now include discussion of how to prevent the acquisition of problematic data at the point of transfer, or through early intervention in the creator's recordkeeping process.

The learning program for the processing of the NRTEE records required this sort of dynamic response as learning needs shifted. The portfolio archivist became frustrated by the manual nature of the archival processing of the RDIMS content. Though incredibly engaged in the work, the individual review of over 20,000 files was an undeniably time-consuming and tedious task. It was therefore essential that our collaboration seek out new solutions in order to avoid discouragement, or the development of the belief that this sort of work would be required for all digital archival records. Thus, I changed one of our weekly meetings to a brainstorming session in which we began thinking about alternate approaches.

How could we proactively engage with departments upstream to prevent LAC receiving such poorly structured data? Could such discussions make LAC's preferred macroappraisal approach feasible for digital records? These questions were

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

serendipitously timed. Digital Integration was actively researching the Producer-Archive Interface Methodology Abstract Standard (PAIMAS). At the same time, the portfolio archivist was engaged in a disposition process with four other Government of Canada institutions. We decided to workshop the standard as part of this process, in an attempt to understand the digital recordkeeping contexts of each of these entities. We sought to understand whether the proposed transfer of their digital archival records to LAC would be both trustworthy and feasible. The relationships we developed through a PAIMAS-driven approach led to the transfer of digital content that was not only well structured, but also transferred in preferred preservation formats. Macroappraisal of these digital records was easily achieved.

Perhaps the most outstanding result of this collaboration was the portfolio archivist's newfound ability to engage directly with some of the basic assumptions of the "paper mind." Macroappraisal strategy is the theoretical underpinning of government archives work at LAC. It was developed, however, in the context of analog records, and the ways in which it must be adapted to cope with digital realities is not yet well understood. Through our collaboration, the portfolio archivist realized that having an intellectual understanding of government program functions is no longer sufficient; archivists must also understand the context of digital recordkeeping systems for macroappraisal to continue to be viable. This demands significant change in the skill sets of archivists, and as such, represents a very deep disruption to professional identity that cannot be effectively explained through a large-scale lecture or presentation. A true realization of the implications of digital disruption can only emerge when a learner self-identifies digital challenges as a learning need, and has that need addressed in a knowledge transfer program that concretely identifies areas where new approaches are necessary. The resulting sense of ownership over the solution leads to true acceptance of the need for change. Digital expertise will then emerge slowly, over time. The collaborative model for digital knowledge transfer, therefore, is an important component in changing our understanding of the very nature of archival work.

The ability to respond deftly to changes in a learner's self-identified needs is a difficult but very important component in the development of such a successful digital knowledge transfer collaboration. Malcolm Knowles states that "the truly artistic teacher of adults perceives the locus of responsibility for learning to be a learner; he conscientiously suppresses his own compulsion to teach what he knows his students ought to learn in favor of helping his students learn for themselves what they want to learn" [11]. It is crucial to listen carefully to our collaborators. Are they finding their current work discouraging? Are there previously unidentified digital issues with which they are struggling, that they may not be able to articulate? Can we help identify these issues? If so, can we reprioritize our learning programs to address the new challenges? Being able to identify such issues, and being able to respond to them quickly and effectively, is an important skill that we must develop within ourselves to enable successful collaboration.

The collaborative model also requires that the digital preservation professional relinquish some control over the deliverables of the learning program. If our collaborators would like to discuss acquisition, for example, can we put aside the processing of backlog content to explore this new path? There is of course a balance to be struck between meeting identified deliverables and institutional objectives and encouraging the self-identified digital capacity growth needs of our collaborators. Finding this balance will depend significantly on institutional context, but flexibility and responsiveness, in my experience, leads to better outcomes.

III. Current Challenges

A significant challenge to the success of the collaborative model is how our potential collaborators perceive digital archival and digital preservation work. Digital capacity cannot be developed under the model until our "paper-minded" colleagues self-identify the acquisition of digital knowledge as a learning priority. This accounts for the slow adoption of engagement with such issues that we are seeing in our work as digital preservation professionals. If our colleagues believe that they do not require any new skills to appraise, acquire, describe, and make accessible digital content, it will be impossible to help them self-identify specific learning needs and

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

develop learning programs. How can we address such entrenched ideas?

Shifting the culture of an institution at a grassroots level may be a potential solution. As early adopters at LAC move through the collaborative model, they are beginning to discuss its benefits with their peers; while these conversations are in their infancy, there is potential in a "teach the teacher" concept. If the collaborative model can engage the attention of a few previously "paper-minded" colleagues, and help them meaningfully engage with how digital archival work is different from analog archival work, they may pass this understanding to their colleagues organically. Those receiving this information may, in turn, begin to identify digital skills amongst their own learning needs. I am hopeful that LAC's early adopters will continue to disseminate this thinking, so that the collaborative model's digital capacity building can engage more and more of our staff.

IV. Conclusion

In early 2019, a co-collaborator that had been working with me on digital knowledge transfer since mid-2018 suggested that "all archivists should become digital archivists." This attitude is incredibly promising; it represents an acknowledgement that all archivists must develop the skills we might now attribute only to "digital" archivists. As records become almost exclusively born digital, digital preservation needs and concerns will be even more integral to managing archives. I see collaboration and knowledge transfer between digital preservation professionals and their more "paper-minded" colleagues as the path forward. We cannot do this alone; we need our colleagues to help us implement end-to-end digital preservation programs. Through a collaborative model built on adult education theory and techniques, I have witnessed outstanding results in digital capacity building amongst those who self-identify digital as a learning priority. It remains to be seen whether such gains can be brought to scale within LAC, and whether the model might be useful for others in the profession.

Aknowledgment

I would like to thank all of the individuals who have participated in the collaborative model for digital knowledge transfer with me through our work at Library and Archives Canada. A special thanks to Roddy McFall for his unending enthusiasm as the collaborative model's first participant. I would also like to thank Tom Smyth for his encouragement and support of this work.

References

[1] T. Cook, "Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-Custodial and Post-Modernist Era," *Archives & Social Studies: A Journal of Interdisciplinary Research*, vol. 1, no. 0, p. 403, March 2007.

[2] A-S. Klareld and K. L. Gidlund, "Rethinking Archives as Digital: The Consequences of 'Paper Minds' in Illustrations and Definitions of e-Archives," *Archivaria*, vol. 83, pp. 81-108, Spring 2017.

[3] M. Knowles, "Andragogy: An Emerging Technology for Adult Learning," p. 55, http://www.nationalcollege.org.uk/cm-andragogy.pdf

[4] Knowles, pp. 55-56.

[5] Knowles, p. 62.

[6] Knowles, p. 58, and J.A. Alegarbes, F.G. Ruaya, and J.E. Cordero, "Diagnosis of Needs in Curriculum Development," https://slideplayer.com/slide/9974599/

[7] Knowles, p. 58.

[8] Knowles, p. 65.

[9] Library and Archives of Canada Act (S.C. 2004, c. 11), 13 (4), https://laws-lois.justice.gc.ca/eng/acts/L-7.7/

[10] Knowles, p. 61-62.

[11] Knowles, 68.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

One repository solution doesn't fit all

Towards a distributed and collaborative approach to digital preservation services at the Amsterdam City Archives

Ana van Meegen Silva
Stadsarchief Amsterdam
The Netherlands
A.van.Meegen@amsterdam.nl

Abstract – Different digital objects have different preservation requirements involving strategies, policies and practices that go beyond the capacity of a single archival institution. The variety and complexity of digital objects requires specific knowledge for ingesting and the preservation of file formats. Large data files call for a different storage set-up than the approach required for the preservation and distribution of small objects. Building an organizational and technological infrastructure that can cope with the diversity and complexity of digital objects involves an effort that far exceeds what a middle-sized institution can do alone. This paper describes how the Amsterdam City Archives aims to collaborate with other non-profit institutions and partners to improve the quality of the preservation of and access to digital objects while reducing costs.

Keywords – digital preservation services, collaboration, distributed approach, archives repository.

Conference Topics – Collaboration; Exploring New Horizons.

I. Introduction

The Amsterdam City Archives (*Stadsarchief Amsterdam*, SAA) is the historical documentation center of the city of Amsterdam and forms an integrated part of the Amsterdam City Council. The institution is responsible for the City Council's archives and holds records of private institutions and citizens of Amsterdam. The collection is based on two perspectives: information about the city and records formed by the citizens of Amsterdam. The institution currently has approximately 120 terabytes of digital objects, including 28.5 million digitized documents and images, 8 million digital-born objects, and 15,000 moving images and sound files.

In 2020, the SAA will be implementing a whole new digital infrastructure that will renew its archives repository, collection management system and discovery platform. These changes are needed to comply with European regulation and modern technology. SAA uses the Open Archival Information System (OAIS) model as conceptual framework for the information architecture of it's infrastructure. In the period leading up to the roll out of the infrastructure, a comprehensive evaluation of different solutions for the management of digital objects was carried out. The main challenges for the archives repository were the accessibility performance of digital objects versus a low cost storage solution that can provide the preservation functionalities required. During this preparation period, SAA worked with different use cases that reflected particular situations to which the new infrastructure should provide a solution. After investigations, two of the use cases were put aside and new solutions through collaboration are now being explored. In this paper I will explain why these two use cases cannot be integrated in the new infrastructure and how collaboration with other organizations can be a solution for digital assets that do not fit an archival infrastructure based on small-sized digital objects.

II. Exploring collaborations

SAA's digital infrastructure is a standard structure that supports elementary digital object management, from preservation to use. The diversity of digital objects, however, does not always fit in this infrastructure. SAA is working with two cases that go beyond their current capabilities: a cost reduction storage solution for large-sized objects and an integrated solution for preservation and

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

discovery of Council meetings. For both cases, SAA is exploring collaborative solutions, each with their own distinctive features. Different institutions use a shared digital infrastructure, assigning different components of the OAIS model for each institution [1]. In the collaborative cases of SAA however, the collaboration is not a shared digital infrastructure but a service purchase where the providers facilitate parts of the digital infrastructure.

A. Preserving large-sized media files

The Migrant Television Netherlands [2] (MTNL) was a foundation that created television programs with a focus on topics that concerned migrants and migration in the Netherlands. MTNL produced programs from 1984 to 2013, and broadcasted in the four major Dutch cities through local broadcasters. The collection contains a total of 4,000 hours of material. The organization stopped their activities and donated their archives to the City Archives of Amsterdam, Rotterdam, The Hague and Utrecht. During this transition to the archives it was decided to digitize all of the collection's VHS tapes [3]. The digitization was carried out by the Netherlands Institute for Sound and Vision (*Nederland Instituut voor Beeld en Geluid*, NIBG) and for each tape they delivered two files: an HD master file and a low resolution file meant for online purposes. The low resolution files are stored and preserved at the Amsterdam City Archives. It was decided to store the master files temporarily at the NIBG until SAA implements a new repository. The master files can be accessed through the NIBG website [4]. SAA is still working on a solution for the presentation of the low resolution files on its website.

During the preparation period for the renewed digital infrastructure project at the Amsterdam City Archives, the institution investigated possibilities to archive the MTNL master files at the organization. The purpose of the analysis was an integrated approach for all SAA digital objects, knowing that more archives such as those of the MTNL are likely to follow and a solution will need to be in place to accommodate those collections.

For the preservation and metadata management of the files the organization will not require extra functionality; the currently available and renewed systems are also suitable for large files. The organization would, however, need a differentiate workflow for accessibility; these online requests could be easily arranged. The real problem was the storage solution. The analysis showed that a tape storage solution is, at present, the best method to house large-sized object files with lower storage costs. This solution, however, is not suitable for small-sized objects requiring a quick response time and daily use. For an integrated in-house storage approach the organization would need to host two different storage solutions and would therefore not be able to achieve the intended cost reduction.

The best solution remains the storage of the master files at the NIBG. The NIBG is responsible for the archiving of the collections of Dutch broadcasting corporations and is one of the largest media archives in Europe. With more than 17 petabyte of stored collections and more than 3 million online media assets, the NIBG is one of the main experts on media archives in the Netherlands and they have a repository solution that is suitable for the archiving and preservation of large-sized files. The institution has been using tape storage solutions in combination with OAIS-compliant functionalities for preservation and has a Date Seal of Approval certificate [5]. A management storage hierarchy functionality was put in place for the different storage workflows maximizing the data monitoring and the performance capacity [6]. These qualities and storage solutions make the NIBG the best collaborative partner for the preservation of the MTNL master files. There are, however, issues that need to be addressed. Besides the arrangements made for the storage and availability of the files through NIBG's website, there are no arrangements for usability of content, metadata enrichment, statistics on usage frequency and incidents/problems, etc. The roles and responsibilities beyond the preservation of the files have not yet been defined; neither organization knows what they can expect of the other. A better collaboration agreement that defines the mutual objectives with regard to the collection, the future improvements in both organizations and the service level expectations should be discussed and implemented

B. Preservation of and access to council information

Decisions made by the board of Amsterdam City Council are publicly available. Every citizen can follow

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

live streams of the council's meetings and can access meeting agendas and preparatory documents. This rich content information on Amsterdam's decision making is made available through an online information platform [7] hosted by NotuBiz, the vendor that records the video minutes. NotuBiz offers an integrated platform solution that follows the whole political decision-making process. Citizens can view the agenda of upcoming meetings and the submitted documents beforehand. The archive module enables the user to search or browse through different categories such as date, subject, the name of a councilor or document, etc.

In 2017, the council's clerk contacted the City Archives. The eight-year contract period with NotuBiz was coming to an end, a new European tendering process was starting and it was unsure whether NotuBiz would be the new facilitator of the video minutes and online platform. SAA is responsible for the archiving and preservation of all the council's digital assets, including the video minutes. A problem analysis was conducted and the key conclusion was that SAA is very capable of archiving and preserving all the digital objects available on the platform, but that it does not have a platform that can integrate the different digital objects in a user-friendly way. A solution could be to provide a platform that has the same functionalities as the NotuBiz website, but that would mean that Amsterdam City Council would pay twice for the same functionalities. Another solution was to archive the video minutes at SAA and make them available through the new councilor's platform. This solution had the most chance of being implemented until the news was announced that NotuBiz had won the tender and had signed a new eight-year contract with the City of Amsterdam.

During the search for an archival solution for the video minutes, NotuBiz proposed that they facilitate the preservation management of the digital objects in collaboration with SAA. SAA would give advice on and monitor the preservation management implementation at NotuBiz. If the implementation proves successful, the NotuBiz services will be expanded to include the preservation module. This solution is not yet definitive, as both organizations need to think it through and investigate the impact that this model can have on the organization. The organization focusses on creating and sharing content and prior to the discussions with the Archives it had no idea of the problems and concerns of preservation management. The content is stored correctly and the organization takes care of the conversion of objects for accessibility purposes but there are no preservation policies and regulations in place. An external consultant conducted an OAIS analysis on the products of NotuBiz and made recommendations for improvements. The organization is now seriously considering the next steps and how these improvements will fit into its business model. For the Archives it means that they will start a collaboration with an organization that is not OAIS-compliant and that both organizations will work together to achieve the intended preservation goals.

III. Opportunities and challenges

Digital archives are diverse and therefore complex material resulting in a variety of possibilities and choices for storing, describing, preserving, distributing, discovering and accessing the information. Different choices entail many different opportunities and flexible solutions. The Amsterdam City Archives will be switching from an integrated in-house solution to a more flexible distributed approach for the preservation of digital objects, where collaboration is the key to success. But we are not there yet; there are various issues that need to be further analyzed and discussed.

Firstly, it is important to determine what the purpose of the organizations is and what results they wish to achieve for the collection and whether this fits into a collaborative partnership. The choice for collaboration was not a strategic one but more a pragmatic one, that fit the solution required for the arisen problem. As part of the project for renewing the digital infrastructure, SAA investigated whether there were other solutions for these use cases at the technical level and concluded that it was best to move towards a collaborative approach. A storage solution is, however, not the whole picture and a lot of work will still have to be done. In addition, the other organizations will also need to decide which future steps they want to take.

Preservation of digital objects is the core business of SAA and needs to be conducted and documented well. NIBG is a non-profit OAIS-compliant

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

organization that shares the same preservation principles as the Archives. This is not the case for NotuBiz, but that should not be an insurmountable problem. In a collaborative setting, the OAIS compliancy is a joint venture between both partners. For both organizations this means that they need only be partly compliant, but together they will form a complete match and will accommodate all the requirements needed for the preservation governance. There will also be overlap in the governance structure, which is needed to guarantee that both organizations see digital preservation as an important activity that ensures the durability and accessibility of digital objects.

Another issue that needs further analysis for preservation is continuity. Continuity of services for the long term is one of the core requisites for an organization working with sustainable archives. In a collaborative setting you can only guarantee continuity for as long as the collaboration lasts. This issue is no different than when an organization chooses for standard third-party services. It is important that expectations and legal requirements are explained and documented. In addition, an exit strategy should be agreed between the partners beforehand. In the governance structure it is important to mention that the archives institute will always be responsible for the continuity of the services if and when the partnership comes to an end.

Ensuring content usability and access is part of the core business of the Amsterdam City Archives. For both use cases the other organization is responsible for the discovery and access of the collection that they manage. In both cases, the Archives have until now had no influence on the presentation and access of information. For the Archives it is important to determine how much influence the organization wants to have in decisions about the description and usability of the collection. It is also important that users who start their search on the Archive's platform are also able to find the collections available on the other platforms. The exchange of metadata and/or content is a topic that should be further explored.

There are many differences in collaborations with a commercial partner or a non-profit institution. Both have their advantages and disadvantages, and for both it is important that the roles, responsibilities and expectations are well documented and that both partners know what to expect of each other. An evaluation should take place from time to time, where users' reports, technological developments and future perspectives are assessed. These topics will be on the agenda over the coming period so as to intensify collaboration and improve the collection's preservation and usability.

IV. Conclusion

The Amsterdam City Archives is convinced that a distributed collaborative approach for parts of the collection is a welcome solution for specific problems entailed by some collections. By making use of the knowledge and infrastructure of others the organization can improve preservation management, reduce costs and improve the collection's usability and user experience. SAA is working with two cases where collaboration seems to be the best outcome. However, there are still some challenges that need to be further explored.

In a collaborative setting there are several topics that should be discussed to avoid disappointments and miscommunication. Firstly, the organization should identify internally which goals and results they want to achieve within the collaboration and how they want to manage gaps and employ strategies to address them. Secondly, and most importantly in a collaborative setting, expectations should be managed and documented. In a partnership with a commercial organization it is not unusual to talk about service level management and agreements but for non-profit organizations this could lead to an uncomfortable situation.

Collaborations can help non-profit organizations professionalize their services for a larger community. For SAA, it is also important that they can make it clear to their partners which expertise or services, as well as financial compensation, they will contribute to the collaboration. In both cases, the collaboration with SAA is not a regular service and there is therefore no standard solution or service in place. SAA and its partners are still working on implementing and further improving the collaboration, with each case having its own challenges and opportunities.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

References

[1] G. Knight and Mark Hedges "Modeling OAIS compliance for disaggregated preservation services" The International Journal of Digital Curation, volume 2, no. 1, pp. 62-72, 2007. M. Day "Toward distributed infrastructures for digital preservation: the roles of collaboration and trust." The International Journal of Digital Curation, volume 3, no. 1, pp. 15-28, 2008.

[2] See the collection at the website of the Stadsarchief Amsterdam https://archief.amsterdam/inventarissen/inventaris/20002.nl.html

[3] The description of the MTNL Amsterdam Archive Collection is available at https://archief.amsterdam/inventarissen/inventaris/20002.nl.html.

[4] The collection is available at NIBG's website https://zoeken.beeldengeluid.nl/search/program?limit=10&offset=0&q=mtnl%20amsterdam&sort=-nisv.sortdate&x=false

[5] For more information see the "Digital Preservation Policy 2016" http://publications.beeldengeluid.nl/pub/388

[6] D. Steinmeier "Storage binnen OAIS: normatief model en GAP analysis voor Beeld en Geluid", 2013. Online available at: http://publications.beeldengeluid.nl/pub/408. Please note that a major renewing infrastructure project ran in the organization in 2018 and that this publication does not reflect the current situation, but the storage hierarchy functionality is still in place.

[7] See the platform at: https://amsterdam.raadsinformatie.nl.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Building Network Capacity Among Memory Institutions

A Multi-strand Development Approach

Melinda Haunton
The National Archives
United Kingdom
melinda.haunton@nationalarchives.gov.uk

Jo Pugh
The National Archives
United Kingdom
Jo.pugh@nationalarchives.gov.uk

James Travers
The National Archives
United Kingdom
James.travers@nationalarchives.gov.uk

Abstract – The National Archives of the United Kingdom has a leadership role for archives in England, representing a very broad range of memory institutions, most with limited digital capacity. This paper outlines the multi-strand approach underway to build capacity across this network of archives. The paper focuses on the aims of different intervention approaches and the involvement of the archives community and its stakeholders in different elements of delivery. The paper finishes with preliminary analysis of impact and known risks of this approach.

Keywords – networking, capacity-building, national archives, archives, memory institutions

Conference Topics – Building Capacity, Capability and Community; Collaboration: a Necessity, an Opportunity or a Luxury?

I. Introduction

The National Archives of the United Kingdom (TNA) has long undertaken collaborative activity with the wider UK archives sector, but in 2011 it received a specific leadership mandate from government to develop the archive sector, within the nation of England [1]. This leadership role has required a strategic approach to addressing key challenges which inhibit the development of archives in a modern context.

The current sector vision Archives Unlocked identifies sector digital capacity as one of three key development themes, to support both preservation and discoverability of archive holdings [2]. Previous strategic activity, notably Archives for the 21st Century, had similarly identified the challenge of digital capacity, but had set TNA's role as primarily to create guidance and tools for the archive sector to use [3]. Archives Unlocked marks a shift to a more interventive strategy.

A. The archives sector in England

The scale and complex nature of the archives sector across which this leadership role is delivered greatly increases the challenge of delivering change. Depending on definitions, the scope of the leadership role covers several hundred to several thousand archive services [4]. This has influenced the range of approaches currently employed.

1) Types of archive service: Archive services across England are highly heterogeneous. While there is a well-defined network of national and provincial public archives, major archive-holding operations exist among museums, charities, businesses, higher-education institutions, libraries, arts organisations, schools and community-developed archives. Audio-visual archives often exist as separate institutions. Some actively acquire archive collections by transfer or gift from external creators, while others function to deliver long-term preservation for institutional records remaining within the parent body.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The statutory basis for their archive collecting varies from a closely-controlled set of legal obligations for some national record types to a permissive absence of mandate in law [5].

Outside national archive operations, there are few large archive services. Size of staff varies typically from fewer than ten professionals to a service operated by one or two full-time-equivalent posts. Although individual archive services have been addressing the digital transition, this general lack of scale and resources makes fundamental changes such as the shift towards digital particularly challenging. In many services, an individual professional will be expected to master skills across public engagement, analogue preservation, technical cataloguing requirements, depositor liaison, fundraising, and now curation of digital records. It is a formidable challenge. There are unquestionably capacity pressures which have made it possible to postpone tackling digital transformation amidst other, equally urgent, pressures.

2) Types of collections: Whereas the legal basis and scale of archive services varies considerably, the nature of archive collections held is consistent across the vast majority of archive services in England. These are hybrid collections, representing a legacy of records in analogue format, often still actively accruing, and a collecting mandate which includes records in digital format. Acquisition of digital records is inconsistent, but there are few archives whose collecting mandate excludes post-20th century records, so a transition to digital acquisition and preservation is widely required.

B. The leadership role of The National Archives

1. Boundaries and possibilities: TNA's leadership role for the archive sector is only one of its areas of delivery [6]. As the archive for the UK government, management of and access to government digital records is core to the organisation's future. An institutional digital strategy sets its goal to "create the disruptive digital archive" [7]. TNA is also in itself a part of government, and a research institution, in which key archive topics can be explored and interrogated. This provides a vital background for work with the wider archives sector, and scope for innovative practice.

However, it is recognized that solutions suitable to a major national institution are not necessarily scalable. The cutting edge of research may not be what is required for a small archive with very limited resources and low risk appetite, seeking incremental progress. Balancing these roles of TNA is a vital part of fulfilling its work in different areas.

There are further complicating factors. TNA's leadership role is not supported by statutory powers over most classes of archive holdings. Only records which fall into limited categories (public, manorial, tithe, Diocesan) are legally controlled, and TNA's role with each is different. There is no blanket mandate for collecting archives across the nation.

TNA also does not have a statutory governance role over other archive services, nor is it funded to deliver archive services beyond its own collections. Change can only be achieved through partnership and consent. With the scale of the sector and the variety of mandates which support archive services' existence, one single solution to collecting and preserving all digital archives across England is impractically large and complex.

The role of developing archives sector capacity is also shared with other key strategic bodies. The Digital Preservation Coalition, from its UK beginnings, remains a key contributor to development and capacity building. The professional body, Archives and Records Association (UK and Ireland) has a more archives-specific role, and undertakes a range of training and guidance, including through the specialist Section for Archives and Technology. JISC seeks to provide digital solutions in a research context, and provides training and guidance across digital activities. TNA also works closely with sector bodies supporting archives in the three other home nations of the UK. This can make for robust and effective partnerships. It also inevitably adds complexity. TNA is not acting alone in this field, needing to ensure effective partnerships and coordinated effort for most impact. There is however both a government mandate and a regularly-articulated archives sector expectation that TNA will have a strong offer in the area of digital preservation, given its prominence as a challenge across the sector.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

2. Audiences for digital leadership: Development of digital skills throughout the sector requires focused attention on the practitioners who will be working actively with digital systems and collections and those in more senior leadership or strategic roles in which an overview of digital capabilities is crucial but detailed knowledge of tools or approaches is not. Crudely, leaders must know enough to lead. They must be able to evaluate different digital projects, proposals, contractors, tenders and approaches on their merits. If they cannot do this, they cannot adequately direct digital policy within their organisation and risk commissioning poor quality digital offerings and giving weak support to their staff.

But crucially, both archives practitioners and leaders are likely to be working in an environment in which their digital activity is constrained by an organisational IT culture. (If they are not embedded in a wider organisation they are likely to be highly resource constrained.) This is not of itself a problem. Digital humanities activity is by its nature collaborative. Archivists can collaborate with IT peers to achieve delivery of complex digital projects Both can learn to speak the other's language [8]. However where there is no history of successful collaboration with IT within the organisation, this may seem a daunting prospect.

Anxiety over the effort and skills gap required to initiate such collaboration is a key driver of resistance to engagement with digital within some archives. Archivists are used to mastery of their subject. Stepping into a new arena where they must simultaneously argue a strong case while feeling uncertain of their knowledge provokes aversive behaviours such as blaming slow procurement processes for a lack of progress, as if digital work could only be accomplished with a single large or expensive system [9].

II. Sector digital development: a multi-strand approach

This combination of a hybrid-holding archives sector with limited individual capacity, and a weak statutory or funding basis for centralized delivery solutions has guided TNA's thinking and development of a mixed economy of sector digital activity. This is a developing area, and subject to ongoing iterative review.

However, it has become critical to move forward from a passive, guidance-based, approach in previous strategic plans, to a more interventive role in building the sector's strength. Unfamiliar software and high-level technical language in support documentation can make this domain forbidding to archivists with humanities backgrounds. The best documentation, such as the DPC's Digital Preservation Handbook, does a good job of demystifying the discipline and yet has not transformed the confidence and position of the sector. After 15 years of concerted efforts at building excellent guidance and support, barely one third of those working in archives in England describe themselves as confident across a range of digital skills [10]. Stronger drivers are needed, and TNA can contribute at this level.

A. Strategic capacity building

In support of *Archives Unlocked*, a sector workforce strategy was published in 2018. Its first strategic objective: to "empower the archives workforce to adapt to major drivers of change, including digital technologies" [11].

Archives have excelled at certain aspects of digital work, including digitisation and the monetisation of digital assets. But compared to the museum and library sectors they are comparatively poor at a range of digital activities, particularly access and engagement. There has been great success at raising the profile of digital preservation work within archives but this has created a perception that digital archives work is solely preservation work. Overemphasis on digital preservation at the cost of other digital skills is unfortunate because more generic skills are essential precursors to undertaking digital preservation work.

This has also encouraged some institutions to solve a digital capacity gap through preservation software procurement. Unfortunately, purchasing software does not automatically ensure a commensurate increase in an organisation's digital capacity and may indeed weaken the sector, through wasteful cost and a lack of redundant storage, as too many organisations contract with the same provider. In the more mature area of digital cataloguing, we are already seeing unfortunate effects of 'vendor lock in' to outdated legacy systems, in violation of open standards principles [12]. This has also militated against

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

an approach which includes approval of particular software products, and towards an emphasis on understanding individual organizational context and solutions which fit collection needs, budget and technical competencies.

Training and guidance remain an essential part of TNA's activity to address these issues, but the method of approach has changed. Current areas of work include strongly emphasising that digital work is an integral part of archival activity and not an adjunct to it; supporting senior leaders to understand what opportunities digital affords their services; supporting digital skills by developing a package of training and resources to support digital preservation and other digital activities; lobbying for an increased focus on digital content on accredited UK archives courses; and continuing to support open standards and open software. Digital capacity must be built with equal focus on preservation, access and engagement; at an appropriate skill level for the archival workforce of today and deliver clear social benefit to citizens. Increasing the sector's digital capability should be done in such a way as to positively impact the digital literacy of wider society. These approaches are presented in full in the Digital Capacity Building Strategy [13].

B. Supporting networking and collaboration

Archival networks have been invaluable drivers of digital preservation work. The formation of regional consortia, often backed by grants from TNA, to explore issues of mutual concern has led several to develop joint work on digital preservation [14]. Often archives who are ahead in capacity use this as an opportunity to share knowledge and expertise with partners.

In 2018, a digital learning set using action learning techniques recruited 13 members from consortia and services already undertaking digital activity. This provided focused opportunities for knowledge exchange and skills development. In 2019, the set is being widened to form a new Digital Archives Learning Exchange (DALE) for archivists undertaking digital work, with discussion online and at periodic meetings hosted by regional archives. This will be a sector-led to address areas of emerging need. There is a specific aim to keep DALE membership to those who are actively undertaking digital development, moving beyond the model of informational training which is not then implemented in the workplace.

C. Offering new workforce opportunities

If a lack of digital skills and capacity is one of the crucial hindrances for progressing digital archives activity, a possible solution is the creation of new opportunities to draw in people to the archives workforce through different employment routes, broadening the range of individuals who work in archives. TNA has, with the support of the National Lottery Heritage Fund, run a series of traineeship programmes designed to address skills gaps in the archives sector. The most recent of these, Bridging the Digital Gap, specifically aims to bring in technical trainees who already have IT skills and aptitudes, and a background in Science Technology, Engineering and Mathematics. They then receive work-based training in archive principles and implementation of digital acquisition, preservation and access [15]. Among the outputs of the programme is intended to be a job description for a digital archives assistant, to inform recruitment across the sector. This model was built through consultative sector workshops, to ensure a good fit between role and real job requirements.

This approach is further informed by the DigCurV framework, and particularly its Practitioner Lens, which underlines the need for technical skills at operational level, while oversight of policy and strategy may sit with broader roles, in this case with archivists and strategic managers [16]. This is a significant change from the current staffing pattern in archive services, where specialist skills sit primarily with archive professionals. Early response from host archives shows that this has been a valuable innovation.

D. Updating national archive standards

Archive Service Accreditation is a management standard for archives across the UK, originally co-created and published 2012-13. It is supported by a partnership of seven strategic archives bodies, including TNA. The standard is designed to apply to archive services of many types and sizes, following a scalable assessment approach. Reference was made throughout the 2013 Accreditation Standard to "analogue and digital records", aiming to mainstream the management of digital records within collecting practice [17].

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

However, it was clear from the outset that further specific questions on management of digital archives would be required as the archive sector began actively to collect in this area. One benefit of Accreditation should be its role as a lever to improve performance. Leaving questions about digital preservation too open and generic risked weakening this impact. But given the sector's limited capacity, an excessively demanding standard for digital preservation risked limiting the standard to larger, better funded organisations, contrary to its fundamental purpose.

From 2015 a working group of Accreditation representatives collaborated with the Digital Preservation Coalition to address this challenge. After mapping against existing digital preservation standards, it was agreed that no single existing standard should be incorporated as part of Archive Service Accreditation. The group proposed using the NDSA *Levels of Preservation* (v1) as a means of exploring risks to the preservation of digital materials [18]. Further updates to the programme covered topics such as funding, staff capacity, ability to collect material in different formats, and access to born-digital records.

These proposals were agreed by the Archive Service Accreditation Committee and became a live element of assessment from July 2018. The Committee notes that full compliance is not immediately expected, given that this is a developing area, but an awareness of risk and evidence of service planning towards implementing preservation is essential for applicants to succeed [19]. With updates to NDSA Levels since the adoption of this approach, it is intended to continue to refer to a current version of the Levels for the short-term. In the medium-term the fit of the Levels and other standards, maturity models and risk models with the overall Archive Service Accreditation approach will be assessed, and is expected to evolve over time.

E. Evolving policy approaches

TNA's role as sector leader requires consideration of the intersection between policy and delivery, particularly at points of known change.

1) The future of Places of Deposit for Public Records: Among TNA's core legal responsibilities is oversight of a network of Places of Deposit appointed to hold public records [20]. This network allows records of primarily local interest – such as records of health authorities and magistrates' courts – to remain within their communities. However, the transition to digital means some of these records are now generated and maintained centrally. As patterns of record-holding change, so it is time to consider what a future model for holding these records should be. Exploring options here seeks solutions which safeguard the public record and support the sector's capacity to hold all digital records.

2) The impact of digital transition on mechanisms for protection of cultural property: The UK government has a number of mechanisms for protection of cultural property, including archives, such as export controls and tax incentives to donate to public collections. Charitable funders also support acquisition of cultural heritage, including archives. A public market in digital archives is yet to emerge, so that the current mechanisms do not yet have the reassurance of comparator prices to ensure that digital acquisitions can be supported in the way analogue archives have been.

This is an area where colleagues in TNA are undertaking research on the impact of digital archives on these mechanisms in general and specifically on export policy [21]. A range of reviews currently taking place across bodies concerned with archives as cultural property will also be relevant, as will a series of Digital Preservation Coalition events related to digital and value. Initial research suggests there is an appetite for the acquisition of digital archives and that the emergence of a market may be imminent, though its drivers are not yet clear. If the current mechanisms are to continue to work, it is essential to create a healthy digitally-capable archive sector prepared to accept and manage these culturally-valued digital collections. It seems possible that the range of services benefitting from the acquisition of significant archives through these mechanisms, already a relatively narrow group, will narrow further according to digital capacity.

III. Preliminary impacts and observations

These strands of activity have been developing in the first years of *Archives Unlocked*, and cannot be

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

regarded as fully mature. However, we offer preliminary observations for the benefit of others interested in collaborative improvement programmes.

A. Positives of the approach

If this work was not being undertaken, TNA would be failing in its leadership role. Many individual archive services in England have made significant progress towards digital activity through their own efforts. However, were TNA to continue to rely on publishing guidance online and offering generic training, rather than proactive interventions, there is a risk that other services would continue to avoid issue or struggle to secure institutional support to meet this challenge. Gaps in provision would continue to widen. A collaborative, national and networked approach brings services together and allows good practice to be highlighted, supported and shared.

TNA's position within government and with strong links across the archives sector provides a real opportunity to support policy development grounded in evidence of the variety of sector capacity. TNA's comparably dual role within digital institutions and cultural heritage can also be a support. There is some evidence that cultural heritage networks and digital preservation networks are talking among themselves about analogous issues of digital archives and value but are not yet talking to each other. TNA can have a useful role in bringing these networks together.

The flexibility and multiple entry points provided by a faceted approach to sector development mean that individual archive services are able to participate at a level that works for them. Tackling a range of issues offers something to any archive service around this critical area for their future. TNA can play a lead role in some aspects of development, but also act as a valued supporter in other strands of sector development work, incorporating opportunities to pursue its strategic aims.

The visibility of changes to Archive Service Accreditation to include more specific digital preservation content has been a vital part for all partners of moving the conversation across the entire archives sector. Archive services who choose to participate in Accreditation have for the first time assessed risks to the management, acquisition and accessibility of their digital collections alongside more familiar analogue risks. The standard's coverage of all types and sizes of UK archives has made it more difficult for reluctant services to avoid addressing the issue of digital preservation as a core element of their mission.

B. Known risks and weaknesses

The flexibility of the TNA approach has an unavoidable downside: it is more complex than a single, simple programme. As it forms part of multiple aspects of our work, within *Archives Unlocked* and partnership activities, it is harder to communicate than a one-track programme would be. There is a risk of appearing not to be taking action on this critical area, rather than having embedded it across business activities.

The dual nature of TNA's government/sector role also creates points of weakness and risk. As a government body we are not able to advise specifically on individual commercial products, when archives are urgently seeking systems solutions. At a policy level, the statutory duty to protect the public record has to be our priority when looking at the future of devolved management of public records. A solution which empowers the sector's digital capacity-building is desirable, but cannot be the only consideration.

Collaborative and networking activity has weaknesses too. Where no strong leading organisation exists, archives have plainly learned together, but some consortia have moved at the pace of their slowest members and others have fallen victim to the same procurement trap as individual organisations – delaying significant preservation activity until some key partner or vendor can be persuaded to come on board.

Current activity has also focused on developing archive service capacity. It does not address support for researchers to access digital records, thus compounding an issue we have already noted is commonly seen in individual archives. In future, TNA will need to build support across a record's life, from creation to use.

C. Conclusion

Addressing the digital challenge is fundamental to the future health of archives in England, as across

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

the globe. TNA has increasingly focused its strategic sector development in this area, and has moved from enabling and supporting through published guidance, to a more interventive, complex approach. Archives may opt to engage in networks, understand their risks through participation in standards schemes, employ digitally-expert staff or benefit from policy developments which recognize the transformational nature of digital change. While the diffuse nature of TNA's connections with the archives sector creates unavoidable areas of weakness and tension, it also avoids the risks of a monolithic approach which prioritises one single development area, neglecting other opportunities.

Above all, a connected but flexible approach makes change possible. The next challenge will be to move the conversation from regarding digital as a new task towards digital as business as usual. For the archives sector in England to fulfil its mission, we must find ways to move from regarding digital records as a new liability or threat, to recognizing their potential and value as a research resource for the future.

REFERENCES

[1] The National Archives, 'Our leadership role', http://www.nationalarchives.gov.uk/archives-sector/our-archives-sector-role/our-leadership-role/.

[2] The National Archives, *Archives Unlocked* (2017) http://www.nationalarchives.gov.uk/archives-sector/projects-and-programmes/strategic-vision-for-archives/.

[3] See for example The National Archives, *Archives for the 21st Century in Action: refreshed* (2012) https://webarchive.nationalarchives.gov.uk/+/http://www.nationalarchives.gov.uk/documents/archives/archives21centuryrefreshed-final.pdf.

[4] The directory underpinning Discovery lists over 2000 archive-holding institutions in the UK. Not all provide a full, staffed collections management service. They are therefore out of scope of some areas of activity such as Archive Service Accreditation. http://discovery.nationalarchives.gov.uk/find-an-archive.

[5] Guide for Archiving Personal Data, The National Archives. http://www.nationalarchives.gov.uk/documents/information-management/guide-to-archiving-personal-data.pdf pp.26-27.

[6] These multiple roles were specifically articulated in The National Archives' strategic plan, *Archives Inspire* 2015-19 http://www.nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-plans/archives-inspire/.

[7] The National Archives, Digital Strategy 2017-19 (2017) http://www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf.

[8] S. Prater. "How to Talk to IT about Digital Preservation." *Journal of Archival Organization*, vol 18, no. 1, pp. 1-12, 2018.

[9] This pattern has emerged from analysis of sector intelligence and survey responses over time.

[10] Demand for introductory courses run by providers such as the Digital Preservation Coalition remains high. Evidence from Archive Service Accreditation shows a high level of unmanaged digital risk in many archive services. This was confirmed by a 2019 sector survey and supporting research undertaken by JISC and TNA in which only 36% of respondents identified high level commitment to digital preservation in their organization. Confidence in skills across the digital piece was likewise shared by only about one third of respondents; of all the areas, practical implementation of digital preservation showed the weakest confidence and skills. The National Archives/JISC, Digital Capacity Survey, forthcoming 2019.

[11] The National Archives, Workforce Development Strategy (2018) https://www.nationalarchives.gov.uk/documents/archive-sector-workforce-strategy.pdf.

[12] Cabinet Office, 'Open standards principles', https://www.gov.uk/government/publications/open-standards-principles/open-standards-principles.

[13] The National Archives, 'Digital Capacity Building Strategy' (2019) https://www.nationalarchives.gov.uk/documents/digital-capacity-building-strategy.pdf

[14] Examples include an East of England Regional Archives Council Archivematica pilot, captured in a case study http://www.nationalarchives.gov.uk/documents/archives/eerac-case-study.pdf, and Archives First consortium projects on local records life-cycle and minimum preservation requirements: https://www.gloucestershire.gov.uk/archives/digital-curation/digital-curation-projects/archives-first/

[15] The National Archives, 'Bridging the Digital Gap' http://www.nationalarchives.gov.uk/archives-sector/projects-and-programmes/bridging-digital-gap-technical-traineeships-archives/. The project's predecessor, Transforming Archives, won a 2015 Digital Preservation Award, alongside the parallel Opening Up Scotland's Archives project: https://www.dpconline.org/events/digital-preservation-awards/tna-scottish-archives.

[16] DigCurV, Practitioner Lens https://www.digcurv.gla.ac.uk/practitionerLens.html

[17] Archive Service Accreditation Standard (2013) https://webarchive.nationalarchives.gov.uk/20131203151728/https://www.nationalarchives.gov.uk/documents/archives/archive-service-accreditation-standard-june-2013.pdf.

[18] Archive Service Accreditation Committee 2016 item 7.1 https://webarchive.nationalarchives.gov.uk/+/http://www.nationalarchives.gov.uk/documents/archives/archive-service-accreditation-commitee-minutes-12-may-2016.pdf. Publicised as Archive Service Accreditation Digital Developments https://www.slideshare.net/mhaunton/archive-service-accreditation-digital-developments-for-ara-conference-2016.

[19] Archive Service Accreditation Committee 2017 item 8.1 http://www.nationalarchives.gov.uk/documents/archives/archive-service-accreditation-commitee-minutes-11-may-2017.pdf

[20] The National Archives, Approved Places of Deposit http://www.nationalarchives.gov.uk/archives-sector/our-archives-sector-role/legislation/approved-places-of-deposit/

[21] An early output is R. Addison, "Protecting digital cultural assets: a review of the export process and supporting mechanisms", 2019 https://www.nationalarchives.gov.uk/documents/protecting-digital-cultural-assets-report.pdf

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

People Get Ready

Building Sustainability into Digital Preservation Workforce Development

Sharon McMeekin
Digital Preservation Coalition, Scotland
sharon.mcmeekin@dpconline.org
https://orcid.org/0000-0002-1842-611X

Abstract – Sustainability has been a key focus for the digital preservation community in recent years, but we have mostly focused on issues such as the sustainability of systems and funding. We cannot, however, be successful as a community without also developing a robust and skilled workforce. This paper will examine trends in workforce development in digital preservation and suggest steps forward we should take as a community.

Keywords – workforce development, training, education, skills, collaboration

Conference Topics – Building Capacity, Capability and Community; Designing and Delivering Sustainable Digital Preservation.

I. Introduction

Those working in digital preservation are well aware of the need for constant advocacy for what we do, in particular having to repeat the familiar refrain of 'digital preservation is not just storage'. We spend time explaining the importance of policy, documentation, robust systems, and sustainable funding, but do we spend enough time advocating for ourselves? A skilled workforce is essential to digital preservation and should be at the forefront of any advocacy effort as well as included as a key part of strategies for development.

Anecdotal evidence suggests that one of the barriers to this is that we have yet to define ourselves as a cohesive profession, despite attempts by projects such as DigCurV[1] to document a competency framework for a digital preservation professional. Instead we identify as subsets of other groups such as librarians, archivists, IT professionals and more. While this is indicative of the broad range of skills required to successfully undertake digital preservation, it also makes it more difficult to define and advocate for digital preservation roles. In turn, it also causes problems for structuring professional development and identifying and delivering training.

This paper will describe some trends in digital preservation workforce development identified by past projects, provide a brief overview of current issues as seen through the lens of the Digital Preservation Coalition[2] (DPC) membership, highlight work in this area that the DPC is undertaking, and suggest how the digital preservation community could work together to develop a robust approach to workforce development.

II. Some Past Work

One of the largest and most successful projects focused on the topic of digital preservation workforce development to date was DigCurV, completed in 2013. DigCurV surveyed training needs and provision, producing amongst its resources a competency framework for digital preservation practitioners working at three levels defined as "Practitioner", "Manager" and "Executive". The competency framework has proved to be a useful resource used by numerous training providers and University Masters courses to help frame and assess their digital preservation offerings, as well as by projects such as The UK National Archives' Bridging the Digital Gap[3] to help structure a training curriculum for trainees.

[1] https://www.digcurv.gla.ac.uk/

[2] https://www.dpconline.org/

[3] http://www.nationalarchives.gov.uk/archives-sector/projects-and-programmes/bridging-digital-gap-technical-traineeships-archives/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The training survey undertaken as part of the project [1] also produced a rich offering of information about training course provision for digital preservation. Some of the headline figures include:

- All of the courses detailed by responses received were offered in (mostly Western) Europe or the USA
- Training was twice as likely to be provided by subject specialists rather than those with experience in training provision
- Participants required no or only a basic understanding of digital preservation to undertake almost all of the courses (93%)
- The majority of courses (69%) were offered in a large group format
- Most courses did not incorporate any form of assessment (79%) and or offer certification upon completion of the course (60%)

From this it is possible to infer that the majority of digital preservation training, at the time of this survey, privileged those in the West and was at a beginner level. It was also likely to fail to meet a tighter definition of what might be considered training, lacking the high level of interactivity, practical emphasis, and consideration for different types of learners that those skilled in delivering training would consider best practice.

A further analysis of the same survey data by the APARSEN project [2] also raised concerns over the sustainability and suitability of training provision. It identified that only around a quarter (26%) of training courses could be described as being offered on a regular basis, with most (64%) being "one-offs". The report also identified key recommendations for improving digital preservation training:

- Less theory and more practice-based content with hands-on tasks
- More targeted courses (by audience and topic)
- More advanced courses for experienced professionals
- Establish a professional development framework

III. Current Trends

There have been a number of positive steps taken in support of workforce development for digital preservation in the close to six years since the DigCurV and APARSEN reports. For those looking to enter the profession there has been a steady increase in its representation amongst higher education courses, both as stronger elements in existing Master's programs as well as new, dedicated degree programs.

Likewise, there are many exemplars of good practice in training for existing professionals. Leren Preserveren[1] is providing Digital Preservation Award-nominated[2] introductory level training in the Netherlands. The POWRR[3] project and Aus Preserves[4] are making great strides in practical training in their institutes and carpentry sessions, respectively. The Digital Preservation Training Programme[5] has been leading in the provision of online training. The Institute of Sound and Vision[6] offers much in demand training on audio-visual archiving. And the IIPC[7] is developing, in partnership with the DPC, a suite of training resources on web archiving.

While these endeavours and other similar activities are having a real impact for those who attend, their reach in terms of numbers and geography is necessarily limited. This means there is likely still a significant shortfall in the provision of and access to digital preservation training and workforce development opportunities. Evidence of these issues was born-out in a survey the DPC carried out with members in the spring of 2018.

Gaining practical skills was highlighted as the key benefit sought from digital preservation training, with more than two thirds of respondents looking

[1] https://lerenpreserveren.nl/

[2] https://www.dpconline.org/events/digital-preservation-awards/the-finalists

[3] https://digitalpowrr.niu.edu/

[4] https://blogs.unimelb.edu.au/digital-preservation-project/2019/01/15/digital-preservation-carpentry-workshop-at-idcc-2019/

[5] https://dptp.london.ac.uk/

[6] https://www.beeldengeluid.nl/en/visit/events/winter-school-audiovisual-archiving-2019

[7] http://netpreserve.org/about-us/working-groups/training-working-group/

for intermediate or advanced-level training on all topics listed in the survey except "Emulation" (53% requested training at these levels for this topic). When asked for their priorities with regards to topics, practical options gained the most support with "Tools" being the clear leader, followed by "Metadata", "Planning", "Web Archiving" and "AV Preservation". They also overwhelmingly asked for sessions that included learning tools such as demonstrations, hands-on tasks, case studies and discussion exercises. Finally, many respondents noted multiple barriers to attending training: these included a lack of managerial support, no funding to meet costs, no suitable courses available, and an inability to take time away.

Without a larger international survey of training provision, it is hard to compare current digital preservation training with that described by the DigCurV and APARSEN reports, but available evidence suggests that provision does still lag behind demand. This introduces a risk that without action digital preservation skills might stagnate.

IV. What the DPC is Doing

At the DPC, Workforce Development has long been a core part of what we do, and it is one of the six objectives within our strategic plan [3]. We undertake a wide variety of different activities to support Workforce Development with our main outputs being the "Digital Preservation Handbook" (the Handbook)[1], scholarships offered through our Career Development Fund[2] (previously the Leadership Programme), and our training courses "Getting Started with Digital Preservation" and "Making Progress with Digital Preservation"[3] . In recent years we have made a concerted effort to fold the findings of projects such as DigCurV and APARSEN and the needs of the digital preservation community into these activities..

The Handbook underwent a major review and update that was completed in 2015. This resource is freely available via the DPC website and any content can be reused under an Open Government licence. We want to make sure the Handbook does not languish, and that it remains current and authoritative. With this in mind small corrections and updates are made as identified, and work has recently begun on a complete review that will lead to version three. Link integrity checking has been completed and new and updated content will be added, as well as more video content, including tool demos.

Our Career Development Fund undergoes constant review and has been significantly expanded in recent years. The program previously only offered scholarships to events that could be clearly defined as training on digital preservation, but, as described above, much training is introductory and the demand for training currently outstrips what is available. With this in mind we have made two significant updates to the program, guided by the member representatives who make up our Workforce Development Sub-Committee. Firstly, we have expanded the scholarships to cover any kind of professional development activity rather than just training. As an example, in the last three years we have provided ten scholarships for staff from member organizations to attend iPRES. The second expansion now allows members to bring specific scholarship requests to us. These requests do not need to be for development opportunities that are digital preservation specific, as long as they can demonstrate how the knowledge/skills gained will support their organization's digital preservation activities. This approach acknowledges that not all skills required for digital preservation are specific to the domain.

The final major update in recent years has been the reworking of our training outputs "Getting Started in Digital Preservation" and "Making Progress in Digital Preservation". Both courses have been redesigned to more closely align with the content of the Handbook as well as including more interaction through discussion and practical exercises. The key aim was to make sure participants were gaining knowledge and practical skills that would have real impact at their home organizations. Feedback on the changes has been extremely positive and we have maintained a policy of offering these courses to non-members at cost, so they remain accessible.

[1] https://dpconline.org/handbook

[2] https://www.dpconline.org/about/cdfund/

[3] https://www.dpconline.org/knowledge-base/training/training-resources

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Unfortunately, venue and technology constraints mean that the courses still lack hands-on training in the use of tools. Capacity has also meant that they have been primarily restricted to the UK and Ireland and we are only able to regularly deliver three of each course in a calendar year. In an attempt to address both of these issues, the DPC will be investigating online training provision in the next six months and are hoping to run a pilot by Spring 2020. This will hopefully allow for greater geographical coverage and provide the capability to offer new opportunities for learners to work with tools. We are also open to sharing our training materials for those around the world who would like to deliver training locally but may not have the time to develop content.

While this work has all been positive, the DPC continues to seek opportunities further improve and expand our efforts to support the workforce development of our members and the wider digital preservation community. Particular motivation also comes from our commitment to inclusion and diversity in digital preservation as laid out in our 2018 policy, presented during the iPRES 2018 lightning talks. We want to make sure digital preservation training and development opportunities are accessible for all who seek them. With this in mind, there are a number of areas of work we would like to undertake and collaborate on to help promote and improve workforce development for our community.

V. A Call to Arms

Brainstorming and creating a workplan for the DPC's Workforce Development activities generated several suggestions that could help pave the way for a more sustainable and inclusive environment for DP workforce development. They are as follows:

A. Refresh of the DigCurV Competency Framework

The DigCurV competency framework has proven to be an invaluable resource to many since its publication, but as the profession matures there is increasing scope for expanding on the original three lenses described above. This is particularly in response to the depth and breadth of knowledge and skills included even at the most junior level of "Practitioner". It is suggested that the number of lenses is increased, an initial proposal is shown in figure 1. This splits "Practitioner" into junior and senior lenses as well as adding descriptions for "Trainee" and "Graduate". Updates could support development of curricula for education and training, as well as offering a building block mentoring and professional development frameworks. All of these activities could potentially be investigated within a single project. The DPC is working to develop a potential project plan with original DigCurV partners and will be seeking funding. Input from the community will be very much welcomed if funding bids are successful.

B. Skills Audit Framework

Excellent work was undertaken by the Digital Preservation at Oxford and Cambridge (DPOC) Outreach and Training Fellows translating the DigCurV competency framework into a process to audit skills for digital preservation with their organizations [4]. This work would provide an excellent foundation to develop a more generic skills audit framework for the community. Such a framework would be a useful tool for advocacy, planning and development, and ongoing workforce development. This work may be folded into plans for DigCurV but may also become a separate project. Whichever route is chosen, volunteers for a set of international test cases are welcome.

Figure 1 Initial proposal for new structure of lenses for an updated DigCurV Competency Framework.

C. Recruitment Resources

The DPC receives many requests to help its member organizations with recruitment for digital preservation roles, where they are unsure where to start with defining what they need. A generic

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

collection of supporting resources, including role descriptions, salary figures and a guide on where to advertise could aid with this process.

D. *Training for New Audiences*

There is a demand for the development of training resources for audiences beyond the world of information management, and this might also double as an important advocacy tool. Groups for whom an appetite for training content has been expressed include Information Technology professionals, those in management roles, and researchers.

E. *Sharing of Knowledge and Resources*

An essential step towards sustainability, and greater inclusion, in digital preservation workforce development is establishing pathways for sharing knowledge and resources. Like digital preservation itself, collaboration is key! As a first step a DP Training Slack Workspace[1] has been created to facilitate collaboration. This will hopefully provide a central place to share where training resources are available online, as well as for sharing experience and knowledge, and brainstorming new ideas. It is hoped this will in time lead to collaboration and activities such as "train the trainer" events.

VI. Some iPRES 2019 Insights

At iPRES 2019, while this paper was presented, conference attendees were asked to provide answers to three questions on the topic of Digital Preservation Workforce Development using the polling tool Mentimeter. Firstly, they were asked "Are you able to meet your organizations' DP training needs?" (Figure 2) Of 103 responses, 80 (78%) said they were only partially able or unable to fulfil their training needs. This echoes the findings of the projects mentioned earlier, with the trend perhaps even worsening.

The second question asked, "What are your biggest barriers to accessing training?" (Figure 3) Here lack of suitable training was voted the biggest issue, closely followed by difficulties in making time to undertake courses. Again, we see the need for more training, and also training that will fit into increasingly busy work schedules.

Finally, the attendees were asked what topics they would like training to cover (Figure 4). Topics mentioned covered the full range of digital preservation issues, from starting out to 'business as usual' and from organizational issues to those of a more technological/process focus. Overall, training on formats, metadata, tools, and workflows/processes were the most popular. The answers to these questions clearly show the appetite for more and diverse training opportunities in digital preservation.

VII. Conclusion

While there has been much good work in the area of digital preservation workforce development and training, there is still a lot for us to do as a community if we are to establish sustainable, robust and accessible opportunities. We must build on existing resources so that we can clearly define what we need as a profession and give ourselves the tools to get there. The DPC will continue to make workforce development a key strategic priority and will be open to acting as a host, facilitator, supporter, and partner in taking efforts forward.

References

[1] V. Karvelyte, N. Klingaite, J. Kupriene, L. Molloy, K. Snow, and A. Gow, D2.1 Report on baseline survey and evaluation framework. https://doi.org/10.5281/zenodo.438636

[2] C. Cirinnà, S. McMeekin, W. Kilbride, Report on Survey of Training Material/ Assessment of Digital Curation Requirements.

[3] https://www.dpconline.org/docs/knowledge-base/1817-2012-02-21-aparsen-d43-1/file

[4] Digital Preservation Coalition, Strategic Plan for 2018 to 2022. https://www.dpconline.org/docs/miscellaneous/about/1755-dpc-strategic-plan-2018-22/file

[5] S. Mason, and L. Pretlove, Digital Preservation Training Needs Assessment Toolkit.

[6] https://doi.org/10.5287/bodleian:pz64RQmGE

[1] https://dpconline.org/news/dp-training-slack

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Figure 2 Graph showing responses to question on the ability to meet training needs

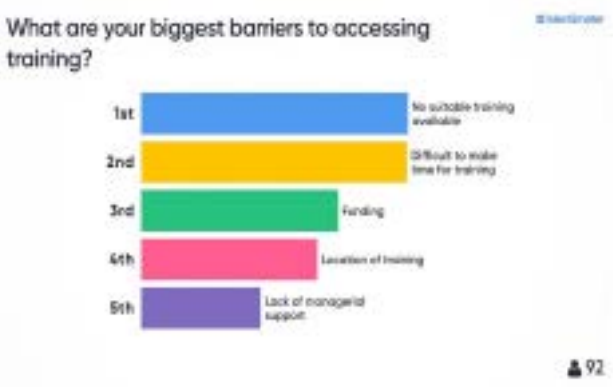


Figure 3 Graph showing responses to question on barriers to accessing training

Figure 4 Word cloud of digital preservation training topics of interest to iPRES 2019 attendees

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Passive Digital Preservation Now & Later

Microfilm, Micr'Olonys and DNA

Vincent Joguin
Eupalia
France
vincent.joguin@eupalia.com
https://orcid.org/0000-0003-0627-8778

Abstract – This paper presents the newly available Micr'Olonys software-on-film digital archiving solution that builds on a multi-layered virtual machine, and microfilm, a proven medium with a life expectancy of 500 years. The passive digital preservation strategy it follows matches the features of a written form contrasting with current active strategies that rather compare to orality. Micr'Olonys stores digital files as 2D barcodes whose content can be restored, using any general-purpose computer and scanner, by implementing the procedure described in a self-contained bootstrap that spans a dozen human-readable pages including a simple 2-page algorithm.

In the future, Micr'Olonys aims to become a powerful preservation tool for complex formats and software, as well as the access key to DNA digital storage to archive massive amounts of data.

Keywords – passive, software, preservation, microfilm, DNA

Conference Topics – Designing and Delivering Sustainable Digital Preservation; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

Digital contents and processing now permeate all human activities. Whereas natural languages take distinct oral and written forms, that of digital manifestations, mediated by computer hardware and software, is unclear. However, digital shares many characteristics with an oral form, with tremendously quick manipulation and transmission, fragile short-lived storage media, and constantly evolving environments and formats. As the Encyclopedia Britannica notes, "both [the Internet and oral tradition] work through rule-governed processes rather than fossilized texts; and both ultimately derive their strength from their ability to change and adapt"[1].Digital lacking to fully qualify as a written form, digital preservation is currently a matter of perpetually migrating media and contents, akin to mechanisms of oral tradition. The latter, whereby information is being preserved and transmitted orally, is characterized by the absence of any fixed form. Nevertheless, in oral cultures, accurate transmission of information may recourse to rhythmic speech (e.g. repetition, alliteration and assonance) which serves a purpose comparable to that of cryptographic hash functions or representation information used in digital preservation to maintain integrity and authenticity.

Indeed, writing digitally requires a series of conditions, from robust media to ensured long term readability. Until now, these conditions have never been met simultaneously, preventing a digital preservation approach similar to traditional paper archiving to emerge.

While oral tradition depends on people to ensure the continued transmission of information, migration-based digital preservation additionally depends on the continued availability of advanced digital technology and sustainability of energy production. However, in the current times of economic uncertainty and global warming, such resource-demanding preservation strategies are at risk of failing to pass our digital heritage on to future generations, which may have far-reaching adverse implications. Or, as the Latin proverb went, "verba volant, scripta manent": speech flies away, writing remains.

II. Digital Preservation Strategies

From the inception of digital preservation in the early 1990s, a dilemma has grown about

[1] https://www.britannica.com/topic/oral-tradition

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

preservation strategies between migration and emulation. While migration – referring in this context to file format conversions over time – appeared as a simpler strategy technology-wise, emulation promised better authenticity and reduced costs. However, they were technological approaches rather than core digital preservation strategies.

Ultimately, only one core strategy came forth, that consisting of "the active management and maintenance of digital objects [...] so they can be accessed and used by future users."[1] Although this definition of Digital Preservation by the United Nations implies a migration approach as it focuses on digital objects, a similar definition implying an emulation approach could be formulated for active digital preservation focusing on digital environments. Active digital preservation strategies not only ensure that digital material can be accessed and used by future users, but more specifically that it is kept alive in future digital environments and can be smoothly accessed in the same or equivalent conditions as then current digital material. Such strategies are obviously desirable in many cases when the material is used frequently, when quick access to it is critical, and/or when it needs to be changed or combined with other material.

But what of digital material that is used infrequently over time, or most often never at all, and which may satisfactorily not be accessed immediately? This kind of material, usually archival records, arguably represents a very large portion of digital material that needs to be kept for the longest periods of time; active strategies, which are costly and prone to both economic and technological risks, do not seem appropriate for such material.

Therefore, a different strategy, passive digital preservation, could better fit the requirements to keep relevant digital material unchanged over time, in fact as is the case when preserving material on paper. Advancing the principles of the emulation approach, passive digital preservation builds on an immutable digital environment, whereas active digital preservation focuses on preserving significant properties in a constantly evolving environment.

While the challenge of active preservation is to accommodate to an evolving environment, that of passive preservation is to comprehensively capture an environment which is always very complex in the digital realm.

III. Reducing the Hardware Environment

A. *The Virtualization Approach*

From 1992, and for 20 years, Jeff Rothenberg has consistently formalized and supported a digital preservation approach focused on the digital environment[2]. His proposed approach [1] has led to the experimental development of a Universal Virtual Computer (UVC) by IBM, in collaboration with the Royal Library and National Archives of the Netherlands [2], which has since been further investigated [3].

Other projects have more recently emerged, both with the very similar approach of a simple dedicated virtual machine such as the iVM project[3] or [4], or using more conventional virtualization or emulation technologies such as the Olive project[4] supported by Vint Cerf who advocated the "digital vellum" concept [5], or the EaaSI project[5]. Some larger-scope initiatives also include similar approaches, in particular the PERSIST program of the UNESCO[6] and the BRASS program initiated by the DARPA[7] .

[1] https://archives.un.org/content/digital-preservation

[2] http://jeffrothenberg.org/Prof/digilong.html https://www.slideshare.net/FuturePerfect_/jeff-rothenberg-digital-preservation-perspective

[3] http://www.dlmforum.eu/jdownloads/eventsarchive/2019Bern/session2/2_1_Rummelhoff_%20stvold_Liab%20_Immortal%20virtual%20machine.pdf

[4] https://olivearchive.org/

[5] https://www.softwarepreservationnetwork.org/eaasi/

[6] https://unescopersist.com/

[7] S. Neema et al., Building Resource Adaptive Software Systems (BRASS), 2015-2019. https://www.darpa.mil/program/building-resource-adaptive-software-systems

B. *The Olonys Virtual Machine*

We started designing and developing the Olonys virtual machine at Eupalia from October 2000 until mid-2003. Work on the virtual machine was resumed in February 2009 within the context of the KEEP project[1].

The main distinctive characteristic of this virtual machine is to replace most of the hardware complexity with software, thereby drastically reducing the hardware environment necessary to instantiate the complete digital environment and to render associated digital content. Olonys achieves this by building on 5 different virtual processor layers, from a full-featured processor on par with the x86-64 or ARM 64-bit processors, down to a very simple processor that supports only two different instruction types (minimal processors are studied in [6]). Compared to IBM's UVC, the middle layer of Olonys (3rd layer) is of comparable complexity with 23 instructions, whereas the simplest layer is almost two orders of magnitude simpler, making it both much more robust, i.e., much less bug-prone, as well as easier and faster to implement.

IV. Materializing the Incorporeal

Reducing the reliance of software on hardware logic gates does not remove the need for software and digital content to reside on a very material substrate, and not in thin air, in clouds, as is still too often believed.

Common recordable digital media – magnetic such as hard disk drives and tapes, optical such as recordable CDs and Blu-ray Disks, electronic such as flash drives – are either not durable, or at least their durability has not sufficiently been proven, especially out of experience since they are all relatively new. Therefore, various digital media, usually optical[2] , have been developed with a defined goal of durability. However, these media are either too expensive (e.g. when using glass) or have not been demonstrated to be more durable than common media (see for example [7]).

Existing analog media used digitally currently stand as the only alternative. Paper and film in particular are trusted carriers, especially when they conform to dedicated standards, for instance permanent paper defined by ISO 9706[3] and LE-500 rated microfilm[4], both of which being designed to last typically 500 years or more when stored appropriately. Moreover, film is both inexpensive and dense when compared with paper, while a magnifying glass is sufficient to retrieve filmed pictures.

Using such optical analog media digitally requires to convert bit streams into visual signals printed as pictures. Beyond the common QR code, Data Matrix and similar two-dimensional barcode standards, dedicated solutions have been developed for both paper (e.g. Optar[5], PaperBack[6] and PaperDisk[7]) and film (e.g. Monolith[8] presented in [8], PreserveOn[9] and Piql[10]). An approach tailored for one medium may also be used with the other, as exemplified by [9].

[1] Keeping Emulation Environments Portable (KEEP), FP7 project, 2009-2012. https://cordis.europa.eu/project/rcn/89496/factsheet/en

[2] See for example:
http://www.arnano.fr/
https://web.archive.org/web/20190401083338/http://www.glazt.com/
https://web.archive.org/web/20190809165627/http://www.mdisc.com/
http://www.datatresordisc.eu/english/introduction-page-dtd.html
http://group47.com/
https://www.southampton.ac.uk/news/2016/02/5d-data-storage-update.page

[3] https://www.iso.org/standard/17562.html

[4] https://www.epminc.com/support/tech-tips/epm-product-information/21-tech-tip-050-life-expectancy-of-microfilm

[5] http://ronja.twibright.com/optar/

[6] http://ollydbg.de/Paperbak/index.html

[7] http://www.paperdisk.com/

[8] https://www.bitsave.ch/eng/index.html

[9] https://web.archive.org/web/20170602102100/http://preserveon.com/

[10] https://www.piql.com/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

V. Optical Digital Bootstrap

While bits-on-paper and bits-on-film solutions overcome the barriers of media durability and independence from a specific technological device to access the preserved signal, a last hurdle remains: ensuring easy decoding of the signal back to exploitable information over the long term, when the necessary hardware and software stack will have long become obsolete and disappeared.

Based on the Olonys virtual machine presented previously, we have created a self-contained bootstrap that is part of Micr'Olonys, a software-on-film solution for long term digital preservation on microfilm. It is meant to be printed at least once per microfilm roll or microfiche so that accessing the preserved content relies on no external resource other than a film scanner and general-purpose (i.e., Turing-complete[1]) computer, both arguably expected to remain readily available over the long term.

This optical digital bootstrap is capable of decoding dense custom two-dimensional barcodes with both intra-frame and inter-frame Reed-Solomon error correction. Using intra-frame correction, the original data will be entirely recovered although up to 7.5% of each barcode may be damaged. Inter-frame correction is able to completely rebuild 3 out of 20 heavily damaged or missing data blocks distributed within 20 different barcodes. This means that some barcodes may have suffered more than 7.5% damage, or up to 3 barcodes per series of 20 may be missing altogether.

The core of the bootstrap consists of 7 pages of worded technical instructions to precisely guide the user of the distant future through the process of setting up the environment necessary for program execution and proper scanning of the barcodes. It includes a simple 2-page algorithm whose implementation is responsible for execution of all necessary software, raw scanned images input and decoded file output.

A user-defined front page, 4 pages of listed letters to be input manually or using OCR upon initial execution of the algorithm, and 5 pages containing system barcodes complete the bootstrap for a total of 17 pages. The system barcodes contain internal Olonys software, including a dynamic recompiler that accelerates execution by a factor of more than 15, and the decoder for the more complex and advanced barcodes that hold preserved data.

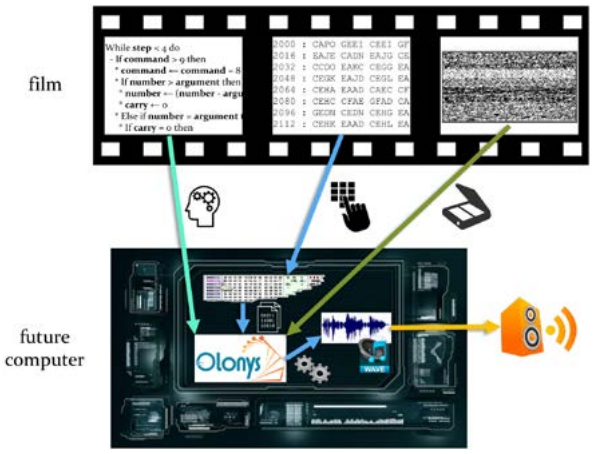


Fig. 1 depicts the example of how the bootstrap is used to restore a sound file preserved on film.

The bootstrap is designed to rely on as few and as simple concepts as possible so as to minimize misunderstandings arising from changes over the course of decades or centuries to the natural language and concepts employed. The bootstrap is currently available in English and French, and both versions may be printed on the same carrier to help recover the intended meaning, similarly to how the Rosetta Stone was key in deciphering ancient Egyptian hieroglyphs.

[1] Incidentally, Turing-completeness is the one property of computational systems that makes emulation and virtualization possible

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

VI. The Future: Software Preservation and DNA Storage

A. *Preserving Complex, Dynamic and Interrelated Content*

Although the Micr'Olonys solution embeds a general-purpose software processor used to automatically decode 2D barcodes, it is currently only able to restore files: their formats therefore need to remain sufficiently simple or normalized to avoid the problem of file format obsolescence. Recommended formats include BMP, uncompressed TIFF, WAV and ASCII text.

Future developments of the solution will include native support for more complex, compressed and composite formats such as PDF, JPEG2000 and 3D models. The bootstrap would be able to restore either the preserved file in its original format, or a conversion to an easily-exploitable uncompressed image or sound stream. Interactive and connected content such as dynamic websites, video games or software-based artworks are also planned to be supported using an additional self-adaptive peripheral management system.

B. *Ultra-high Capacity Passive Digital Preservation*

A major drawback of preserving digital data on microfilm is the relatively low capacity of this medium when compared with current digital storage technologies. For instance, a movie soundtrack can typically fit on digital microfilm, whereas preserving movie frames in lossless quality would require many more reels than an analog film shoot, which is simply impractical.

To qualify for passive digital preservation, any carrier with higher capacity should however fulfill the properties of passive storage and independence from a specific technological device for access. DNA storage matches these criteria: as a natural information carrier, it is proven to preserve information over millennia when stored in appropriate conditions; the ability to sequence DNA, although a highly complex process, can be expected to endure for as long as interest in DNA will exist, i.e., arguably over the very long term. Decoding digital information preserved within DNA will however never be trivial, and therefore an external, simpler carrier will need to be used as an access key. Microfilm or paper would typically be used in this context.

Many actors, including the Paris-based DNA Script company[1] and Microsoft[2], are collaborating to make fast and affordable DNA writing a reality. Significant investments are needed to complete the research and development necessary to reach an operational solution, but the incentive now exists to support the effort. The R&D process is expected to last for another decade and to progress similarly to the process that led to DNA sequencing, i.e., at a pace faster than Moore's law. Simultaneously, digital preservation approaches dedicated to DNA have started to emerge (see for example [10] and [3]). Conforming to acceptable ethics standards may ultimately constitute the main challenge in DNA storage adoption.

If DNA storage becomes a reality, it would stand as an extremely competitive solution to preserve digital information for which immediate access is not required and/or as a fail-safe complement to active preservation solutions.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231954, as well as from Bpifrance, and is currently being supported by the technical industries financial instrument of the Centre national du cinéma et de l'image animée (CNC).

[1] http://dnascript.co/

[2] https://www.microsoft.com/en-us/research/project/dna-storage/

[3] https://oligoarchive.eu/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

References

[1] J. Rothenberg, "Ensuring the Longevity of Digital Information," 1999. http://www.clir.org/pubs/archives/ensuring.pdf

[2] R. Lorie and R. van Diessen, "UVC: A Universal Virtual Computer for Long-term Preservation of Digital Information," 2005. https://domino.research.ibm.com/library/cyberdig.nsf/0/10229b6de0d054c-585256fa900681995?OpenDocument&

[3] N. Krebs, L. Schmitz and U. Borghoff, "Implementing the Universal Virtual Computer", *EUROCAST 2011*, in Lecture Notes in Computer Science, vol. 6927, pp. 153-160, Berlin, Heidelberg: Springer, 2012. https://www.researchgate.net/publication/310576462_Implementing_the_Universal_Virtual_Computer

[4] L. T. Nguyen and A. Kay, "The Cuneiform Tablets of 2015," *SPLASH 2015*, Oct 2015, Pittsburgh, PA. http://www.vpri.org/pdf/tr2015004_cuneiform.pdf

[5] V. Cerf, "Digital Vellum and Archives," *FASTER CoP*, Aug 2016.

[6] https://www.nitrd.gov/nitrdgroups/index.php?title=DigitalVellumAndArchives W. F. Gilreath and P. A. Laplante, *Computer Architecture: A Minimalist Perspective*, The Springer International Series in Engineering and Computer Science, vol. 730, Boston, MA: Springer, 2003.

[7] J. Perdereau, "Investigation on the preservation of digital archives in the framework of GISDON," 2012, page 8. https://www.lne.fr/sites/default/files/inline-files/syylex-glass-dvd-accelerated-aging-report.pdf

[8] P. Fornaro, R. Gschwind, L. Rosenthaler and A. Wassmer, "Monolith: materialised bits, the digital Rosetta film," *DH 2014*, Jul 2014, Lausanne, Switzerland. http://dharchive.org/paper/DH2014/Paper-615.xml

[9] S. W. Schilke and A. Rauber, "Long-term archiving of digital data on microfilm," *Int. J. Electronic Governance*, vol. 3, no. 3, pp. 237–253, 2010. https://publik.tuwien.ac.at/files/PubDat_191964.pdf

[10] M. Dimopoulou, M. Antonini, P. Barbry and R. Appuswamy, "A biologically constrained encoding solution for long-term storage of images onto synthetic DNA," *EUSIPCO 2019*, Sep 2019, A Coruña, Spain. https://hal.archives-ouvertes.fr/hal-02064896

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Building Resilience at the National Geoscience Data Center

Enhancing Digital Data Continuity Through Research Data Management Training

Jaana Pinnick
National Geoscience Data Centre
United Kingdom
jpak@bgs.ac.uk
ORCID 0000-0002-7718-5133

Abstract – The National Geoscience Data Center (NGDC) is the designated repository for the Natural Environment Research Council (NERC) grant-funded Earth science data and holds the CoreTrustSeal certification. The NGDC is hosted by the British Geological Survey (BGS), which co-funds post-graduate research students through the BGS University Funding Initiative (BUFI) program.

This paper describes the research data management training (RDM) course developed and delivered by the NGDC to help instill good data management practices in our students from early on, and to strengthen the long-term quality of research data they generate and deposit with the NGDC. It also looks at how RDM training fits into the wider context of the NGDC modular digital preservation program, currently under development.

This paper is aimed at data repository managers and research data managers who provide user training in data management best practice and digital preservation. It is also suitable for postgraduate students interested in digital continuity and preservation of their research data.

Keywords – Earth Science, Geoscience, Collaboration, Research Data Management (RDM), Post-Graduate Training.

Conference Topics – Collaboration: a Necessity, an Opportunity, or a Luxury?; Building Capacity, Capability and Community.

I. Introduction to BGS and the NGDC

The National Geoscience Data Center (NGDC) is the designated repository for the Natural Environment Research Council (NERC) grant-funded Earth science research data and the guardian for many commercially funded datasets. The NGDC is hosted by the British Geological Survey (BGS), and as one of the NERC Environmental Data Centers [1], it is committed to supporting long-term environmental data management to enable continuing access to these research assets. The NERC data policy [2] requires that all environmental data of long-term value generated through NERC-funded activities must be submitted to the designated repository for long-term management and dissemination.

BGS has existed for almost 185 years and continues to hold some of its original early 19th Century notebooks and maps. However, in this digital age, we recognize that the persistence of digital data is much more precarious than that of these hard-copy materials. It is therefore in our interest that all Earth scientists collecting and generating digital data, which we may end up preserving in the long-term, have a solid understanding of data management processes and best practice.

II. BGS and PhD Studentships

BGS co-funds postgraduate Earth and geoscience research students through the BGS University Funding Initiative (BUFI) program [3]. At least one of a student's supervisors is a BGS scientist, and the students often use the site science facilities for their PhD research projects. This may include access and reuse of data held by NGDC, and exposes them to the operations of the data center as part of their research activities.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

A. *Research Management Training*

BGS identified the need to bridge a skills gap in students' Research Data Management (RDM) training and application and decided to develop tailored guidance to support the data management element of their geoscience research projects. During their time at BGS, the students now participate in a one-day in-house RDM workshop, which has been run for BUFI students and another NERC student-funding program, Central England NERC Training Alliance (CENTA), since 2016. To date, a total of 120 PhD students have participated in the BGS workshop (March 2019). We have also been involved in discussions about delivering this training course to a third national geoscience PhD program in the future.

B. *Aims of the RDM Workshop*

The aim of the workshop is two-fold. Firstly, it aims to build up data management skills and capability amongst the students, and secondly, strengthens the quality of Earth science data that eventually is deposited at the NGDC. The repository is also expected to keep the data for at least ten years after the publication of the research that it underpins. However, the validity of Earth and geoscience data is usually much longer than this [4], and as a Place of Deposit under The UK National Archives, the NGDC is committed to looking after certain data in its care in perpetuity. The informational content of our research data underpins research long into the future because geoscience is an interpretative discipline. As such, the data do not often become obsolete, but the interpretation may change or be superseded as new methodologies and technologies become available.

C. *Benefits of the Workshop*

As part of the collaboration with universities, BGS shares the latest research practices and methodologies with the earth science students, and acts as a link between cutting-edge Earth science research and the extensive data held by the repository. We see the whole UK geoscience research community benefiting from our collaboration with the students in the long run, cultivating relationships, developing mutual trust, exchanging knowledge, and providing professional guidance to them. This collaboration raises awareness of the long-term impact of RDM best practices and its role in the digital continuity and preservation of Earth science data within the research community, including the students and staff at universities, and also those at BGS and NGDC.

III. RDM Workshop Content

The RDM course informs students about the NERC data policy and advises them about the obligations and rights of NERC-funded researchers, such as the requirement to offer a copy of their datasets to a NERC Environmental Data Center, and the possibility to access and reuse existing data. The course includes the basic RDM concepts and gives students the practical skills necessary for them to manage their data in a way that both benefits them and supports the aims of their assigned NERC data center. In the following sections we describe the content of the one-day workshop and what the students can expect to take away from it to develop their individual research data management practices.

After a brief introduction to NERC data policy and its Environmental Data Centers, the *Managing Research Data and Metadata* module demonstrates the value of adding structure to the data, and shows how making data interoperable and discoverable through a number of data portals enables its reuse in the future with tools and by users yet unknown. It also includes guidance on developing robust file naming and versioning strategies, organizing data using a clear file and directory structure, and selecting appropriate file formats before depositing the data in a long-term repository.

The module *Data Management Planning – Completing a DMP* is always popular with the students, as this is often the first time that they have been asked to write a data management plan. By completing a practical exercise students are able to review their research project from beginning to end and to consider the impact of their in-project data management activities on the long-term storage and continuity of their data. It also requires them to consider different types of research assets they are generating, such as new digital data, websites, models, code, software, and so on. These all have their unique requirements that the students have often not thought about at this stage. The session immediately equips the students with data management plans for their own research and with new skills to employ in their research careers.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

During the session on *Ensuring Data Quality and Preparing Data for Depositing* the students learn about the role each link in the PhD data management chain plays in data quality, whether they are the researcher, the data repository, or the project supervisor. We use real life examples from the data center and BGS to demonstrate cases of bad practice, to provide best practice guidance, and to show how to use repository resources to standardize data, to check data documentation and errors, and select what to keep.

In the module *Data Storage and Security and Long-Term Preservation*, we ask the students to consider their data storage and security requirements. We then talk about the main causes of data loss and about how to mitigate them. We also discuss the difference between in-project and long-term sharing and storage of data, why backup is not the same as preservation, and why they should care about preserving their data. In addition, they learn useful practical tips for future-proofing their data, such as using Open Office formats, and creating preservation-ready spreadsheets.

In 2018, we added a module on *Open Science and FAIR Data*, which was well received by the students. According to the feedback, students had not discussed this topic at their universities, and agreed with the benefits of Open Science for researchers, including the need for data to be Findable, Accessible, Interoperable and Reusable (FAIR) [5]. Using persistent identifiers for different elements of their research was also of general interest, with most the students registering for an ORCID unique identifier for researchers before the end of the session.

In the final session, *Data Retrieval and Reuse*, the students investigate research data repositories and data centers, learn to evaluate their trustworthiness, and search for data they may be able to reuse in their PhD projects. This session brings together many of the topics touched upon earlier in the day, including naming and organizing data in order to make it understandable and recoverable, and making it accessible by sharing it and depositing it at an appropriate long-term repository.

IV. The Benchmarking of RDM Training

A. *Developing Feedback Process*

To assess the quality of our RDM workshop, in 2017 we participated in an initiative led by Cambridge University to develop shared benchmarking metrics for RDM training courses delivered across the participating universities and research organizations [6]. The aim of the exercise was to agree on a minimal set of questions as benchmarking criteria to identify what works best for RDM training, and on which questions should be mandatory or optional. The participating members agreed to use six mandatory questions and a five point rating scale where (1) is the worst rating and (5) is the best [7]. We have used this feedback format on four of our workshops now and found it very useful when developing the workshop content and delivery further, establishing which modules and elements are the most useful for the students, and identifying any major gaps in the content.

The participating students often also work in wider NERC-funded research programs. Providing them with the best practices for long-term data management reaches a larger number of our end users because the students share this knowledge with their supervisors and research partners at universities, helping us disseminate the funder requirements for good quality data at the creation stage rather than at the point of deposit. As these practices become a staple part of their professional practice from early on, early career scientists will benefit from these skills over their entire career. This will lead to better transparency and reproducibility of their science and enhance their collaboration opportunities. Earth science discipline as a whole benefits from more robust science and data, which contributes to the development of the national data collections.

B. *Student Feedback Received*

The feedback received from the RDM training course since the introduction of the shared benchmarking metrics indicates that the more tailored the content of the course is to match the needs of the students, the more they benefit from it. BUFI students stated that the course met their

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

expectations on a level of 4.2/4.2 out of 5, whereas CENTA students' rating was slightly lower at 3.2/3.8 respectively. When asked if they would recommend the workshop to their peers, BUFI students gave the course a rating of 4.0/4.3 and CENTA students 2.6/3.7. The figures show that using the feedback loop to enhance the course content and delivery has led to higher satisfaction by the students.

Areas where the course was felt to be particularly useful were learning new data management skills, completing a data management plan, considering what aspects of data management may enhance the continuity of digital data (selection of file formats, providing robust metadata and data documentation alongside the data), and learning consistent file naming and versioning strategies which enable wider data reuse in the future.

Suggestions for improvement included providing more interactive activities and discussions, and spreading the training sessions over a longer period of time. This indicates that it is a challenge to achieve the right balance between providing enough information and skills on one hand, and time for students to practice the learning on their own research projects on the other. To us, collaboration with and communication between universities, students, and their supervisors, is the key to better data management practices, and consequently, to more robust data quality in the long-term.

V. Building digital preservation capability at BGS and the NGDC

Delivering the RDM training workshop is one of the contributing elements to the development of a sustainable and modular digital preservation program for BGS. In the following paragraphs, we give an overview of some of the work we have planned and undertaken following an initial investigation of the NGDC digital preservation requirements [4].

A. Strategic Framework

The BGS digital preservation policy, first introduced in 2017 [8], states that our overall approach is to develop a scalable preservation program, which will be further detailed within the internal preservation strategy (currently under development). Promoting best practice and delivering staff training were identified as key components of our preservation framework. We further evaluated some of the different implementation options available in our internal business case, which also emphasized the role of training and raising awareness of digital preservation.

As a public sector organization, we do not have a large budget to spend on commercial solutions. However, we have extensive in-house data management and developer skills to support the integration of new workflows and procedures as well as training. We therefore decided to pursue a modular solution which allows us to be flexible with our development and implementation, and started by reviewing and enhancing our existing procedures, infrastructure, and digital skills to implement our preservation framework.

We will use our existing discovery metadata schema to create a digital asset register, adding preservation and technical metadata elements to build a complete picture of our digital objects. To develop our digital preservation action plan we will use findings from stakeholder surveys, interviews, and risk assessments. The asset register will provide us crucial information required to make fact-based decisions on our preservation priorities and updates to our data management procedures and strategies.

We will conduct a digital preservation capability assessment to identify the gaps and where our resources are best employed. Implementing our top preservation priorities will be done in collaboration with both the BGS and NGDC data center staff, and with the senior management and the end users. Our digital preservation policy and a flexible strategy – organic yet controlled – will be tailored for the organization and its designated community. Our aim is to use our resources wisely, integrate new relevant digital skills into our existing workflows and practices, and focus our thinking on the long-term preservation and continuity of earth science data.

B. In-House Data Infrastructure

Our data and information infrastructure is largely built and developed in-house. Our corporate digital research data holdings, which are stored on the

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

storage area network (SAN) and the corporate tape archive, currently exceed 1,200 TB.

The key datasets on the SAN are backed up at two other geographical locations. In addition, we hold a legacy magnetic media archive of over 5,600 items on different data cartridges, reels, tapes, CDs/DVDs, cassettes and other media. However, we do not have in-house access to the older technology or the resources needed to rescue most of these legacy data, or to enable informed decision-making on which content to migrate onto new technologies. Even with the necessary resources, we may not have sufficient contextual and rights metadata to allow appropriate reuse of these data.

To avoid this issue from reoccurring in the future we have developed a standardized ingestion and accession procedure, requiring the data depositors to submit all the necessary discovery and rights metadata to accompany their deposits, so that future users have all the information they need to be able to reuse the dataset with confidence. We will build our data preservation capabilities further by adding the function of checksum value creation at ingest and fixity checking for key datasets, and selected PREMIS metadata fields to be maintained alongside our corporate discovery metadata schema.

C. CoreTrustSeal Certification

Reviewing the capabilities of the NGDC was also part of our self-assessment for the CoreTrustSeal (CTS) certification, which we gained in January 2018 [9]. The NGDC wanted to gain the CTS certification to build stakeholder and end use confidence in the repository, and to help benchmark our processes against a recognized methodology.

As part of the CTS submission, we confirmed that all BGS and NGDC staff have access to a comprehensive learning and development program, keeping them up to date with the latest data management techniques. Including PhD students in our training offering is part of this strategy. Engaging with all of our end users from an early stage in their research project lifecycle will help us identify where we can make improvements for the users and streamline the processes to facilitate data management workflows for them.

We are working toward further enhancing our capabilities within the CTS schema, and this will form part of our submission for continued future certification. We see the collaboration with our end users as the way forward to ensure that this certification delivers benefits for both the data center and the users. To achieve this, we will monitor their most up to date requirements and share expertise and experience with other memory and preservation organizations.

VI. Conclusion

Training early career scientists to manage their research data with a view to its long-term preservation accomplishes many important objectives: it raises the students' awareness of digital preservation; it builds their digital preservation capability and professional RDM skills; and it enhances the quality of data and data management skills in the whole Earth science community. To achieve this, NGDC staff must make data preservation relevant to early career scientists, and as easy as possible and automated where this is feasible. We need to communicate our aims and strategies to the next generation of researchers in a way that is useful to them and proves that RDM and digital preservation are a key part of their career progression and wider skills.

Acknowledgment

The author would like to thank Helen Glaves and Barbara Yarusso for their invaluable input and editorial advice.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

References

[1] Natural Environment Research Council. *Data Centres*. Available at: https://nerc.ukri.org/research/sites/data/ (Accessed: February 26, 2019).

[2] Natural Environment Research Council. *Data Policy.* Available at: https://nerc.ukri.org/research/sites/data/policy/ (Accessed: February 26, 2019).

[3] British Geological Survey. *BGS University Funding Initiative (BUFI).* Available at https://www.bgs.ac.uk/research/bufi/home.html (Accessed: February 26, 2019).

[4] Pinnick, J. (2017) 'Exploring Digital Preservation Requirements: A Case Study from the National Geoscience Data Centre (NGDC)'. *Records Management Journal* 27(2), pp.175-191. doi: 10.1108/RMJ-04-2017-0009

[5] FORCE11. *The FAIR Data Principles*. Available at: https://www.force11.org/group/fairgroup/fairprinciples (Accessed March 02, 2019).

[6] University of Cambridge Office of Scholarly Communication. *Unlocking Research.* Available at: https://unlockingresearch-blog.lib.cam.ac.uk/?p=1723 (Accessed February 26, 2019).

[7] RDM Training Benchmarking (2017). Available at: https://osf.io/pgnse/ (Accessed: May 31, 2019)

[8] British Geological Survey. *BGS Digital Preservation Policy.* Available at: https://www.bgs.ac.uk/downloads/start.cfm?id=3173 (Accessed February 26, 2019).

[9] CoreTrustSeal. *National Geoscience Data Centre Implementation of the CoreTrustSeal.* Available at: https://www.coretrustseal.org/wp-content/uploads/2018/01/National-Geoscience-Data-Centre.pdf (Accessed: February 26, 2019).

Building a User Group for a Digital Preservation Product

Lessons from 11 years of User Group History

Jonathan Tilbury
Preservica, UK
jonathan.tilbury@preservica.com

Remke Verdegem
Nationaal Archief, Netherlands
remke.verdegem@nationaalarchief.nl

Euan Cochrane
Yale University Library, United States of America
euan.cochrane@yale.edu

Abstract – Digital Preservation places a strong emphasis on building communities to share experiences and develop solutions. Traditionally these communities were built on geographic or functional alignment or created through external grant funded research activities. As commercial Digital Preservation products have emerged vendors have created User Groups for their customers that seek to fulfil this function. Using the lessons of the 11-year history of the Preservica User Group, this paper explores how these User Groups function and compares this to studies of User Groups in other domains.

Keywords – Community; COTS; User Group;

Conference Topics – Collaboration: a Necessity, an Opportunity or a Luxury; Building Capacity, Capability and Community

I. Introduction

The use of commercially supported products is becoming established as one of the main routes for organizations to build digital preservation programs. The products vary – for example some follow an escrow model and others are open source, and some charge a license fee in return for new product features whereas others rely on project based sponsored feature development. Whatever the business model, all of the current products seek to operate User Groups in different forms.

This paper is intended to enable the digital preservation community to learn from Preservica's [1] experience of operating User Group meetings for 11 years. The lessons are equally applicable to other products and groupings. The paper explores academic studies into this type of group and sees how their conclusions can be applied to the digital preservation world. It also looks at how these groups are likely to evolve in the future.

The authors include the Preservica founder and two long term customers of the system to ensure a balanced and fair view of the User Group is presented.

II. User Group Motivations

A. Contrast to Open Source Communities

Allen (2016) [2] describes how there has been significant research into how Open Source Communities grown and sustain. Whilst initial contribution is based on fulfilling a specific technology need, continued participation results from a personal identification with the ideology, the chance to grow specific skills, and the building of a professional reputation. This is explored more in Skinner (2018) [3].

The relationship with the User Groups for non-Open Source software products is driven by more complex motivations. Allen explores how the market economics in which a service is delivered for a fee contrasts with commons-based modalities of community exchange. Product User Groups are subject to both models and thus can be considered to be a hybrid-economic software community. This drives not only "hard" issues like licensing and support but also "soft" social contexts defining the rules and norms under which users participate.

B. Product Influence

There are many motivations to become involved in a User Group based around the roadmap of the product it supports. Lapham (2006) [4] sees it as key to sustaining software-intensive systems. Users can

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

get visibility of the product roadmap, work together to push vendors to deliver specific functionality, and get early warning of potentially problematic changes to the product or vendor.

C. *Knowledge Sharing*

Allen (2016) conducted an extensive study of the motivations and dynamics of product User Groups based on SharePoint communities in the US. With no chance to interact on the product roadmap, this showed that the principle drivers were as follows:

1. Learning and gaining access to knowledge, for example "free" consulting
2. Connecting to others, creating personal relationships
3. Commercial opportunities to gain access to services and make contacts that could lead to job opportunities
4. Improvement of personal reputation within the community
5. The duty to reciprocate support to others and to the vendor

Allen also observed that face to face meetings are more able to fulfil the higher priority needs than online communities and more likely to have active participants. Online communities tend to be dominated by a few leaders and contributors as observed by Nielsen (2006) [5].

D. *Vendor Motivations*

Vendors participate in, and in many cases fund, User Group to help them achieve specific business goals. These can be characterized as product, support and commercial benefits, and encourage the vendor to make considerable investments into the success of the meetings.

On the product side, it is critical for the product managers, designers and developers to understand the specific needs of their user community. As Zemke (1998) [6] described "If employees have not been taught how to identify customers and understand their expectations, it will be all but impossible for them to meet these expectations, far less exceed them". The ability of users of the technology and the people responsible for creating it to interact increases the chances of the system being able to fulfil the user's needs.

The involvement of the vendor's support team can be pragmatic, allowing users to resolve specific issues, and can be pre-emptive, allowing personal relationships to be built before problems occur.

Commercial motivations include the ability to sell further features to existing customers. The principle benefit however is to create motivated users who express their satisfaction to the wider community and are willing to act as references for future opportunities.

E. *Digital Preservation Perspective*

The digital preservation community has specific characteristics and context that drive community participation. Some of this is explored in Kwon (2006) [7] looking at collaboration in US State Government. This includes

1. Digital Preservation spans several roles including Archivists, Librarians, Records Managers and Information Technology. The lack of a shared language poses problems and leads to silos of information and battles to support agency "turf" hinder communication.
2. Despite this, practitioners "showed a strong willingness to gather together on a regular basis and network with one another" but more formal partnerships were required to secure ongoing participation.
3. As Digital Preservation is on the boundary between traditional communities it can become a community of practice in its own right.

Since Kwon (2006) the growth of member organizations such as The Digital Preservation Coalition [8] suggests the growth of Digital Preservation as a community of practice. Higgins (2017) [9] argues it is becoming its own profession rather than an add-on to archives, libraries and records management with its own disciplines and professional bodies. This will drive community participation and exchange of information.

III. Preservica User Group History

A. *Initial meeting (2008)*

The Preservica product started as the "Safety Deposit Box" developed by Tessella as a series of custom projects with an overlapping code base purchased via large government tenders. The

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

foundation User Group meeting in London, UK gathered three organizations together to compare their projects. Whilst interesting, this delivered little practical benefit as there was little overlapping functionality.

This meeting did however establish the willingness for customers to meet and discuss the Digital Preservation space with organizations using broadly similar technology. The discussions provided a template for the topics that would be discussed at later meetings.

B. Project funded development (2009-2014)

The Safety Deposit Box (SDB) was finally released as a supported product, but development was largely funded by new sales and requirements for specific features were funded by sponsored development.

The first full User Group meeting was in London in 2009. It was attended by 8 different organizations and the agenda covered the vision and roadmap, customer introductions, discussions on how to work as a community, an update on support, and more general discussions. Although not formally recorded, feedback was generally positive, especially for new users who wanted to learn about the system.

In the following years meetings were held in Bern, London, Vienna, Budapest and The Hague. The pattern of the meetings was broadly similar, with the addition of a feedback session to compare what was asked for at the previous meeting compared to what was delivered.

Time was added for users to have a private discussion without the vendor present on the first day to agree their priorities and to ensure they were able to raise their concerns effectively. This was then presented to the vendor team who responded on the second day. These sessions were useful in giving users the freedom to raise any issue they wished but their input was uneven and tended to feature more comments from certain expert individuals.

The challenge during this period was that as much of the development of the Safety Deposit Box technology was conducted as a result of requirements to fulfil new sales or by sponsored development, the development team had little control over the roadmap. Reporting back showed that many of the requirements were being fulfilled but this was not by a managed process.

By the time of the last meeting in this period there were 14 user organizations attending plus 3 partner organizations that provided services to go with the Preservica product. Topics had extended to include operations activities such as whether to trust the cloud and what is backup best practice, and governance of a shared linked data registry.

A series of changes during the period were drivers for changes in the User Group meetings. In 2015 Preservica became an independent company as it left Tessella Group. It also appointed a Product Manager and set up a formal Product Management process with a structured roadmap management system. Most significantly, in 2012 Preservica launched a cloud hosted preservation service with initial customers in the US followed by a small number in the UK. These customers tended to be smaller, to be paying significantly less and to have fewer resources to sponsor new features. The challenge for the User Group was to change to support these new types of user.

C. Oxford User Group (2015-)

From 2015 the International User Group meeting has taken place at one of the colleges in Oxford, a short distance from the Preservica offices. This allows more of the Preservica team to attend and interact with the users and allows UK users of the cloud hosted version of Preservica to attend without the need for air travel. Users from Europe and Australia have found Oxford to be easy to visit via the London transport hubs. The meetings take place over two days.

Over the four User Group meetings in Oxford there has been gradual formalization of the interactions between Preservica and the group. The main area this has changed is the way the roadmap features are presented and the gathering of feedback from users on their priorities. Initially this was conducted using post-it notes and voting but from April 2019 this will be conducted electronically to allow accurate and immediate feedback to be gathered. These are then reviewed next year to assess how many were delivered.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Another change is the user discussions forums. These have changed from a whole-group discussion into smaller sector-based groups who can discuss issues specific to their situation. These "birds of a feather" groups have allowed much more focused feedback to be generated.

During the meeting there is significant time in lunch, coffee breaks and at the evening social event for users to talk with other attendees to share information and build personal networks. This is one of the most valued aspects of the event.

The first Oxford meeting hosted 17 organizations, growing to 41 in 2018. Each year a customer survey is conducted that explores User Group satisfaction and suggestions for improvements. The most recent meeting in April 2018 had good satisfaction levels with 98% likely to recommend that their organization continues to attend. However, it did also contain several suggestions for improvements, including more workshops and more user project discussions.

It is worth observing that the feedback and roadmap voting is conducted by Preservica and presented back to the User Group. Whilst no one has queried this, there remains the option to move to independent assessment of the feedback in the future either by the community themselves or via a third party.

D. North American User Group (2014-)

The growth in the number of North America users of Preservica, especially the cloud edition, created the need to host a dedicated US User Group session. This has been run as a side meeting at the Society of American Archivists event. The initial meeting was an end of day presentation and social event attended by a small number of customers. After four years there was a call to run a full day meeting as a pre-conference event. This was first run in 2018 with 52 organizations attending.

The North American User Group has experimented with allowing users to attend online. This has had some success, with users able to vote for features using online polls. However, it has also presented significant technical challenges, especially as it is often hosted by a conference venue that struggles with this type of remote participation.

The topics covered at the North American User Group have been largely the same as at the Oxford event, covering new features, voting on the roadmap and user stories. Electronic voting on the roadmap was introduced in August 2018 and proved successful.

Feedback from the latest one-day event mirrored the Oxford meeting with 87% saying they were very likely to recommend attendance next year. As in Oxford the users asked for more time in future meetings on User Projects and Workshops.

E. Online user interactions

Preservica users are also able to contribute to the user portal, and online forum. There are also monthly webinars and the possibility to participate in special interest groups on specific product features. These interactions can be seen to follow the participation model observed by Neilson (2006).

IV. User Group Participation Observations

A. Type and range of active participation

Left unmoderated, participation at the talks of the User Group meetings will follow the observations of Neilson (2006) with input dominated by a few leaders, with some active contributors and a large number of passive participants. However, the organizers can intervene, encouraging specific users to present to ensure a wider range of contributors are heard.

The workshop sessions in smaller groups of around twenty participants also require careful moderation to encourage contribution from all attendees. These sessions are better at getting feedback for a wider range of participants but can be dominated by vocal contributors if left unmanaged.

The participation levels in networking time are much more evenly spread. It appears that all attendees participate in network building, maybe driven by the personal benefit they gain and the large number of attendees with highly aligned motivations.

B. Alignment with SharePoint studies

At the April 2019 meeting attendees were asked to score the observations of Allen (2016) out of 10

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

and the results were as follows:

Question	Allen Position	Attendee score
Find out about roadmap	N/A	9.1
Influence roadmap	N/A	8.9
Product learning	1	8.0
Connecting to others	2	8.3
Commercial opportunities	3	6.8
Personal reputation	4	5.0
Sharing my experience	5	6.9

Users want more roadmap information and influence, product knowledge and the opportunity to grow their personal networks but are less interested in sharing their own personal experience or gaining personal recognition. This shows very good alignment with alignment with the observations of Allen (2016) and is being used to guide future User Group priorities.

V. Future Directions

A. *User Participation in Organization*

The User Group has been organized by Preservica staff that also request and analyze the feedback. The involvements of user representatives in these activities would be welcomed by both Preservica and the users but requires volunteers to step forward. This was initiated in April 2019 with volunteers identified to help organize the 2020 meetings.

B. *Independent User Groups*

As product usage grows, totally independent User Groups are formed. These are often driven by geographical or functional groupings and vendor involvement can be minimal or absent. Preservica has already seen such groupings in New England and BENELUX. These meet annually to compare experiences and align their input to the User Group.

The development of such independent groups is expected to grow and may follow the more formal pattern of the Independent Oracle User Group observed by Malcher (2016) [10]. As these groups grow, they require more formal governance, active leadership, and variety of volunteers. They can be delivered online or face to face and can facilitate a wide variety of interactions, for example conferences, workshops, social events, lunches and other meet ups. Membership can be individual or corporate and may be free or paid, as can the events.

One of the main strengths of these groups is that they can provide a united voice to the vendor. This is beneficial for the users as their voice has more strength and beneficial to the vendor as the input is analyzed and prioritized.

C. *Tiered User Groups*

Currently the User Groups in Oxford and the US are paid for by Preservica, reflecting the level of subscriptions provided by all organizations that use the system. It is possible that in the future lower price subscriptions are offered for a lower specification system without free User Group participation.

As the user base becomes more geographically dispersed it is likely to be more necessary to add more remote participation via video streams and remote voting. This will also fulfil the needs of customers with restricted travel budgets.

Lastly, as users outside of cultural heritage and academic sectors start to use Digital Preservation, it is likely they will not want to invest the time in attending a User Group for something that is less core to their mission. It may be that online participation via forums is sufficient for this community.

D. *Cross-product groups*

The Preservation Action Registries initiative described at iPres 2018 by Addis et al (2018) [11] shows how vendors may also cooperate with their competitors, exchanging information in order to better support their user communities. As users become involved in these activities, cross product user communities will establish themselves. The governance of these communities presents an

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

interesting challenge as participation grows that could be informed by Preservica's User Group experience.

VI. Conclusions

Digital Preservation Product User Groups offer an valuable addition to established user communities. They can help users benefit from the technology more effectively and create an effective dialogue between users and vendors. The patterns they follow are common with other software industries.

The Preservica User Community has matured over many years but still has areas it can improve, specifically increasing the opportunity for knowledge sharing and increasing user involvement with its execution. In the future new approaches will have to be explored as the characteristics of the user community change.

References

[1] Preservica www.preservica.com

[2] Allen, W.S. "Participation and Knowledge Exchange in a Hybrid-Economic Software Community" J Knowl Econ (2018) 9: 767. https://doi.org/10.1007/s13132-016-0366-8

[3] Skinner, Katharine "Community Cultivation, A Field Guide", Educopia https://educopia.org/wp-content/uploads/2018/11/CommunityCultivationFieldGuide.pdf

[4] Lapham. Mary Ann, and Woody. Carol, "Sustaining Software-Intensive Systems," Software Engineering Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, Technical Note CMU/SEI-2006-TN-007, 2006. http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=7865

[5] Nielsen, J. (2006). Participation inequality: the 90-9-1 rule for social features. Nielsen Norman Group. https://www.nngroup.com/articles/participation-inequality/. Accessed 8 Mar 2019

[6] Zemke, Ron "Best Practices for Customer Service" HRD Press (1998) ISBN 978-0874254433

[7] Kwon, Hyuckbin, Pardo, Theresa A. and Burke, G. Brian "Digital Preservation Community: Lessons on Interorganizational Collaboration", Proceedings of the 7th Annual International Conference on Digital Government Research, DG.O 2006, San Diego, California, USA, May 21-24, 2006

[8] Digital Preservation Coalition www.dpconline.org

[9] Higgins, Sarah "Time to become our own profession?", https://www.dpconline.org/blog/time-to-become-our-own-profession

[10] Malcher, Michelle "User Group Leadership", Apress 2016, http://doi.org/10.1007/978-1-4842-1115-1

[11] Digital Preservation Interoperability through Preservation Action Registries https://figshare.com/articles/Digital_Preservation_Interoperability_through_Preservation_Actions_Registries/6628418

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Building Sustainable Digital Preservation

A Typology of Digital Collections

Helen Hockx-Yu
University of Notre Dame, USA
hyu3@nd.edu

Donald Brower
University of Notre Dame, USA
dbrower@nd.edu
https://orcid.org/0000-0001-7347-546X

Abstract – The digital preservation function in an academic research library is chiefly concerned with the ongoing access to its digital collections. This paper describes the work in developing a typology of digital collections at Hesburgh Libraries, University of Notre Dame, which serves as a framework for planning and building a sustainable Digital Preservation program.

Keywords – Libraries, Digital Collections, Digital Preservation, Typology, Framework

Conference Topics – Designing and Delivering Sustainable Digital Preservation

I. Introduction

The Hesburgh Library at the University of Notre Dame has a digitization program and an institutional repository with bit-level preservation commitment. [1] Yet, some of library's digital collections fall outside the scope of these. We desired a holistic digital preservation strategy mindful of our complete holdings.

The Digital Collections Typology Project took place between April and December 2018. The project developed a Typology of Digital Collections that can be used as a framework to plan and guide digital preservation.

The nature of libraries' collections has evolved over the last twenty years or so, from physically owned and locally stored collections to what Dempsey coined "facilitated collections", which include a broad range of local, external and collaborative resources organized around user needs. [2] As many academic and research libraries, Collections of the Hesburgh Library spread across the spectrum as illustrated in the diagram below. Each collection in the spectrum may require a different approach to long term preservation.

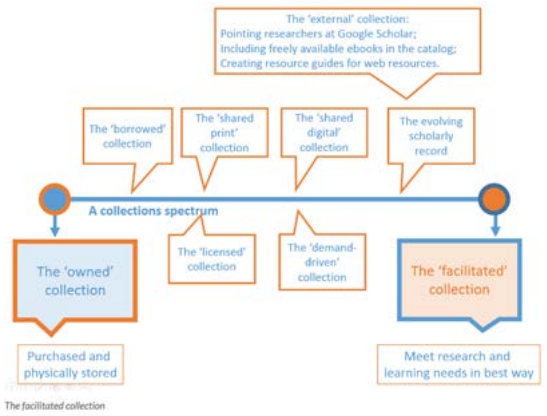


Figure 1 Collections Spectrum. [3]

Hesburgh Libraries' Digital Collections can be in any of the formats below:

- Entire collections in digital format
- Items in digital format that are a part of a collection which also has analog items
- Stand-alone items in digital format
- Digital surrogates of analog collection items
- Digitally encoded content on physical media

The scope of the project also extends to physical items that require digitization to preserve access, for example VHS tape and audio cassettes (which are both analog encoded content on physical media).

II. Relevant works

While Dempsey's Collections Spectrum provided a useful context for understanding the evolution of collections, we needed something more detailed to approach digital preservation holistically based on common characteristics.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The field of digital preservation and curation has developed a number of additional frameworks that provide useful context:

- The DCC Lifecycle, an iterative, high-level overview of the stages required for successful curation and preservation of data. [4] While our work focuses on data (collections) and preservation action, the DCC model provides the"big picture"and reminds us that actions are required across all stages of the lifecycle, and cannot only be limited to the stage of "preservation action."
- The generic, appraisal/selection framework for digital curation by Jinfang Niu. [5]. Although our work mainly concerns items that have already been appraised, selected and ingested as part of our collections, the framework still offers useful criteria (i.e. mission alignment, value and cost) that we can apply in future to decide which type of collection items should be prioritized for preservation.
- The SPOT model for risk assessment. [6] identifies threats to digital preservation. This is relevant to our work and the typology of preservation threat can be mapped to our typology of digital collections to help with prioritization.
- The Digital Content Review Process developed by the Digital Preservation Management Workshop [7]. Our work shares the same goals but is much less comprehensive as it is intended as a starting point to understand the high-level, broad scope of digital preservation at the Hesburgh Library. We expect to include many of the useful details and even adopt parts the framework when carrying out the next level of assessment and prioritization.

We also considered inventory management, a common method that libraries use to assess, examine and track the condition of their collections. This would be a valid starting point for digital preservation, too, but over a certain size inventories become hard to work with. Moreover, we have a few special collections where just performing an inventory would be a significant undertaking. We therefore deferred inventory to the next stage and decided to start with something more general.

III. The typology

A typology is a general classification of items that provides a structure for understanding items by highlighting the properties either shared or not shared between them. A typology can be contrasted with an inventory where for any sizable number of items an inventory becomes unwieldy and is not useful for understanding the items as a whole. We see typology as being in the middle ground of detail between the conceptual Collections Spectrum and an inventory.

Our focus was preservation, so the typology was organized to assist with the preservation (and non-preservation) of digital content. The attributes were chosen to be those that are significant to deciding on preservation actions.

The types are each defined by unique attributes. Some attributes relate to the control and access of the types, such as whether an item is managed by the Library or vendors; whether an item is a physical object, e.g. a VHS tape; or whether an item is digitized, i.e. a digital surrogate of a physical item. In case a digital asset is the result of digitization then we would need to maintain the knowledge of the link between the physical item and the digital surrogate.

Our typology identified three broad types for the Library's digital collections, listed below. These are high-level and likely to be applicable to other institutions. More detailed types are expected to vary between institutions and lead to different preservation approaches, depending on resources and constraints.

Vended collections are resources that the Library is given permission to use for a limited purpose or timeframe. Access to vended collections by patrons generally takes place online at platforms provided by copyright holders or licensors. Examples are electronic journals, books and databases. Unlike physical purchases, vended content may "disappear" from the Library's collections once the license is contractually terminated.

Library-managed collections are resources the Library chooses to or is obliged to exercise stewardship over. This content is mostly kept on university

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

operated equipment, but occasionally external services are used to host content and provide access. This category contains two subtypes: digital surrogates (or digitized) and born-digital. Digital surrogates are the electronic captures of physical items in our collections. "Born-digital"are items that do not have a corresponding physical item in the collection.

"Physical media" is a "related type", which includes physical items that have digital information or that we expect to be digitized, e.g. computer disks, CDs, VHS tapes, and audio recordings.

In-house digital creation are resources produced by the Library for various purposes that are not collection items, including marketing and instructional material, materials related to contests and student awards, source code, websites, blogs, access copies, LibGuides, etc.

IV. Methodology

To gain an understanding on the digital material the Library currently deals with, the project team interviewed 37 faculty and staff members. The majority were subject librarians and staff who work directly with digital content. The interviews were free-form, but there was a written prompt to help guide it.

The questions were over the kinds of digital content interviewees either handled, purchased, or had in their collections. If interviewees knew any item counts or storage sizes, we recorded that as well. The answers pointed not only to large collections hidden-in-plain-sight, such as nearly 8000 VHS tapes, and more than 15k DVDs, but also to unique one-offs, and comments on processes and policies related to digital content. The interviewees were often unable to provide exact counts or indicate where to obtain the numbers. This confirmed our assumption that more detailed inventorization is required, allowing us to plan and prioritize the collection items that are in scope for preservation.

From the interview notes, the team then compiled a list of every content or file type mentioned. Since the content names were transcribed directly from the interview notes, there were many variants of the same name, so the content names were normalized and then grouped into similar kinds. The team reviewed the resulting list, extracted and organized attributes based on how they aligned with preservation decisions, and developed recommended preservation actions for each type.

V. Using the typology

The main application of the typology is the definition of the scope of digital preservation for the Hesburgh Library, followed by a broad approach for each type, that either builds on existing workflows or points to new work. By having a comprehensive typology, we could make intentional preservation decisions for all of our content.

It became clear that not every type or all instances of a type need to or can be preserved. The Library's digital preservation program should focus on the long term stewardship of collections managed by the library, as well as digital surrogates produced from physical media. Digital surrogates produced for access, for example images from a book scanned for use on a poster, are outside the scope of long-term preservation.

Similarly, instances of In-house digital creation are generally not intended for long term preservation. They may have significant short term value or be required to support the Library's operations, so need to be kept safely, as with any other operational data. Some instances within this type may become a part of the **Library-managed collections**, for example Doctoral dissertation or Master's thesis. Formal process should be followed to accept the material which conceptually moves the content from one type to another. Some instances may be considered University Records, and should be dealt with in accordance with the University's Records Management and Archives Policy.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

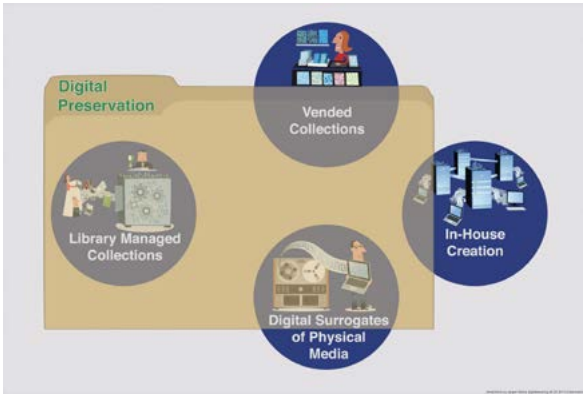


Figure 2 Typology and Scope of Digital Preservation.

For Vended collections, there is relatively little that the Library can do to preserve the content pro-actively. One area where the Library can exercise some influence is in the license negotiation process. We therefore recommended to continue the current practice of asking explicit questions about continued access, the possibility for archiving, and the content providers' preservation commitment. We also expressed a word of caution with regard to local archival copies, which do not guarantee perpetual access. They become the Library's responsibility and should be treated as part of Library-managed collections.

Library-managed collections are the focus of digital preservation. We recommended short term projects, as well as longer term, more strategic work:

1. Develop guidelines and tools to help curators determine where to store and expose digital content - this is essentially an communications effort, intended to make sure that the available storage and access options are well understood.

2. Organize immediate effort to mitigate the preservation risks related to obsolete physical media items in Rare Books and Special Collections, digitizing and moving these to stable and managed storage.

3. Utilize existing infrastructure including the institutional repository and campus archival storage service as (interim) preservation storage for Library-managed digital collections. The goal is to accommodate as many of our collection items as possible, especially those that for various reasons are not benefiting the current digital preservation care in place.

4. Devise a process (supported by the necessary tools and systems) that systematically identifies Library-managed collection, and maintains an ongoing overview. Such overview of our holdings is essential to digital preservation: you simply cannot protect your data if you don't know your data.

V. Discussion and Conclusion

Overall, our typology project has been very useful. Constructing the typology gave us an opportunity to consider our collections holistically and record the various places content had been stashed over time.

The typology has created a common terminology between different groups in the Library facilitating discussion and planning. The typology also raised awareness of the specific characteristics leading to the different approaches required to preserve our digital collections.

A caveat is that our typology provides a content-centric view. The broad approaches recommended based on the typology are action-oriented and should not be seen as a digital preservation strategy. Our work will certainly inform the development of such a strategy, but it does not take into account the organizational component, addressing issues such as staffing and training needs, and financial requirements. In this sense, the typology is only a starting point, eliminating what is irrelevant at a high-level and pointing us to a way forward.

While our focus was the big picture, interviewees did mention tools and systems that they used to handle digital collections. In hindsight we should have been more systematic with collecting this data and aggregating it so that we would have a more detailed list of place to examine for follow-up inventories and analyses.

The typology is a high-level framework that helped us describe and categorize the Library's collections with enough details to understand the broad patterns yet prevented us from not seeing the forest for trees, too early in the process. These patterns were used to define the scope of broad

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

approaches for digital preservation, and allowed us to arrive at the priority areas rapidly.

Acknowledgement

The authors are grateful to Patrick Rader for creating Figure 2 and to all other project members and interviewees for their contributions.

References

[1] "CurateND". https://curate.nd.edu/.

[2] [2] L. Dempsey. "Library collections in the life of the user: two directions." https://www.liberquarterly.eu/articles/10.18352/lq.10170/.

[3] [3] Ibid.

[4] [4] "DCC Curation Lifecycle". http://www.dcc.ac.uk/resources/curation-lifecycle-model.

[5] [5] J. Liu. "Appraisal and Selection for Digital Curation." *International Journal of Digital Curation*. VOL 9 NO 2 (2014). https://doi.org/10.2218/ijdc.v9i2.272.

[6] [6] S. Vermaaten, B. Lavoie and P. Caplan. "Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment." *D-Lib Magazine*. Volume 18, Number 9/10 (2012). http://mirror.dlib.org/dlib/september12/vermaaten/09vermaaten.html.

[7] [7] Digital Preservation Management. "Digital Content Review: Process and Results." https://dpworkshop.org/workshops/management-tools/process-results.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Development and improvement of image processing scheme for archiving inscription

Hideyuki Uesugi
National Institute of Japanese Literature
Japan
uesugihd@me.com
https://orcid.org/0000-0002-0787-3723

Masayuki Uesugi
Japan Synchrotron Radiation Research Institute
Japan
uesugi@spring8.or.jp
https://orcid.org/0000-0001-6261-9034

Teruhiro Tani
Wakasa-ji Culture Research Institute
Japan
tkatani@icloud.com

Abstract – This paper shows improvement of image processing scheme for digitally archiving of inscription on stone monuments. The scheme succeeded in speeding up the image processing, and also improved the readability of the inscription text. The method developed in this study is important for the future development of automatic recognition of characters in inscription from photographs. Keywords – epigraph, inscription, image processing,

digital archive, citizen science,

Conference Topics – Designing and Delivering Sustainable Digital Preservation,

I. Introduction

Numerous stone monuments have been erected across Japan over 1,400 years, since the seventh century. Although it is estimated that the total number of monuments exceeds one million, the actual number remains unknown. Recently, the importance of these stone monuments has been recognized widely. For example, by heeding the warning engraved on a stone monument, "Do not build a house lower than this altitude," indicating where a tsunami had reached in the past, people in that area could escape from the tsunami caused by the Great East Japan Earthquake. The Tohoku Regional Bureau of the Ministry of Land, Infrastructure and Transport released an archive titled, "Tsunami damage and tsunami stone monument information archive" to the public [1]. In fact, there are many stone monuments in the areas affected by the Tokai and Nankai earthquakes in the past, where several earthquakes of magnitudes exceeding that of the Great East Japan Earthquake have occurred. However, most stone monuments exist outdoors and have deteriorated because of weathering. Furthermore, accurate information is no longer available if the stone monuments themselves were lost because of urbanization or road work. Therefore, it is urgently necessary to archive as many stone monuments as possible.

It is difficult to decipher weathered inscription from ordinary photographs. Their surfaces are often covered by moss, and characters on the surface are unclear (Fig. 1). Owing to the ambiguity of the photo images of inscription, researchers have difficulty in the data from photographs of stone monuments. The accuracy of reading the text on the stone monuments from photo images largely depends on the skill of the investigator and field conditions, weather, direction of sunlight, and accessibility to the samples. The quality of the images taken under these different conditions changes largely, making it also decrease the quality of the archive.

In recent year, Software and methods have also been developed to obtain the three-dimensional shape of stone monuments by using Photometric stereo, such as RTI [2]. Furthermore, methods that can restore the shape of stone monuments based on the motion of the camera have dramatically advanced, and it is possible to automatically restore the three dimensional shape from a large number

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

of digital photographs [3-4]. However, it was difficult to apply those methods to the fieldwork of the huge number of stone monuments because of time-consuming of image processing.

Figure 1 Sample images of unclear inscriptions.

We are currently developing a new method of digital archiving of inscriptions [5-6]. Our method emphasizes three features, “Speedy”, “Simple”, and “Lightweight” for the fieldwork. In Japan, the investigation of stone monuments is normally undertaken by retired people, who are instructed by the museum curator. For this reason, the photography step should be simple, to facilitate public participation. Another reason is that condition of stone monuments is diverse. Photographic instruments should be made lightweight to make it possible to visit several places with diverse conditions, as shown in the figures. For this reason, we make the photography step as simple as possible.

II. Method

In our image processing scheme, we require shadow images of the whole text on the stone monument illuminated by oblique light. A non- shadow image of the surface of the stone monument, called as the background image, is also required. After the acquisition of these images and registration of archive data in our system, the system applies basic image processing to all the images acquired through field work, as a general procedure to prepare for advanced image processing, which aims to automatically extract characters according to a programmed procedure.

The “basic image processing” scheme developed in this study consists of the following steps:

(1) monochromatizing all shadow and non-shadow images into 8-bit grayscale,
(2) subtracting the background image (non-shadow image) from all shadow images to remove non-shadow contrast from the image and to enhance the shadow of characters (Fig.2-(a)),
(3) applying a Gaussian filter to (2) to create a mask image of the area illuminated by oblique light, by blurring the light (Fig.2-(b)),
(4) subtracting (2) from (3) to enhance the shadow of the character and mask the area illuminated by oblique light in (2),
(5) adjusting and unifying the brightness and contrast of images from (4) using the upper and lower limits of gray value derived from the histogram of each image.

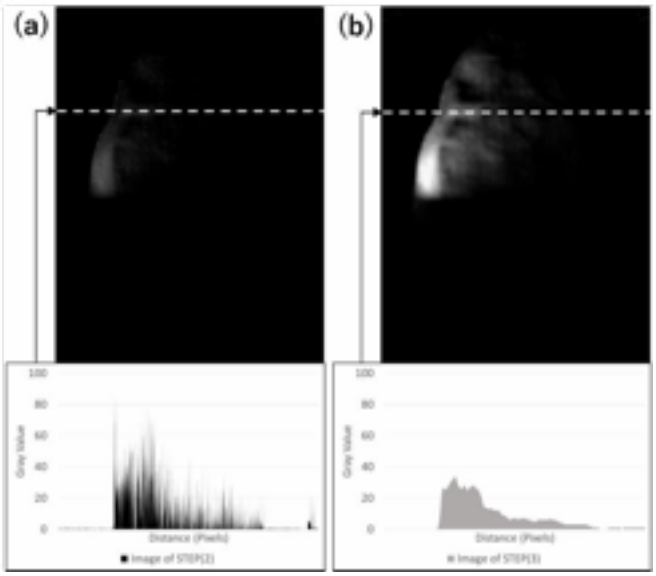


Figure 2 (a): Image of step (2) and a line profile of gray value along the dashed line, (b): Gaussian Blur image of step (3).

Considering steps (3) and (4), there are two purposes of the image processing. First, we aim to enhance the shadow of the character on the stone monument. Second, we aim to mask the image only for the area illuminated by oblique light. Between steps (3) and (4), the contrast of the image was inverted, and a negative contrast image was obtained. If the gray value of a pixel becomes negative during subtraction, we set the gray value to 0.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

In Fig. 3, we can clearly see that the contrast of the shade (i.e., contrast of the text) is enhanced compared to surrounding noise. In addition, the intensity of the area outside the illumination by the oblique light is still low. Originally, the "shadow" of the engraved character on the surface of the monuments and the luminance of the "shade" part outside the irradiation area had similar gray values. However, by reversing the gray value of the shadow by this method, "shade" becomes having much lower gray values than "shadow", and we can create an image in which only the shadow of the character in the irradiation area is present (Fig.3-4). The advantage of this method is that even if we cannot irradiate the entire area of the inscription at once because of larger text area (ROI, Region of Interest) than irradiation area, all of the inscription can be extracted by photographing the stone monuments with changing the irradiation area and combining all images after the acquisition (Fig.5).

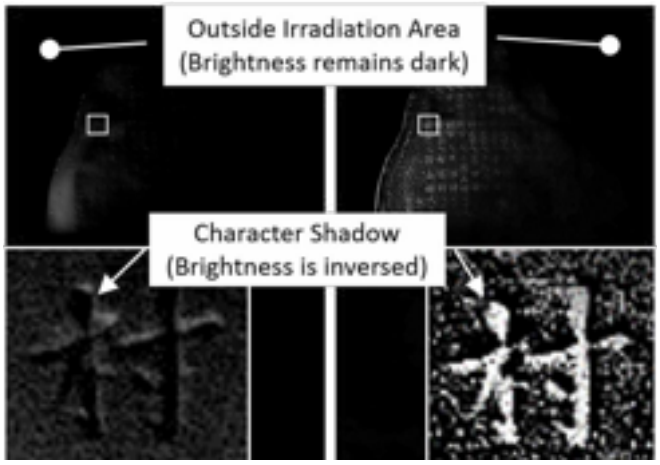


Figure 3 Image in which only the shadow of characters in the irradiation area remains high gray value

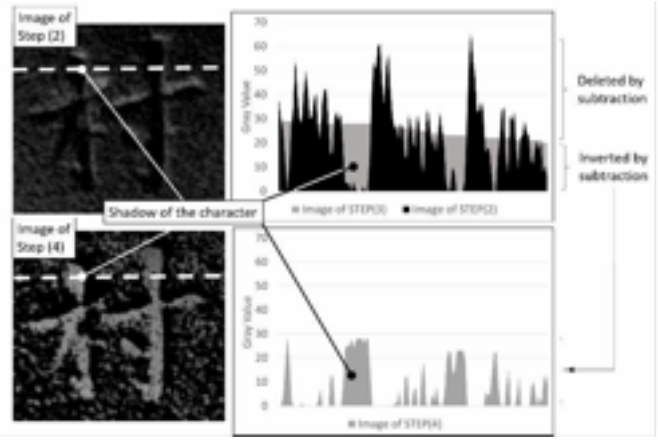


Figure 4 Comparison of images of step (2) (upper left) and step (4) (lower left) around a character. Line profiles along the dashed line are also shown on the right-hand side of the figure.

We had applied a low-pass filter at step (3) in previous studies, but used Gaussian Blur in this study. As a result, we succeeded almost the same result in significantly speeding up the image processing while obtaining the same result (Table I).

Figure 5 Result of Fig.1

TABLE I

Processing speed comparison

	Number of Image	Image Size	File Format	Processing Time	
				Low-Pass	Gaussian
Sample1 (Fig.3-6)	30	60.9MB	tiff	06'06''67	00'50''68
Sample2	6	60.9MB	tiff	01'31''45	00'10''80
Sample3	27	5.0MB	jpg	03'27''48	01'11''24
Sample4	13	5.8MB	jpg	01'47''73	00'31''01

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

III. Results

Although we succeeded in increasing the inscription readability significantly, the data are still insufficient for the automatic recognition of the characters on the surface of stone monuments, because of low signal-to-noise ratio of the image in cases of Fig.6.

Figure 6 Original image (left) and combining all images (Right)

Therefore, we applied cluster labeling method to reduce noise level [7]. Cluster labeling is a method in which all adjacent of a pixel those having a specified range of pixel values are regarded as one chunk, i.e. a cluster. After the recognition of all clusters on the image, gray values of the pixels in a cluster were changed according to the cluster size. Pixels in the largest cluster have gray value 1, and pixels in the second largest cluster have gray value 2. Because gray value of all clusters shows the size of clusters, we can remove small clusters (showing noise) by thresholding the image and extract only large clusters (showing characters) easily (Fig. 7).

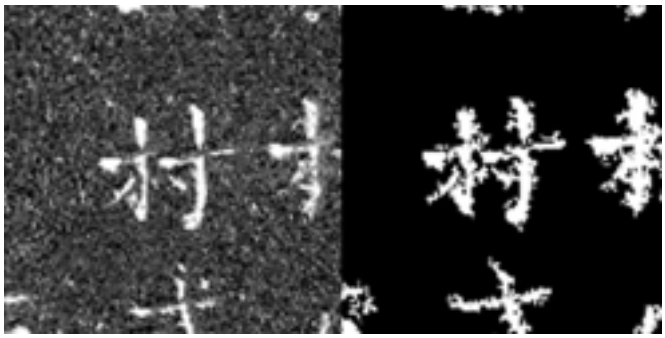


Figure 7 Result of cluster labeling

However, these noise reduction steps cannot be automatically processed with the programmed flow. In future work, we may be able to develop automatic processing of this step by accumulating the result of image processing along with the important parameters of the samples, such as the base material of the stone monument, shape, and tendency in a region, in a database.

IV. Discussion

This report mainly focuses on inscription extraction technology. In future work, we will evaluate the improvement for the efficiency of character recognition quantitatively, by calculating signal to noise ratio and doing an experiment of automatic recognition of inscription by machine learning content using Dataset of PMJT Character Shapes [8]. We will also develop a description schema for inscription with EpiDoc [9]. Furthermore, since this method can acquire images for multiple samples at once, we can obtain a large number of data with considerably small processing time. Figure 10 shows a demonstration for the application of our method to multiple coins those having much faint irregularity compared to inscriptions on stone monuments. Thus, the method can applicable not only inscriptions, but also divers samples of the historical and Archaeological materials, to digitize their surface in speedy.

Figure 8 Result of Coins

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Acknowledgment

HU was supported by JSPS KAKENHI Grant Numbers 25730201 and 16K21522. MU was supported by JSPS KAKENHI Grant Number 15H03755.

References

[1] Tsunami Damage • Tsunami Stone Information Database, Tohoku Regional Bureau of the Ministry of Land, http://www.thr.mlit.go.jp/road/sekihijouhou/archive/top.pdf

[2] G. Palma, M. Corsini, P. Cignoni, R. Scopigno, and M. Mudge, "Dynamic Shading Enhancement for Reflectance Transformation Imaging," Journal on Computing and Cultural Heritage (JOCCH), 3 (2), September 2010.

[3] A. Koutsoudis, B. Vidmar, G. Ioannakis, F. Arnaoutoglou, G.Pavlis, C. Chamzas : Multi-image 3D reconstruction data evaluation, Journal of Cultural Heritage, 15 (1), pp. 73-79, 2014.

[4] A. Koutsoudis, B. Vidmar, G. Ioannakis, F. Arnaoutoglou, G.Pavlis, C. Chamzas : Development and Disintegration of Maya Political Systems in Response to Climate Change, Science, 09 Nov 2012 338, pp. 788-791, 2012.

[5] H. Uesugi, and M. Uesugi, "Development of the Database for Images of the Text on the Stone Monuments," Culture and Computing (Culture Computing) 2013 International Conference on, pp. 149-150.

[6] H. Uesugi, and M. Uesugi, "Stone Monument Text Image Database," Theory and Practice in Modern Computing, 2015, pp. 203-207.

[7] "Slice" -Softwares for basic 3-D analysis-, Slice Home Page, Japan Synchrotron Radiation Research Institute. http://www-bl20.spring8.or.jp/slice/

[8] Dataset of PMJT Character Shapes, http://codh.rois.ac.jp/char- shape/.

[9] S. Orlandi, R. Santucci, F. Mambrini, and PM. Liuzzo, "Digital and Traditional Epigraphy in Context", Proceedings of the EAGLE 2016 International Conference, 2017.

Early Exit Strategies in Digital Preservation

Ashley Adair
University of Texas Libraries,
University of Texas at Austin
United States
a.adair@austin.utexas.edu

Maria Esteva
Texas Advanced Computing Center,
University of Texas at Austin
United States
maria@tacc.utexas.edu

Benn Chang
University of Texas Libraries,
University of Texas at Austin
United States
changtb@austin.utexas.edu

Abstract – Digital preservation is a continuous activity requiring long-term effort, the lack of which presents risks for data falling behind in maintenance, representation, functionalities, and long-term safeguarding. However, contingencies in a preservation pathway can change quickly. Going to the rescue of data at preservation risk requires potentially costly and time consuming strategies. The ability to respond successfully is enhanced by planning an exit strategy for the data. We present two scenarios enacted in response to the closure of a distributed data preservation initiative and stress the importance of a prior "plan B" to digital preservation plans.

Keywords – exit strategy, at-risk data, distributed digital preservation

Conference Topics – Designing and Delivering Sustainable Digital Preservation; Collaboration: a Necessity, an Opportunity or a Luxury?

I. Introduction

Because digital preservation efforts exist on extended time scales, conditions surrounding their context are bound to change. The ongoing nature of digital preservation has been extensively stressed. Administrative tools such as cost sustainability calculators [1] [2] and decision-making matrices [3], and technical approaches such as auditing [4], migration [5] [6], and virtualization or emulation [7] [8], allow institutions to select and maintain a preservation pathway. However, when conditions for preservation change, the pathway is disrupted. Responding to data at risk requires implementing another set of measures, often developed on the spot. Depending on the context and status of the data, and on the possibilities of the institutions that support them, the approaches may entail significant challenges, particularly if not considered and codified in advance.

In this paper, we discuss two different scenarios enacted due to the abrupt closure of a large distributed data preservation initiative [9]. While our approach to depositing two sets of data in this network included several strategies that supported exit efforts, failure to outline a comprehensive early exit strategy in each case led to extra effort and decision-making following news of the closure. Based on this experience we identify what worked, what could have been improved, and provide recommendations.

II. Depositing Data Into a Distributed Preservation Initiative

The case study we present concerns a large distributed digital preservation initiative that opened in 2016. It was comprised of nodes at academic institutions geographically dispersed throughout the United States, each using a different storage architecture. Members of the initiative bought a data allocation for deposit in the network. They worked with an ingest node, which used a centralized suite of tools to deposit data and replicate it to additional nodes for long term storage. The transfer mechanism for

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

the initiative was BagIt [10], a widely adopted specification for grouping files in a standardized directory structure (a "bag") and attaching "tag files," plain text files containing descriptive and administrative metadata, a file manifest, information about the version of the bagging tool used, and checksums for each file in the bag.

UT Austin served as a network node, receiving content from other member institutions for storage at the University of Texas at Austin's Texas Advanced Computing Center (TACC) via ingest tool implementation and hosting by the Texas Digital Library [11]. Data deposited by UT Austin in the initiative would be copied to TACC and two additional geographically dispersed nodes.

III. Data Deposit: DesignSafe-CI

In 2015 DesignSafe, hosted at TACC, become the awardee of a National Science Foundation cyberinfrastructure (CI) grant to build an end-to-end data management and analysis portal for natural hazards engineering [12]. The grant required taking custody of data from the previous iteration of the project, which had been hosted at two other institutions for more than a decade [16]. The legacy data, composed of ~2000 datasets and their metadata, were migrated into the new web-based portal for distribution and access. While the metadata for each dataset in the collection followed a logical model, it was not translated into a standard schema. The new system involves a second copy of the data on a geographically replicated file system.

In late 2016, we began preparing this legacy data for ingest into the distributed digital preservation initiative. The goal was to explore a long term preservation proof of concept by creating a subset of static data and its metadata as a third dark archival copy. The cyclical nature of funding for the CI meant that special care had to be taken to make the data and knowledge of it and its whereabouts portable, anticipating when the next host institution would take custody in 5-10 years.

To prepare for deposit, the data were grouped per data publication (research project) and packaged according to the BagIt specification. When possible we enclosed each project in one bag according to the 200 GB limit for the distributed initiative's ingest tool. For projects over 200 GB we enclosed data in sequenced bags. In each bag we also placed descriptive metadata, which was scraped from the legacy site interface. Multiple attempts to recover the metadata directly from the legacy database were unsuccessful. An oversight on our part was not pursuing extracting the metadata from the new system as a JSON file.

Due to the expected changes data ownership, we needed an identifier system to track the preservation network data packages over time. Each bag was given an ARK identifier [13] through a global identifier service before deposit. The ARK pointed to the new location of the dataset so that information about the project was maintained. Using this strategy, upon changes in data stewardship, the identifiers could be updated to show new custody.

For our own recordkeeping, and to provide future custodians information about the preservation network packages, we created metadata packages for each bag to retain locally. We stored a copy of each bag's tag files and copies of the network's ingest and replication tool reports in a directory named according to bag identifiers. We placed a copy of these within the cyberinfrastructure for transmission to future awardees.

IV. Data Deposit: UT Libraries

At the same time, the UT Libraries were preparing their own data for ingest into the network. These were archival master TIFF images of content digitized from library collections, primarily representing items such as rare books, University theses and dissertations, maps, and government reports.

Copies of the flies were stored in bags in the Libraries' LTO tape archive, largely organized only in relation to their date of creation, and without descriptive and in some cases technical metadata. The online projects arising from these digitization efforts feature descriptive metadata for the files, but asynchronous legacy workflows meant that metadata were not ready for vaulting at the time that files needed to move to tape to free processing space on disk.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Because purchasing storage in the distributed digital preservation initiative represented a significant cost to our organization, we wanted to prepare our data to a higher degree of preservation quality for ingest than we had been storing it locally. To prepare, we restored a copy from tape, reorganized files in logical content units, generated FITS technical metadata [14], and re-bagged, making use of bag-info.txt files to add basic descriptive metadata for each package. This metadata came from various sources, such as project web portals, digitization records, and in some cases institutional memory.

Bags were ingested into the network in the same manner as the natural hazards legacy data, with bag tag files and ingest reports retained locally. The initiative marketed very long data retention goals, meaning that staff creating these initial ingest bags could be retired by the end of the service terms. This reality stressed the importance of local recordkeeping regarding our deposits that could be persisted in our organization over time. Notably, the enhanced data packages were not re-written to tape locally, since we assumed they would be preserved in the distributed network and the data were sizable by our local storage standards. The content file-only bags were retained as originally stored.

V. Exiting the network: DesignSafe-CI

In early 2019 the distributed digital preservation initiative announced that it would shutter. Because we had no formalized exit strategy to turn to, quick action was needed to decide the disposition of the data stored within it.

We first investigated which network nodes received copies of our data and began conversations with staff there to determine options. In the end, we found that full copies of all UT Austin data, both DesignSafe's and UT Libraries', had been replicated to a file system at the TACC network node. Because we are campus partners with an existing collaborative relationship, this offered us some time and flexibility to move forward.

With the DesignSafe data, we initiated testing on the CI to ensure that the data we placed in the network had been effectively ported to the new CI for access. We searched the cyberinfrastructure for legacy project numbers that we had embedded in the network bag identifiers and found that all were present. Because the data was ported and includes the geographically replicated copy, we decided not to recall the copies that were at the other three national nodes. These copies will be deleted. If we decide to make a third copy of the data, it can be sent to TACC's tape archive.

A simultaneous development was our university's adoption of a new global identifier service that does not support ARKs. With this change, the DesignSafe preservation bag ARKs were decommissioned. We did not anticipate at the time of creating the ARKs, which were central to our preservation plan, that this service would be disrupted. Had the distributed initiative continued we would have needed a new strategy for identifiers, illustrating how many preservation services and systems can change in a short period of time within one preservation pathway. Risks for each dependency in a plan, especially regarding services and systems outside of one's immediate control, should be taken into account at the outset. Risk management is not well represented in current digital preservation literature but would be a fruitful area for future work [15] [16] [17].

VI. Exiting the network: UT Libraries

UT Libraries' data took another path. Since we knew that the deposited data packages were superior to our local copies, we wanted to retrieve them. We first collected bag identifiers applied by the Libraries while preparing the data for ingest, using a client that was part of the technology stack of the distributed network. Interacting with TACC storage node was via iRODS iCommands, an open source data management software [18]. After copying the data to local storage, a post copy verification computed SHA2 values on both ends for comparison. Each copied tarball was then extracted and had bagit-python validation run. Since the ingests into the distributed initiative were an early proof of concept using new technology, this time consuming validation assured us that the bag contents were an exact match to what had been originally placed into the network.

The UT Libraries are now exploring alternative options for storage duplication. For the time being,

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

we write two copies of all data for preservation to tape, with one being stored in an off-site vendor facility.

VII. Conclusions

In each of these cases, staff at TACC and the UT Libraries worked together to expend considerable effort strategizing an approach to preservation packages for ingest into the distributed digital preservation initiative, along with even more time and effort spent actually creating the data packages. We then meticulously tracked and recorded ingests of the packages into the network. We did not, however, spend enough time creating a plan that could be enacted quickly if the network failed or we needed to leave it for our own reasons.

In the case of DesignSafe, we took the continuation of the initiative for granted and concerned ourselves primarily with how we would let new CI awardees know about the packages that we deposited into the network. At the UT Libraries, we wanted to take the best advantage of our financial investment in the network by depositing the best-organized, most fully-described copy of our data possible. Because we were aware that some technical aspects of the network were still in development when our ingests started, we had a degree of skepticism about how we or the receiving nodes would keep track of our bags over time. And, as previously described, we were mindful of potential staff turnover in the long term. These led us to make decisions about preparing archival packages that would be fully self-describing. We wanted our data, once out of our hands, to be understandable to anyone encountering it without the staff who prepared it needing to be available for explanations over the long term. These strategies all addressed aspects of data's persistence in the initiative over the long term, but not what we would do in the event of closure.

Our lack of a fully formed exit strategy cost us a good deal of staff time and effort. For DesignSafe, had we kept records for each bag that the corresponding project was safely ported into the new cyberinfrastructure, we could have notified the partner nodes immediately that they could delete the preservation network bags, rather than use valuable time tracking bag and project whereabouts on news of the closure. We expended significant staff time and computational resources at the UT Libraries pulling down and verifying a copy of all of our network bags from TACC storage when the network closed. In the end it would have been much more efficient for us to have written the enhanced copies to tape locally as the new copy of record at the time of their creation. On closure of the network we then could have simply agreed to delete the distributed copies.

One positive outcome for the UT Libraries is that since we were able to retrieve and verify these higher quality packages when the distributed initiative closed, we can supplant the lower quality packages in our tape archive right as we are planning a tape migration. Another is that the exercise of creating the superior preservation bags for the distributed network transformed our ongoing local work. We now treat all preservation data with the same approach that we devised for participating the distributed initiative. We are also developing a Digital Asset Management System (DAMS), which will help automate much of the work involved in creating these enhanced preservation packages and supply us with means for including more robust structured descriptive metadata.

In summary, our efforts in DesignSafe and the UT Libraries to prepare data for the distributed preservation initiative should have been matched by equally careful early exit strategy planning, risk analysis, and risk management. This came into sharp view when the initiative closed and we needed to respond quickly. However, the experience presented an opportunity to improve on previous shortcomings in the projects involved, ended with successful retrieval of data, and pushed us to make point-forward changes in existing practices so that we would not repeat mistakes of the past.

Our recommendations for exit strategies in digital preservation include:

- Pay equal attention not just to how to best use a system or tool but also how to stop using it, possibly very abruptly. We were careful in planning our ingest packages and process, but then caught off guard by needing to exit the initiative on a relatively short timeline.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

- Consider the goals of an exit strategy. With one in place, what will you be able to do? What is most important: Efficiency? Ease? Technical considerations? Had we planned for how abruptly the network might dissolve we would have devised a strategy that made data deletion a quick and easy decision. The network bags would only have represented an additional replication.
- Assess dependencies early in the planning process. If we had done this, we might have foreseen how lack of support for ARKs could cause issues later in the switch to a new identifier system.
- Include metadata in preservation packages, not just data. Without metadata files may become meaningless over time. UT Libraries enhanced packages became valuable in the network exit because they were the only copies with metadata alongside the content.
- Preferably include structured metadata to allow interoperability with future systems. In our examples, lack of structured metadata will make pushing preservation packages back into a repository a problem.
- Include identifiers that link replicated data with the projects to which they belong so that provenance can be retraced. This helped us track the DesignSafe data, assuring safety to delete network bags.
- Keep careful local records of what data have been sent for replication, where, and when. Again, this helped us verify our decisions at exit.
- Select tools that offer hash checking at both ends of transfers for data integrity. This is well-established in digital preservation but bears repeating.
- Carefully consider contractual language and technical documentation when selecting a preservation approach, but proceed with caution knowing that even with written terms in place conditions may change over time.

Acknowledgement

This work was partially funded by the National Science Foundation grant number 1520817

References

[1] J. Morley, "Storage cost modeling," presentation, figshare, 2019.
[2] https://doi.org/10.6084/m9.figshare.7795829.v1
[3] K. Dohe, D. Durden, "The cost of keeping it: toward effective cost-modeling for digital preservation," presentation, open science framework, 2018.
[4] https://doi.org/10.17605/OSF.IO/HVD5F
[5] N. Tallman, L. Work, "Approaching appraisal: framing criteria for selecting digital content for preservation," presentation, Open Science Framework, 2018.
[6] https://doi.org/10.17605/OSF.IO/8Y6DC
[7] S. Marks, *Becoming a Trusted Digital Repository*, Chicago, IL: Society of American Archivists, 2015, pp. 46-49.
[8] S. Marks, *Becoming a Trusted Digital Repository*, Chicago, IL: Society of American Archivists, 2015, pp. 50.
[9] K. Green, K. Niven, G. Field, "Migrating 2 and 3D datasets: preserving AutoCAD at the Archeology Data Service," *ISPRS International Journal of Geo-Information*, vol. 5, no. 4, pp. 44-56, April 2016.
[10] D. Rosenthal, "Emulation & virtualization as preservation strategies: a report commissioned by The Andrew W. Mellon Foundation," New York, 2015.
[11] https://mellon.org/Rosenthal-Emulation-2015
[12] D. Anderson, J. Delve, D. Pinchbeck, "Toward a workable emulation-based preservation strategy: rationale and metadata," *New Review of Information Networking*, vol. 15, no. 2, pp.110-131, November 2010.
[13] D. Minor, "The Digital Preservation Network," presentation, 2014.
[14] http://web.stanford.edu/group/dlss/pasig/PASIG_September2014/20140918_Presentations/20140918_04_DigitalPreservationNetwork_DavidMinor.pdf
[15] J. Kunze, J. Littman, E. Madden, J. Scancella, C. Adams, "The BagIt File Packaging Format (V1.0)."
[16] https://tools.ietf.org/html/rfc8493
[17] The Texas Digital Library.
[18] https://www.tdl.org/about-tdl
[19] E. Rathje, C. Dawson, J.E. Padgett, J.P. Pinelli, D. Stanzione, A. Adair, P. Arduino, S.J. Brandenberg, T. Cockerill, C. Dey, M. Esteva, F.L. Jr. Haan, M. Hanlon, A. Kareem, L. Lowes, S. Mock, G. Mosqueda, "DesignSafe: a new cyberinfrastructure for natural hazards engineering," *ASCE Natural Hazards Review*, 2016.
[20] doi:10.1061/(ASCE)NH.1527-6996.0000246.
[21] H. Tarver, M. Phillips, "Identifier usage and maintenance in the UNT Libraries' digital collections," presentation, DCMI International Conference on Dublin Core and Metadata Applications, 2016.

[22] http://dcpapers.dublincore.org/pubs/article/view/3846
[23] File Information Tool Set (FITS).
[24] https://projects.iq.harvard.edu/fits/home
[25] [15] S. Hein, K. Schmitt, "Risk management for digital long-term preservation services," *Proceedings of the 10th International Conference on Digital Preservation (iPRes 2013)*, 2013. https://services.phaidra.univie.ac.at/api/object/o:378059/diss/Content/get
[26] [16] A. McHugh, P. Innocenti, S. Ross, R. Ruusalepp, "Risk management foundations for digital libraries: DRAMBORA (Digital Repository Audit Method Based on Risk Assessment)," Second Workshop on Foundations of Digital Libraries, 2007.
[27] https://www.researchgate.net/publication/45456621_Risk_management_foundations_for_DLs_DRAMBORA_digital_repository_audit_method_based_on_risk_assessment
[28] [17] S. Ross, A. McHugh, "Preservation pressure points: evaluating diverse evidence for risk management," presentation, International Conference on Digital Preservation (iPres 2006), 2006. https://services.phaidra.univie.ac.at/api/object/o:294550/diss/Content/get
[29] [18] Integrated Rule-Oriented Data System (iRODS).
[30] https://irods.org/

The case for a standard that's old news

Recommendation of PDF/A for digitized newspaper preservation

Anna Oates
Federal Reserve Bank of St. Louis
United States of America
anna.oates@stls.frb.org
0000-0002-1848-6382

William Schlaack
University of Illinois at Urbana-Champaign
United States of America
schlaac2@illinois.edu
0000-0002-8658-9885

Abstract – Since 2004, the Libraryof Congress, abeholden stakeholder in the risk assessment of and consideration for file formats, has supported the preservation of and access to digitized historic newspapers through the National Digital Newspaper Program (NDNP), a distributed, mass digitization program. This paper evaluates the implementation and validation of PDF as specified for NDNP, explores the benefits of PDF/A, and analyzes the adverse effects for digital preservation as realized in current digitization workflows.

Keywords – file formats, file validation, newspaper digitization, PDF, PDF/A

Conference Topics – The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

The Portable Document Format (PDF) plays a vital role in the continued access of digitized newspaper. This is particularly true under the aegis of the National Digital Newspaper Program (NDNP), a collaboration between the National Endowment for the Humanities (NEH) and the Library of Congress (LC), which enables access to and preservation of digitized historic newspapers. The NDNP specification includes recommendations for PDF profiles, which, since the introduction of Version 2.1 in 2006, recommends PDF/A-1 where no conflict exists with the NDNP PDF Profile.1[1] In addition to the recommendation for PDF/A, the standard requires Type1 encoding of embedded fonts, Flate compression for text streams, PDF linearization for optimized load performance, embedded XMP metadata, and 13 other requirements forconformance to the NDNP PDF technical specifications (Version 2.4) [1]. This paper explores what features of existing PDFs in the NDNP collection do not conform to PDF/A, identifies challenges effacing PDF validation, and offers recommendations for investigating alternative validation workflows for NDNP PDF Profiles.

II. Background

A. National Digital Newspaper Program

In 2003, NDNP originated from a formal agreement between the NEH and the LC with the stated goal to provide permanent digital access to historic newspapers published in the United States. The project began in 2004 with NEH funding multiple state-level institutions ranging from public universities to state historical societies, and LC supporting awardees throughout the digitization process and hosting Chronicling America,2[2] a website dedicated to providing free and open access to newspapers digitized through the NDNP. NDNP builds upon the success of the 29-year NEH and LC collaboration, the United States Newspaper Program (1982-2011), which saw all fifty U.S. states, the District of Columbia, Puerto Rico, and the U.S. Virgin Islands contribute bibliographic data for over 140,000 newspaper titles, and preserve approximately 65 million newspaper pages on microfilm [2]. Through NDNP, grant awardees from 46 states, the District

[1] The mostrecent version of NDNPtechnical specifications removes specific PDF/Arequirements,replacing themwithgeneral PDF/Arecommendations with the goal to minimize any conflicts with the current NDNP PDF specification (NDNP PDF Profile (Version 2.4)

[2] Chronicling America: https://chroniclingamerica.loc.gov/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

of Columbia, and one U.S. territory (Puerto Rico) have contributed over 14.8 million newspaper pages representing 24 languages published between 1789 and 1963. As the content steward for NDNP, the LC works to establish and maintain imaging and bibliographic standards to manage the large-scale preservation of newspapers. When selecting titles, awardees must meet the "Technical Guidelines for Applicants" specification.[1] These technical specifications describe an "extendable, scalable, and sustainable workflow" for awardee institutions [3]. As part of theproject deliverables under the current specification, awardees provide an information package that includes a TIFF, JPEG2000, ALTO XML, and PDF file for each newspaper page. Since 2016, the Federal Agencies Digital Guidelines Initiative (FADGI) Still Images Working Group has explored the possibility to use JPEG2000 and PDF/A as master file formats [4]. LCcontributes to FADGIandinturn uses FADGI to inform best practices for digitization. The current recommendations in the guideline demonstrates an increasing emphasis on PDF/A as a format of importance for newspaper digitization.

B. Portable Document Format—Archival

Since 2006, the NDNP technical specifications have asserted that, "Except where conflicting with any of the other requirements of [the NDNP PDF] profile, conforming to PDF/A (ISO 19005-1) is recommended" [1]. In 2005, the International Organization for Standardization introduced ISO 19005, a standard supporting the long-term preservation of electronic documents. The sustainability is ensured by excluding or requiring document aspects, such as prohibiting embedding of Javascript or requiring font embedding, respectively. Based upon PDF 1.4, ISO 19005- 1:2005 specifies the use of PDF 1.4 for long-term preservation as the Portable Document Format—Archival. Since the first publication, two additional specifications have been released as ISO 19005-2:2011 and ISO 19005- 3:2012, which specify how to implement PDF 1.7, standardized as ISO 32000, as PDF/A. Despitechronological ordering, the later releases of the standard do not indicate a "better or more advanced" format [5]. They are simply different approaches to creating a sustainable PDF. The implementation of ISO 19005 as PDF/A permits three levels of conformance: Level A (accessible); Level B (basic), and Level U (unicode). Table I. provides an overview of the conformance levels relevant to each ISO 19005 publication. Each conformance level enables the use or disuse of requirements in the specifications "to prevent the onerous requirements for full conformance presenting a barrier to software developers" [6].

TABLE I

	ISO 19005-1:2005	ISO 19005-2:2011	ISO 19005-3: 2012
Level A	PDF/A-1a	PDF/A-2a	PDF/A-3a
Level B	PDF/A-1b	PDF/A2b	PDF/A-3b
Level U	NA	PDF/A-2u	PDF/A-3u

Amongthe features introduced with ISO 19005-2:2011, PDF/A-2 permits embedding of JPEG2000 images. Since its release, JPEG2000 has been subject to scrutiny for long- term preservation [7]. Despite its criticism, JPEG2000 is ubiquitous to digitization projects and is included as a required format for the NDNP file package. ISO 19005- 3:2012 differs from ISO 19005-2:2011 in that the specification permits embedding of any file type, a provision which, in response to widespread criticism, resulted in a report from the National Digital Stewardship Alliance on the risks and benefits of the format [8].

C. Validation

1. Digital Viewer and Validator

Before the LC accepts a submission package, the content creator must validate the submission package through the Digital Viewer and Validator (DVV), Version 2.2.1 [9]. Developed by the Library of Congress for NDNP, DVV "'wraps' JHOVE [(JSTOR/Harvard Object Validation Environment), Version 1.0] and extends JHOVE's existing TIFF, PDF, and JPEG2000 modules with the NDNP-specific validation rules" [10]. For example, JHOVE validates whether a PDF is well-formed, and the NDNP extension validates that the PDF is "grayscale, downsampled to 150dpi and encoded using JPEG, using a medium (or 40) quality setting" [1]. The NDNP extension for DVV only validates the NDNP file requirements, which means that PDF/A compliance is not included in the validation process. In addition to validating the file image and

[1] Both current and previous guidelines are located here: https://www.loc.gov/ndnp/guidelines/

container formats against JHOVE and additional rules implemented in the NDNP specification, DVV validates METS and ALTO records using Schematron [10].

As part of validation, DVV generates and embeds in the METS file a SHA-1 value as the digital signature, or fixity. Successful appendage of the original object's fixity indicates that a file is valid, and "proves that the technical metadata . . . was created by the NDNP Validation Library" (i.e., DVV) [10]. With multiple stakeholders of the content, including the LC, the awardee, and in many cases, a digitization vendor, frequent fixity checks are essential. In an effort to ensure that no changes occurred during file transfer, DVV's verification function enables users to verify the file fixity by checking the digital signature values generated during validation [10].

2. *Validation Challenges*

With myriad PDF versions, substandards, and versions of substandards, PDF validation is challenging. As found in Lindlar, Tunnat, and Wilson's paper on "A PDF Test-Set for Well-Formedness Validation in JHOVE - The Good, the Bad and the Ugly," JHOVE fails to meet a ground truth requirement for PDF validation [11]. Thus, while DVV includes an extension of the JHOVE PDF module, a validation module specific to the validation of the NDNP PDF Profile may be better suited for accurate assessment and to avoid constraints imposed by the capacity of JHOVE's PDF module.

Given that the DVV's validation profile is not testing for compliance with the ISO 19005-1 specification for PDF/A, awardees who wish to submit PDF files which conform to the NDNP PDF Profile in addition to ISO 19005-1 must rely upon workflows outside of the formal NDNP validation process to ensure PDA/A compliance.

In evaluating PDF/A converter software4[1]for the Florida Virtual Campus digital repository workflow, Jamin Koo and Carol Chou identified that the pdfaPilot, 3-Heights, and PDF/A Manager occasionally failed to identify non- conformance, resulting in 90%, 74%, and 91.30% accuracy rates, respectively [12]. As with the PDFlib "Bavaria Report on PDF/A Validation Accuracy," Koo and Chou limited their evaluation of tools against the PDF/A-1b (ISO 19005, conformance Level B) profile [13]. With an imperfect accuracy rate for PDF/A conversion across tools, PDF/A- specific validation is requisite.

3. *veraPDF*

As a result of the EU PREFORMA (PREservation FORMats) project, the veraPDF consortium released a validation tool dedicated to validation of PDF/A: veraPDF [14].5[2] In addition to validating against each PDF/A profile (PDF/A-1a, PDF/A-1b, PDF/A-2a, PDF/A-2b, PDF/A-2u, PDF/A-3a, PDF/A-3b, PDF/A-3u), veraPDF allows users to create a profile unique to their institutional needs. For example, NDNP might extend the validation to validate the NDNP PDF Profile specification that, "The PDF will open to single page layout," a rule not required for a valid PDF/A [1].

III. Method

In this section, the authors introduce the approach for selecting a testing corpus, define testing processes, and outline a metric for analysis of results.

A. *File Selection and Retrieval*

The pages from 69 newspaper issues were identified and downloaded from the Chronicling America website, totaling to 382 unique PDFs (i.e., pages). In an effort to establish a diverse testing corpus, one issue was selected from each contributing awardee, totaling to 45 issues.6[3] From the awardees' corporea, the authors selected the first issue from the most voluminous English-language

[1] here are two methods of converting or generating a PDF/A: create or conform. To "create" a PDF/A file, a user begins with a source file (e.g., Microsoft Word Document or LaTeX). To "conform" to PDF/A, a user begins with another PDF.

[2] Seealsothehomepage forveraPDF forfurther information and resources: https://verapdf.org/

[3] As noted in the background section above, NDNP has 46 participating awardees, of which only 45 have contributed content to dat

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

newspaper.[1] Following the same selection process, one issue representing each non-English language was selected, totaling to 24 newspaper issues.

B. File Validation

The authors tested against all PDF/A versions (PDF/A-1, PDF/A-2, and PDF/A-3) for a more comprehensive assessment of possible compliance to ISO 19005 across the corpora, and they limited the scope to conformance Level B for its suitability for digitized documents.8[2] While NDNP PDFs contain embedded OCR, the logical structural requirements for PDF/A are difficult to achieve. Using the veraPDF (Version 1.12.1) GUI on a Windows 10 OS, the authors batch processed the 382 PDF files against the PDF/A-1, PDF/A-2, and PDF/A-3 modules. Results were saved as XML for ease of data manipulation. The authors then parsed theveraPDF XML output toa CSVforincreased ease of analysis.

C. Validation Evaluation

Upon identifying the rules failed between the PDF/A-1, PDF/A-2, and PDF/A-3 validation profiles, rules were categorized by type. Four overarching types were identified as encompassing considerations fundamental to newspaper digitization: XMPMetadata, Embedded Images, Embedded Fonts, and Object Streams (in order of importance). The authors have identified and ranked categories in accordance to their salience for access and preservation of digitized newspapers within the NDNP framework.

IV. Discussion

As expected, all files conformed to the NDNP PDF specification, in that the files are PDF version 1.4, the PDF base required by both PDF/A-1 and the NDNP PDF specification. Also as anticipated, all files failed validation against the PDF/A-1b,9[3] PDF/A-2b, and PDF/A-3b veraPDF modules.

A. XMP Metadata

Perhaps the most ubiquitous element of access and preservation, metadata supports robust contextualization of a digital object. The XMP header metadata embedded in PDFs and other image and container files supports descriptive and administrative elements. ISO 19005-1 requires that, if metadata properties are contained in the document information dictionary, XMP "analogous properties . . . shall also be embedded . . . with equivalent values" [15]. The document information dictionary "contains the creation and modification dates of the file, together with some simple . . . metadata," which are not included in the XMP serialization but may be embedded in XMP extensions [16]. While non-conformance to this rule does not impact the NDNP workflow, it does suggest poor metadata practices. (See CosDocument, Appendix 1.) Poor metadata practice is further exemplified in the content objects' failure to meet the XMPPackage and XMPProperty requirements for ISO 19005, which require valid XMP serialization, stating that "XMP form shall use predefined schemas defined in XMP Specification, or extension schemas that comply with XMP Specification" [15, 17]. An essential feature included in the NDNP Technical Specification, best practice suggests that XMP metadata should conform to the XMP Specification.

B. Embedded Images

Validation failure for images included undefined OutputIntents of color profiles and use of interpolation, both which introduce potential harm of long-term sustainability. The NDNP Technical Specification requires that images be digitized at grayscale but does not impose requirements upon the colorspace. As seen in Appendix 1, results indicate that content creators fulfilled the grayscale requirement but did not define the colorspace (e.g., DeviceRGB, DeviceCYMK, DeviceGray) in the OutputIntent. The inclusion of image interpolation suggests that artifacts were generated through the image conversion

[1] Puerto Rico has contributed only Spanish-language newspapers. An issue from the most voluminous newspaper was selected to represent their awardeeship; an issue from the second most voluminous newspaper was selected as the overall most voluminous Spanish-language newspaper.

[2] As cited in "Preservation with PDF/A (2nd Edition)", minimal compliance with ISO 19005 in Level B conformance is suitable for digitized documents to "[render] visual appearance" [6]

[3] 9 PDF/A-1b profile defaulted for all files in the auto-detect module

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

process. The interpolation algorithm is specific to rendering software and thus cannot be embedded in a file. Use of interpolation impacts access by imposing an appearance which may not be rendered in every implementation.

C. *Embedded Fonts*

While font object streams render underneath the image object stream, fonts are captured during the OCR process and therein are embedded in the PDF. Thus, while it may seem that the visual appearance of a font is unnecessary, font information is captured in the OCR. All validation profiles indicated 137 failures to PDF/A-1 Rule 6.3.4-1, and PDF/A-2–PDF/A-3 Rule 6.2.11.4-1, which suggests that awardees are using Type 3 fonts. This further suggests that the DVV did not identify when PDF files contained Type 3 fonts, despite the requirements of the NDNP specification, which states that, "Only the 14 standard Type 1 fonts [may] be used. These fonts will not be embedded" [1]. This suggests that there are inconsistencies between the NDNP Specification and the digitization output. Increased validation not supported in the DVV's current validation profile, Version 2.1.1, may enable content creators and stewards to adhere to the NDNP Specification, as well as general recommendations for long-term preservation as identified in ISO 19005.

Perhaps the most notable difference between the PDF/A-1b and PDF/A-2b–PDF/A-3b validation profiles as seen in the results of this test is the relationship between Rule 6.3.5-2 (PDF/A-1b) and 6.2.11.4-4 (PDF/A-2b–PDF/A-3). The validation output for PDF/A-1b and PDF/A-2b–PDF/A- 3b differed in that PDF/A-1b validation identified 28 failed instances of Rule 6.3.5-2, which requires that, "For all Type 1 font subsets referenced within a conforming file, the font descriptor dictionary shall include a CharSet string listing the character names defined in the font subset" [18]. PDF/A-3b validation identified 34 failed instances of Rule 6.2.11.4-4, which requires that CID fonts be completely identified, whether or not all fonts are referenced. While no correlation has been identified between the rules, the fact that the PDF/A-1b validation profile did not detect failure for CID fonts in Rule 6.3.3-2 and vice versa such that the PDF/A-2b and PDF/A-3b validation profiles did not detect failure for CharSet with Rule 6.2.11.4-3 is of note.

Although evaluation of files which contained errors was not added to the methodology for this preliminary research, the authors are interested in the validation of fonts for non-English newspaper content. In skimming the results, it was found that Fraktur, a prominent glyph set used in many early 20th century German-language texts, did not return any font errors.

D. *Object Streams*

Object streams are necessary to long-term sustainability of content by exacting, for example, the beginning and end of an indirect object. However, the logical structure requirements for PDF have not been included in the NDNP PDF Profile, and as such has been considered out of scope for evaluation of PDF in the context of newspaper digitization under the current aegis of NDNP.

V. Future research

As digital collections experience exponential growth, repository managers seek alternatives to economize file storage space. Excluding TIFF from the newspaper archival information package reduces storage requirements, supporting long-term preservation of content produced at scale. The authors recommend additional investigation of PDF/A as a master file format for both preservation and access to minimize potential risks.10[1]

VI. Conclusion

The authors posit that, for institutions digitizing newspapers using a framework that places emphasis on PDF, the validation tool employed for quality assurance must ensure that the requirements for PDF are actually met. The discussion demonstrates that, without adequate validation software, non-conforming files may be approved unnoticed. Additional research is necessary to provide recommendations for implementation of veraPDF validation within newspaper digitization workflows, to remediate existing errors in PDFs as found in the non- conformance of fonts to the NDNP PDF Profile.

[1] Yan Han recommends PDF/A for its suitability asany information package throughoutthe preservation cycle—submission, archiving, dissemination— renderingthe requirement of an archival TIFF redundant [19]

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Acknowledgment

The authors thank the Library of Congress for making newspaper content freely accessible for all scholarly purposes, including validation of PDFs.

References

[1] Library of Congress, "The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants 2019-2021," October 24, 2018. www.loc.gov/ndnp/guidelines/NDNP_201921TechNotes.pdf

[2] M. Sweeney (Contributor), "The National Digital Newspaper Program: Building ona FirmFoundation,"Serials Review,vol. 33 no. 3, pp. 188-189, https://doi.org/10.1080/00987913.2007.10765120

[3] D. Thomas and M. Sweeney,"Sustainability inthe United States National Digital Newspaper Program," IFLA International Preservation News, No. 56, May 2012, pp.12-20. http://www.ifla.org/files/pac/ipn/IPN%2056.indd.def.pdf

[4] Federal Agencies Digital Guidelines Initiative, "FADGI Program: Impacts and Benefits,"April 23 2019. http://www.digitizationguidelines.gov/about/FADGI-impacts_20190423.pdf

[5] Y. Han, "Beyond TIFF and JPEG2000: PDF/A as an OAIS submission information package container." Library Hi Tech, vol. 33, no. 3, 2015, pp. 409-423. https://doi.org/10.1108/LHT-06-2015-0068

[6] B. Fanning, "Preservation with PDF/A (2nd Edition)" DPC Technology Watch Report 17-01, July 2017. http://doi.org/10.7207/twr17-01

[7] C. Adams, "Is JPEG-2000 a preservation risk?" [Blog post], January 28, 2013. http://blogs.loc.gov/thesignal/2013/01/is-jpeg-2000-a- preservation-risk/

[8] C. Arms, et al,. "The Benefits and Risks of the PDF/A-3 File Format for Archival Institutions: An NDSA Report," February 2014. http://www.digitalpreservation.gov/documents/NDSA_PDF_A3_report_ final022014.pdf

[9] Library of Congress,"National DigitalNewspaperProgramDigitalViewer and Validator (DVV) User's Guide (DVV Version 2.2.1)," November 2, 2016. http://www.loc.gov/extranet/wiki/library_services/ndnp/images/9/92/D VV2_userguide_11_02_16.pdf

[10] J. Littman, "A Technical Approach and Distributed Model for Validation of Digital Objects," D-Lib Magazine, vol. 12, no. 5, May 2006. http://www.dlib.org/dlib/may06/littman/05littman.html

[11] M. Lindlar, Y. Tunnat, and C. Wilson, "A PDF Test-Set for Well- Formedness Validation in JHOVE - The Good, the Bad and the Ugly," Paper presented at the iPRES 2017: Keeping Cultural Diversity for the Futureinthe DigitalSpace—FromPop CulturetoScholarlyInformation, Kyoto,Japan, September 27, 2017. http://files.dnb.de/nestor/weitere/ipres2017.pdf

[12] J. Kooand C.H.C. Chou, "PDFto PDF/A: Evaluation of Converter Software for Implementation in Digital Repository Workflow," New Review of Information Networking vol. 18, no. 1, 2013, pp. 1-15. https://doi.org/10.1080/13614576.2013.771989

[13] PDFlib GmbH, "Bavaria Report on PDF/A Validation Accuracy," May 4, 2009. https://web.archive.org/web/20170103204840/http://www.pdflib.com/ fileadmin/pdflib/pdf/pdfa/2009-05-04-Bavaria-report-on-PDFA- validation-accuracy.pdf

[14] C. Wilson,R. McGuinness,and J. Jung,"veraPDF: buildinganopensource, industry supported PDF/A validator for cultural heritage institutions," Digital Library Perspectives vol. 33, no. 2, 2017, pp. 156-165. https://doi.org/10.1108/DLP-08-2016-0031

[15] Document management—Electronic document file format for long- term preservation—Part 1: Use of PDF 1.4 (PDF/A-1), International Organization for Standardization (ISO), Standard 19005-1:2005.

[16] J. Whitington, "Chapter 4. Document Structure," PDF Explained. https://www.oreilly.com/library/view/pdf-explained/9781449321581/ch04.html

[17] Adobe Systems Incorporated, XMP Specification Part 1: Data Model, Serialization, and Core Properties, April 2012. https://wwwimages2.adobe.com/content/dam/acom/en/devnet/xmp/ pdfs/XMP%20SDK%20Release%20cc-2016-08/XMPSpecificationPart1.pdf

[18] veraPDF, PDF/A-1 validation rules. http://docs.verapdf.org/validation/pdfa-part1/

[19] Y. Han and X. Wan, "Digitization of Text Documents Using PDF/A," Information Technology and Libraries, vol. 37, no.1, 2018, pp. 52-64. https://doi.org/10.6017/ital.v37i1.9878ology and Libraries, vol. 37, no.1, 2018, pp. 52-64.

Appendices

A. *Appendix 1: PDF/A Failed Validation Rules11[1]*

[1] Not all tests were equivalent between the PDF/-1 and PDF/A-2–PDF/A-3 validation profiles. In particular, the PDF/A-2–PDF/A-3 validation profile does not contain rules for testing the CosDocument and PDExGState objects. Furthermore, the PDCIDFontand PDType1Fontobjects, whichfailed against

See Appendix 2 for relevant veraPDF validation rules.

Object Type	PDF/A-1b Files Failed	PDF/A-2b Files Failed	PDF/A-3b Files Failed
CosDocument Appendix 2: 16	235		
CosIndirect Appendix 2: 3,4	100	100	100
CosStream Appendix 2:1	40	40	40
CosStream Appendix 2:2	16	16	16
MainXMP Package Appendix 2:19	382	382	382
PDCIDFont Appendix 2:9	-	34	34
PDDeviceGray Appendix 2:6	344	344	344
PDDeviceRGB Appendix 2:	19	19	19
PDExtGState Appendix 2:14	120	-	-
PDExtGState Appendix 2:15	120	-	-
PDFont Appendix 2:8, 11	137	137	137
PDTrueType Font Appendix 2:10, 13	16	16	16
PDType1Font Appendix 2:12	16		
PDXImage Appendix 2:7	105	105	105
XMPPackage Appendix 2:17	12	12	12

B. Appendix 2: Relevant veraPDF Validation Rules

1. **Rule 6.1.7-1** (ISO 19005-1:2005, ISO 19005-2:2011, ISO19005-3:2012) "The value of the Length key specified in the stream dictionary shall match the number of bytes in the file following the LINE FEED character after the stream keyword andpreceding the EOLmarker before the endstream keyword."
2 **Rule 6.1.7-2** (ISO 19005-1:2005, ISO 19005-2:2011, ISO 19005-3:2012) "The stream keyword shall be followed either by a CARRIAGE RETURN (0Dh) and LINE FEED (0Ah) character sequence or by a single LINE FEED character. The endstream keyword shall be preceded by an EOL marker."
3. **Rule 6.1.8-1** (ISO 19005-1:2005) "The object number and generation number shall be separated by a single white-space character. The generation number and obj keyword shall be separated by a single white-space character. The object number and endobj keyword shall each be preceded by an EOL marker. The obj and endobj keywords shall each be followed by an EOL marker."
4. **Rule 6.1.9-1** (ISO 19005-2:2011, ISO 19005-3:2012) "The object number and generation number shall be separated by a single white-space character. The generation number and obj keyword shall be separated by a single white-space character. The object number and endobj keyword shall each be preceded by an EOL marker. The obj and endobj keywords shall each be followed by an EOL marker.
5. **Rule 6.2.3-2** (ISO 19005-1:2005) "DeviceRGB may be used only if the file has a PDF/A-1 OutputIntent that uses an RGB colour space." Same test condition as **Rule 6.2.4.3-2** (ISO 19005- 2:2011, ISO 19005-3:2012) "DeviceRGB shall only be used if a device independent DefaultRGB colour space has been set when the DeviceRGB colour space is used, or if the file has a PDF/A OutputIntent that contains an RGB destination profile."
6. **Rule 6.2.3-4** (ISO 19005-1:2005) "If an uncalibrated colour space is used in a file then that file shall contain a PDF/A-1 OutputIntent, as defined in 6.2.2."12[1] PDF/A-2–PDF/A-3 and PDF/A-1b, respectively, contain different rules in the validation profile, thus resulting in inconsistent

[1] 12 Rule 6.2.2-2 (ISO 19005-1:2005) "If a file's OutputIntents array contains more than one entry, then all entries that contain a DestOutputProfile key shall have as the value of that key the same indirect object, which shall be a valid ICC profile stream."

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

validation output. Same test condition as Rule **6.2.4.3-4** (ISO 19005- 2:2011, ISO 19005-3:2012) "DeviceGray shall only be used if a device independent DefaultGray colour space has been set when the DeviceGray colour space is used, or if a PDF/A OutputIntent is present."

7. **Rule 6.2.4-3** (ISO 19005-1:2005) / **6.2.8-3** (ISO 19005-2:2011, ISO 19005-3:2012) "If an Image dictionary contains the Interpolate key, its value shall be false."
8. **Rule 6.2.11.4-1** (ISO 19005-2:2011, ISO 19005-3:2012) "The font programs for all fonts used for rendering within a conforming file shall be embedded within that file, as defined in ISO 32000-1:2008, 9.9."
9. **Rule 6.2.11.4-4** (ISO 19005-2:2011, ISO 19005-3:2012) "If the FontDescriptor dictionary of an embedded CID font contains a CIDSet stream, then it shall identify all CIDs which are present in the font program, regardless of whether a CID in the font is referenced or used by the PDF or not."
10. **Rule 6.2.11.6-2** (ISO 19005-2:2011, ISO 19005-3:2012) "No non-symbolic TrueType font shall define a Differences array unless all of the glyph names in the Differences array are listed in the Adobe Glyph List and the embedded font program contains at least the Microsoft Unicode (3,1 - Platform ID=3, Encoding ID=1) encoding in the 'cmap' table."
11. R**ule 6.3.4-1** (ISO 19005-1:2005) "The font programs for all fonts used within a conforming file shall be embedded within that file, asdefined in PDF Reference 5.8, except when the fonts are used exclusively with text rendering mode 3."
12. **Rule 6.3.5-2** (ISO 19005-1:2005) "For all Type 1 font subsets referenced within a conforming file, the font descriptor dictionary shall include a CharSet string listing the character names defined in the font subset, as described in PDF Reference Table 5.18."
13. **Rule 6.3.7-1** (ISO 19005-1:2005) "All non-symbolic TrueType fonts shall specify MacRomanEncoding or WinAnsiEncoding, either as the value of the Encoding entry in the font dictionary or as the value of the BaseEncoding entry in the dictionary that is the value of the Encoding entry in the font dictionary. If the value of the Encoding entry is a dictionary, it shall not contain a Differences entry."
14. **Rule 6.4-4** (ISO 19005-1:2005) "The following keys, if present in an ExtGState object, shall have the values shown: BM - Normal or Compatible."
15. **Rule 6.4-5** (ISO 19005-1:2005) "The following keys, if present in an ExtGState object, shall have the values shown: CA - 1.0."
16. R**ule 6.7.3-1** (ISO 19005-1:2005) "If [a document information dictionary appears in a document], then all of its entries that have analogous properties in predefined XMP schemas ...shall also be embedded in the file in XMP form with equivalent values."
17. **Rule 6.7.9-1** (ISO 19005-1:2005) "The metadata stream shall conform to XMP Specification and well formed PDFAExtension Schema for all extensions." Same test condition as Rule **6.6.2.1-4** (ISO 19005- 2:2011, ISO 19005-3:2012) "All metadata streams present in the PDF shall conform to the XMP Specification. All content of all XMP packets shall be well-formed, as defined by Extensible Markup Language (XML) 1.0 (Third Edition), 2.1, and the RDF/XML Syntax Specification (Revised)."
18. **Rule 6.7.9-2** (ISO 19005-1:2005) "Properties specified in XMP form shall use either the predefined schemas defined in XMP Specification, or extension schemas that comply with XMP Specification." Same test condition as **Rule 6.6.2.3-7** (ISO 19005- 2:2011, ISO 19005-3:2012) "All properties specified in XMP form shall use either the predefined schemas defined in the XMP Specification, ISO 19005-1 or this part of ISO 19005, or any extension schemas that comply with 6.6.2.3.2."
19. **Rule 6.7.11-1** (ISO 19005-1:2005) / **6.6.4-1** (ISO 19005-2:2011, ISO 19005-3:2012) "The PDF/A version and conformance level of a file shall be specified using the PDF/A Identification extension schema."

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Significant Properties Of Spreadsheets

An Update On The Work Of The Open Preservation Foundation's Archives Interest Group

Remco van Veenendaal
National Archives of the Netherlands, The Netherlands
Remco.van.Veenendaal@nationaalarchief.nl
https://orcid.org/0000-0002-2351-1677

Frederik Holmelund Kjærskov
Danish National Archives, Denmark
fhk@sa.dk

Kati Sein
National Archives of Estonia
Estonia
kati.sein@ra.ee

Jack O'Sullivan
Preservica
United Kingdom
jack.osullivan@preservica.com

Anders Bo Nielsen
Danish National Archives, Denmark
abn@sa.dk

Phillip Mike Tømmerholt
Danish National Archives, Denmark
pmt@sa.dk

Jacob Takema
National Archives of the Netherlands, The Netherlands
jacob.takema@nationaalarchief.nl

Extended abstract for poster proposal – In this extended abstract, the Open Preservation Foundation's Archives Interest Group reports on our ongoing investigation of significant properties of spreadsheets. Using the InSPECT methodology for investigating significant properties of electronic content, our goal is to get hands-on experience in investigating the significant properties of deposited spreadsheets by adding a Spreadsheet Testing Report to the InSPECT Testing Reports lore. An additional result of the AIG investigation is a Spreadsheet Complexity Analyser tool that extracts spreadsheet-specific properties and can be used to calculate the complexity of a spreadsheet based on the values of those properties.

I. Introduction

The Open Preservation Foundation's Archives Interest Group (AIG) started in July 2016. In order to establish the work priorities for the AIG, we used the elements of the SCAPE Catalogue of Preservation Policy Elements [1] as a long list of priorities. Each AIG member prioritised their copy of the list. Combining the lists resulted in our work priorities. One priority is an investigation of the significant properties of spreadsheets.

II. Why Investigate The Significant Properties Of Spreadsheets?

The AIG chose to investigate significant properties of spreadsheets, because (a) we wanted to get hands-on experience in investigating significant properties as a means of understanding the original deposited object, and how to preserve it and (b) as national archives, we receive more and more spreadsheets that are eligible for long-term preservation, but are faced with the current shortcomings of ensuring long-term accessibility of the spreadsheets while still preserving their significant properties. The Danish National Archives in particular had been asked to add suitable formats for preserving spreadsheets to their list of accepted formats, and in order to choose a format, needed to know which properties the format should be able to preserve.

The digital preservation community has

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

investigated significant properties in general and those of spreadsheets in particular, but there have been few significant properties of spreadsheets studies in recent years, while spreadsheet technology keeps changing. The few significant properties of spreadsheets resources available have e.g. been collected in the list of Significant Significant Properties [3], with 18 properties for spreadsheets, stemming from 2 resources. As AIG we found this too meagre a basis for decision-making and decided to start our own investigation of significant properties of spreadsheets.

III. Using The InSPECT Methodology

The AIG members created a recommended reading list about significant properties, collected spreadsheet example files and spreadsheet (file format) specification documentation as a knowledge base. We looked for significant property investigation methodologies and decided to use the InSPECT methodology for investigating significant properties of electronic content [2] for our investigation.

The InSPECT methodology is a well-documented formalized methodology that has been used and re-used in significant property investigations and resulted in a collection of Testing Reports[1]. We want to add our work to this lore.

IV. Object Analysis And Spreadsheet Complexity Analyser

The AIG followed the activities defined by the InSPECT methodology. At the time of writing, we are performing the Object analysis set of activities. We have selected spreadsheets as our object type, analysed the structure of spreadsheets by using (property extraction) tools and studying (file format) specifications and identified the purpose of spreadsheet properties by classifying them as one of the categories Content, Context, Appearance, Structure or Behaviour. We are currently working on steps 4, 5, and 6: linking behaviours to functions and structure.

One task of the InSPECT methodology is to get a list of tools that can be used to extract (technical) properties of electronic content. While listing and testing tools for extracting properties of spreadsheets (including Apache Tika[2], Dependency Discovery Tool[3] and the New-Zealand Metadata Extraction Tool[4]), we noticed that there were hardly any tools for extracting spreadsheet-specific properties, like used cells and worksheets, hyperlinks, formulas and scripts, embedded objects, pivot tables, etc.

Another challenge arose when we discussed possible subtypes of spreadsheets. Our initial thoughts were to have 'simple/static' spreadsheets vs. 'complex/dynamic' ones, where the former are mainly meant for pretty-printing tabular data on a single worksheet, and the latter for more complex calculations across more than one worksheet.

The combination of these two issues resulted in the need for a tool that can analyse the complexity of spreadsheets based on the values of extracted spreadsheet properties. This tool did not exist. We therefore developed a 'Spreadsheet Complexity Analyser', voted on which properties this analyser should be able to extract, and decided when a spreadsheet would be deemed 'simple/static' or 'complex/dynamic'. One test of the tool showed that 99% of a 180,000 Microsoft Office Excel file test set from the National Library of the Netherlands Ejournal collection were 'complex/dynamic'.

Even after revisiting our decision rules, we noticed that a categorisation in 'simple/static' and 'complex/dynamic' may be too simplistic. And that pre-programmed decision rules limit the possible uses of the tool. Users should e.g. be able to configure their own decision rules, as different organisations may have different decision criteria.

But even if we were to drop the idea of using the tool for distinguishing between different sub-types of spreadsheets, the tool seems to be a candidate for filling a gap in the property extraction and migration quality assessment tool ecosystem. If accepted, we will use the poster opportunity to discuss uses and improvements of the tool with our audience.

[1] See e.g. https://web.archive.org/web/20160416031256/http://www.significantproperties.org.uk/testingreports.html.

[2] Available from http://tika.apache.org/, accessed March 15, 2019.

[3] Available from https://sourceforge.net/projects/officeddt/, accessed March 15, 2019.

[4] Available from http://meta-extractor.sourceforge.net/, accessed March 15, 2019.

V. Conclusion

In this extended abstract, we presented the state of affairs of the AIG's ongoing investigation of significant properties of spreadsheets. We have chosen the InSPECT methodology for investigating significant properties of electronic content and are using that methodology to get hands-on experience in investigating the significant properties of spreadsheets. We are performing the Object analysis set of activities, and are in the process of linking spreadsheet behaviours to functions and structure. As a result of our work, a Spreadsheet Testing Report will be added to the InSPECT Testing Report lore.

Our preliminary conclusions from the Object analysis support earlier findings of significant property studies: the complexity and context-sensitivity of and degree of freedom inherent in spreadsheets makes creating an exhaustive list of significant spreadsheet properties practically impossible. But a list of (technical) significant properties does already help choose suitable file formats for preserving spreadsheet information. Further stakeholder analysis is required for fine-tuning our work.

An additional result of the AIG investigation is a Spreadsheet Complexity Analyser tool that extracts spreadsheet-specific properties and can be used to calculate the complexity of a spreadsheet based on the values of those properties.

Acknowledgements

The AIG would like to thank the OPF team: Becky McGuinness and Charlotte Armstrong for facilitating AIG meetings, and Carl Wilson for an introduction to GitHub and giving feedback on the Spreadsheet Complexity Analyser.

References

[1] Bechhofer, S., Sierman, B., Jones, C., Elstrøm, G., Kulovits, H., Becker, C.: Final version of policy specification model. http://www.scape-project.eu/deliverable/d13-2-catalogue-of-preservation-policy-elements (2014). Accessed March 15, 2019

[2] Knight, G.: InSPECT Framework Report. https://web.archive.org/web/20160520083956/http://www.significantproperties.org.uk/inspect-framework.html (2009). Accessed March 15, 2019

[3] Lucker, P., Sijtsma, C., van Veenendaal, R.: Significant Significant Properties. https://osf.io/rtjw3 (2018). Accessed March 15, 201

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Preferred, Obsolete Or In-Between? Developing A Criteria Catalogue For AV-Material

Preservation planning at the German National Library of Science and Technology (tib)

Merle Friedrich
German National Library of Science and Technology, Germany
merle.friedrich@tib.eu
0000-0001-7158-8583

Abstract – The born-digital audio-visual (AV) holdings of the German National Library of Science and Technology are analyzed regarding the present file formats. The most frequent AV file formats are examined in terms of suitability as preservation format based on a catalogue of criteria. Furthermore their risk of obsolescence is evaluated using view paths. The examined file formats are not preferred as preservation formats, but they are not obsolete either.

Keywords – Obsolescence; Audio-Visual Material; Preservation Planning; Technology Watch

Conference Topics – The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

The German National Library of Science and Technology (TIB) has a collection of audio-visual (AV) material. TIB acquires new content from different producers. This multitude of producers leads to a variety of different file formats which are preserved in TIB's digital archive. The poster will describe how the risk of obsolescence of born-digital AV files is determined for the three file formats in which the majority of born-digital AV material in TIB's holdings are available.

A. Background

AV material usually consists of a container (e.g. mp4), which wraps one or more content streams. The content streams are video/ audio stream(s). They can be encoded by different audio and video codecs [1, p. 137] (e.g. Pulse Code Modulation, FFv1). I will refer to file format as the combination of container, video codec and audio codec and differentiate on the level of format version as indicated by MediaInfo [2]. The examined file formats are *MPEG-4/AVC/AAC, Version 4,* as well as *WebM/VP8/Vorbis* and *MPEG-PS/ MPEG Video, Version 2/MPEG Audio, Version 1.*

B. Research question

Preservation Planning as described in the OAIS covers questions regarding the obsolescence of file formats [3, 4.15]. File formats can have three different states: they are either *preferred as preservation format, or not preferred, but not obsolete*, or they can be *obsolete*. Which state are the examined file formats in? The poster describes the verification of the following hypotheses:

1. The examined file formats are not preferred as preservation formats.
2. The examined file formats are not obsolete.

II. Attributes of Preferred Formats

A literature study reveals different approaches in order to assess preferences regarding a file format. Todd combines the findings of different studies and concludes the most common criteria for file format selection are adoption, platform independence, disclosure or documentation, transparency, and metadata support [4, p. 10].

A. Criteria for Suitability as Preservation Format

Each of the five criteria is broken down into preferred, acceptable and critical factors. For each of the factors measurements were developed. These measurements consider the requirements of TIB's

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

designated community. E.g. metadata support is rated acceptable if technical metadata is embedded in the format. This was tested through the extraction of technical metadata with MediaInfo [2].

B. Classification of TIB's holdings

Each container, video and audio codec is weighted according to the developed measurements. An overall weighting reveals if a container or codec was considered preferred, accepted or critical. The three examined file formats are rated critical for preservation purposes, which verifies the first hypothesis.

III. Obsolescence

A file format is obsolete if it is at risk to become inaccessible [5, p. 93] to our designated community. According to Ryan there is one factor to measure obsolescence: if no rendering software is available a file format is obsolete [6, p. 14].

A. Availability of Rendering Software

The National Library of the Netherlands uses view paths as a formal approach of evaluating the availability of rendering software. A View path consists of the information on the hardware platform, the operating system and the viewer application (incl. version) which enables the designated community to access the content of the file [7, p. 48].

B. Classification of TIB's holdings

For each of the examined formats two view paths were documented. These view paths work independent from each other, meaning that no component is used twice. With two view paths for each file format the second hypothesis - the examined file formats are not obsolete - is verified.

IV. Results & Conclusion

The majority of TIB's born-digital AV holdings come in a file format which is not preferred for preservation, but not obsolete. Both hypotheses are verified.

Although this is true for the file format, it is not necessarily true for each file. Not all files are implemented according to the file format specification and therefore valid. There is a critical lack of (open source) validation software for AV files. AV playback software is tolerant to implementation errors so that testing the render ability cannot replace validation [8, p. 28].

V. Future Work

Regarding the view paths TIB must evaluate if testing and documenting view paths for all (AV) file formats add a crucial value in order to determine the right point in time for migration. A regular check if the view paths still apply to the designated community should be scheduled. This presumes a deep insight of the equipment and requirements of the designated community.

It could be evaluated, if the lack of validation software can be counterbalanced by tentatively migrating into a file format which is preferred for preservation purposes. Further research should bring into focus the automated evaluation of (digital to digital) migration of AV content.

References

[1] B. E. Koenig and D. S. Lacey, "Forensic Authentication of Digital Audio and Video Files," in Handbook of digital forensics of multimedia data and devices, A. T. S. Ho, Ed., Hoboken: Wiley, 2015, 133-181.

[2] MediaArea, MediaInfo. [Online] Available: https://mediaarea.net/de/MediaInfo. Accessed on: Nov. 01 2018.

[3] Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-M-2, 2012.

[4] M. Todd, "File formats for preservation: Technology Watch Report," https://www.dpconline.org/component/docman/?task=doc_download&gid=375, 2009.

[5] D. Pearson and C. Webb, "Defining File Format Obsolescence: A Risky Journey," IJDC, vol. 3, no. 1, pp. 89–106, https://doi.org/10.2218/ijdc.v3i1.44, 2008.

[6] H. M. Ryan, "Who's afraid of File Format Obsolescence?: Evaluating File Format Endangerment Levels and Factors for the Creation of a File Format Endangerment Index," School of Information and Library Science, University of North Carolina, Chapel Hill, 2014.

[7] J. F. Steenbakkers, "Digital Archiving in the Twenty-First Century: Practice at the National Library of the Netherlands," Library Trends, vol. 54, no. 1, pp. 33–56, http://muse.jhu.edu/article/193231, 2005.

[8] J. Houpert, P. Melas, W. Bailer, and P. Walland, "Recommendations and techniques for content in a 'born robust' form," 2015. Accessed on: Jun. 25 2018.

Engaging Decision Makers

An Executive Guide on Digital Preservation

Sarah Middleton
Digital Preservation Coalition, UK
sarah.middleton@dpconline.org

Sharon McMeekin
Digital Preservation Coalition, UK
sharon.mcmeekin@dpconline.org
0000-0002-1842-611X

Abstract: The Executive Guide on Digital Preservation provides practitioners with a combination of generic and specific messages and motivators designed to communicate with senior executives, legislators and budget holders, as well as decision and policy makers with a view to embedding the value of digital preservation at the core of every organization.

Keywords: advocacy, engagement, awareness, executives, senior management

Conference Topic: 4. Building Capacity, Capability and Community

I. Significance of the Work

While, as a practice, digital preservation is growing and becoming more widely understood, within any organization this understanding is likely to be limited to pockets of practitioners and the colleagues with whom they regularly interact. More frequently, the same practitioners will find themselves lacking resources and funds to support their activities, because awareness of or support for digital preservation is not present at executive level. Advocacy is, therefore, an essential part of digital preservation.

The Digital Preservation Coalition (DPC) and United Nations Educational, Scientific and Cultural Organization's (UNESCO) Executive Guide on Digital Preservation [1] (the Guide) provides practitioners with a combination of generic and specific messages designed to support communications with senior executives, legislators, budget holders, decision and policy makers with a view to embedding the value of digital preservation at the core of every organization.

II. Objectives

A senior level advocate in an organization will champion digital preservation on behalf of the practitioner, but engaging executives is a challenge.

The Guide is a set of resources to help those responsible for the preservation of digital materials inform senior members of their organizations about the importance of digital preservation, the risks faced from inaction and the opportunities preserving digital materials can create.

Users of the Guide may belong to memory and heritage institutions, commercial organizations, government bodies and not-for-profits. The development of the Guide is supported by (UNESCO) whose member states each have a role in implementing the UNESCO *Recommendation concerning the Preservation of, and Access to, Documentary Heritage Including in Digital Form* [2]. In order to support the UNESCO member states, and all organizations, in their mission to preserve our digital heritage the Guide aims to be as broad reaching as possible. It does, however, recognize that even within the same sector or state, every organization is different, has different priorities, risk factors and motivators. It therefore presents a set of generic and sector specific statements, which may be selected and tailored by individual organizations to assist in their internal advocacy work and in the application of the UNESCO recommendations.

III. Structure

Interactive and customizable, the Guide is an online and accessible collection of resources, available to all free of charge, and contains:

- Generic statements defining digital preservation in a range of clear, non-jargon terms which may be understood by various organization types

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

- Information relating to digital preservation in specific sectors and organization types
- A summary of the importance of digital preservation generally and specifically
- Key motivators for digital preservation which may be relevant to different organization types
- Risks and opportunities related to digital preservation
- Downloadable and customizable templates
- Useful links to related and relevant resources
- Statistics and evidence to be used in supporting the case for digital preservation

Each statement is associated with an organization type for which it might be most relevant. Currently the Guide contains statements which have been created for (and by):

- Archives
- Businesses
- Higher Education and Research
- Libraries
- Museums and Galleries

Each statement is also associated with a set of organizational 'motivators.' Based on the DigCurV Executive Lens on Digital Skills [3], the motivators are issues that are important to an organization, and likely to be the things Senior Executives are most concerned about. The Guide identifies the following motivators:

- Accountability
- Authenticity
- Business Continuity
- Compliance
- Corporate / Cultural Memory
- Costs
- Reputation
- Revenue
- Security
- Technology

Users may interrogate the content by organization type, motivator or by browsing all statements before inserting into customizable PowerPoint or Word templates alongside evidence to support digital preservation, by way of statistics and case studies.

IV. Development and use

The Guide is a community developed and community owned resource. The statements which make up the Guide have been contributed by volunteer members of the DPC which represent just a small portion of the global digital preservation community.

Nevertheless, it is hoped that the Guide will be relevant to an international and cross-sector audience as much of the information is widely applicable and spans organization types and geography.

Information contained within the Guide may be used to create and deliver the message about digital preservation in the most relevant and appropriate format for a user's organization.

Some of the statements contained within the Guide may be applicable to more than one organization type. These statements are presented as just one way of communicating the importance of, and risks, opportunities, and needs associated with digital preservation. However, as each organization and the content it manages is different, it is expected that users will need to customize the messages, selecting the most appropriate statements presented.

Once the messages contained within the Guide have been tailored for use within a specific organization, these customized messages are welcomed as updates for inclusion in future iterations of the resource, and their re-submission into the Guide is encouraged.

V. Conclusion

The Guide is a living and evolving resource which, with the submission of subsequent statements tailored to other organisation types, will continue to grow and be relevant to the digital preservation community. This evolution is aligned with the fact that advocacy is not one *single* action, it is a lifecycle of actions and activities which must be repeated continuously in order be successful.

Recognizing the nature of this challenge, the DPC sees the benefit in sharing resources like the Guide which forms part of a suite of resources which are available free of charge and for the entire community – members and non-members alike. The Guide, and these other resources, support the entire advocacy lifecycle and aim to relieve some of the burden in engaging with decision-makers and executives on digital preservation.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

REFERENCES

[1] Executive Guide on Digital Preservation, DPC: https://www.dpconline.org/our-work/dpeg-home

[2] Recommendation concerning the preservation of, and access to, documentary heritage including in digital form, UNESCO: http://portal.unesco.org/en/ev.php-URL_ID=49358&URL_DO=DO_TOPIC&URL_SECTION=201.html

[3] DigCurV Executive Lens on Digital Skills, DigCurV: https://www.digcurv.gla.ac.uk/executiveLens.html

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

Safeguarding the nation's digital memory

A bayesian model of digital preservation risk for the disruptive digital archive

Alec Mulinder
The National Archives
United Kingdom
alec.mulinder@nationalarchives.gov.uk
0000-0002-8900-3798

Sonia Ranade
The National Archives
United Kingdom
sonia.ranade@nationalarchives.gov.uk

David H Underdown
The National Archives
United Kingdom
david.underdown@nationalarchives.gov.uk
0000-0002-8123-4655

Abstract – This poster offers the iPRES community an interim update on work at The National Archives (UK) to develop a Dynamic Bayesian Network which describes and explains the complexity of the digital preservation risk environment. The intention of this project is to evaluate the applicability of a Bayesian statistical approach to support a risk-based approach to digital preservation. This would complement our existing standards-based approaches and support evidence led decision making by the archive.

This poster focuses on the technical work undertaken so far during the project, highlighting the potential benefits of the approach and identifying areas for further investigation.

Keywords – risk management, decision support, Bayesian Network, elicitation protocol, evidence base

Conference Topics – Exploring New Horizons.

I. EXTENDED ABSTRACT

This poster offers the iPRES community an interim update on work at The National Archives (UK) to create a Bayesian statistical model of the digital preservation risk environment [1]. We discuss the requirements and potential benefits of the approach; describe our investigation of techniques for combining hard data with a knowledge base distilled from the experience of our archivists; present and review our initial experimental Bayesian network; identify areas for further development and highlight our plans for further work, in terms of technical development, building partnerships and growing the capability of the digital archive.

A. Benefits of the Approach

We identify gaps in our current approaches to digital preservation and highlight how these may be addressed through application of the evidence based methods being proposed. The benefits of the method are discussed:

1. *Complexity and Transparency:* The digital preservation environment is highly complex with multiple interdependent factors at work. It is difficult to understand and harder still to explain to stakeholders. The proposed model is well suited to expression as a graphical map of dependencies, events, actions and impact, which is accessible and relatively easy to communicate, understand and critique.
2. *Prioritization:* The model quantifies risk exposure. We anticipate that this will allow us to compare and prioritize very different types of threats to the digital archive with potential impact in different areas.
3. *Evidence Based:* The approach is data-driven, incorporating evidence from our observations and experience in the form of conditional probability tables within the Bayesian network. The model will also operate in areas where we lack hard data, via techniques for the elicitation of structured expert judgement to populate prior values for our probability distributions.

B. Eliciting Expert Judgement

We offer an overview of our trial of a structured protocol for expert elicitation [2]. This is designed

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

to compensate for bias, be transparent, open to challenge and well documented. We describe our experience of working with the protocol to improve the quality of expert judgements being incorporated into the experimental model.

C. *The Experimental Bayesian Network*

We present a graphical representation of our experimental network as it currently stands. This highlights areas for further development, including the need to better reflect risk mitigation actions in the model; handling complexity; modelling failure events that arise from the same underlying factors; incorporating factors relating to the structure and information density of the records being preserved.

D. *New Horizons*

We highlight the potential of the tool for building a more inclusive digital archival practice for the future.

1. *A New Approach to Trust:* This 'bottom up' data driven approach offers an alternative to our current standards-based model of digital preservation. It is flexible enough to accommodate diverse contexts and different priorities, giving greater control to the archivist.
2. *An Inclusive Digital Archive:* This is primarily a structured and evidence based model, which also makes space for expert judgement for a more nuanced approach to decision making. It can help us analyze our experience, decisions and reasoning to reduce bias and improve transparency and accountability. This will start to narrow the gap between the technical work of the digital preservation function and the archive's wider role within our society.
3. *New Partnerships:* The approach is potentially greater than the sum of its parts. It will operate most effectively when digital preservation specialists pool our knowledge, expertise and data to share insight and experience across archival institutions and the wider digital heritage community.
4. *Building Capability:* This project starts to put quantitative modelling techniques into the hands of archivists, helping us to develop the skills and thinking we will need to build the next generation of disruptive digital archives [3].

References

[1] S. Wright, M Barons, R Wilkerson, J Smith, "Report of National Archives workshop for modelling digital preservation using Bayesian Networks", University of Warwick, TNA internal report, November 2018.

[2] M Burgman, A Hanea, M McBride, B Wintle, "A practical guide to structured expert elicitation using the IDEA protocol", Methods in Ecology and Evolution, 9, 1, 2018.

[3] The National Archives, "Digital Strategy", March 2017, www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Enhancing Services to Preserve New Forms of Scholarship

Kate Wittenberg
Portico, United States
kate.wittenberg@ithaka.org

David Millman
New York University, United States
dsm@nyu.edu
https://orcid.org/0000-0001-5028-1101

Karen Hanson
Portico, United States
karen.hanson@ithaka.org
https://orcid.org/0000-0002-9354-8328

Craig Van Dyck
CLOCKSS, United States
cvandyck@clockss.org
https://orcid.org/0000-0001-6277-6013

Susan Doerr
University of Minnesota Press, United States
doer0012@umn.edu

Abstract – The advance in technologies for publishing digital scholarship has outpaced the development of technologies for reliably preserving it. Authors and publishers are creating increasingly sophisticated products without realizing that some of their enhancement choices might put preservability--and valuable scholarship--at risk. The poster describes the in-progress work and findings of a collaboration between preservation organizations, libraries, and publishers that are creating enhanced digital publications. The work aims to identify what can be effectively preserved with existing technologies, and to produce a recommended set of practices to help authors and publishers prioritize and plan their enhanced digital products for maximum preservability

Keywords – monographs, publishing, collaboration, preservation

Conference Topics – Collaboration: a Necessity, an Opportunity or a Luxury?; Building Capacity, Capability and Community.

I. Background

The advance in technologies for publishing digital scholarship has outpaced the development of technologies for reliably preserving it. Authors and publishers are creating increasingly sophisticated products without realizing that some of their enhancement choices might put preservability--and valuable scholarship--at risk.

As scholarly communication becomes more complex, publishers are creating digital products that are more than mere proxies for a print original; they offer new types of content and user experience. Features of these new forms of scholarship might include audio and video content embedded within text; interactive, high-resolution images, maps, and visualizations; annotations that can be shared, saved, and further annotated, privately or publicly; non-linear paths of engagement; and custom-built digital publications. Specific examples of this form of scholarship can be found in the Library of Open Access Titles of the University of Minnesota Press [1], which uses the open source Manifold platform, Fulcrum of the University of Michigan Press [2], the digital projects of the Stanford University Press [3], and Open Square of the NYU Press [4].

A single publication on these platforms can potentially include many of the innovations mentioned. Each of these innovations presents preservation challenges; their combination creates an even greater challenge: the need to maintain multiple formats and the connections among them.

While there has been remarkable growth within the digital preservation community in developing shared strategies, practices, and tools, this knowledge has matured around a limited scope of file formats--text, audio, video, and image. With the

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

exception of Web archiving, efforts to preserve complex digital objects have been fewer and smaller in scale. Preservation workflows at scale were designed for simpler objects.

II. Project Scope

The poster presents the in–progress work and findings of a project funded by the Andrew W. Mellon Foundation. The *Enhancing Services to Preserve New Forms of Scholarship* project consists of two key sets of activities that will be performed over 18 months, beginning in April 2019.

A. Identify what can be preserved

The first activity is to determine the degree to which the enhanced forms of scholarship *can* be preserved using existing technologies. This will focus on scalable solutions that can be incorporated into the workflows of the participating preservation institutions. The ideal outcome would be a seamless transfer of enhanced publications to the preservation organizations with sufficient actionable information for recreating the user experience, as well as ensuring that their component bits can be preserved.

B. Develop strategies and guidelines

Iteratively and in parallel with the first activity, the project partners will develop a set of guidelines and current best practices identifying the classes of materials and workflows more likely to lead to successful preservation. The guidelines will describe a set of conventions that can serve as a model for other publishers and, potentially, create an environment in which digital scholarship has a better chance of being effectively and sustainably preserved at scale

III. Partners

The project involves collaboration among a group of university presses, libraries, and preservation services who are looking at innovative, effective, and sustainable approaches to preservation of these new forms of scholarship.

The publishing organizations include NYU Press, Michigan Publishing, the University of Minnesota Press, Stanford University Press, and Vega Academic Publishing at Wayne State University.

The preservation service organizations include CLOCKSS, Portico, and the libraries of the University of Michigan, Duke University, and NYU. The team working on the Emulation as a Service Infrastructure (EaaSI) project at Yale University will be consulted for materials demanding either a particular "original experience" or their original software context.

IV. Formats and Features

After considering the composition of a variety of enhanced works, a list of formats and common features was identified for analysis. First, the project will include 3 major formats:

- EPUB3s with embedded multimedia
- Self-contained HTML5 projects that are not dependent on remote resources
- Web-based publications dependent on external content, code, or systems.

In addition, several features that are typical of these works will be given separate attention:

- Works with supporting files -- specifically video, audio, software, code, or data -- that are linked or embedded, local or remote.
- Works with dynamic features more complex than hyperlinks that require content to be retrieved from a remote server based on user interactions. These features could have: (a) a finite number of states e.g. click play to stream video, hover-over for a pop-up; or (b) an open-ended number of states – it's likely that a replica of the live service would be required for the full experience e.g. full text search engine, map navigation, live feed.

Analysis of each work will include consideration of versioning support and the ownership and rights of supporting material. Portico and CLOCKSS will receive or collect representative samples and attempt to ingest them into the archive. The results will contribute to the final guidelines documentation.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

References

[1] Library of Open-Access Titles, University of Minnesota Press. https://manifold.umn.edu

[2] Fulcrum, University of Michigan Press. https://www.fulcrum.org

[3] Stanford Digital Projects, Stanford University Press. https://www.sup.org/digital

[4] Open Square, NYU Press. https://opensquare.nyu.edu

The Australasia Preserves Story

Building a digital preservation community of practice in the Australasian region

Jaye Weatherburn
University of Melbourne Australia
jaye.weatherburn@unimelb.edu.au
orcid.org/0000-0002-2325-0331

Abstract – Building capacity through collaboration is essential to drive successful ongoing digital curation and digital preservation practice. This poster highlights the growth of the Australasia Preserves digital preservation community of practice, an initiative aiming to increase collaborative opportunities for varied institutions and individuals.

Keywords – Community of practice, Collaboration
Conference Topics – Building Capacity, Capability, and Community; Collaboration: a Necessity, an Opportunity, or a Luxury?

I. Introduction

Australasia Preserves is an initiative in the Australasian Region (currently comprising Australia and New Zealand) that initially evolved from the University of Melbourne's Digital Preservation Project work [1]. Throughout 2018 and into 2019 the community has evolved and thrived, due to the enthusiasm and resourcing offered by many individuals and organizations (including state and national libraries and archives, universities, and cultural heritage organizations.)

This poster seeks to generate discussion and solicit input from the iPRES community about how to build on the initial progress and success of this community initiative, particularly with regard to the question: how do grassroots communities of practice continue to grow and provide value to a wide range of varied domains, especially when facing considerable organizational and geographic boundaries?

II. The birth of a community of practice

To enable and support collaborative knowledge sharing and to build a community of practice in Australasia, irrespective of organizational affiliation, the Digital Scholarship team within Scholarly Services at the University of Melbourne organized a half-day inaugural event in Melbourne, Australia, in February 2018 [2]. This event featured speakers from across Australia and New Zealand to launch "Australasia Preserves" as an open community of practice. This community evolved, by attendee demand, into an online forum to maintain and expand the connections made. The result is an active forum of over 200 members across a range of domains.

III. Enabling community growth

Collaborations and partnerships are increasingly essential to enable sharing of digital preservation knowledge, expertise, and technology infrastructure (such as storage, or shared services), especially when facing issues around resourcing. To carry on the momentum generated from the inaugural launch event of Australasia Preserves, during 2018 the University of Melbourne devoted time and resources to host monthly online meet- ups for the community, with a variety of speakers on relevant digital preservation topics.

In addition, two local face-to-face events were organized in partnership with other organizations. In July 2018, the State Library of New South Wales organized and ran the monthly meet-up in Sydney, providing resources and support to live-stream and record the event for those who could not attend in person [3]. In November 2018, the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) organized and hosted the monthly

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

meet- up for World Digital Preservation Day, titled "Digital Preservation for Everyone," in conjunction with the Australian Academic and Research Network (AARNet) who provided the live streaming and recording support for this event [4]. The ability to live stream and record events is essential to contribute to growing a community of practice for digital preservation in Australasia, particularly due to the considerable geographical dispersion in our region.

Based on a very active year of community events, a 2018 "highlights" document was produced, to further share and promote the benefits of engaging in this community initiative [5].

IV. Community engagement and outputs

At the inaugural launch event, there was discussion on the idea of developing "digital preservation carpentry" lessons. These lessons would be taught in the Library Carpentries style [6], focussing on hands-on activities and experimentation with tools to manage content for long-term preservation.

Throughout 2018 more discussion and activities [7] revealed the importance of ensuring that digital preservation concepts were tightly coupled with tool experimentation, so that essential elements such as authenticity and integrity could be clearly demonstrated. We also created a survey [8] to find out what people were most interested in learning

A small group of Australasia Preserves members put together a trial workshop for digital preservation carpentry [9] at the 14th International Digital Curation Conference in early February 2019 [10]. This workshop was well received, with many suggestions for further improvements.

After reflections on the trial workshop were shared at the February 2019 monthly meet-up for the community, the idea to create a working group for digital preservation education arose from the discussion. This working group has now been formed (with members from Australia and New Zealand) to finalise the work undertaken to date on digital preservation carpentry, and to more broadly explore how to develop openly accessible Australasian-focussed training materials.

Also at the 14th International Digital Curation Conference, the Australasia Preserves community participated in the IDCC Unconference [11], running two sessions. The aim of these sessions were to brainstorm topics and ideas for the community to engage with during 2019. We also discussed the challenge of sustainability for communities of practice, and various issues in digital preservation.

In 2018 Australasia Preserves established an annual international award, "Best Digital Preservation Song and Dance for World Digital Preservation Day" [12], with the reigning champions from 2018 [13] preparing to defend their title in 2019.

Also particularly of relevance to the information professions (professions that can on occasion be overly excited by the provision of baked goods), the Australasia Preserves community is honoured to have among its ranks the (unofficial) baking champion of the world, Kirsten Wright, who has produced various spectacular delicacies to celebrate Australasia Preserves milestones – evidence that community growth can be greatly enhanced by coming together over cake [14].

As the Australasia Preserves community has grown, we put in place a community code of conduct to ensure everyone feels welcome to join and participate [15]. We have also developed a Briefing Pack to better share the value of what this community is bringing to digital preservation practice in our region [16], to help advocate for ongoing support for the community, and to drive membership so that more people can come together to share digital preservation knowledge and experience.

Acknowledgments

I would like to thank the University of Melbourne for the support and resourcing that is helping to ensure the ongoing success of the Australasia Preserves community of practice. Special thanks to Matthew Burgess (State Library of New South Wales), Fiona Blackburn and Anthony McLaughlin (AIATSIS) and Adam Bell (AARNet) for organising events for the community throughout 2018. Thanks also to the Australasia Preserves co-organizer team: Rachel Tropea, Lyle Winton, Peter Neish, Matthew Burgess, Lachlan Glanville, Carly Lenz, Jan Hutar, Carey Garvie,

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Elise Bradhaw and Ali Hayes-Brady, and to the wider community members too numerous to name here, who devote time and energy to the online forum and the monthly meetups.

References

[1] J. Weatherburn, "Establishing Digital Preservation at the University of Melbourne," September 2016 [Proceedings of the 13th International Conference on Digital Preservation] http://hdl.handle.net/11343/143364

[2] Australasia Preserves: Establishing a Digital Preservation Community of Practice https://blogs.unimelb.edu.au/digital-preservation- project/2018/03/06/australasia-preserves-establishing-a-digital- preservation-community-of-practice

[3] Australasia Preserves, State Library of New South Wales https://www.youtube.com/playlist?list=PL8vqPRPhXZB_4YcbG-F30sCh932mOFj09

[4] Australasia Preserves at AIATSIS https://www.youtube.com/watch?v=zRgDdkHDadI

[5] Australasia Preserves 2018 highlights https://blogs.unimelb.edu.au/digital-preservation- project/2018/12/18/australasia-preserves-2018-highlights

[6] Library Carpentry https://librarycarpentry.org

[7] Digital Preservation Carpentry https://blogs.unimelb.edu.au/digital- preservation-project/2018/05/14/digital-preservation-sprint

[8] Digital Preservation Carpentry Survey https://bit.ly/2W2PlGj

[9] Digital Preservation Carpentry Workshop Trial http://www.dcc.ac.uk/events/workshops/digital-preservation- carpentry

[10] 14th International Digital Curation Conference http://www.dcc.ac.uk/events/idcc19

[11] Australasia Preserves at IDCC https://blogs.unimelb.edu.au/digital-preservation- project/2019/02/14/australasia-preserves-at-idcc-2019

[12] Awards Ceremony: Best Digital Preservation Song and Dance for World Digital Preservation Day 2018 https://blogs.unimelb.edu.au/digital-preservation-project/2018/12/11/awards-ceremony-best-digital-preservation-song- and-dance-for-world-digital-preservation-day-2018

[13] Digital Preservation Song Challenge https://dpconline.org/blog/idpd/digital-preservation-song-challenge

[14] SpectacularBaked Goods Offering https://twitter.com/jayechats/status/1067868188854018048

[15] Developing a community code of conduct for Australasia Preserves https://blogs.unimelb.edu.au/digital-preservation- project/2018/11/07/developing-a-community-code-of-conduct-for- australasia-preserves

[16] Australasia PreservesBriefing Pack 2019 https://blogs.unimelb.edu.au/digital-preservation- project/2019/02/27/australasia-preserves-briefing-pack-2019

The Web Curator Tool Relaunch

The Next Generation of Web Crawling

Jeffrey van der Hoeven
National Library of the Netherlands
The Netherlands
Jeffrey.vanderhoeven@kb.nl
https://orcid.org/0000-0002-2534-8017

Ben O'Brien
National Library of New Zealand
New Zealand
Ben.O'Brien@dia.govt.nz
https://orcid.org/0000-0002-4290-2972

Abstract – This poster will highlight new features of the Web Curator Tool (WCT), added from January 2018 onwards through a collaboration between the National Library of New Zealand (NLNZ) and the National Library of the Netherlands (KB-NL). One of the themes from the collaboration has been to develop a modern fit-for-purpose WCT. This involves learning the lessons from previous developments, responding to recent trends in the web archiving community and completing a technical uplift. On this foundation a new, revamped WCT has been developed and released as version 2.x. As well as highlighting the latest developments the poster outlines the roadmap and community building planned for the WCT in the coming years.

I. Introduction & background

In 2006 the NLNZ and the British Library developed the WCT, a collaborative open-source software project conducted under the auspices of the IIPC. The WCT managed the web harvesting workflow, from selecting, scoping and scheduling crawls, through to harvesting, quality assurance and archiving to a preservation repository. The NLNZ has used the WCT for its selective web archiving programme since January 2007. However, the software had fallen into a period of neglect, with mounting technical debt: most notably its tight integration with an out-dated version of the Heritrix web crawler. While the WCT is still used day-to-day in various institutions such as the KB-NL, it had essentially reached its end-of-life as it has fallen further and further behind the requirements for harvesting the modern web. The community of users have echoed these sentiments over the last few years.

During 2016/17 the NLNZ conducted a review of the WCT and how it met their business requirements, and compared the WCT to alternative software/services. The NLNZ concluded that the WCT was still the closest solution to meeting its requirements - provided the necessary upgrades could be made to it, including upgrading to the Heritrix 3 web crawler. Serendipitously, the NLNZ discovered that another long-time WCT user, the KB-NL, was going through a similar review process and had reached the same conclusions. This led to collaborative development between the two institutions to uplift the WCT technically and functionally to be a fit for purpose tool within these institutions' respective web archiving programmes.

II. WCT next gen uncovered: version 2.x

The objective of the joint effort of NLNZ and KB-NL is to get the WCT to a platform where it can keep pace with the requirements of archiving the modern web. The first step in that process was to decouple the integration with the old Heritrix 1.x web crawler, and upgrade to the more modern Heritrix 3.x version. Improved ability to configure the crawling variables were also realised. Apart from the technical side, the documentation has been given a major uplift, including updated instructions on installing the new WCT and migrating from older versions to the latest 2.x. The new version of WCT was launched at the end of 2018 [1] and is available on Github[1] as open source.

[1] Web Curator Tool made available on Github, https://github.com/DIA-NZ/webcurator

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

III. Continuing the work, together

Both NLNZ and KB-NL are working to jointly improve the WCT even further and have drawn up a roadmap with further milestones to be delivered in 2019 and beyond. This includes better support for various ways of using the WCT in web archiving by adding predefined user journeys, better support for quality assurance and making WCT suitable for crawlers other than Heritrix. Virtualising WCT by containerizing it is also on the agenda.

With our effort in revamping the WCT we hope to encourage existing WCT-users to upgrade their install base to the latest version and inspire others to start using it and take part in a growing upcoming community dedicated to improving the way we archive the web for generations to come.

References

[1] K. Teszelszky, "Web Archiving Down Under: Relaunch of the Web Curator Tool at the IIPC conference, Wellington, New Zealand," 2018. https://www.kb.nl/en/news/2018/web-archiving-down-under-relaunch-of-the-web-curator-tool-at-the-iipc-conference-wellington-new

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

ARCHIVER

Archiving and Preservation for Research Environments

João Fernandes
CERN
Switzerland
joao.fernandes@cern.ch
https://orcid.org/0000-0002-0445-7038

Jamie Shiers
CERN
Switzerland
Jamie.Shiers@cern.ch
https://orcid.org/0000-0001-8404-3750

Bob Jones
CERN
Switzerland
Bob.Jones@cern.ch
https://orcid.org/0000-0001-9092-4589

Sara Pitonnet Gaiarin
TRUST-IT
Italy
s.pittonet@trust-itservices.com

Abstract – Do you need to acquire standards-based, cost-effective archiving and preservation services? Are ingest rates, data volume and long-term support important to you? The ARCHIVER project aims to introduce significant improvements in these areas of archiving and digital preservation services, supporting the IT requirements of European scientists developing end-to-end archival and preservation services for data generated in the context of scientific research. With a total procurement budget of 3.4 million euros, the project will use a Pre-Commercial Procurement (PCP) approach to competitively procure R&D services from firms in three stages covering design, prototyping and pilot, over a 3-year period (Jan 2019 - Dec 2021). The resulting services will become part of the catalogue of the European Open Science Cloud (EOSC) initiative funded by the European Commission (EC). This contribution will showcase the results obtained during the project phases up to iPRES2019, providing an overview of the PCP process for the supply side and how the wider demand side community can benefit from the ARCHIVER results through the ARCHIVER Early Adopters program.

Keywords – Archiving, Data Preservation, Pre-Commercial Procurement, Tender, EOSC

Conference Topics - Designing and Delivering Sustainable Digital Preservation, Exploring New Horizons

I. Introduction

Currently, many research projects struggle to manage their data, as the archiving and preservation services are inadequate and fall below expectations while data stewardship costs are frequently underestimated.

Using the EC Pre-Commercial Procurement (PCP) [1] instrument, the ARCHIVER project's goal is to fulfil these data management promises in a multi-disciplinary environment, allowing each research group to retain ownership of their data whilst leveraging best practices, standards and economies of scale. The objective is to address the critical gaps between what is increasingly required by funding agencies, requested by data creators and eventual (re-)users and what is currently commercially available.

ARCHIVER will procure R&D services that address the archiving and long-term preservation needs of multiple scientific domains, implemented under the OAIS reference model and related standards [2] supporting FAIR [3] data principles.

II. ARCHIVER approach

Acting as a collective of procurers, the ARCHIVER consortium aims to create an eco-system for specialist ICT companies active in archiving and

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

digital preservation, willing to introduce innovative services capable of supporting the expanding needs of research communities, under a common innovative procurement activity for the advanced stewardship of publicly funded data in Europe.

These innovative services will be ready to be commercialized, by the end of the project (December 2021). The project is coordinated by CERN [5], with a consortium of procurer research organisations (CERN, DESY [6], EMBL-EBI [7] and PIC [8]) and experts (ADDESTINO [9] and Trust-IT [10]) and receives funding from the European Union's Horizon 2020 research and innovation programme [11].

A. *ARCHIVER PCP phases*

The invitation to tender of ARCHIVER will follow the implementation R&D phases foreseen in the PCP instrument:

1. Phase 1 - Solution Design: Provision of a design report including architecture and technical design of components. The activity during this phase, will produce the results to be taken into account in the selection process that allows a contractor to proceed to the subsequent project phase. The expectation is to select multiple designs in order to promote competition and prevent vendor lock-in.
2. Phase 2 - Prototype Development: selected contractors from the Design Phase will build prototypes of the designed solutions and make them available to the procurer organisations forming the buyers group. During the Prototype Phase, basic functionality testing will be performed by specialists from the Procurer organisations. The results of these tests will drive the selection process for the last project phase (Pilot Phase).
3. Phase 3 – Pilot Deployment: Selected contractors will deploy expanded prototype services. These pilots will be essentially pre-production services to be tested in aspects such as performance, scalability and robustness. These services will potentially be exposed to end-users and early adopters, in order to determine if the resulting services are suitable for their needs. ARCHIVER will promote the development of the business models of the resulting services across all phases of the project, requesting contractors to provide total cost of ownership (TCO) and commercialisation plans for their solutions. This aspect is fundamental in order to provide a clear cost perspective to organisations that will purchase the resulting services at the end of the project.

B. *Requirements and R&D scoping*

Before project execution, a preparation phase takes place both for requirement assessment of the demand side and public consultation of the supply side. As announced in the corresponding Prior Information Notice (PIN) [12] published in the Official Journal of the European Union (OJEU) [13], an open market consultation process (OMC) will be carried out prior to the invitation to tender, in order to engage firms in an open dialog and exchange of information with the procurer organisations forming the buyers group. The OMC is organised as an evolutionary consultation process, formed by several sessions either targeted to the demand side or to suppliers, in order to estimate value, complexity and required R&D effort to support the foreseen use cases. The main goal of the OMC is the identification of the innovation potential from a technological point of view. As an additional benefit, the suppliers will also acquire important information about the project use cases, the tender process and the different phases across the project. The results obtained during the OMC will be used as a basis for the Request for Tender specification to be published in October 2019.

In addition to the OMC itself during this period, training sessions will be organized on OAIS conceptual framework for members of the ARCHIVER consortium. The main objective is to promote the reference model and to provide consortium members with the skills and knowledge necessary to respond to emerging issues in digital preservation from an organisational perspective. The Digital Preservation Coalition (DPC) [15], an external expert collaborating with ARCHIVER of which CERN is a member, will provide a range of opportunities for digital preservation training during several phases of the project.

IV. ARCHIVER Innovation Potential

ARCHIVER will contract ICT companies to perform the research and development of innovative services

to archive and preserve scientific data of multiple scientific domains. Although OAIS conformant solutions already exist in the market, it remains to be demonstrated functionality at the petabyte scale of scientific data from multiple research domains, with very high data ingestion rates (10Gbps/day), preferentially via commons solutions, interconnected via the GEANT network. These aspects combined with federated identity management services to support authentication and authorization to the archives and, the exploitation of a hybrid deployment model combining the resources of the scientific organisations on-premises and commercial archiving and preservation services, can create the opportunity to innovate in a number of promising but un-tested aspects in order to develop the next generation of archiving and preservation services. In parallel, these services will be required to be compliant with the current regulations and legislations (e.g. GDPR [16]) and that the associated business models are transparent and well defined taking into account several factors: archives lifetime, maximum ingestion rates, data volumes, number of copies, data portability and exit strategies. The ARCHIVER project will also launch an Early Adopter Programme. Through this initiative, ARCHIVER wants to encourage wider testing of the R&D produced during the project co-funded by the ARCHIVER Buyers Group organisations to any public organisation having a need for digital archiving and preservation services. In order to ensure that the resulting solutions are as widely applicable as possible, ARCHIVER is seeking Early Adopters with the ability and enthusiasm to explore their use in other contexts. TThe ARCHIVER project will also launch an Early Adopter Programme. Through this initiative, ARCHIVER wants to encourage wider testing of the R&D produced during the project co-funded by the ARCHIVER Buyers Group organisations to any public organisation having a need for digital archiving and preservation services. In order to ensure that the resulting solutions are as widely applicable as possible, ARCHIVER is seeking Early Adopters with the ability and enthusiasm to explore their use in other contexts.

The initiative is part of ARCHIVER strategy to make its results have wider impact, in the context of European Open Science Cloud (EOSC).

The EOSC is a component of the 'European Cloud Initiative', as a key part of the Digital Single Market Strategy. The vision of the EOSC is to give the European Union a global lead in research data management and ensure that European scientists reap the full benefits of data-driven science, by providing an environment with open services for data storage, management, analysis and re-use across disciplines.

As part of the EOSC context, ARCHIVER intends to fill the gap by providing a set of services with an aligned set of core requirements for research data archiving and preservation, as well as a list of criteria for the selection of high quality trustworthy repositories where researchers can store their data for sharing, a key building block of the research process.

1. Acknowledgements

The authors would like to thank the ARCHIVER consortium members for the production of this contribution.

References

[1] Pre-Commercial Procurement process: https://ec.europa.eu/digital-single-market/en/pre-commercial-procurement
[2] OAIS Reference Model: http://www.oais.info/
[3] FAIR Principles: https://www.go-fair.org/fair-principles/
[4] EC Directorate-General of Research and Innovation, Prompting an EOSC in practice, DOI: 10.2777/112658 (2018)
[5] CERN: https://home.cern/
[6] DESY: http://www.desy.de/
[7] EMBL-EBI: https://www.ebi.ac.uk/
[8] PIC: https://www.pic.es/
[9] ADDESTINO: https://www.addestino.be/
[10] TRUST-IT: https://www.trust-itservices.com/
[11] Horizon 2020: https://ec.europa.eu/programmes/horizon2020/en
[12] ARCHIVER Prior Information Notice:
[13] https://ted.europa.eu/udl?uri=TED:NOTICE:87071-2019:TEXT:EN:HTML&src=0
[14] OJEU: https://www.ojeu.eu/
[15] Planning poker: https://en.wikipedia.org/wiki/Planning_poker
[16] Digital Preservation Coalition: https://dpconline.org/
[17] GDPR Data Retention Guide: https://www.dpnetwork.org.uk/gdpr-data-retention-guide/

Software Preservation Services in Cultural Heritage Organizations:

Mapping the Landscape

Wendy Hagenmaier
Georgia Institute of Technology, USA
wendy.hagenmaier@library.gatech.edu
ORCID: 0000-0003-4045-0068

Christa Williford
Council on Library and Information Resources, USA
cwilliford@clir.org
ORCID: 0000-0001-6273-3793

Monique Lassere
University of Arizona, USA
mlassere@email.arizona.edu
ORCID: 0000-0001-5395-8974

Lauren Work
University of Virginia, USA
lw2cd@virginia.edu
ORCID: 0000-0002-0941-6921

Jessica G. Benner
Carnegie Mellon University, USA
jbenner@andrew.cmu.edu
ORCID: 0000-0002-2427-1222

Seth Erickson
Pennsylvania State University, USA
sre53@psu.edu
ORCID: 0000-0002-5570-7201

Abstract – Preserving software is a prerequisite for preserving and providing access to digital cultural heritage and research. The recent formation of the Software Preservation Network (SPN) has provided momentum for a better understanding of the landscape of software preservation activities. This poster discusses preliminary results from a study undertaken by SPN's Research Working Group. Our specific research questions are: What software preservation services are cultural heritage professionals currently providing? What are the gaps in services? What are the opportunities for future service provision? Our Service Provider Study focuses on software preservation activities happening in libraries, archives and museums. This study will inform a foundational agenda that SPN members and other cultural heritage professionals can use to conduct further research on sustainable software preservation services.

Keywords – software preservation, software, libraries, museums, archives

Conference Topics – Exploring New Horizons; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

Preserving software is a prerequisite for preserving and providing access to digital cultural heritage and research, and software is increasingly considered a research product or artifact in itself. For decades, researchers and practitioners in information science, digital preservation, and allied fields have discussed the necessity of software preservation.

The recent formation of multiple groups focused on software preservation--including the Software Sustainability Institute, which focuses on research software [1], Software Heritage, which aims to preserve software as cultural artifacts [2], and the Software Preservation Network (SPN) [3], which is an alliance of cultural heritage professionals and others interested in software preservation--has provided momentum for better understanding the landscape of software preservation activities.

SPN's mission is to preserve software through community engagement, infrastructure support, and knowledge generation. Our mission as the

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Research working group is to facilitate research projects that bring individuals and organizations with diverse perspectives and interests together to document and analyze the landscape of software preservation and access.

This year we, the members of the SPN Research Working Group, launched a study of software preservation service providers in libraries, archives, museums, and other organizations who work to preserve software, in order to identify (1) the services that are currently being provided, (2) gaps in services, and (3) opportunities for future service provision. Ideally, it will serve as the first in a longitudinal series of studies that will track the landscape of software preservation over time.

II. Service provider

In 2015, Meyerson and Vowell conducted a survey [4] to better understand cultural heritage practices surrounding long-term preservation and access to digital primary resources stored in proprietary file formats. While this study produced information on who is collecting software, it didn't solicit detailed data about what they are collecting and how they are collecting it. Furthermore, the landscape of software preservation activities has evolved since this initial research.

The Research Working Group's Service Provider Study seeks to dig into details about how software preservation work is currently being done. The study includes a survey and set of follow-up interviews aimed at documenting how cultural heritage organizations are approaching software preservation services. Our specific research questions are: What software preservation services are cultural heritage professionals currently providing? What are the gaps in services? What are the opportunities for future service provision?

The survey questions were designed to gather data to address each of these research questions. Our target population are individuals working in cultural heritage organizations such as libraries, archives, and museums. We received Institutional Review Board approval for the study and adhered to GDPR protocols for all European responses. The survey includes multiple choice and free-response questions with the goal of obtaining a high-level understanding of the types of software preservation services participants are providing, as well as services participants plan or hope to provide in the future. The interview protocol includes a set of questions that address the same topics covered in the survey but enable the participant to answer in more detail.

III. Preliminary Insights

We launched the survey in January of 2019, and as of June 2019, we are in the early stages of analysis and have conducted most interviews. Our survey received 124 responses from a variety of institutions, including academic research universities (57.3%), government entities (19.4%), and a variety of "other" organizations including public libraries, entertainment companies, museums, and commercial entities (12.1%). Fifty-five respondents (44%) affirmed that they were currently providing software preservation services. To date, we have also completed fourteen interviews with survey respondents who indicated their willingness to do so.

Survey responses reflect perspectives from a wide variety of job roles including: librarian (24.2%), archivist (24.2%), information technologist (13.7%), curator (8.1%), administrator (8.1%), conservator (4.0%), and others (17.7%), including digital preservationist, software preservationist, professor, program manager, digital asset management specialist, archaeologist, scientist, data specialist, data preservationist, photographer, and collections manager. The top two motivations cited by respondents who preserve software are to provide access to other collections (58%) and because the software itself is a part of the cultural record (51%), although preserving software so it can be reused was also a significant concern (47%).

While a range of institutions and professionals are considering or providing services related to collecting, describing, preserving, and providing access to software, our study suggests that only a small minority are engaging in these activities on a frequent basis. Of respondents who self identified as currently providing services, only small numbers chose the word "frequently" to characterize how often they:

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

- provide consultations about software preservation (8.1%)
- actually preserve software (10.5%)
- create metadata for preserved software (7.3%)
- provide access to software (6.5%).

Several professional development and capacity gaps were identified in both the survey responses and interview conversations. Many respondents felt a need for knowledge about the history of computing to do their work, wanted access to best practices for the field, and indicated they lacked sufficient inventories to know what software is included in their collections. Those with more established software preservation programs expressed an interest in developing capacity for providing emulation environments for researchers. Given the variety of needs reflected in the preliminary results, the study suggests that community organizations like SPN can play an important role in addressing gaps for both new and experienced service providers. This study will help inform a foundational agenda that SPN members and other cultural heritage professionals can use to conduct further research as well as to develop, refine, and promulgate good practices for sustainable software preservation.

References

[1] Software Sustainability Institute, 2019, Homepage, https://www.software.ac.uk

[2] Software Heritage, 2019, Homepage, https://www.softwareheritage.org

[3] Software Preservation Network, 2019, Homepage, http://www.softwarepreservationnetwork.org/

[4] Meyerson, J. and Vowell, Z. (2015). Software preservation in cultural heritage survey. https://docs.google.com/document/d/1l70lIFKpdh-vItKx8SwgNwGP8pm8cdwb1kvCMc1CKYUg/edit

Creating Continuity for Digital Preservation Projects

Steps for archiving project outputs when Business-As-Usual has not yet been achieved

Edith Halvarsson
Bodleian Libraries
University of Oxford United Kingdom
edith.halvarsson@bodleian.ox.ac.uk
orcid.org/0000-0003-4634-3140

Sarah Mason
Bodleian Libraries University ofOxford
UnitedKingdom
sarah.mason@bodleian.ox.ac.uk
orcid.org/0000-0002-3156-1586

Abstract – This poster abstract summarises how theDigitalPreservationatOxfordandCambridgeProjectself- archived its research outputs, with the aim of extending the impact of digital preservation activities at the end of theproject.Keywords – digital preservation, research data man- agement, digital archiving, project outputs

Conference Topics – Designing and Delivering Sustain- able Digital Preservation; Building Capacity, Capability and Community

I. Introduction

Digital preservation requires ongoing commitment to providing stewardship for digital collections over time [1,p. 128]. However, digital preservation is still not business- as-usual (BAU) for many organisations. One-off project funds is a common model for supporting digital preser- vation activities, but one which comes with risks to con- tinuity. Digital preservation research outputs can eas- ily become orphaned as project governance boards are dissolved and fixed-term staff move on to new roles [2,p. 242]. In this way, digital preservation projects are exposed to exactly the same risks as other academic projects. Questions of research output sustainability of- ten does not come to the forefront until project funding streams end [3, p. 62].

This is an issue which should be taken seriously by organisations running digital preservation projects, as they are not immune. Issues such as web technology obsolescence, linkrot, as well as institutional memory loss has affected past digital preservation projects.[1] The Internet Archive is in some cases the only method of accessing these resources when there is no remaining digital preservation staff at an organisation.

The Digital Preservation at Oxford and Cambridge (DPOC) project ran for 2.5 years (from July 2016 to De- cember 2018). DPOC aimed to break the project cycle for digital preservation at Bodleian Libraries (Oxford) and Cambridge University Library and establish BAU pro- grammes [5]. Despite successfully acquiring further fund- ing for digital preservation through the development of local business cases, the funding still sits within a project based framework. Acknowledging the risks which the project model poses to the continued stewardship of research outputs, the DPOC project undertook a self- archiving component in October-December 2018

II. Purpose

The purpose of the DPOC self-archiving exercise was to improve the usability, findability, and citability of digital preservation research and knowledge gathered during the project. The intent was to make future

[1] Examples of this type of loss includes the CASPAR project and DigCurV project. CASPAR's website http://www.casparpreserves.eu/ has been inaccessible since 2015. DigCurV reports were published via http://www.digcur-education.org, which now redirects to a simplified WordPress site [4]. Some of the reports are now not available on the new site or from the Internet Archive's snapshots.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

stewardship and decision making about project resources easier for staff at the libraries by making content 'ready for preser- vation'. By undertaking the exercise, the libraries can also use it as a practical applied example when speaking to academics about digital preservation and end-of-life scenarios for research projects.

This poster outlines the steps which the DPOC project took in order to self-archive project outputs. It is the hope of the authors that presenting this process will spark discussion around how the digital preservation field can manage continued access to its intellectual out- puts in an environment where digital preservation BAU is still not the norm for many organisations.

A. Step 1: appraisal of internal project outputs

In the 2.5 years that DPOC ran it created a large body of internal documentation as an outcome of its research activities. The appraisal of research outputs was recognised as a key deliverable of the self-archiving component to ensure that staff who are not familiar with the DPOC project can easily navigate and make use of project outputs. The basis of all appraisal decisions was the project's Communication Plan developed at the start of DPOC. The Communication Plan outlined the intended use of all communication platforms (including email, in- stant messaging applications, task management tools, and SharePoint). It also outlined backup schedules, file naming, and versioning methods.

This was the most time consuming aspect of the self- archiving component as each document and tool was individually assessed. However, as file naming and ver- sioning rules had been largely adhered to the prepara- tion of content was made simple.

B. Step 2: preparing content

The second step was to create additional metadata and prepare content for the Libraries' RMS or institutional repositories – Cambridge University's Apollo and the Oxford University Research Archive (ORA).

Apollo and ORA were chosen as the natural homes of any externally facing project outputs as each individual dataset could be assigned a persistent identifier and be included in any future digital preservation actions performed on content in the repositories.

The Libraries' internal Wiki was used as the main source of internal information regarding the DPOC project. An about page regarding key deliverables of the project was added to the Wiki, which integrates with the Libraries' SharePoint instance and also links out to any code stored on the Libraries' internal GitLab. Addi- tional information regarding each document was added to SharePoint to ensure that documents remain under- standable even if the link between the Wiki and Share- Point breaks.

C. Step 3: social media and blog

The DPOC blog (www.dpoc.ac.uk) was the main method of exposing the DPOC research outputs before they were moved into the institutional repositories. www.dpoc.ac.uk is a WordPress site maintained by Bodleian Libraries. While the WordPress platform is among the less complex applications for Bodleian Li- braries to maintain, it is still an application based plat- form which requires ongoing maintenance that may alter the functionality, look and feel of the DPOC blog. It can- not be guaranteed that files uploaded to the blog remain accessible and persistently citable over time.

For this reason, a snapshot was taken of www.dpoc.ac.uk using Archive.it. Once the snapshot is complete a static version of the WordPress will be created and the interactive features (such as blog comment functionality and Twitter integration) will be removed.

A social media dataset was also created in ORA which contains statistics, individual blog downloads, and Twitter identifiers for content covering the period July 2016 to March 2018. If www.dpoc.ac.uk needs to be closed in the future due to obsolescence or security concerns, the libraries can instead direct users to the social media dataset without having to undertake further work.

D. Step 4: stewardship

Finally - ownership of documents, repository datasets and the WordPress was formalised and assigned to a se- nior member of library staff. This role, or a successor to the role, will make curatorial and preservation decisions about any DPOC project outputs (as well as other follow on digital preservation project outputs) going forward.

III. Conclusion

Continuity of research outputs is a challenge when digital preservation still operates within a funding envi- ronment heavily reliant on project one-off monies [3, p. 62]. Although the goal of digital preservation is ongoing stewardship, applying research data management and self-archiving approaches may be necessary. Applying these methods may help an organisation retain some institutional memory around digital preservation, so that it does not have to start from scratch, even where there are gaps in project funding.

Acknowledgment

The authors would like to thank team members and project managers on the DPOC project, in particular Somaya Langley for her wealth of knowledge on digital preservation projects.

References

[1] E. M. Corrado and H. M. Sandy, Digital Preservation for Libraries, Archives, and Museums, 2nd ed. Lanham: Rowman & Littlefield, 2017.

[2] E. Halvarsson, S. Mason, L. Pretlove, D. Gerrard,

[3] S. Langley, and J. Mooney, "Could collaborative re- search between two major libraries help consoli- date digital preservation and break the "project cy- cle"?" In Digital Preservation in Libraries: Preparing for a Sustainable Future, J. Myntti and J. Zoom, Eds., Chicago: ALA Editions, 2019, pp. 241–261.

[4] S.Langley and E. Halvarsson, "Planning for the end from the start: An argument for digital steward- ship, long-term thinking and alternative capture ap- proaches," in Proceedings of DCH 2o17, 2017, pp. 62– 63.

[5] Digcurv: Digital curator vocational education europe. [Online]. Available: http : / / schreibman . eu / digcurv/.

[6] Digital preservation at oxford and cambridge. [Online]. Available: https://www.dpoc.ac.uk.

Long-term Preservation of PDF files in institutional repositories in Japan

Teru Agata
Asia University, Japan
agata@asia-u.ac.jp
0000-0002-1593-9323

Yosuke Miyata
Teikyo University, Japan
m@miyay.org
0000-0002-5239-5396

Atsushi Ikeuchi
University of Tsukuba, Japan
atsushi@slis.tsukuba.ac.jp
0000-0002-2601-3892

Abstract – In the open access environment, many textual resources have become available in the PDF format on the Web. This research aims to survey PDF files in Japanese institutional repositories (IRs) to address the problems encountered during their long-term preservation. With that aim, 1.5 million PDF files collected from Japanese IRs were analyzed with regard to file format, encryption, and metadata. Most PDF files did not conform to PDF/A. A total of 30.5% of PDFs were encrypted and many PDFs did not have embedded metadata. These results imply that PDF files in Japanese IRs have several serious problems for their long-term preservation.

Keywords – PDF, Metadata, Institutional Repositories, Long-term preservation

Conference Topics – Building Capacity, Capability and Community

I. Introduction

In the open access environment, many textual resources have become available in PDF format on the Web. It is desirable for science to sustain an open access environment. If IRs were impermanent, it becomes essential that PDF files are unencrypted and have accurate metadata. For example, PDF/A is regarded as the preferred format for digital text preservation in several recommendations including [1]. Encryption and poor metadata make the document less accessible and less searchable.

It has been pointed out that some PDF producing software create invalid files and research indicates that the quality of the PDF files is heterogeneous [2]. Termens et al. explored file types and PDF handling in two Spanish IRs and found many PDF files to be encrypted [3]. As the two repositories showed different tendencies for file handling, analyzing more repositories can potentially reveal a general approach for PDF handling.

The heterogeneity of PDF files is also an issue with PDF-embedded metadata. For example, reference manager services usually extract metadata embedded in PDF files. Serious problems can arise when the embedded metadata is different from the actual content. Sometimes PDF-embedded metadata can cause security issues. For example, Society for Industrial and Applied Mathematics releases guidelines for protecting referee personal information [4]. When PDF gets released on the Web, differences between published metadata and PDF-embedded metadata may create confusion.

This research aims to address the problem of the PDF's preservability by surveying PDF files in Japanese IRs. We investigate the following research questions:

RQ1 How many PDF files are prepared in preservable format?
RQ2 How many PDF files are encrypted in preservable manner?
RQ3 How many PDF files have good enough embedded metadata to go alone on the Web?

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

II. Method

We collected and analyzed the PDF files as follows:

1. **Collecting Metadata Records:** Metadata records were collected from IRs in Japan. In February 2019, we harvested metadata records from 582 IRs via ListMetadataRecords of Open Archives Initiative Protocol for Metadata Harvesting. All metadata was in the junii2 format that was prepared for JAIRO, which in turn is a federated search system of Japanese IRs. This format contains "fulltextURL" element for URL of full text files. We collected 2,103,600 metadata records and 1,556,390 of these records had URLs. According to the statistics from JAIRO, this accounted for 74% of all metadata records in Japanese IRs in February 2019.

2. **Collecting PDF Files:** After collecting metadata records, PDF files were collected. We extracted 1,556,390 URLs from the fulltextURL element and tried to download all of them. Factors such as change of IRs' URL and invalid URLs prevented downloading. A total of 1,509,767 files were downloaded. Most of the files were PDFs, but some were non-PDF file formats.

3. **Analyzing PDF Files:** Different types of information was extracted from PDF files using iText 7.1.0 library: PDF file information for RQ1, security information for RQ2, and metadata for RQ3. Some files failed to parse as PDF for various reasons including an invalid string in the PDF file. Finally, we analyzed 1,411,082 PDF files. TABLE I show the basic statistics of the files examined.

III. Results

Only 0.9% of PDF files conformed to PDF/A (or PDF/B or PDF/UA), which is specialized for long-term preservation of electronic documents. 11.2% were tagged (structured) PDFs making them accessible to the visually impaired.

As shown in TABLE III, 30.5% of PDF files were encrypted. The "Printing not allowed" setting prevents open access. The "Screen Readers not allowed" setting prevents visually impaired people from extracting text from PDF file. Furthermore, encrypted PDF files cannot be converted to any other format in future.

To allow a PDF file to be distributed separately from the metadata of the IR, it is desirable that the PDF file have enough and accurate embedded metadata. Many PDF files did not have metadata embedded in "Document Information" part (48.9% had creator, 17.9% had title, and 1.5% had key words). In contrast, 35.7% of PDFs had creator metadata embedded in XMPMetadata part. Various types of PDF producing software were used as in a previous study [2]. Even if the metadata was embedded, the metadata was not necessarily accurate. Some PDF files still had their original file's metadata and had not been updated to match the current file's metadata. However, PDF files created by academic commercial publishers like Elsevier contained a lot of rich and accurate metadata.

IV. Summary

The answers to our research questions are as follows:

RQ1 Few PDF files were made in a format that was appropriate for long-term preservation.
RQ2 30.5% of PDF files are encrypted and they might prevent future reuse.
RQ3 Many PDF files did not have sufficient metadata for distributing them separately from the metadata of the IRs.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

Our results show that PDF files in Japanese IRs have several serious problems for their long-term preservation.

TABLE I
Basic Statistics

Metadata records in Japanese IRs	2,103,600
Full text URLs	1,556,390
Downloaded files	1,509,767
Downloaded PDF Files	1,509,470
Parsed PDF Files	1,411,082

TABLE II
PDF File Information for Long-Term Preservation

PDF/A	0.9%
Tagged PDF	11.2%

TABLE III
PDF Security Information

Encrypted	30.5%
Printing not allowed	0.6%
Screen Readers not Allowed	1.3%

TABLE IV
PDF Document Information and XMPMetadata

Creator in PDF Document Information	48.9%
Title in PDF Document Information	17.9%
Keywords in PDF Document Information	1.5%
Creator in PDF XMPMetadata	35.7%

References

[1] Smithsonian Institution Archives. "Recommended Preservation Formats for Electronic Records." Internet: https://siarchives.si.edu/what-we-do/digital-curation/recommended-preservation-formats-electronic-records [Mar. 16, 2019].

[2] H. Helin, K. Koivunen, J. Kylander, and J. Lehtonen, "402.2 PDF Mayhem: Is Broken Really Broken?," 27-Dec-2018. [Online]. Available: osf.io/fzxc9.

[3] M. Termens, M. Ribera and A. Locher. "An analysis of file format control in institutional repositories", *Library Hi Tech, Vol. 33, pp.162-174, 2015.*

[4] Society for Industrial and Applied Mathematics. "Protecting Referee Personal Information." Internet: https://www.siam.org/Publications/Journals/Related/Journal-Policies/Detail/protecting-referee-personal-information*, [Mar. 16, 2019].*

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

Concept of a Preservation System for Scientific Experiments in HPC

Kyryll Udod
Ulm University
Germany
kyryll.udod@uni-ulm.de
https://orcid.org/0000-0002-3506-7980

Volodymyr Kushnarenko
Ulm University
Germany
volodymyr.kushnarenko@uni-ulm.de
https://orcid.org/0000-0001-7427-2410

Stefan Wesner
Ulm University
Germany
stefan.wesner@uni-ulm.de
https://orcid.org/0000-0002-7270-7959

Abstract – this poster presents a concept of a preservation system for computations on High Performance Computing (HPC) resources. It covers some important challenges and possible solutions related to the preservation of scientific experiments on HPC systems for their further reproduction. Storage of the experiment as only a code with some related data is not completely enough for its future reproduction, especially in the long term. Preservation of the whole experiment's environment (operating system, used libraries, environment variables, input data, etc.) using containerization technology (e.g. Docker, Singularity) is proposed as a suitable solution for that. This approach allows to preserve an entire environment, but leaves a problem, how to deal with the commercial software that was used within the experiment. As a solution authors propose to replace during the preservation procedure all commercial software with their open source analogues, what should allow future reproduction of the experiment without any legal issues. The prototype of such a system was developed, the poster provides a scheme of the system and the first experimental results.

Keywords – HPC, reproducible research, containerization, research experiments preservation

Conference Topics – What is emerging practice in software preservation and in emulation/ virtualization?

I. Introduction

High Performance Computing plays an important role in almost every research area providing to the users always suitable hardware and software resources to solve complex scientific problems. Because of the continuously growing community of HPC users and amount of research experiments, a question of research data management on HPC starts to play a significant role [1] including reproducibility of the research results as a major aspect for the scientists [2]. Containerization technology (e.g. Docker, Singularity) could be used to preserve a complete environment of the scientific experiment. This approach works well, especially when within the experiment only open source software is used. But in the case of commercial libraries some legal issues can come by the future reproduction procedure. To make a preservation of the commercial software easy and free of any legal problems, some special solution is needed.

This poster begins with generally available preservation options and related to them problems, potential solutions are discussed. Authors touch a question, what type and scope of the preserved information is needed to allow further reproduction of the experiment and how this information could be taken. In the final part the poster presents a scheme of the system prototype that was developed to solve the mentioned above problems related to the preservation and further reproduction of scientific experiments, where also commercial software was involved. The first experimental results and further investigation steps are discussed as well.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

II. Mechanisms for Research Preservation on HPC Systems

To make a scientific experiment reproducible, also software and hardware information should be preserved [3] - information about the operating system and all used within the experiment software libraries, as well as information about the hardware components and configuration aspects of the current HPC system.

Related to the experiment information usually can be extended from the job-script, what can include the used in the experiment software libraries with specific versions, environment variables, etc. Reproduction of the experiment in this case could be possible, but only on the same machine with the same (not changed) configuration of the system, what is practically not possible for the long term because of continuous system updates.

That's why for the long term preservation not only information about the experiment, but also about the whole system components (in a specific for the experiment state) should be stored. In this case containerization technology (e.g. Docker, Singularity) could be used. It allows to create a full copy of the system with all related and used within the experiment software components

III. Problems with the Containerization Approach

With the containerization approach often unlimited access to the preserved components is needed, what can be difficult in the case of HPC systems, where not all software components can be copied or even accessed by the user.

Even when software components are accessible, they could be not open source and some license could be needed for their further reuse. These two issues stay as a central point of the poster's topic

IV. Proposed Approach for Preservation of Software Components with Limited Access

To preserve not fully accessible for the user (because of the account rights) or some commercial software components, authors propose an approach, that all such components should be replaced with their open source alternatives, which are always accessible, free to use and can be stored and reused later without any legal issues.

To collect all needed for the preservation information about the experiment, some available for the user mechanisms should be used, e.g. status request for the current computational job, which represents the experiment (e.g. via "checkjob" or similar command that is traditionally available on HPC systems as a part of the job scheduler).

A proposed preservation system consists of two components. One component is a special script that requests the job status and collects all needed information about the experiment locally on the computing machine (HPC cluster). The second component represents an external server. The server provides a REST-API that can be used for the communication with the client part - the first component on the side of the HPC cluster. Information from the cluster is sent via the POST request. The server is responsible for the containerization procedure and replacement of the commercial libraries with their open source alternatives. The process works automatically, but the user can also steer it via the web-interface (e.g. to replace some software or choose the most suitable version of it). Further running procedure for the created containers can be performed via the EaaS system (Emulation as a Service) [4]. Publication and referencing of the preserved containers are foreseen.

For the current moment the proposed preservation system is in an early prototypical phase, as test experiments some molecular dynamic simulations with "SIESTA" [5] are used.

V. Acknowledgements

The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK).

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

References

[1] B. Schembera and T. Bönisch, “Challenges of Research Data Management for High Performance Computing,” in Research and Advanced Technology for Digital Libraries, 2017, pp. 140–151.

[2] S. Hunold, “A Survey on Reproducibility in Parallel Computing,” arXiv:1511.04217 [cs], Nov. 2015.

[3] P. F. Klaus Rechert and Tom Ensom, “Towards a Risk Model for Emulation-based Preservation Strategies: A Case Study from the Software-based Art Domain”, 13th International Conference on Digital Preservation (iPRES2016), Bern, Switzerland, 3-6 October 2016.

[4] http://citar.eaas.uni-freiburg.de/

[5] https://departments.icmab.es/leem/siesta/

Malware threats in digital preservation

Extending the evidence base

Maureen Pennock
The British Library
United Kingdom
maureen.pennock@bl.uk
https://orcid.org/0000-0002-7521-8536

Michael Day
The British Library
United Kingdom
michael.day@bl.uk
https://orcid.org/0000-0002-1443-5334

Evanthia Samaras
University of Technology Sydney, Australia
evanthia.samaras@student.uts.edu.au
https://orcid.org/0000-0001-5683-722X

Abstract – Virus checking is an established process in most pre-ingest digital preservation workflows. It is typically included as part of a general threat model response and there has to date been relatively little research into the virus checking function specifically within a long term context. The British Library recently began a small research project to explore this issue, using data from a legacy digital collection established by the 'Flashback' project and supplementary data provided by the UK Web Archive. Our poster presents this research and findings to date, raising questions about the overhead of virus checking at scale, when organizations should virus-check content, and the legacy capabilities of anti-virus software.

Keywords – digital preservation, malware, Flashback, virus checking

Conference Topics – Exploring New Horizons; The Cutting Edge – Technical Infrastructure & Implementation.

I. Introduction

Memory organizations typically conduct malware checks on collections as a pre-ingest step in digital preservation workflows, so that 'infected' materials can be identified and dealt with before they would be able to inflict damage on users' computers or the organization's network. Such a step is widely considered to be best practice; however, it is generally employed in response to a general threat model and there is less clarity about the actuality of malware risks over the longer-term, particularly with legacy digital content that is rarely accessed. Many institutions thus incur virus-checking overheads during ingest processing, when it is less than clear exactly how malware risks might manifest. Malware is, effectively, something of an underexplored question in evidence-based digital preservation practices.

Acknowledging this, in late 2018 the British Library initiated a small research project that focused specifically on the virus-checking function within digital preservation and processing workflows, including a three month PhD research placement. This poster introduces that work, provides a high level overview of malware in the Library's legacy 'Flashback' corpus and in the UK Web Archive, and explores some of the research questions that the project seeks to address

II. What is malware?

Malware is a catch-all term that refers to various types of 'malicious software,' including viruses, worms, Trojan horses, spyware and ransomware. It is often designed to exploit specific computing environments or applications. Viruses and other types of malware can not only disrupt or damage these environments, they may also provide a means for unauthorized access to computer environments and/or information. According to Farbowitz [1], the "two characteristics common to most malware are that it operates without the consent of the computer user or network administrator, and that much of it self-replicates in order to spread."

Malware and viruses can be identified and

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

removed by anti-virus (AV) software. These use multiple techniques to counter the threat of malware. For example, AV software providers identify and collect 'signatures' that correlate to specific instances of malware. The signature databases for the products grow over time and are regularly updated with new definition files. AV software also uses heuristic analysis to identify malware, examining code for suspicious properties.

III. Malware at the British Library

While the British Library does not actively collect malware as collection items, malware is known to exist as part of other items acquired by the Library for preservation. This research is based on malware identified in two key datasets: the Flashback corpus and the UK Web Archive.

The Flashback corpus consists of disk images of legacy digital content that was originally acquired by the British Library on handheld media such as floppy disk or CD. The corpus dates from between approximately 1980 to 2010 [2]. The UK Web Archive is a more recent dataset, comprised of millions of UK-relevant websites harvested annually under the UK Non-Print Legal Deposit Regulations (2013). Collectively these two sources provide a sample of both legacy and contemporary malware, delivered and acquired through common malware distribution mechanisms for their respective periods.

The Flashback corpus has been the main focus of the analysis to date. At the time that this analysis was undertaken (February 2019), a total of 22,061 disks had been imaged, roughly half of which were CD-ROM or CD-R, just under a third were 3.5" floppy disks, and the remainder were either 5.25" floppies or Digital Versatile Discs (DVDs). All of the disks were virus-checked prior to being imaged. Of these, 120 of the discs were identified as containing some kind of malware. While this would inevitably include an unquantified number of false-positives, this means that 0.54% of the Flashback disk images have so far been classified as containing malware.

The majority of malware-positive files identified within the Flashback corpus to date have been found on disks previously attached to consumer computer magazines, mainly CD-ROMs and DVD-ROMs published during the 2000s, although some have also been located on non-magazine-based items. Not all of these files, however, represent the same level of risk. For example, three of the magazine cover disks contained joke programs, which seem to have been intentionally distributed with the disks. However, a further item, a disk acquired with PC Gamer No. 58 (July 1998, Disk B), was found to contain three files infected with the Marburg virus, designed to run on Windows 95/98. The visual effect of the Marburg virus is repeated instances of the Windows error icon overlaid on screen content, though it was also found to delete integrity bases of known AV products [3]. In addition, a large number of files contained macro viruses written for MS Word, although there are also many examples of files identified as Trojans and, at the other extreme, a file containing a fragment of a boot-sector virus that was apparently frequently encountered 'in the wild' in the floppy disk era.

Rates of malware in the UK Web Archive are calculated differently from those in the Flashback corpus. The Flashback concept of an 'item' does not transfer to the UK Web Archive, as the boundaries of any given website are fluid, due to the inherent 'linking' system upon which the web is reliant. Malware data for the UK Web Archive is thus currently only available as a percentage of the overall size of the data. We refer to this as the percentage of the stored data flagged as malware-infected, rather than the collection. Up until the end of 2017, malware in the UK Web Archive accounted for less than 0.01% of the stored data.

IV. Research questions

The percentage of stored data from the UK Web Archive is significantly lower than the percentage of malware infected items in the Flashback corpus. That may be explained by the difference in measurement processes. Further work is planned in coming months to address this, as is further identification of the various malware and their function. Nonetheless, the relatively high percentage of malware-flagged items found in the Flashback corpus indicates that virus checking remains necessary. Given the inevitable processing overheads of virus checking at scale, and considering that access rates for content are relatively low, we intend to explore whether it might be more efficient to virus check primarily at

the point of access, rather than at ingest. To answer this question we need a better understanding of at least two things: a) the threat model for storing malware in a long-term digital repository; and, b) the performance of virus checking software over time.

Threat-model analysis will consider, for example, the functionality of the malware found in our samples and the likelihood of malware activation within a controlled read-only storage environment. Regarding performance of AV software over time, we expect that the capabilities of AV software to identify malware should improve as it encounters more samples: this would suggest that delaying the virus-scanning process may result in a more comprehensive output (including a reduction in the number of 'false positives'). On the other hand there is a gap in the literature to definitively confirm that legacy signatures relating to obsolete malware are never removed from providers' databases: this may result in a limit as to how long it is viable to delay the virus checking function. Additional evidence is needed either way.

The poster will reflect up-to-date findings for these research questions and will prompt valuable debate and discussion with conference attendees to help drive this currently under-researched area forwards.

REFERENCES

[1] J. Farbowitz, *More than digital dirt: preserving malware in archives, museums, and libraries*. MA Thesis, New York University, 2016, https://archive.org/details/16sThesisFarbowitzFinal

[2] M. Day, et al., "The preservation of disk-based content at the British Library: Lessons from the Flashback project," *Alexandria: The Journal of National and International Library and Information Issues*, vol. 26, no. 3, pp. 216-234, 2016, https://doi.org/10.1177/0955749016669775

[3] F-Secure Knowledge Base: Marburg, https://www.f-secure.com/v-descs/marburg.shtml

Digital Preservation in a High Security Environment

Student Records, Encryption, and Preservation

Annalise Berdini
Princeton University, USA
aberdini@princeton.edu
https://orcid.org/0000-0001-5385-7356

Abstract - For the past five years, Princeton University Library – specifically the University Archives – has striven to create a robust digital preservation program for its born-digital and digitized records. Due to lack of time and available staff, the Library decided that a third party digital preservation service would be the best solution. It could be acquired relatively quickly, and it wouldn't require asking for funding to hire multiple dedicated staff to build a home-grown digital preservation system. Obtaining buy-in from stakeholders and finding a service that met Princeton's needs proved to be a challenge, especially due to the sensitivity of student records in the collections, which would require a high level of privacy and encryption key maintenance in addition to standard data integrity and preservation processing tools. Throughout 2018, Princeton worked to partner with a service previously unavailable due to University data sovereignty requirements – Arkivum's Perpetua – and work began to develop a Princeton-specific solution that met the needs of the University: Most importantly, geographically dispersed cloud storage locations, Princeton-based control over data integrity and authenticity checks, an encryption key management system for student records maintained by the vendor but managed by repository staff, and a reliable and quick exit strategy.

Keywords – Encryption, student records, privacy, security, preservation

Conference Topics – Exploring New Horizons; The Cutting Edge: Technical Infrastructure and Implementation

Putting the case together for obtaining a digital preservation system was key to the success of the project. Outreach and advocacy for investment in digital preservation is challenging even for institutions where funding is available, and is dependent on more than whether or not the need can be proven and demonstrated. While Princeton is a well-funded institution, proving the case for digital preservation as a concept – and what that preservation system would look like – was still necessary. This process required archivists and IT staff to work together with the product vendor and a special campus-wide digital architecture and security review board, which reviews any digital product or service that pairs with University digital servers or content, and whose approval is required for new services and processing tools.

Student records and their requirements drove much of the preservation system review process and helped bolster the argument for a digital preservation program. Although the most recent push for a digital preservation program began in full force in 2018, Princeton University Library staff had been advocating for a digital preservation system for over five years. Princeton University Archives is responsible for the preservation and accessibility of Princeton student records, which may include anything from student organization records to student academic files and transcripts. Many of these records are mandated collecting materials, designated by the institution as essential to preservation of the history of the business of the University, and in the case of the academic or disciplinary files, are subject to strict confidentiality and security requirements -- some federally imposed. These restrictions are relatively easy to apply to analogue records, which may be physically secured in a pass-coded vault or

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

restricted area in the stacks, with little chance of an outside user finding and being able to gain access to large swaths of the records.

However, the challenges of maintaining that level of security for born-digital records are higher, as they include serious concerns about hacking or unintentional leaks of large quantities of data. Many archivists and repositories have reservations about data encryption, which effectively alters the original data so as to be unreadable. It is (usually) secure, but will it endanger the authenticity and integrity of the data? Who will have access to the keys, and how will they be managed for ever-increasing blocks of data? Working with Arkivum to find an acceptable solution to these questions while also meeting the security concerns of the institution was crucial to the acquisition of the service, and to the long-term preservation of Princeton's student records. The solutions Princeton found may be useful to other institutions searching for ways to protect their content while addressing long-term encrypted data preservation concerns. Additionally, the process of gaining buy-in at Princeton for digital preservation by pointing to some of its high-security records will inform other repositories of methods they can use at their own institutions, especially in the case of institutions that must also collect records requiring high levels of security.

This poster will discuss the process and strategies used to gain support from University administration for digital preservation of highly sensitive records, how to work with a vendor to develop a repository-specific solution for digital preservation, and the process of investigating and developing options for an encryption key management system that protects student records while maintaining preservation goals. It will be useful to other institutions and practitioners seeking buy-in for their own systems, whether in-house or third-party, and will address the question of long-term encryption key management. This poster will be relevant to practitioners interested in preservation of highly sensitive records, encryption and key management systems, and third party digital preservation systems like Arkivum.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Videotex Art Restoration

Technical and Conceptual Challenges

John Durno
University of Victoria
Canada
jdurno@uvic.ca
0000-0003-0419-9742

Abstract – This poster will discuss the technical and conceptual challenges associated with achieving an authentic restoration of videotex art, in the context of a project currently underway to recover Canadian Telidon videotex artwork from the early 1980s. Strengths and weaknesses of various strategies will be discussed, including emulation, format migration, software reconstruction, and the use of period hardware. Goals of the poster include showcasing the strategies employed to date, and inviting criticism and comment from others with relevant experience to share, so as to refine and improve our methodology going forward.

Keywords – Videotex, artwork, emulation, authenticity

Conference Topics – The Cutting Edge: Technical Infrastructure and Implementation.

I. Introduction

The Digital Preservation Coalition has identified videotex materials as "practically extinct," requiring urgent action if these important early examples of our emergent digital culture are to be preserved [1]. This poster will discuss the challenges associated with achieving an authentic reconstruction of videotex art in the context of a project to recover Canadian Telidon videotex artworks from the early-mid 1980s.

The project to recover and restore Telidon art began in 2015 with the restoration of several Telidon artworks in the collection of University of Victoria Archives. The project has since grown to involve artists and arts organizations across Canada, as more digital files and video recordings have since come to light, and news of the project has spread. Participants in this collaborative project now include the University of Victoria Libraries and Archives, the Toronto-based arts centres InterAccess and Trinity Square Video, the Artexte Archives in Montreal, and a group of independent artists and curators, working together to achieve our goals of developing web and gallery exhibitions, and a permanent digital archive of Telidon Art and related materials [2].

II. Background

From 1978 through 1985 Telidon was a project of the Canadian federal Department of Communications to create interactive computer-based information networks using a made-in-Canada version of videotex. Although Telidon was only one of several international efforts to create videotex networks, the Canadian version was capable of presenting more sophisticated graphics than its competitors and for a time was seen as having a strong potential to become the dominant standard [3]. .

There was considerable interest in Telidon from across the cultural sector, as corporations, technologists, educators and artists all sought to harness its potential in various ways [3]. Several hubs of Telidon content production formed, such as at the Creative Writing Department at the University of Victoria, Video Inn in Vancouver, Trinity Square Video and Toronto Community Videotex (later InterAccess) arts centres in Toronto, the University of Quebec in Montreal and NSCAD University on the east coast. These hubs provided access to the tools of production to a broad spectrum of cultural producers, including artists. .

During the heyday of Telidon art production (roughly 1981 through 1987), Canadian Telidon artworks were exhibited regionally on networks of

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

public kiosks as part of Telidon field trials; at national events such as Expo '86 in Vancouver; and at international arts exhibitions, representing Canada in the 1983 Bienal de São Paulo, and in the "Technology and Informatica" section of the 1986 Venice Biennale..

As some of the earliest examples of interactive, networked art in Canada, these works stand as an important tributary in the development of Canadian digital art. However that significance that has been impossible to fully assess or appreciate due to the near total disappearance of the artworks as their supporting hardware and software obsolesced following the end of the Telidon project.

III. Current Status

Information services built upon Telidon were not ultimately commercially viable, and federal funding was withdrawn from the project in 1985. Although considerable content was created for Telidon systems during the time it was active, little of it has survived into the present in an intelligible form. Until recently, the remaining digital files known to exist were thought not to be viewable due to their dependence on long-obsolete Telidon terminals, special-purpose devices for interpreting and displaying vector-encoded Telidon files [4]. .

As Telidon was never widely adopted, a relatively small number of Telidon terminals were produced and very few are still known to exist. Alternative, approaches to rendering Telidon graphics have been developed for this project, combining emulation, format-migration and software reconstruction, but none perfectly recreates the experience of Telidon on period hardware, due to differences in timings, colours, fonts and visual textures. One may question whether a "perfect" restoration is even possible given the variation that existed even among hardware implementations in the 1980s.

IV. Challenges

Telidon art recovery is complicated by the fact its data encoding protocol went through two distinct iterations. The first, Telidon 699, was a draft specification deployed in early field trials. The second version, NAPLPS, was more widely deployed across North America. It served as the graphics encoding for several networked information services of which the IBM/Sears joint venture Prodigy was the best known and most successful.[5, pp.142-144] Although software that can display NAPLPS graphics still exists, the same cannot be said for Telidon 699, which was entirely dependent on dedicated hardware terminals. NAPLPS was by design not backwards compatible with Telidon 699. [3, pp.395-400].

Restoring NAPLPS graphics is therefore easier than restoring Telidon 699, though not without challenges. A standards compliant rendering is not always enough to achieve fidelity to the original works, as an accurate rendering is also generally dependent upon emulating the clock speeds and data transmission rates of early 1980s computing hardware. For the more complex, interactive works, the software and database that controlled the sequencing of the images must be recreated, and new software was written for this purpose. In addition, encoding problems with the original files are fairly common, requiring manual correction to repair visible defects in the rendered works. .

Recovery of Telidon 699 artwork has so far been achieved by recording the output of one of the few functioning Telidon 699 hardware terminals still in existence. In addition to the ever-present possibility of hardware failure prior to the completion of the project, problems with this approach include the lack of interactive features in the resulting recordings, and a noticeable loss of visual clarity in the recorded output. To address these issues, an alternative strategy of forward-migrating Telidon 699 to NAPLPS is being actively explored. .

Finally, there is a small but important group of works for which only period video recordings remain. Interventions in these cases have been limited to digitizing the video and using HTML5 and javascript to rebuild the interactivity that was lost when the works were transferred to videotape. This work goes beyond restoration into the realm of reconstruction, and the active participation of the artists is critical to ensuring the result remains true to their original intent. .

This poster will use examples to illustrate the challenges identified above, discuss the interventions

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

we have found to be necessary in the restoration of works of Telidon videotex art, and consider to what extent such interventions can be made without compromising the authenticity of the original works.

References

[1] Digital Preservation Coalition, "Bit List of Digitally Endangered Species, Revision 2," November 2018. [Online]. Available: https://www.dpconline.org/docs/miscellaneous/advocacy/1932-bitlist2018-final. [Accessed: Mar. 18, 2019].

[2] C. Hampton, "This Canadian dial-up art is older than the internet, and was long thought to be lost — until now," May 15 2018. CBC. [Online]. Available: https://www.cbc.ca/arts/this-canadian-dial-up-art-is-older-than-the-internet-and-was-long-thought-to-be-lost-until-now-1.4662086. [Accessed: Mar. 18, 2019].

[3] D. Parkhill, *The Beginning of a Beginning*. Ottawa: Department of Communications, 1987.

[4] J. Pearson, "The Original Net Artists: The Artists of Telidon, Canada's Doomed Pre-Internet Web", Jul. 21,2015. Motherboard. [Online]. Available: https://motherboard.vice.com/en_us/article/ezveak/the-original-net-artists. [Accessed: Mar. 18, 2019].

[5] M. Banks, *On the Way to the Web: The Secret History of the Internet and Its Founders*, Berkeley: Apress, 2008.

CD-Ark

A Tool for Cooperative Processing of Optical Discs

Zdeněk Hruška
The Moravian Library
Czech Republic
zdenek.hruska@mzk.cz
0000-0001-7606-5693

Abstract – A valuable and unique part of library collections, certain types of optical discs are becoming obsolete and face a variety of risks ranging from physical damage to loss or theft. In order to protect the content of optical data discs in Czech libraries, a collaborative software tool was developed. The CD-Ark software is designed to store data from the discs, creating data packages that include an ISO disc image alongside with a checksums, as well as technical and bibliographic metadata; the whole package is uploaded onto a joint server, ensuring bit-level preservation.

Keywords – Optical discs, Cooperation, Iso image, Bit-level preservation, Libraries.

Conference Topics – Collaboration; Building Capacity, Capability and Community

I. Introduction

Though in the past they seemed an attractive option to store information, nowadays it is obvious that CD and DVD discs do not age well. They are prone to physical damage due to poor storage conditions and easily lost or stolen. It is mainly the data stored on the disks that needs to be preserved for future generations, but the carrier itself has a certain cultural value and its image, too, should be conserved. However, up until very recently little attention has been paid to the issue of their protection and long-term preservation in spite of the large number of optical disks in the collections of libraries across Czech Republic. The number is not accurate due to imprecise catalogue records, but out of tens of thousands optical discs throughout Czech Republic, the Moravian library alone stores around 25,000. Therefore a project CD-Ark was launched which developed a software that would process their content and ensure its long-term protection.

II. CD-Ark

CD-Ark is a set of tools which enables cooperative processing of optical data discs (CD-ROMs and DVD-ROMs in particular) and the storage of the bit copy on the central server. It is designed mainly for large libraries that can collectively process optical data disks they hold in their collections. Duplicity is avoided at the moment by the system checking the bibliographic metadata and the disc image checksum.

As far as legal issues are concerned, the project was carried out in accordance with the Copyright Act as well as the Library Act. According to Article 18 of the Library act, libraries are under obligation to ensure the protection of the library collection, even by the conversion of library documents to another type of carrier, if necessary. Therefore, as the law stands, libraries can make copies of carries for archival use.

CD-Ark consists of two software applications: the CDArk-client and the CDArk-server which are described below in greater detail.

III. Workflow

In collaborative effort, libraries can achieve better results and the processing of optical disks is made more efficient. This is the very premise on which CD-Ark is built. With more libraries involved, the data producer is identified by a unique identification number that all Czech libraries have.

How does processing optical discs work with this set of tools? The CD-Ark-client (being a desktop application) is installed on a workstation in institutions

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

that take part in the project who also have disk space allocated on the central server. First, a metadata record is downloaded from the electronic catalogue, followed by the CD-Ark-client application software creating a data image of the optical disc (the so-called ISO image based on ISO 9660). As the carrier itself is a cultural artefact, the top of the disk alongside with the booklet and cover are scanned in order to preserve the optical disc's visual appearance. Once all is in one place, a package of data along with the checksums is sent to the central server for further processing.

Here, on the CD-Ark server, output from DROID (a file format identification tool) is added into the data package. An OCR (in TXT and ALTO) is created from the scans and finally everything is wrapped in BagIt (the standard data package format created by the Library of Congress for data storage and transfer). Subsequently, all this data is stored on the server and the checksums enable regular automatic checks for data corruption.

IV. Conclusion

Thus at least bit-level data protection is solved. That being said, the problem of logical long-term preservation still needs to be addressed. Now, however, it is crucial that we transfer the data to a more reliable medium; not only are the optical discs at risk of physical damage, loss or theft, but some are inevitably approaching the end of their lifespan. For future long term preservation, an open source LTP system ARCLib, is currently being developed in collaboration of various Czech libraries. Hopefully having tackled problems arising from the logical preservation of certain formats (such as software and games), the system is planned to ensure logical data preservation.

A long and arduous journey lies in front of us; extracting the data from the discs was no more than the first step leading towards the long-term protection of optical disc contents.

References

[1] Act No. 121/2000 Coll., on Copyright and Rights Related to Copyright and on Amendment to Certain Acts. 2006 [Accessed: 17-Sep.-2019]. Available: https://wipolex.wipo.int/en/text/137175.

[2] ARCLib: Popis projektu, c2019 [Accessed: 17-Sep.-2019]. Available: https://arclib.cz/popis-projektu/.

[3] DROID: (Digital Record and Object Identification). GitHub [online]. 2018 [Accessed: 17-Sep.-2019]. Available: https://digital-preservation.github.io/droid/.

[4] CDArk-client. GitHub [online]. c2019 [Accessed: 17-Sep.-2019]. Available: https://github.com/moravianlibrary/CDArcha-client.

[5] CDArk-server. GitHub [online]. c2019 [Accessed: 17-Sep.-2019]. Available: https://github.com/moravianlibrary/CDArcha-server.

[6] ISO 9660. The File Format Wiki [online]. 2019 [Accessed: 17-Sep.-2019]. Available: http://fileformats.archiveteam.org/wiki/ISO_9660.

[7] J. Kunze, J. Littman, L. Madden, E. Summers, A. Boyko, and B. Vargas, The BagIt File Packaging Format (V0.97) [online]. 2016 [Accessed: 17-Sep.-2019]. Available: https://tools.ietf.org/id/draft-kunze-bagit-14.txt.

[8] Law No. 257/2001 Coll. of 29 June 2001 on Libraries and Terms of Operating Public Library and Information Services (Library Act) [online]. 2001 [Accessed: 17-Sep.-2019]. Available: https://ipk.nkp.cz/ipk-en/legislation/02_LibAct.htm.

Safe Havens For Archives At Risk

Guidelines, Principles and Approaches

Afelonne Doek
International Institute of Social History
Netherlands
ado@iisg.nl
ORCID ooo-ooo2-9373-717X

Tim Gollins
National Records of Scotland Scotland
United Kingdom
tim.gollins@nrscotland.gov.uk
ORCID oooo-ooo3-2685-14o4

Abstract – The main aim of our poster is to raise awareness among the Digital Preservation community of the need for, and particular considerations of, (digital) Safe Havens For Archives at Risk. The document "Guiding Principles for Safe havens for archives at risk" [1], has been recently published and endorsed by the International Council of Archives [2]. The authors are involved in continuing work to develop detailed commentaries on each of the Guiding Principles. While the guiding principles are format neutral, we highlight specific digital considerations that are emerging in development of the commentaries.

Safe-Havens, Archives, Risk

Collaboration; Exploring New Horizons; Building Capacity, Capability and Community

I. Background To The Guiding Principles

Immediately following iPRES2016 in Bern, Swisspeace [3] , in collaboration with the Swiss Federal Department of Foreign Affairs [4] and the Swiss Federal Archives [5], arranged an International Expert Working Meeting on "*Safe Havens for Archives at Risk*" [6]. The meeting convened 37 representatives from international, governmental, and non-governmental institutions. The goal was to bring together institutions that had found or were looking for a safe repository for their archives (sending institutions), with institutions interested in, or already offering, a safe repository for archives at risk (hosting institutions), as well as experts and practitioners on the subject matter. The aim was to jointly discuss the needs, challenges, good practice and the way forward of safe havens for archives at risk. The participants acknowledged the importance of continuing their discussions and decided to create a working group. This group subsequently developed the "*Guiding Principles for Safe Havens for Archives at Risk*", now published and endorsed by the ICA [1].

II. Why Are Guiding Principles Needed?

The Principles provide guidance for both sending and hosting institutions for situations in which the safeguarding of originals or security copies of records through relocation can contribute to "*Dealing With The Past*" processes. The principles deal with preserving such archives and records requiring immediate action to protect them from armed conflict, military occupation, repressive actions by government or non-government actors, and recurrent, unavoidable natural risks.

Dealing With The Past processes address the rights of victims and societies as a whole, to truth, justice, reparation, and guarantees of non-recurrence in the aftermath of grave human rights violations, breaches of international humanitarian law, and related grave forms of corruption that facilitated these crimes. Records and archives of all types are central to these processes [7].

III. Particular Considerations

In the context of records relating to dealing with the past processes, both sending and hosting institutions should consider two further concepts: "*Do No Harm*" and "*Conflict Sensitivity*".

Do *no harm* is a concept geared to recognizing that intervention, of any kind, in a (conflict) context becomes part of that context. It aims at minimizing

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

the harm programs and activities may inadvertently cause by simply providing assistance in a conflict situation.

Conflict sensitivity refers to the practice of understanding how activities interact with conflict in a particular context, to mitigate the unintended negative effects of the activities on the context, and to exert a positive influence on the conflict wherever possible.

IV. THE PRINCIPLES

There are 18 Principles divided into four groups.

A. General

1. *Dealing with the Past Principle*
2. *Last Resort Principle*
3. Transparency Principle.
4. Legality and Agreement Principle
5. *Main Goal Principle*
6. Ethics Principle
7. *Fair Agreement Principle*
8. No Financial Profit Principle

B. The Substance of The Agreement

9. Processes in Agreement Principle
10. Ownership Principle
11. *Duration Principle*
12. *Anticipating Succession Principle*
13. Constituent Spirit Principle

C. Characteristics of Hosting Institutions

14. Legal Environment Principle
15. *Control of Material Principle*
16. Physical Characteristics Principle
17. *Professional Standards Principle*

D. Rapid Response

18. *Rapid Response Principle*

V. PRINCIPLES WITH SIGNIFICANT DIGITAL PRESERVATION ASPECTS

The following principles have significant Digital preservation aspects. We have included the text of the principle for reference.

1. Dealing With the Past.
"Safe haven solutions should always be implemented if archives/records that contribute to Dealing with the Past are at risk of destruction or alteration."

2. Last Resort
"A safe haven solution abroad should only be implemented when it is deemed impossible to store, protect and preserve the information safely within the country of origin, especially when transferring originals."

5. Main Goal Principle
"The goals of the sending institution in seeking a safe haven for archives/records should always be paramount in determining how they are treated by the hosting institution."

7. Fair Agreement Principle
"Safe haven arrangements should always be based on a fair agreement, acknowledging the potential asymmetrical nature of the relationship, including language barriers, mitigating the risks deriving therefrom and not taking advantage of the asymmetry."

11. Duration
"When negotiating the terms of the agreement, the parties should take into account the diffculty of planning the duration of the arrangement and include options for extension and closure. Precautions should be taken regarding possible changes of governments in both jurisdictions that could endanand capabilities used to manage the hosted archives/records should always be demonstrably under the direct control of the hosting institution, unless stated otherwise in the agreement."

17. Professional Standards
"Hosting Institutions should work in accordance with internationally recognized professional standards."

VI. DIGITAL PRESERVATION CONSIDERATIONS

A. "Dealing with the Past" and "last Resort".

Records that contribute to dealing with the past may come in all types. Often the content of these records, and their source, makes them especially vulnerable to risks from armed conflict, military occupation, repressive actions by government or non-government actors, and recurrent, unavoidable natural risks. This complex risk space can mean that unconventional approaches to digital preservation are needed to mitigate the risks. A technical approach to storing and/or preserving duplicates

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

of either of born-digital archives or archives that have been digitized may appear straightforward. However, the socio-technical nature of the risks bearing on archives seeking a safe haven, mean a wider range of approaches than would normally be considered appropriate for digital preservation may bring benefit, including:

- Full scale digital preservation in the host institution or in the cloud.
- Web Archiving in an existing public (or private) web archive.
- Remote storage only in the cloud.
- Storage only on a local server of the host institution. Storage only on duplicated o[ine media held in secure physical storage.

If a Cloud or Web archiving approach is chosen, the security, jurisdictional, and regulatory issues of these systems must be carefully considered. In the same way that raids on physical premises can be unpredictable, the hacking of a digital systems can occur equally without warning and from anywhere in the world. In every case it is important that both the technical and practical risks introduced by any safe haven approach be balanced alongside the risks that are attempting to be being mitigated. The particular circumstances of the archives at risk as well as the sending and hosting institutions must be considered paramount.

B. "Main Goal Principle" and "Fair Agreement Principle"

These two principles address the inevitable and significant power asymmetry between sending and hosting institutions in safe havens. In terms of *Dealing with the Past, Do No Harm* and *Conflict sensitivity* this power asymmetry may also have the additional context of colonial, or post-colonial, power dynamics between nations and their respective national institutions.

In the digital context, great care should be taken to avoid inadvertent and inappropriate assumptions of *preservation value* to govern the treatment of digital Safe Haven material, for example sending and hosting institutions may not necessarily have the same view of *Significant Properties*. The iPRES2019 keynote by Michele Caswell [8] and the associated journal publication *Dusting for Fingerprints: Introducing Feminist Standpoint Appraisal* [9] may be especially relevant in understanding preservation value or archival value in situations of oppression or power asymmetry.

C. "Duration" and "Anticipating Succession".

Succession planning [10] and the eventual need to export material from any system are common challenges in the digital preservation field. However, the particular uncertainties introduced in the context of safe havens bring further focus on this aspect. Flexibility of approach and responsiveness are important and we suggest that simpler workflows, more manual processes and less sophisticated technical mechanisms will be appropriate in many digital safe havens. In safe havens situations the ability to demonstrably eliminate all of the hosted archival content from any preservation environment is likely to a critical consideration once the materiel has been returned to the sender. Many automated preservation systems may find this an unusual and challenging requirement.

D. "Control of Material" and "Professional Standards".

In some jurisdictions [11] the holding of public archives beyond state boundaries is unlawful. This could significantly limit the technical solutions available in a safe haven context for either the sending of hosing institutions. In terms of professional standards, while archivists and digital preservationists skills are clearly needed, the particular risks associated with the security of digital safe haven material may mean that cybersecurity expertise may be required at a much greater level than usual in the hosting institution.

Acknowledgment

The authors thank their colleagues on the Working Group for Safe Haven for archives at Risk [6], Swisspeace [3], the Swiss Federal Department of Foreign Affairs [4], the International Institute of Social History [12], and National Records of Scotland [13] for their help and support.

References

[1] *Guiding principles for safe havens for archives at risk*, International Council on Archives, Feb. 2019. [Online]. Available: https://www.ica.org/en/guiding-principles-for-safe-havens-for-archives-at-risk.

[2] *International council on archives website*, International Council on Archives, 2019. [Online]. Available: https://www.ica.org/en.

[3] *Swisspeace website*, Swisspeace, 2019. [Online]. Available: https://www.swisspeace.ch/.

[4] *Swiss federal department of foreign affairs website*, Swiss Federal Department of Foreign Affairs, 2019. [Online]. Available: https://www.eda.admin.ch/ eda/en/home.html.

[5] *Swiss federal archives website*, Swiss Federal Archives, 2019. [Online]. Available: https://www. bar.admin.ch/bar/en/home.html.

[6] *Outcome paper: Safe havens for archives at risk*, Swisspeace, Oct. 2016. [Online]. Available: http://archivesproject.swisspeace.ch/fileadmin/user_upload/archivesproject/Publications/Save_havens_for_archives_Final_outcome_paper_2016.pdf.

[7] *Annual report of the united nations high commissioner for human rights and reports of the oice of the high commissioner and the secretary-general, Office of the High Commissioner for Human Rights*, Aug. 2009. [Online]. Available: https://www2.ohchr.org/english/bodies/hrcouncil/docs/12session/A-HRC-12-19.pdf.

[8] M. Caswell, "Keynote 2 : Whose digital preservation? locating our standpoints to reallocate resources," in *iPRES2019, 16th International Conference on Digital Preservation*, M. Ras and B. Sierman, Eds., Amsterdam, 2019. [Online]. Available: https://vimeo.com/362491244/b934a7afad.

[9] ——, "Dusting for fingerprints: Introducing feminist standpoint appraisal," in *Radical Empathy in Archival Practice*, E. Arroyo-Ramirez, J. Jones, S. O'Neill, and H. Smith, Eds., ser. Special issue, Journal of Critical Library and Information Studies, vol. 3, Litwin Books, 2019. [Online]. Available: https://journals.litwinbooks.com/index.php/jclis/article/view/113.

[10] *Vienna convention on succession of states in respect of state property, archives and debts*, United Nations Treaty Collection, Apr. 1983. [Online]. Available: https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=III-12&chapter=3&clang=_en.

[11] *Union with england act 17o7 section xxiv*, Queen's Printer for Scotland, Acts of the Old Scottish Parliament, 1707. [Online]. Available: http://www.legislation.gov.uk/aosp/1707/7/section/XXIV.

[12] *International institute of social history website*, International Institute of Social History, 2019. [Online]. Available: https://iisg.amsterdam/en.

[13] *National records of scotland website*, National Records of Scotland, 2019. [Online]. Available: https://www.nrscotland.gov.uk/.

Archiving the Scholarly Git Experience

An Environmental Scan of Git Preservation Efforts

Vicky Steeves
New York University, USA
vicky.steeves@nyu.edu
0000-0003-4298-168X

Genevieve Milliken
New York University, USA
genevieve.milliken@nyu.edu
0000-0002-3057-0659

Abstract – Our poster will reflect our recent efforts to understand the workflows and policies needed for the long-term preservation of code, annotations, and other scholarly ephemera from Git hosting platforms. We undertook an environmental scan of the existing processes and tools for capturing and actively archiving Git data and their associated, supplemental materials. We will present the results of this broad environmental scan, covering a wide variety of approaches, organizations, and workflows that could possibly be used to create a baseline on which to build and expand archival tools. Our efforts are geared toward acquiring, archiving, and providing permanent access to source code, and the materials around it, and argue that the whole should be considered part of the scholarly record.

Keywords – digital preservation, software preservation, git, digital archiving, github

Conference Topics – Exploring New Horizons

I. Introduction

Library and archives professionals are actively addressing software preservation via legal channels[1], partnerships/consortia (e.g. the Software Preservation Network) and technical efforts (e.g. Emulation-as-a-Service). These projects currently focus on access and emulation, as well as establishing best practices for fair use for preservation. One unexplored horizon, however, is the complexity of capturing the entirety of a repository. Git hosting platforms introduce features, histories, and ephemera not in the Git data format. As a result, the whole repository--including these other elements--is not capturable simply by cloning (copying) the source code.

This issue is important because the open source community broadly use repository hosting platforms such as GitHub, GitLab, and Bitbucket to publish their software, often with their complete histories tracked via a source-code versioning system. Scholars, too, are using these hosting platforms, because they allow for sharing, collaboration, and even scholarly transactions such as discussions, reviews, and annotations of code. For example, the Journal of Open Source Software and the Journal of Open Source Education are both fully hosted and run on GitHub, complete with reviews, discussions, and publishing efforts. Another example is GitHub Classroom, which was made specifically to merge a learning management system (LMS) with source code hosting features so that students and teachers could track assignments, grades, etc.

In most cases, research code and pedagogical materials, as well as the annotations and discussions within each repository, are currently not being captured for long-term preservation. While there are projects and tools that can capture just the Git data (e.g. copying the source code) or just the ephemera (e.g. annotations on commits), no tools or workflows currently can archive all of it together.

Git provides a way to capture a repository in a single file, including branches, commit history, tags, heads, remote heads, and files -- i.e. a Git bundle[2]. This, however, misses the rich materials only available on the hosting platforms. GH Archive ameliorates this by capturing the ephemera, but not the source code. It records the public GitHub timeline, archives it, and makes it freely accessible via Google BigQuery (free) for analysis. The decision not to include source code is a critical gap in their collection methods. In contrast to GH Archive, the Software Heritage Foundation only preserves source code. This makes sense given that their mission is to "collect, preserve, and make accessible source code for the benefits of present and future generations"[3].

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

One initial solution is to think in terms of web archiving each host repository. This solution, however, becomes problematic when we consider that web archiving is presently an imperfect tool. For example, the Wayback Machine from the Internet Archive--the largest web archiving organization to date--has archived pages within the github.com domain. Between May 14, 2008 and March 20, 2019 there were 35,140 captures of Github.com. Since the captured pages are often incomplete, and represent a small fraction of the total existing GitHub repositories, it may not be a reliable resource without substantial oversight and intervention.

In terms of research in this area, there is a large body of scholarship on Git hosting platforms. This notwithstanding, there is currently no archival solution that captures both code and its contextual elements simultaneously. As a result, the repositories available on Git hosting platforms are, and will continue to be, at risk from a preservation standpoint.

II. Our Investigation

Our poster will reflect our recent efforts to understand the workflows and policies needed for the long-term preservation of code, annotations, and other scholarly ephemera from Git hosting platforms. We undertook an environmental scan of the existing processes and tools for capturing and actively archiving Git data and their associated, supplemental materials.

We will present the results of this broad environmental scan, covering a wide variety of approaches, organizations, and workflows that could possibly be used to create a baseline on which to build and expand archival tools. Our efforts are geared toward acquiring, archiving, and providing permanent access to source code, and the materials around it, and argue that the whole should be considered part of the scholarly record. To facilitate this effort, we are also engaged in a scan of state-of-the-art web archiving and source code preservation standards, techniques, tools, and workflows.

Ultimately, we hope to fill gaps in the current digital archiving landscape by gathering and interpreting a broad range of scholarship on these, and allied, topics in order to think more strategically about the future.

III. Next Steps

In the second phase of our project, we will build upon the foundational work of the environmental scan. In particular, we will write a production-ready archival specification for capturing and preserving Git repositories and scholarly transaction data (e.g. annotations on pull requests), as well as a specification for aligning Git hosting environments with scholarly needs. We will then solicit feedback on the archival specification from the greater digital archiving community and iterate until a cohesive version is established.

Selection criteria and other guidance around which repositories to capture is not in scope for this project, as it should be made on an institutional level and at the discretion of the web archivist(s).

However, this resulting specification can be employed by web archivists to ensure their captures are complete and representative of the scholarship currently in the Git data format, on Git hosting platforms.

References

[1] Patricia Aufderheide, Brandon Butler, and Krista Cox, "Code of Best Practices in Fair Use for Software Preservation," Association of Research Libraries, Center for Media & Social Impact, Program on Information Justice and Intellectual Property, Sep. 2018.

[2] Scott Chacon, "git-bundle." [Online]. Available: https://git-scm.com/docs/git-bundle. [Accessed: 20-Mar-2019].

[3] Software Heritage Foundation, "Mission." [Online]. Available: https://www.softwareheritage.org/mission/. [Accessed: 20-Mar-2019].

A VERSATILE SOLUTION FOR LONG-TERM PRESERVATION OF RESEARCH DATA

Data Life-Cycle Management: The Swiss Way

Pierre-Yves Burgi
IT department
University of Geneva, Switzerland
pierre-yves.burgi@unige.ch
0000-0002-4956-9279

Hugues Cazeaux
IT department
University of Geneva, Switzerland
hugues.cazeaux@unige.ch
0000-0002-5618-2670

Lydie Echernier
Library and IT departments
University of Geneva, Switzerland
lydie.echernier@unige.ch
0000-0002-3932-8404

Abstract – Developed in the context of the 2017-2020 Swiss national programme "Scientific information: Access, processing and safeguarding", the DLCM solution (dlcm.ch) consists of an open and modular architecture for long-term preservation of research data. While being compliant with the OAIS standard, the proposed solution makes the deposit of research data straightforward, yet allows the qualification of the submitted formats so as to inform users of the risks incurred on the long term.

Keywords – research data repository, OAIS compliant, web services, preservation planning, user-friendly portal, OLOS

Conference Topics – Designing and Delivering Sustainable Digital Preservation; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

Data is the foundation for research. From laboratory results to population statistics or satellite images, researchers rely on data to formulate hypotheses, review the work of others, and ultimately advance the understanding of the world. But for researchers, managing increasingly large datasets can be a burden, and a distraction from their actual research work. Each step of the data life-cycle, from acquisition, storage to publication and long-time archiving, requires the use of different tools and providers. Regulations, in Switzerland and abroad, also affect how research data may be shared and stored. The complexity and cost of managing research data is an unnecessary interference for researchers and their home institutions. A solution to this problem wouldn't just benefit researchers, but society at large.

A. Context

The Swiss DLCM project [1] was initially launched in September 2015 and regrouped library and IT teams from eight Swiss higher education institutions (HEI). The first phase of the project focused on the identification of the researchers' needs, which led to the documentation of main use cases and basic services, such as: access and contact point for information, training, and personalized advice, as well as active data management solutions, including long-term preservation, and data publication according to international standards [2]. In August 2018, the phase 2 of the project (which regroups 3 Swiss HEI) was initiated (and is due to end in December 2020) with the goal to extend the prototype developed in the previous phase into a national service, available through a Web portal, or directly imbedded in researcher's existing software ecosystem, and grounded in realistic business models to ensure financial sustainability [3]. A multilingual National coordination desk offering tailored support, consulting and training to the academic community through a network of experts is another main outcome of the project.

B. DLCM Long-Term Preservation Solution

The outcome of the DLCM project consists of a highly integrated data management system for researchers. Based in Switzerland, and backed by multiple Swiss data centers, this solution allows researchers to safely manage, publish and preserve

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

on the long-term their data. Designed to be compatible with most data creation tools and formats, the DLCM solution works across all disciplines. Its modular interface allows users to build a plan suited to their needs, compliant with data management regulations and standards. Crucially, it enables researchers to retain full control over their data, allowing them to monitor and limit access to their datasets to a restricted circle of researchers.

II. A VERSATILE ARCHITECTURE

The DLCM solution (aka "OLOS") consists of an open and modular architecture for the long-term preservation of research data, compliant with the OAIS standard (ISO 14721) [4] and FAIR principles [5], and compatible with cloud deployment. The various modules of the solution offer a range of services that allow researchers to prepare their data for preservation, namely: to submit them with a pre-ingest step followed by ingest, to archive them physically, to index the metadata and to be able to access them according to specific rights. For uploading large datasets (TB and more), which still represent real challenges for transferring them via Internet, CERN's technologies (Xrootd, FTS, etc.) are being considered.

This set of services, which includes DOI attribution to the dataset, are available via RESTful APIs (Representational State Transfer) Web services, which guarantees the implementation of best practices in the domain: virus scanning, file format detection, checksum calculation, integrity check, replication, etc. In other words, they are standardized and therefore agnostic to the technology. For instance, laboratory tools can be connected to DLCM through the deposit/file upload/validation APIs with minimum adaptations. In addition to APIs, an intuitive portal offers an alternative path to submit and access data, while hiding the complexity underlying the OAIS mechanisms.

A. Metadata

In the DLCM solution, archiving metadata is a key and mandatory feature of the system to preserve the dataset with the right level of information. This feature is important to search and share datasets, but also to enforce the correct action of preservation in the future, i.e. format conversion. Users provide mandatory and recommended Datacite fields: identifier, creator, title, description, publisher, publication year, and resource type; and optional information about the uploaded items (in advanced mode). The solution thus offers the capability to complete a dataset with specific metadata, which are research domain dependent and embedded within the METS container. Furthermore, to ensure integrity of the preserved objects, all administrative information is contained within the PREMIS schema.

B. Preservation policy

Another outcome of the DLCM project consists of a preservation planning module (preservation-centric workflows), which handles replication and synchronization of data, stored in different data centres. This feature represents a major capability to address the long-term preservation objective. This notion defines the preservation strategy to be enforced by organizational units or by institution. This policy will set: the number of copies, the geographic location and physical multi-tiering support (spinning disks, tapes, etc.), the qualification of the data format compliance, the frequency of integrity check, the migration plan, and the data disposal at the term of the life-cycle, if applicable. Service invoicing will depend on those selected options.

C. Data Preservation

In principle, before ingesting data into a repository, researchers should comply with recommended formats so that migration cycles can be ensured on the long term. Such a compliance is however not always possible for various reasons, which include performance and volumetry, and also simply because commercial solutions often force the use of proprietary formats. To support the preservation of any data regardless of format, data are preserved at the bit-level only, providing researchers with the possibility of storing data at minimal cost while complying with the publishers' and/or funders' requirements. Obviously, this is not what information professionals are seeking, but it responds to field needs. Conversely, with more standard formats, preservation mechanisms are applied to ensure that data remain intelligible throughout the migration cycles, yet this necessitates a more intensive preparatory work from the data producers. In both cases, preserved data remain as accessible as possible for further uses, and is not kept as dark archives. Access is consequently facilitated, opened as much as possible following standard formats, with the aim of

serving other researchers in the near and far future.

Value assessments must be performed so that the costs for preserving bitstreams match the intrinsic value of the preserved datasets, given that this value can lessen with time. Such assessments will be done by the preservation planning module, through a workflow which will rise a warning to the data steward, either at the term of the data's declared preservation duration, or on a periodic basis.

III. Conclusions

The information contained in research data constitutes a valuable asset for the researchers. Yet researchers too often ignore the importance of the data life-cycle, a concept that encompasses a multitude of facets. The DLCM solution was designed to facilitate such a cycle while ensuring best practices in long-term preservation are applied. Key features of the proposed solution are the modularity of the architecture and the possibility to either ingest data directly from the researcher's working environments through APIs or to deposit them manually through a user-friendly portal.

References

[1] Swissuniversities. Data Life-Cycle Management (DLCM) project. https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Organisation/SUK-P/SUK_P-2/181-011_DLCM-Phase2_Abstract_21.02.17.pdf Retrieved 27 June 2019.

[2] Burgi, P.-Y., Blumer E., Makhlouf-Shabou, B. (2017). Research data management in Switzerland: National efforts to guarantee the sustainability of research outputs. IFLA Journal, 43, pp. 5-21. doi:10.1177/0340035216678238

[3] Burgi, P.-Y., Blumer, E., Jelicic, A., Dieude, A., Sesartic, A. (2017). Innovative approach for project viability: From a diversity of business models to harmonized and scalable national services. https://archive-ouverte.unige.ch/unige:102820 Retrieved 28 June 2019.

[4] CCSDS Secretariat. Reference Model For An Open Archival Information System (OAIS): Recommended Practice, CCSDS 650.0-M-2 (Magenta Book). (2012). https://public.ccsds.org/pubs/650x0m2.pdf Retrieved 27 June 2019.

[5] Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1): 160018. doi:10.1038/sdata.2016.18.

Introduction on authorized preservation practice of the National Digital Preservation Program in China

Chao Wang, Zhenxin Wu, Jiancheng Zheng
National Science Library,
Chinese Academy of Sciences 33 Beisihuan Xilu, Zhongguancun
Beijing P.R.China ,10019
wangch,wuzx,zhengjc{@mail.las.ac.cn}

Abstract –This work will introduce the structure of the National Digital Preservation Program in China , and why do we establish the form of authorized preservation.

Keywords – Digital Preservation, Authorized Preservation, National Program, China

Conference Topics –Collaboration: a Necessity

I. Introduction

Supported by the Ministry of Science and Technology (hereinafter referred to as MoST) of China, the National Science and Technology Library Consortium (hereinafter referred to as NSTL) has been implementing the National Digital Preservation Program (hereinafter referred to as NDPP) since 2013, aiming to promote systematic, comprehensive and reliable preservation of major databases of scientific literature published inside and outside of China.

NDPP is a national-level preservation program. As a public service infrastructure, its construction and operation are supported by state funds. At the same time, NDPP is a cooperative preservation network composed of domestic libraries or institutions subscribing digital scientific and technological resources. Several preservation institutions undertake resource preservation tasks.

II. Organization structure

A. Management Institution

As the administration of NDPP, NSTL is a scientific and technological literature service institution established by MoST of China in 2000. Its main responsibilities are to organize national libraries and institutions to issue joint statement on long-term preservation of digital resources, to coordinate and plan for preserving target resources and preservation centers, to support and assist preservation centers conducting long-term preservation negotiations, to organize public certification and auditing for the mechanism and effect of long-term preservation, to supervise and audit public services when necessary, to coordinate necessary backup and inheritance preservation, and to provide stable funding for NDPP.

B. Preservation Centers

NDPP selected a limited number of institutions with mature technical conditions and perfect operation mechanisms to undertake the tasks of preservation. NDPP has gradually constructed three preservation centers: National Science Library, Chinese Academy of Sciences (hereinafter referred to as NSL), Institute of Scientific and Technical Information of China(hereinafter referred to as ISTIC), and Peking University Library.

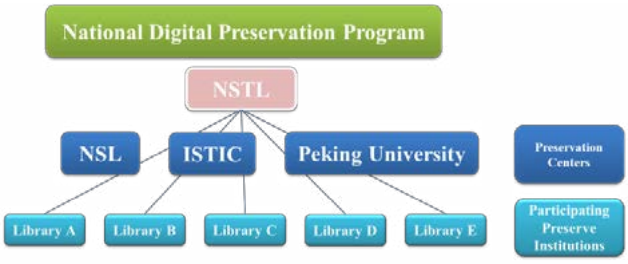
Through signing legally long-term preservation agreement with publishers, preservation centers receive and preserve provided data, provide public services for participating institutions when resources cannot be accessed, and accept public certification and auditing.

C. Participating Institutionsu

Libraries that signed the joint statement of long- term preservation are considered to be the participating institutions of NDPP. As of January 2019, 216 domestic libraries participate in NDPP, including academic libraries, university libraries, public libraries and professional libraries. They all

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

recognize that the right of long-term preservation is an integral part of library's rights of purchased knowledge content. They will proactively advocate and maintain the right in negotiations, and authorize NDPP to take the tasks of long-term preservation of subscription resources.



III. What is authorized preservation

Due to the limited number of preservation centers, plenty of participating institutions of NDPP which want to realize the long-term preservation of subscription resources need to authorize NDPP to take the tasks of preservation.

For NDPP, the number of resources preserved by preservation centers is limited, while the number of resources subscribed by participating institutions is large. The form of authorized preservation will greatly expands the scope of NDPP's preserved resources. For publishers and participating institutions of NDPP, the program is a completely social service that does not charge any deposit fees and does not impose any additional financial burden. At the same time, as a national- level preservation program, NDPP's long-term preservation planning, equity management mechanism, technical system, organizational management and economic input, public service policy, and public certification of long-term preservation have reliable management mechanisms, can provide reliable and sustainable long-term preservation service for authorizing preservation institutions.

IV. Case

NSL, as one of the preservation center of NDPP, was authorized by Medical Library to have long- term preservation of the Primal Pictures database.

V. Conclusions

Authorized preservation is a more flexible and universe long-term preservation mechanism, compared with preservation centers directly issuing and achieving resources preservation.

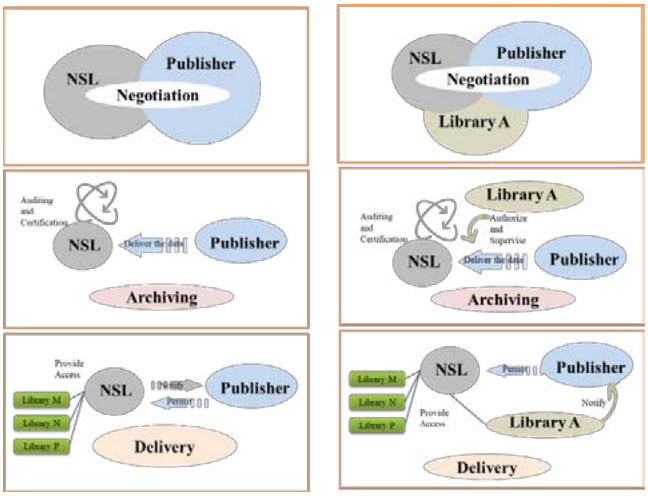


Figure 2 Different Preservation models

Acknowledgments

This work was supported by Foundation of National Digital Preservation Program.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Preservation Metadata Dictionary

PREMIS implementation in practice

Marjolein Steeman
Netherlands Institute for Sound and Vision
Netherlands
msteeman@beeldengeluid.nl

Yvette Hollander
Netherlands Institute for Sound and Vision
Netherlands
yhollander@beeldengeluid.nl

Abstract – This poster tells the story of designing the PMD in a way that is fully conformant with PREMIS, the leading standard on preservation metadata. It will give insight in the main structure of the PMD and it will illustrate its practical use with some examples.

Keywords – Preservation Metadata, PREMIS, OAIS information model

Conference Topics – Designing and Delivering Sustainable Digital Preservation.

I. Introduction

This Preservation Metadata Dictionary (PMD) of the Netherlands Institute for Sound and Vision, combines multiple object levels and perspectives including technical metadata on the file and bitstream level, event metadata, and rights metadata. The dictionary plays an important role in operational decision making, for instance on designing a new workflow on digitization or on implementing a new ingest workflow.

II. PREMIS conformant

International standards on metadata in general offer a generic and conceptual framework: a set of ideas and rules, flexible and broadly applicable. Each standard does so for its own perspective or domain.

The OAIS Information Package [1] is composed of several information objects. The PREMIS data model [2] consists of four core entities that in a way relate to the Information Package.

Implementing standards like these requires translating its set of ideas and rules to an organisation's own practices [3]. This means refining, itemizing and omitting where applicable. It also implies constantly making choices along the way, taking into account the principles of conformance and the principles of use [4]. This journey will be illustrated for the implementation of the PREMIS standard at Sound and Vision.

III. Practical use

A. Daily Ingest: the MXF as the source

The MXF that is imported daily is born digital. But as such it is the end product of the creation of a broadcast. The case will show how Sound and Vision perceives the audit-trail and what will be documented as significant properties.

B. Legacy, digitization: a digibeta from BCN

For the MXF that is the result of a digitization project from tape, the documentation of significant properties will be quite different.

C. Oldest Legacy: the film as the source

Finally the MXF that is created as a mezzanine for the DPX. Yet another case that had to be represented by the PREMIS model. The outcome is shown for the MXF as well as for the DPX.

IV. References

[1] OAIS2012, CCSDS, "Reference Model for an Open Archival Information System (OAIS)", Magenta book,
[2] http://public.ccsds.org/publications/archive/650x0m2.pdf, 2012.
[3] PREMIS Editorial Committee (2015). PREMIS Data Dictionary for Preservation Metadata, Version 3.0. http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf
[4] Digital Preservation Metadata for Practitioners, Angela Dappert, Rebecca Squire Guenther, Sébastien Peyrard, editors, 2016
[5] PREMIS Conformance, dd. 20-11-2017. http://www.loc.gov/standards/premis/premis-conformance-20150429.pdf

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The Integrated Preservation Suite

Demonstrating a scalable preservation planning toolset for diverse digital collections

Peter May
British Library, UK
peter.may@bl.uk
https://orcid.org/0000-0001-8625-9176

Maureen Pennock
British Library, UK
maureen.pennock@bl.uk
https://orcid.org/0000-0002-7521-8536

David Russo
British Library, UK
david.russo@bl.uk
https://orcid.org/0000-0003-2829-3936

Abstract – The Integrated Preservation Suite is an internally funded project at the British Library to develop automated and scalable preservation planning capability for a highly diverse and growing digital collection. Core components include a technical knowledge base, a software repository, a policy and planning repository, and a preservation watch function, all accessed via a web-based Preservation Workbench and designed to integrate with any repository software through a modular API-based architecture. We will demonstrate work to date, showing how preservation plans can be constructed and populated for the formats currently supported.

Keywords – digital preservation strategies, knowledge base, preservation planning, software preservation, preservation watch.

Conference Topics – The Cutting Edge – Technical Infrastructure & Implementation

I. Introduction

The Integrated Preservation Suite is an internally funded project at British Library that builds upon several years of preservation activities to develop and enhance the Library's preservation planning capability, largely focused on automation and addressing the risks and opportunities specific to the Library's heterogeneous digital collections. It aims to achieve this through the development and integration of several components – including a knowledge base, a software repository, a policy and planning repository, and a web-based workbench - designed to meet separate but complementary goals, all combined with the population of these components with content required for the infrastructure to work in a business environment.

This demonstration will showcase the functionality and contents of the Integrated Preservation Suite to date.

II. Overview

The Integrated Preservation Suite is intended to help us with format-based risk mitigation at scale and across all of our collections. We have developed the architecture and associated data models recursively through the course of the project, building components from the ground up to meet our needs where necessary. The project is a three-year initiative and we are currently in the final year of development, after which the suite will be deployed, maintained, and further populated as a business-as-usual concern.

Components reviewed in the demonstration will include:

Knowledge Base (KB): a graph-based curated knowledge base containing information from multiple different sources regarding formats, software, and wider technical environments relevant to the Library's digital collections;

- Preservation Software Repository (SR): a digital repository containing requisite current

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

and legacy software for rendering files stored in our digital repository and implementing preservation plans;
- Policy and Planning Repository (PPR): a document repository for storing collection-specific data including collection profiles, preservation policies, and collection-specific preservation plans;
- Preservation Workbench (PW): a web-based graphical user interface providing functionality for searching and curating the Knowledge Base, the Software Repository, and the Policy and Planning Repository, as well as for managing and developing format-specific preservation plans.

We will also discuss plans for the Preservation Watch - a component that monitors the archival store and our other IPS components in order to provide notifications to users regarding potential preservation risks, thereby instigating preservation planning.

Functions we expect to demonstrate will include:
- Preservation planning: Currently, our preservation planning approach is broadly following a SCAPE/Plato planning methodology, [1] bringing together various facets of information about a collection at risk to define the plan requirements, evaluate different strategies to mitigate any risks, analyze the results, make a recommendation, and construct an executable plan.
- Curating incoming data for the Knowledge Base: Data in the Knowledge Base is aggregated from several different sources; the Workbench provides a web-based interface to allow an appropriate user to compare incoming data with existing data and make decisions about how to proceed with each datum.
- Searching the Knowledge Base: Keyword searching is supported, e.g., a user can search for "PDF" or "Adobe" or any other term. This matches on key properties within the Knowledge Base, such as the file format/software name or an extension. To facilitate more in-depth queries, such as for identifying software that can migrate file formats, we provide a set of search labels with which to tailor queries.

Information on the overall architecture, data models, and development process will also be available. Staff will be on hand to discuss these and non-technical elements of the project including licensing challenges (and progress) for content in the software repository.

III. Required infrastructure and/or resources

We will need to use our own laptops for the demo. Wifi will be required. A projector may also be required depending on the room set up.

The duration of the demo can be tailored depending on how long you make available. We are not planning a structured session.

Goals of the demo:
- Showcase the integration element of the Suite as a pre-requisite for scalable preservation planning across diverse collections
- Engage with attendees to gather feedback that might inform future development or maintenance work

Acknowledgment

We would like to thank the other IPS team members whose dedicated work has helped shape and deliver the project: Michael Whitmore, Kevin Davies, Michael Day, Akiko Kimura, Simon Whibley, and John Beaman. We would also like to thank the National Library of Australia for kindly supplying us with a spreadsheet of their collected data, and the maintainers of File-Extensions.org for giving us permission to scrape their website.

References

[1] M. Kraxner, M. Plangg, K. Duretec, C. Becker, and L. Faria, "The SCAPE planning and watch suite," in Proceedings of the 6th International Conference on Digital Preservation, iPRES 2013., Lisbon, Portugal, September 2 – 6th, 2013 [Online]. Available: http://hdl.handle.net/11353/10.378091.

Review, Appraisal, And Triage of Mail:

A Tutorial

Christopher A. Lee
University of North Carolina
USA
callee@ils.unc.edu

Kam Woods
University of North Carolina
USA
kamwoods@email.unc.edu

Abstract – This full-day tutorial will focus on open-source software (OSS) tools and methods for libraries, archives and museums (LAMs) to identify email in born-digital collections, review email sources for sensitive or restricted materials, and perform appraisal and triage tasks to identify and annotate records. We will specifically explore products of the Review, Appraisal and Triage of Mail (RATOM) project's use of machine learning to separate records from non-records, along with natural language processing methods to identify entities of interest within those records. In addition to gaining hands-on experience using the tools, participants will also learn about the rationale for their development, how they relate to other available software, and how processing of email can fit into larger digital curation workflows. We will conclude with a discussion of implications for participants in their own institutions.

Keywords – email curation, appraisal, natural language processing, machine learning, timelines

Conference Topics – Exploring New Horizons; The Cutting Edge: Technical Infrastructure and Implementation

I. BACKGROUND

Despite progress on various technologies to support both digital preservation and archival description, we have still seen relatively little progress on software support for the core library, archives and museum (LAM) activities of selection and appraisal.

The BitCurator environment allows LAMs to engage in acquisition of data from media, as well as carrying out characterization and triage of the data, exposing numerous data points that can inform selection and appraisal decisions, including file types, file sizes, timestamps, original directory structures, potentially sensitive features. However, these various forms of output are generally static. Users have expressed interest in additional ways to iteratively make judgements. Software developed by the BitCurator Access and BitCurator NLP projects – including topic modelling and named entity extraction - can further facilitate such decision-making. By incorporating timeline representations and visualizations, LAM professionals also can get useful, high-level views of the materials that they are investigating.

Email is often part of collections and acquisitions that include other types of materials, and LAMs frequently want to take actions that reflect those contextual relationships. The Transforming Online Mail with Embedded Semantics (TOMES) project has developed software to identify email accounts of public officials with enduring value in order to capture, preserve and provide access to important government records.

The Review, Appraisal and Triage of Mail (RATOM) project (2019-2020), funded by the Andrew W. Mellon Foundation, is a partnership between the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill (UNC-CH) and the Division of Archives and Records (DAR) of the North Carolina Department of Natural and Cultural Resources (NC DNCR). RATOM is developing capabilities for interactive selection and appraisal in the BitCurator environment in ways that can be directly connected to the output of TOMES to support iterative processing, in which information discovered at various points in the processing workflow can support further selection, redaction or description actions. Among other activities, this involves mapping of timestamp, entity, sensitive features and other elements across the tools.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

II. TUTORIAL OVERVIEW

This full-day tutorial will focus on open-source software (OSS) tools and methods for libraries, archives and museums (LAMs) to identify email in born-digital collections, review email sources for sensitive or restricted materials, and perform appraisal and triage tasks to identify and annotate records. We will specifically explore products of the Review, Appraisal and Triage of Mail (RATOM) project's use of machine learning to separate records from non-records, along with natural language processing methods to identify entities of interest within those records. In addition to gaining hands-on experience using the tools, participants will also learn about the rationale for their development, how they relate to other available software, and how processing of email can fit into larger digital curation workflows. We will conclude with a discussion of implications for participants in their own institutions.

III. INTENDED AUDIENCE

This tutorial should be of interest to information professionals who are responsible for curation of digital collections that are composed – in whole or in part – of email. Another intended audience is individuals involved in digital preservation research, development and IT management, who will learn how to incorporate various open-source tools and services into institutional workflows (including through combinations with other proprietary software).

IV. PREPARATION AND PREREQUISITE KNOWLEDGE

Participants are not required to have prerequisite knowledge in any specific machine learning or NLP tools or methods. However, some previous knowledge of email structure and some command line experience will both be helpful. The tutorial will make use of a combination of hosted and local software. Participants will be expected to download and install software in advance.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

From Preservation To Access In One Step

Aviary: An Access Platform For Audiovisual Content - Demo Session

Pamela Vízner
Consultant, AVP, Chile - United States
pamela@weareavp.com

Bertram Lyons
Senior Consultant, AVP, United States
bertram@weareavp.com

Abstract – Access to audiovisual content continues to lag behind that of textual and still image content. Its time-based nature means it is tremendously underserved. However, current technology makes it possible to create automated transcriptions from audio, which can be used to implement granular search tools with high degrees of accuracy. This, together with building visualization tools that give the user the power to navigate audiovisual content easily, constitutes a unique opportunity to enhance access to audiovisual collections.

Additionally, although some preservation systems offer access platforms, these are not designed to display audiovisual content in its full potential.

Aviary offers specialized access to audiovisual content with granular permissions, and fully searchable indexes and transcripts. Generic integrations with preservation systems such as Preservica and Archivematica are in being developed currently.

This proposal is to offer a live demonstration of Aviary focused on integrations with digital preservation systems. We expect to be able to provide a demonstration of the system and this specific feature at iPRES 2019.

Keywords – audiovisual, access, integrations, discoverability, permissions management.

Conference Topics – Designing and Delivering Sustainable Digital Preservation ; The Cutting Edge: Technical Infrastructure and Implementation

I. Why Aviary?

Access to audiovisual content continues to lag behind that of textual and still image content. Its time-based nature means it is tremendously underserved. Typically access is only offered as a static description of the entire audiovisual object, forcing users to spend large amounts of time scrolling through content in order to find what they need. In spite of this, people more and more rely on audiovisual content as a means to access and consume information, e.g. videos on social media, podcasts, audiobooks, etc. [1][2]

In addition to the searchability issues, organizations in the cultural domain often struggle to provide access to materials that have complex intellectual property statuses or have special access restrictions, such as location, type of user, or sensitive content.

Some existing digital preservation systems offer access platforms that directly interact with the preservation system to provide quick access to materials being preserved [3]. These access systems, while useful as a first layer of search and retrieval, are not designed to allow users to deeply navigate time-based content.

II. The Solution

Current technology makes it possible to create automated transcriptions from audio, which can be used to implement granular search tools with high degrees of accuracy. Speech-to-text (STT) technologies can output structured transcript files that can be consumed by other applications to offer searchability and enhanced display. This allows users to quickly find phrases and keywords in the transcripts that point directly to the part of the audio or video where these are mentioned, saving user's time.

The ability to extract this information as transcripts, together with building visualization tools that give the user the power to navigate audiovisual content easily, constitutes a unique opportunity to enhance access to audiovisual collections.

Moreover, the possibility to interact directly with digital preservation systems removes the burden many organizations have to implement access

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

systems that fit the needs and capabilities of cultural organizations while meeting users' expectations.

III. Aviary: True Access To Audiovisual Content

Aviary, a multi-tenant, web-based audiovisual access platform takes advantage of STT technologies to offer a smooth and efficient user experience in the navigation of time-based media. Transcripts generated by these automated tools are linked to the media files, allowing users to playback exact points in the audio or video where the subject of interest is mentioned. These transcripts are immediately indexed to provide full searchability.

Aviary also allows publishers to have granular control over access to the different elements of the published content, including media, transcripts, indexes, users, user groups, location, and email domains, which gives the opportunity to optimize access to content and related metadata for materials with complex access restrictions.

IV. Integrating With Existing Tools

As a tool specialized in access to time-based media, Aviary has been designed to easily integrate with tools currently used by archives, museums, and libraries to provide an application that can be integrated to a variety of workflows.

At the time of writing, Aviary offers integration with popular streaming and access platforms, including: YouTube, Vimeo, Avalon, and SoundCloud. Additionally, Aviary offers custom integrations with Kaltura [4] and ArchivesSpace [5]. WebVTT [6], OHMS XML [7], and plain text transcripts and indexes can be uploaded to provide immediate granular search and retrieval. A bulk import feature is available for OHMS XML files and content formatted using the Import Aviary Package [8]. Closed caption is supported for WebVTT transcripts.

The following features are in development for the first half of 2019:

1) Generic integration with Preservica and Archivematica.
2) Annotations (using W3C Web Annotation Model)
3) Ability to attach supplemental content
4) IIIF integration and support
5) Additional bulk import formats, such as MARC XML and MODS.
6) Automated transcription integrations.
7) In-app transcript and index editing.
8) Integrations with learning management systems.

V. Aviary Demo

The demo proposed in this document aims to focus on the general features of Aviary while focusing on our integrations with digital preservation tools. For the first part of the demo (15 minutes), the purpose is to familiarize attendees with the tool and its structure, user and permission levels, search, and navigation. The second part (15 minutes) will focus on integrations with digital preservation systems, followed by a 15-minute Q&A section for a total duration of 45 minutes. The goal of the demonstration is to introduce Aviary and how it can be integrated to digital preservation and access workflows. As a tool in development, another goal is to get feedback from potential users in order to improve this specific feature.

Acknowledgement

The development of Aviary has been possible thanks to a multidisciplinary team of professionals. From vision, to design, to development and implementation, Aviary is a group effort from start to finish. The authors want to thank everyone who has been involved in this project.

We also want to thank our early adopters and people who have served as testers; you have all helped shape this application.

References

[1] Data gathered by companies such as Edison Research, Nielsen, and The Interactive Advertising Bureau (IAB) around market and user behavior in the podcast industry indicate that production and consumption continues to increase. https://musicoomph.com/podcast-statistics/

[2] Digital video content is one of the strongest economic drivers in the global entertainment industry. As a correlation, digital video is an increasingly large percentage of born-digital collections materials. https://www.forbes.com/sites/nelsongranados/2018/12/18/digital-video-and-social-media-will-drive-entertainment-industry-growth-in-2019/#487c87284661.

[3] An example of such platform is Preservica's Universal Access, which provides web access to audiovisual materials. https://preservica.com/digital-archive-software/document-access

[4] Kaltura is a company that develops applications for video content, among them, Kaltura Video Platform. https://corp.kaltura.com/

[5] ArchivesSpace is an open source archival management system supported by Lyrasis. https://archivesspace.org/

[6] WebVTT is a W3C standard for displaying timed text in connection with the HTML5 <track> element. https://www.w3.org/TR/webvtt1/

[7] OHMS XML is a metadata schema generated by the Oral History Metadata Synchronizer (OHMS) tool for the description of audio transcripts, developed by the Louie B. Nunn Center for Oral History of the University of Kentucky, USA. http://nunncenter.org/ohms-info/

[8] For organizations or publishers that do not have their data stored in one of the supported standards, Aviary offers a package that allows to structure data and files for bulk imports.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Jisc Open Research Hub

Demonstration—from Repository to Preservation

Paul Stokes
Jisc, UK
Paul.stokes@jisc.ac.uk
0000-0002-7333-4998

John Kaye
Jisc, UK
John.kaye@jisc.ac.uk
0000-0002-4400-4252

Tamsin Burland
Jisc, UK
Tamsin.burland@jisc.ac.uk
0000-0002-5129-979X

Dom Fripp
Jisc, UK
dom.fripp@jisc.ac.uk
0000-0001-5352-4666

Tom Davey
Jisc, UK
tom.davey@jisc.ac.uk
0000-0002-0319-5574

Abstract – Jisc's Open Research Hub (JORH) integrates a number of repository, preservation, reporting and storage platforms as a one stop shop for researchers and research managers. The service offers both open source and proprietary systems and allows data and metadata to be shared openly if required. The platform has been developed through years-long consultation with the UK HE research sector and sector bodies, along with contributions from both in-house Jisc and third-party experts. And was launched in July 2018.

The need for such a solution has arisen from the sector's desires to achieve several, shared aims, including: greater collaboration; tackling the reproducibility crisis; enabling better research; enabling a better user experience; and meeting funder requirements.

Jisc's custom-built repository—the Open Research Repository—is part of the Jisc Open Research Hub. It's built upon an extensive data model and rich messaging layer, providing users with a clean, simple, and easy-to-learn interface for the deposit, approval, and discovery of a range of outputs. In particular it allows for a seamless end to end experience for the user; from deposit straight through to preservation.

Jisc's position in the UK higher education / research sector, as well as the scale of the service provides us with many domain-specific insights to share with iPRES delegates, ranging from the broad methods mentioned above, down to individual design decisions informed by our research and domain expertise.

Keywords – Research Data, Integration, Preservation, Shared Services, Repository to Preservation

Conference Topics – The Cutting Edge: Technical Infrastructure and Implementation.

I. Introduction

Jisc's Open Research Hub, integrates a number of repository, preservation, reporting and storage platforms as a one stop shop for all users of repository and preservation systems. The service offers a range of component systems and a range of potential integrations, both in terms of data (and meta data) sources and endpoints. It allows data and metadata to be managed, preserved and shared as openly as possible and as securely as needed.

This demonstration is intended to show the ease by which data can be deposited, along with semi-automatic selection and entry of meta data from a variety of systems through to preservation (also in a variety of systems).

It will be of interest to content generators, developers, integrators, vendors, repository managers, curators, research data managers, support staff, and data end-users.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

II. Demonstration content

The demonstration will illustrate the core functionality and features of the open research hub. It will highlight the motivations, and methods applied to achieve a compelling user experience in the repository and beyond through the creation of user-friendly, automated workflows between various systems.

As part of the demonstration, the presenter will highlight development leading up to the launch in July of last year (and development since then) and will highlight our custom repository development as well as how we worked with suppliers of two preservation systems (Archivematica and Preservica). The demonstration will show how Jisc has created a technical and storage architecture that automates many traditionally manual tasks. This includes the 'messaging layer' that allows new products and services to be easily integrated. Time permitting, examples of additional integrations that are currently being developed (such as Pure CRIS System and Haplo repository) will also be shown.

As the UK Higher Education sector's trusted experts in 'digital', Jisc strives to deliver high quality services. The demonstration will highlight our recognition that, in the research domain, the viability of any new platform is greatly affected by the usability of the systems put before users. Platforms have a greater chance of gaining traction if they meet or exceed user's expectations in terms of ease-of-use. With the best will in the world from institutions and policy makers, users have a powerful tendency to use the most frictionless solution to fulfil their given needs; it is therefore vital to ensure we encourage good practice and engagement by delivering services that are intuitive and un-demanding to use. In this sense the demonstration will frame its presentation of user experience as a strategically important part of the wider move to open science.

It will also focus on the need for interoperability. In this day and age, no repository or preservation system can expect to be embraced by the community if it can't be integrated with extant systems (and future systems and standards as they emerge) to allow for the creation of a totally seamless and as pain free as possible user experience.

The demonstration will feature potential solutions to some of the challenges faced in developing such a system, for example creating simple workflows in complex areas and the challenge of preserving diverse research data file formats that aren't currently recognized by existing preservation registries and services.

The demonstration will also show some of the non-technical side of Open Research Hub; such as being informed by and informing good practice in the sector through hosting community events; and creating information, advice and guidance content based on good practice and findings realised through the development of the Open Research Hub. It will also, importantly, cover how Jisc has approached creating a sustainable business model for the service and how we can support institutions in exploring its potential.

III. Demonstration Requirements

This demonstration will need a wired internet connection, a PC, a screen & projector or display screen.

The demonstration will be up to 20 minutes with an additional 5 to 10 minutes for questions/ discussion.

IV. Goals and Conclusion

The take-home messages from this demonstration are:

- These shared services provide opportunities to enhance and integrate existing products and services in the sector.
- The potential for a well-designed, integrated system to transform the user experience. Data deposit doesn't need to be hard.
- The importance of a compelling user experience in the adoption of services among users.
- That systems can be produced that allow disparate systems to interact to provide a solution that is greater than the sum of its parts.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Score Model for the preservation of digital collections

A case for another digital preservation self-evaluation tool

Bert Lemmens
Staff Member PACKED, Belgium
bert@packed.be

Robert Gillesse
Digital Archivist, International Institute of Social History, Netherlands
Robert.gillesse@iisg.nl
https://orcid.org/0000-0002-5021-837

Abstract – The Score Model for the Preservation of Digital Collections is an online (www.scoremodel.org) self-evaluation tool for cultural heritage institutions which aims to help them with getting a grip on digital preservation. The idea and rationale behind the Score Model is that it can be easily used by smaller and less technically advanced organizations. It is therefore as jargon-free as possible and gives organizations hands-on advise on which issues should be tackled first. The Score Model has the form of an online tool in which the user has to register for an account. They then have to answer 56 questions - separated in seven ISO 16363 inspired clusters - which result in a final report. This report gives a comprehensive overview of how the user has scored for the different clusters and which issues the organization should tackle first. Additionally, the report gives an impression of how far the organization has advanced in becoming a Trusted Digital Repository in the form of qualifying for a Core Trust Seal.

Keywords – self-evaluation, risk assessment, certification.

Conference Topics – What steps are needed to build capacity and skills for organizations of all sizes?

I. Introduction

The Score Model for the Preservation of Digital Collections [1] is an easy to use, as much as possible jargon-free, self-evaluation tool that aims to help cultural heritage institutions with getting a grip on digital preservation. It is meant for organizations who take long-term digital access seriously but which are still developing policies and practices in this area. Practically the Score Model comes in the form of an online tool where the user answers 56 questions (or criteria), which result in a final report that gives an indication of how far the organization is advanced in digital preservation and offers suggestions for improvement in the form of an action plan.

II. Motivation and aim

The wish to facilitate an easy-to-use, online self-evaluation digital preservation tool for less technically oriented institutions, was the motivation behind the creation of (yet) another self-evaluation instrument. The existing evaluation or certification tools (see below) are mostly documents with a high level of complexity, full of technical and (OAIS) jargon and therefore not very accessible (in terms of use) to less experienced users. Clearly these tools are often meant as the basis for (external) audits, mostly executed by experts in the field.

However justified this complexity may be in regard of the arduous domain of digital preservation, these tools carry the risk of resulting in the opposite effect, in which inexperienced users drop the case for digital preservation altogether.

The choice for an easy-to-use self-evaluation tool was made on the assumption that by using it, institutions would be more inclined to critically view their own digital preservation policy, expertise, systems and workflows. The format of an online score model was chosen because of its easily accessible and interactive nature, because it can be easily maintained and the because user results can be easily published and (if wished for) shared.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

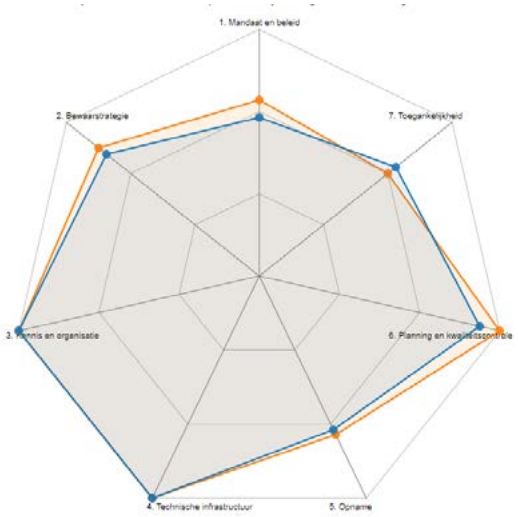
III. How the Score model works

The user of the Score Model has to create an account. This makes it possible to fill in the questions in more sessions (which will often be the case) and revisit earlier made reports. Score Model end reports are explicitly not shared with the outside world and are not used for benchmarking purposes. This was considered essential as the data produced is in many cases of a confidential nature.

After logging in, the user has to create a new report or they can go back to earlier made one. The user then starts filling in the 56 criteria - divided over seven clusters - of which Score Model consist. Each question is to be answered by a deliberately simple yes or no. As in some cases the answer will not always be so clear-cut, the user is urged to answer the questions in a consequent way. Each criterion comes with some contextual information, i.e. an explanation of the risk when no action is undertaken and an example.

As stated already, the criteria are divided into seven sections: organization and policy, preservation strategy, expertise and organization, storage management, ingest, planning and control, and access. These follow more or less the logical order used in ISO 16363 and other evaluation tools. Criteria can have three risk levels: high, normal or low. The idea behind this is that the high-risk criteria ideally have to be solved first, followed by the medium- and low-risk criteria. Also the risk levels have an effect on the score.

When the user has filled in all criteria they will receive a final report which contains a spider graph that shows how they have scored for each cluster. Also they receive an action plan which contains the five most urgent criteria the organisation should ideally solve first. The complete final report contains all criteria and what the user has answered.



Example of the spider graph that is part of the Score Model final report. It shows the scores for the seven clusters of which the Score Model consists. The dark area is the score of the organisation. The light area gives an indication of how the organisation scores in relation to the Core Trust Seal.

IV. Relation with other evaluation tools

The model is based on several well-known audit tools and evaluation checklists for digital preservation: the Trusted Digital Repository (TDR) Checklist (ISO 16363), DRAMBORA and fore mostly the Core Trust Seal. The Score Model was inspired by a concept of the Digital Preservation Capability Maturity Model by Charles Dollar. Like most of these tools, the Score model focuses on both technical and policy/organization-related considerations.

One of the big challenges of building the Score model was translating the often complex and interlinked terminology of these tools and checklists into understandable, but not simplistic criteria. Also for the sake of conciseness the amount of criteria had to stay within reasonable bounds.

The real difference with the aforementioned evaluation tools is thus that the Score Model aims to assist organizations in prioritizing what should ideally be done first to come to a level of 'trustworthiness'. Also the Score Model is unique in that it gives the user an indication of how far they are advanced in gaining TDR status in the form of a Core Trust Seal certificate. This can easily be seen in the spider graph from the final report.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

V. Broader implications

As jargon-free and technically understandable as the Score Model tries to be, some parts of the model may still be hard to be fill out. As the model covers the whole spectrum from institutional policy, financial and organizational practices, preservation planning, ICT strategies and dissemination, it might be a challenge for one person to give answers to all criteria.

This is of course a bigger problem than the Score Model itself. The broad implications of implementing digital preservation policies in an organization are exactly why digital preservation is still very problematic for a lot of cultural heritage institutions. The feeling of urgency and the (costly) investments needed in people, knowledge, soft- and hardware may, in the daily battle for other priorities, be snowed under or even (willfully) ignored. Of course the Score Model as an instrument cannot solve this lack of commitment, means and/or feeling of urgency. However, it can help by giving insight into where the major obstacles lay and where the first steps towards digital preservation "trustworthiness" might be taken.

VI. Conclusions

The Score Model has its own place among other digital preservation self-evaluation and certification tools. Its added value lies in its jargon free, easy to use and practical, hands-on approach to digital preservation challenges.

References

[1] Homepage Scoremodel: https://scoremodel.org. The Score Model was created and is maintained by Flemish and Dutch cultural heritage supporting organizations PACKED (Centre of Expertise in Digital Heritage https://www.packed.be/en/), DEN Digital Heritage Netherlands, https://www.den.nl/ and the International Institute of Social History (https://socialhistory.org) on behalf of the Dutch Heritage Network (https://www.netwerkdigitaalerfgoed.nl/en/).

[2] ISO 16363: Audit and certification of trustworthy digital repositories https://www.iso.org/standard/56510.html, Core Trust Seal, https://www.coretrustseal.org/, DRAMBORA https://www.repositoryaudit.eu/.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Dawn of Digital Repository Certification Under ISO 16363 Exploring the Horizon and Beyond

Perspectives From Three Institutions

David Giaretta
Primary Trustworthy Authorisation Body Ltd, England
david@giaretta.org
ORCID: 0000-0001-8414-7509

Lisa LaPlant
Government Publishing Office , United States
llaplant@gpo.gov
ORCID: 0000-0002-4924-1361

Jamie Shiers
CERN
France, Switzerland
jamie.shiers@cern.ch
ORCID: 0000-0001-8404-3750

Jessica Tieman
Panel Facilitator
Government Publishing Office
United States
jtieman@gpo.gov
ORCID: 0000-0002-9547-0448

Irfan Zuberi
Indira Gandhi National Centre for the Arts, India
irfan.zuberi@ignca.nic.in

Maureen Pennock
Panel Facilitator
British Library
London maureen.pennock@bl.uk
ORCID: 0000-0002-7521-8536

Abstract – The dawn of Trustworthy Digital Repository Certification under the ISO 16363:2012 standard is on the horizon. Across the digital preservation community, institutions are eager to learn more about the processes of preparing for and undergoing an ISO 16363 audit from an accredited third-party organization. As the first ISO 16363 audits in the world have been performed, repositories want to learn value and benefit that certification provides. This panel features representatives from three different repositories representing three countries with distinct collections, designated communities, organizational infrastructures, and unique challenges. Institutions represented on the panel have either recently achieved certified or are currently undergoing an ISO 16363 audit. This panel will explore each repository's experience during, leading up to, and following certification. The panel will include a representative from the accredited external auditing body who has performed these audits to respond to audience questions about the audit process. Panelists from repositories will present varying perspectives on the future of digital repository certification, the role of digital preservation standards, and approaches to implementation. All panelists will present arguments, concerns, and criticisms regarding the ISO 16363 standard and existing methods of repository assessment.

Keywords – repository, trustworthy, standards, audit, certification

Conference Topics – Exploring New Horizons

I. Introduction

As of March 2019, two digital repositories have received certification under the ISO 16363:2012 Audit and Certification of Trustworthy Digital Repositories standard. Only one auditing body, the Primary Trustworthy Digital Repository Authorisation Body, Ltd. (PTAB) has publically announced their accreditation to perform such audits[1]. Though the digital preservation community has recognized the importance of standards and best practices for over two decades, now that an ISO 16363 certification process is in place, a relatively small number of repositories have initiated an external audit. Those repositories which have pursued certification represent diverse collections and designated communities, and this diversity presents unique challenges for the audit and audit preparation activities. Based on their experience with the audit and certification process, these early adopters are eager to share their perspective about the horizon of digital repository certification and standards including a discussion of the feasibility and value of an ISO 16363 for various types of digital repositories and organizations.

II. Institutional Perspectives

This panel features representatives from the United States Government Publishing Office (GPO),

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

the Indira Gandhi National Centre for the Arts National Cultural AudioVisual Archives (NCAA), and the European Organization for Nuclear Research (CERN). Each of these repositories have either received certification or are in the process of being certified under the ISO 16363:2012 standard. Also represented on the panel is the British Library, an institution which has performed a self-assessment against repository standards but currently does not intend to pursue ISO 16363 certification. Each of these organizations have specific institutional missions and stakeholders. These varying institutional contexts prompted different, yet similar, reasons for pursuing or not ISO 16363:2012 certification. Representatives from each organization will discuss which assessment methodology was the most appropriate for their institution and goals. Panel questions will be presented from two alternating members of the panel in order to allow for full participation in responding to prepared questions, as well as facilitated questions from the audience.

For each of these institutions, the implementation and operation of a standards-based digital repository is one of their key responsibilities. Preparing for the ISO 16363:2012 audits and self-assessments entailed significant preparation and training. Each panelist will share information about their preparatory activities including the benefits of attending an ISO 16363 training course. Panelists will also reflect on challenges to performing the audit, the criticality of institution-wide support, and audit planning.

Panelists will discuss how their repository benefitted from their certification experience. Representatives will also explore the ways in which their organization has improved or changed workflows and processes as an outcome of the audit, how certification has impacted stakeholders or designated community members regarding the repository and its role, and how each repository measures the success of their audit. In addition, for those repositories that have achieved certification, the panel will also discuss their next steps to maintain certification and how they view the future for their repository post-certification.

III. Exploring the Direction of ISO 16363:2012

In addition to the three institutional repositories represented on this panel, PTAB, an auditing body and contributing author to the ISO 16363:2012 standard, will provide perspectives on the historical development of the standard and the ISO certification process. PTAB will reflect on feedback received in regards to the standards and the certification process and how, as auditors, they believe that formal repository certification will impact digital preservation and broader institutional missions, communities, and nations.

Each repository will have the opportunity to share assumptions they had prior to the certification process and how their experience working with the auditing body might inform their preparation for future surveillance and re-certification audits in order to maintain certification. Repository representatives will share their perspectives on ways in which ISO certification and the implementation of these standards may be impacted by major policy directives, such as national efforts for open access, accessibility, transparency laws or other legal mandates.

Through a series of facilitated questions and responses, the repositories will respond to frequent arguments against ISO 16363 certification. Participants on the panel will be prompted to consider the appropriateness of ISO 16363 certification as compared to tiered approaches to assessment, such as the WDS/RDA repository audit method, or peer-to-peer assessment models. Panelists will respond to whether or not the certification will be able to maintain relevance if more repositories are not participating in ISO 16363 audits and if the certification is ultimately successful in establishing reputation across the digital preservation professional community. Is the ISO standard too complex or comprehensive to be practicable for most institutions? Are ISO 16363 audits truly effective and rigorous enough to evaluate long-term trust in institutions with organizational infrastructure and funding models which are complex, interdependent, and changing? Panelists will consider the costs of certification versus the costs of risk associated with

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

unreliable preservation systems. How would the ISO audit be impacted by reliance on third-party software services, enterprise technology dependences, or even geographically distributed systems? The audience will have the opportunity to question panelists, including representation from PTAB, about the audit experience, or more broadly, the suitability of assessment for their individual situations. Representation from the British Library allows the audience to gain perspective on ways in which repository managers can evaluate and improve their institutional practices successfully without pursuing ISO certification. Additionally, panelists will review the certification process and share feedback on how digital preservation standards could be more applicable and approachable to other repositories interested in pursuing assessment and certification under ISO 16363.

References

[1] Audit & Certification, Primary Trustworthy Digital Repository Authorisation Body Ltd. http://www.iso16363.org/iso-certification/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Preserving eBooks: Past, Present and Future

A Series of National Library Perspectives

Trevor Owens
Library of Congress, USA
trow@loc.gov
https://orcid.org/0000-0001-8857-388X

Maureen Pennock
British Library, UK
maureen.pennock@bl.uk
https://orcid.org/0000-0002-7521-8536

Faye Lemay
Library & Archives Canada
Canada
faye.lemay@canada.ca

Tobias Steinke
Deutsche Nationalbibliothek
Germany
t.steinke@dnb.de
https://orcid.org/0000-0002-3999-1687

Abstract – This panel will present and discuss different eBook workflows and challenges from four national libraries, considering a range of issues from technical complexities to evolution of the content type and changes in the publishing/collecting landscape.

Keywords – digital preservation, ebooks, ingest, formats, scale, access

Conference Topics – The Cutting Edge: Technical Infrastructure & Implementation; Exploring New Horizons

I. Overview

eBooks are the backbone of many a National Library collection, constituting a substantial proportion of the digital content our readers expect to be able to access and consult. Our digital preservation activities reflect this, with established infrastructures and workflows for eBook acquisition, ingest, management and access, all at scale. Yet the eBook as a content type is evolving, and user expectations for access are evolving alongside. Dealing with this requires both a responsive framework and an eye on the horizon.

This panel brings together experts from leading national libraries to openly discuss various elements of their respective eBook preservation activities and research programs, and explore where similarities and differences may lie. Below we summarize the eBook collections at each organization, existing challenges, and research activities.

A. eBooks at the British Library

Since 2013 The British Library has collected eBooks under the UK's Non-Print Legal Deposit (NPLD) Regulations. Our preferred formats are EPUB and PDF though we also have a small number of MOBI files. There are around 400,000 NPLD eBooks in the collection with access rates at around 5,500 per month. We also have a substantial number of digitized books published under commercial partnerships with Google and Microsoft. Going forwards, we have an interest in Open Access eBooks published outside of the UK and eBooks published as mobile apps.

Current challenges include ensuring an uninterrupted supply to readers during a forthcoming repository migration, and delivering access to all six UK Legal Deposit Libraries in line with regulation requirements for single sequential access. Active research areas include collection and preservation of mobile apps and evolution of the EPUB format.

B. eBooks at the Library of Congress

The U.S Library of Congress has acquired eBooks through a wide range of different programs and initiatives. For years, the institution has received and acquired eBooks through its Cataloging in Publication Program, special relief agreements for copyright deposit, web archiving, and other routine transfer methods for acquisition.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

In support of the digital collecting plan, staff across the institution are currently working to expand these efforts and to pilot acquiring, preserving, and delivering selected open access eBooks. The majority of this content is in PDF and EPUB formats, but the institution has copies of eBooks in a much wider range of formats as well. As outlined in the Library of Congress Digital Strategy, it is necessary to plan for work around eBooks in terms of exponential collection growth. To that end, a key area of focus for the institution is working to scale up and enhance workflows and processes.

C. eBooks at the Deutsche Nationalbibliothek

The German National Library has currently around 1 million eBooks in the formats PDF and EPUB, equating to approx. 16% of all collected digital publications (excluding digitized objects). The German legal deposit collection has included eBooks since 2006. eBooks are ingested in the digital preservation system of the German National Library. All eBooks are analyzed and validated, resulting in generation of a risk analysis 'ingest level'. Checks include tests on copy protection especially in PDF files. There is a separate repository for giving access.

In an ongoing internal project all aspects of the digital workflows are currently being optimized for a better performance. This includes using a common workflow engine, replacing the repository for access with something more fitting and consolidating the different workflows for digital objects including eBooks.

D. eBooks at Library & Archives Canada

LAC has been acquiring eBooks of various different formats since the 1990's. Digital legal deposit legislation came into effect in 2006, though participation in the legal deposit program varies with commercial/retail publishers and scholarly communities lagging behind government and self-published content.

The current technical platform for eBook acquisition is based on a pilot project created in 1994. In 2018, LAC embarked on an initiative to modernize its systems and, as part of that, procured Preservica as a DAM and a Digital Preservation Solution. New information package specifications for published heritage collections are currently being developed for use within Preservica. In addition, LAC's Published Acquisitions sector is working to implement a collection gap analysis and monitoring framework in order to measure and expand participation in the Legal Deposit program. Another key activity is the development of a seamless platform for publishers and authors to transfer digital content and metadata to LAC. One of the desirable outcomes is to ensure that streamlined workflows from acquisition to preservation are developed.

II. Panel Structure

Following short introductions on the state of the practice to acquire, preserve, and deliver eBooks at each institution, panelists will then move on to discuss a range of questions such as:

- How does your organization staff and support eBook acquisition, preservation and access?
- How have you embedded preservation support into your end to end workflows?
- Do you have preferred formats for eBook preservation; if so, what are they and why?
- What are the biggest challenges you have encountered in collecting, preserving and providing access to eBooks?
- What changes have you seen in your eBook collection over the past decade and how have you responded?
- How are you monitoring the publishing landscape for more changes going forwards?

Panelists will discuss answers in advance of the session to ensure answers are representative of the variety in our approaches, thus ensuring we provide sufficient conflicting perspectives to create interesting discussion. Attendees will be encouraged to ask additional questions of the panelists during an open-ended Q&A session.

III. Panelists

Maureen Pennock is Head of Digital Preservation at the British Library. She sits on the Digital Preservation Coalition Board of Directors and co-chairs the DPC Special Interest Group for Digital Preservation in National Libraries, Archives and Museums. She is also Chair of the UK Legal Deposit Libraries' Digital Preservation Committee and a member of the UNESCO PERSIST initiative.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Dr. Trevor Owens serves as the first Head of Digital Content Management at the U.S. Library of Congress. In addition, he teaches graduate seminars in digital history for American University's History Department and graduate seminars and digital preservation for the University of Maryland's College of Information, where he is also a Research Affiliate with the Digital Curation Innovation Center

Tobias Steinke works at the German National Library on the conceptual development of digital preservation and is responsible for the web archiving project of the library. He has been involved in several national and international projects about digital preservation and standardization.

Faye Lemay has been the Manager of Digital Preservation at Library and Archives Canada for nearly a decade and has been the driving force in the development and deployment of a comprehensive digital preservation program. She oversees the long-term preservation of Canada's digital documentary heritage comprised of published heritage, government records and private archives.

The panel will be moderated by Paul Wheatley, Head of Research & Practice at the Digital Preservation Coalition. Paul is an experienced panelist and moderator with many years of experience working with digital collections and in digital preservation.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

(I Can't Get No) Satisfaction

Measuring Success in Digital Stewardship Programs

Karl-Rainer Blumenthal
Internet Archive, USA
karlb@archive.org
0000-0001-5686-9950

Peggy Griesinger
University of Notre Dame, USA
mgriesi2@nd.edu
0000-0003-2771-5367

Julia Kim
Library of Congress, USA
juliakim@loc.gov
0000-0002-5229-1191

Shira Peltzman
University of California, Los Angeles, USA
speltzman@library.ucla.edu
0000-0003-0067-2782

Vicky Steeves
New York University, USA
vicky.steeves@nyu.edu
0000-0003-4298-168X

Erwin Verbruggen
Institute for Sound and Vision, Netherlands
everbruggen@beeldengeluid.nl
0000-0003-2004-1945

Abstract – Approaches to digital stewardship vary from institution to institution. Given the substantial differences among organizational models and program maturity, what indicates successful organization of the long term work of digital preservation to practitioners? Panelists will introduce and contextualize their ongoing research into the shared characteristics of successful digital preservation programs throughout the field. They will identify the emergent themes articulated by research subjects thus far and engage attendees to discuss challenges and opportunities of digital preservation at their respective institutions.

Keywords – maintenance, sustainability, policies, organizational culture

Conference Topics – Designing and Delivering Sustainable Digital Preservation; Building Capacity, Capability and Community

I. Background

Recent publications indicate that practitioners are increasingly discontent with how their institutions organize digital preservation responsibilities and duties. The 2017 staffing survey conducted by the National Digital Stewardship Alliance (NDSA) asked respondents whether or not they agreed with the following statement: "The way our digital preservation function is currently organized (staffing levels, expertise, where they are placed within the larger organization) works well." Of the 133 people who took the survey, roughly 46% either disagreed or strongly disagreed with that statement[1]. This represents a significant increase from the 34% who responded similarly to the same question in the report's 2012 iteration[2]. The increasing discontent with the status quo suggests that there is ample room for improvement, but stops short of indicating where or how changes should be made. Additionally, Oya Rieger's The State of Digital Preservation in 2018: A Snapshot of Challenges and Gaps articulates some areas in need of improvement as reported by senior management and "thought leaders" in the field, among them: ambiguity of responsibilities, misalignment of expectations, and the need to attend to inclusivity, diversity, and social justice[3]. The opportunity remains to engage a diverse spectrum of digital preservation practitioners to solicit perspectives on what works well, what does not, and what organizational improvements might address existing shortcomings and concerns. In this panel, members of the project team will discuss their research and engage audience members in a discussion that addresses key themes that have emerged thus far.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

II. Study

Panelists are researching how these gaps impact the experience of practitioners throughout the field by conducting and analyzing a series of in-depth interviews with digital preservation professionals who represent diverse backgrounds, organization types, career stages, and managerial responsibilities. Participants in the study were asked to reflect on and evaluate how digital preservation is organized at their institutions. The research seeks to identify the metrics and to analyze the organizational and cultural factors upon which digital preservation practitioners base their judgments of program design. The goal of this study is to identify specific areas and benchmarks for improvement in digital preservation program design.

III. Panel

A. Panel Structure

The first 10 minutes will provide an overview of the research project and its methodology. Following this brief introduction, audience members will be asked to respond to several anonymous multiple choice polls (administered either digitally via Twitter or manually via notecards) that will be designed to provide a jumping off point for an interactive discussion that centers on the challenges and opportunities of digital preservation at their respective institutions. (E.g.: "If you could make one change at your organization to improve digital stewardship where would you focus? A) Leadership, B) Staffing, C) Policy, D) Funding). The poll will serve as an ice-breaker to get the audience members thinking about this subject concretely. The moderator will make observations about the results and ask both panelists and audience members alike a series of questions designed to spark discussion and debate around the research project's major themes.

This will have two benefits: it will allow the panelists to discuss their individual and tentative conclusions, conjectures, and reflections based on their existing research. It will also enable attendees to share their insight and interpretations, which will, in turn, increase the diversity of perspectives considered in shaping the project's conclusions and outcomes. As this research is ongoing, questions are incompletely resolved. There is ambiguity among panelists' interpretation of the sometimes contradictory responses to structured interview questions on how to best address the growing dissatisfaction, both expressed in the project's literature review and in the ongoing research. Audience input will enrich the panelists' understanding of their existing data and emergent themes.

B. Emergent Themes

Themes that have already emerged in this research and which will guide the majority of the discussion include: the role of leadership in articulating and supporting a strategic vision and/or mission statements; the effects of different internal decision-making paradigms; communication and digital preservation comprehension needs among coworkers, managers, donors, and peers; immediate priorities for programmatic change; and other factors that contribute to low morale and burnout among practitioners.

C. Impact

This guided discussion will encourage discussion and debate about how digital preservation can be done well and how it can be done poorly, regardless of an institution's size or capacity. The themes discussed in this panel will reveal key signifiers of success or failure in digital preservation programs that transcend variations in institutional type or funding model. They will inform efforts to improve the overall functionality of digital preservation programs and will have practical implications for practitioners themselves, the middle managers who often directly oversee their work, and the senior-level administrators charged with leading their organizations. Ultimately a vision for a successful model of digital preservation practice will emerge -- one that sustains its workforce as well as its mission.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

References

[1] W. Atkins, C. Ghering, M. Kidd, C. Kussmann, J. M. Perrin, M. Phillips, S.Schaefer, Staffing for Effective Digital Preservation 2017: An NDSA Report. The National Digital Stewardship Alliance. 2017. https://ndsa.org/documents/Report_2017DigitalPreservationStaffingSurvey.pdf

[2] W. Atkins, A. Goethals, C. Kussmann, M. Phillips, M. Vardigan, Staffing for Effective Digital Preservation: An NDSA Report. The National Digital Stewardship Alliance. December 2013. https://ndsa.org/documents/NDSA-Staffing-Survey-Report-Final122013.pdf

[3] O. Rieger. The State of Digital Preservation in 2018: A Snapshot of Challenges and Gaps. Ithaka S+R. October 2018. https://doi.org/10.18665/sr.310626

[4] Rémi Rampin, Vicky Steeves, & Sarah DeMott. (2019, March 23). Taguette (Version 0.5). Zenodo. http://doi.org/10.5281/zenodo.2605647

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

The People And Processes of Digital Preservation

International organizations leveraging internal wisdom to build support for digital records

Jeanne Kramer-Smyth
World Bank Group
USA
jkramersmyth@worldbankgroup.org

Montserrat Canela
UNHCR
Switzerland
canelaga@unhcr.org

Ineke Deserno
NATO
Belgium
deserno.jozefina@hq.nato.int

Abstract: Most initiatives to build digital preservation programs focus almost exclusively on technical issues: hardware, software, and information architecture. Our panel will instead discuss the organizational, functional, and staff-related questions that needed to be answered in order for three archives at international multilateral organizations to evolve their programs to support digital records. A significant portion of the panel will be dedicated to comparing and contrasting our challenges, approaches, and successes.Keywords – digital preservation, capacity planning, organizational issues, leveraging expertise, international organizations

Conference Topics – Building Capacity, Capability and Community: Designing and Delivering Sustainable Digital Preservation

I. Introduction

Staff from the archives of three international organizations: the World Bank Group (WBG); the United Nations High Commissioner for Refugees (UNHCR); and the North Atlantic Treaty Organization (NATO), will discuss lessons learned and approaches to the preservation of digital records. The juridical environment of each international organization creates a unique situation for their archives and for the staff who care for them. The archives of international organizations are not subject to the laws or government directives that normally provide a mandate for compliance in records management and long-term preservation. These archives typically have a role to play throughout the records lifecycle: from birth, to death, to the archival afterlife. The knowledge, expertise and experience of staff working in the archives of international organizations are the bedrock of the records management and archival work for these institutions. With born-digital records the functions remain the same, but the way the functions are executed will be different. By engaging the diverse talents of the people in our teams and valuing our shared expertise within our organizations, we have the best chance of creating procedures and methods that carry our archives forward into a future dedicated to preserving and providing access to born-digital records.

A. At the World Bank Group (WBG)

1. Challenges

In launching the WBG's digital preservation program, we identified the people and process investments needed to evolve into an archives that can handle both analog and born-digital records. The new solution needed support from both our existing archives team as well as from teams across our broader organization. We also needed to leverage our existing infrastructures and archival expertise.

2. Approaches

The implementation of our 'Digital Vault' has relied on giving agency to the staff of our archives. We identified six focus areas around which we created teams to evaluate and tackle all the related issues: Transfer

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

& Ingest; Description, Metadata, & Intellectual Control; Access & Reference; Selection; Strategy & Planning; and Formats, Preservation, & Long-Term Access. The focus teams have been given great latitude in how they organize their work - prioritizing and assigning deliverables to answer their subject area's questions. Since our staff have in-depth knowledge of WBG records throughout the lifecycle (and the technology at play), important efficiencies and opportunities in digital records preservation have been revealed..

3. Successes

Rather than attempting to propose and implement changes to our existing processes, workflows, staffing model, and standards from the top down, this approach of engaging the full team has given them greater agency and personal investment in this new endeavour. It has also enabled us to build bridges to other teams across the WBG.

B. At the United Nations High Commissioner for Refugees (UNHCR)

1. Challenges

Critical to our work has been finding ways to support our organization's core mandate to protect refugees and displaced people. With 68.5 million refugees worldwide, over 16,000 staff stationed all over the world in some of the world's most challenging situations, complex data and formats, and outsourced IT services, UNHCR faces many challenges. How can we justify the investment in such a system over, for example, the investment in things that directly save lives? How, can an outsider understand UNHCR well enough to gain acceptance within the organization and argue for the necessity of such a solution? Where to start?

2. Approaches

Our solutions seeks to fit UNHCR's particular organizational, technological and (of course) resource limitations while giving back benefits to our stakeholders: refugees/displaced people, staff and the community of practice. It looks at how to solve and support, not drain existing resources.

3. Successes

UNHCR's work toward digital preservation started long before it embarked on the Digital Preservation Project. Records Management in the paper and the electronic environment has proven to be the foundation for the success of DPP.

C. At the North Atlantic Treaty Organization (NATO)

1. Challenges

The complexity of military and political interrelationships at the international level and the wide spectrum of situations in which NATO digital information is generated requires specific procedures and approaches for their management and preservation. The multinational character of NATO missions demands a solid policy and procedural framework for all stakeholders addressing reliability, authenticity, confidentiality, integrity, ownership and availability of digital information.

2. Approaches

The NATO Archives acquired a tool to enable the preservation of digital information of permanent value and to establish a trusted digital repository. Currently all efforts aim to develop procedures and working practices for the preservation of NATO's digital information of permanent value. We include all stakeholders throughout NATO in the process. The stakeholders included members of the NATO information, archival and data community as well as experts in NATO member countries. This covers the pre-ingest process of digital information, the development of a metadata schema and in the organization of its collections. The goal is to share the expertise with the NATO information and data management community and to incorporate the procedures into information and knowledge management policies and systems throughout the Organisation.

3. Successes

The Archives Committee, composed of archival and information management experts from member countries of NATO, provided guidance and direction for digital preservation at NATO. A solid policy framework for the preservation of NATO's digital heritage was established. This framework has been fundamental in raising awareness throughout the Organisation about both the complexity of preserving digital information and the importance of addressing digital preservation early in the lifecycle.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

D. Discussion

Much of the panel time will be reserved for discussion of the contrasts among our approaches and answering questions from the audience. These are some sample questions we plan to address:

- What is your best advice to an archives starting this journey?
- What were the biggest organizational, functional, or staff-related challenges you had to overcome?
- What allies from across the information management & technology communities within your organization were most supportive?
- In what ways did your changes to handle hybrid records change performance of existing tasks?
- How did you build understanding and support for your digital preservation program, often in the shadow of limited resources and demands of our organization's primary missions?

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Achieving Criticality of Preservation Knowledge

Sustaining Digital Preservation in the Nuclear Field

Elizabeth Kata
International Atomic Energy Agency
Austria
E.Kata@iaea.org
https://orcid.org/0000-0002-9916-5854

William Kilbride
Digital Preservation Coalition
United Kingdom
william.kilbride@dpconline.org
https://orcid.org/

Jenny Mitcham
Digital Preservation Coalition
United Kingdom
jenny.mitcham@dpconline.org
https://orcid.org/0000-0003-2884-542X

Jim Moye
J&A Preservation
United States
moye2@LLNL.gov
https://orcid.org/0000-0002-4905-1445

Abstract – The concerns of the digital preservation community are shared by those working in the nuclear field - preserving information for as long as it is required, taking into account the needs of diverse stakeholders, explaining needs to IT professionals, preserving data from legacy systems and obsolete media, managing accessibility and information security. Yet the length of time knowledge needs to be accessible and understandable, the accountability for public safety, and non-proliferation concerns present further challenges in the preservation of information related to nuclear science and its applications. Preservation efforts in this field must be approached and executed with these challenges in mind while also aligning with community good practices.

Practitioners from the nuclear field will present their experiences and. the panel discussion will address collaboration and impediments, appraisal and records management in light of the length of time material must be available, and preservation approaches. The panel will demonstrate the ways in which those in the nuclear field benefit from and apply good preservation practices and highlight how practitioners in the nuclear field can contribute to the broader digital preservation community thus opening a dialogue on how we can progress together.

Keywords – digital preservation, nuclear knowledge management, decommissioning, information security, digitization.

Conference Topics – Exploring New Horizons; Building Capacity, Capability and Community.

I. Breaking Down The Silos

Knowledge management has been a topic of discussion in the nuclear field since the outset of the 21st century, yet digital preservation of nuclear knowledge may be relegated to a side note or addressed only cursorily[1]. Preservation experts, archivists, and records and information managers in the nuclear field only come into limited contact with one another or the larger digital preservation community. The sometimes classified nature of our holdings can make it more difficult to exchange information and experiences. Without making claims to cover the full range of issues in preserving nuclear knowledge, in particular research data management aspects, this panel seeks to bring representatives working on digital preservation issues in the nuclear field together to come into dialogue with one another and with a wider digital preservation public.

Practitioners working on the preservation of nuclear knowledge from the International Atomic Energy Agency (IAEA), the Digital Preservation Coalition (DPC), and J&A Preservation will each address, in 8-10 minute presentations, challenges in the field. They will discuss how they relate to digital preservation challenges in general, how the approaches in the nuclear field could further inform the efforts of the digital preservation community, and how through sustained dialogue we might achieve criticality in our preservation knowledge.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

II. Long-Term Accountability To Many Stakeholders

Elizabeth Kata will introduce some of the digital preservation challenges in the nuclear field she has encountered in her work at the IAEA. With a dual mission to promote the peaceful uses of atomic energy while preventing the proliferation of nuclear weapons, the IAEA leads knowledge sharing projects on the one hand and tightly guards certain information on the other. Stakeholders are not only the Agency itself, but rather all Member States, the public, and commercial industries, which requires a heightened awareness for building and maintaining trust across the spectrum of stakeholders.

The need to balance access to authorized users with information security over time (such as a timeframe of 100 years used for nuclear power plants) presents a specific set of problems. Due to the complex IT environment, some aspects of commercial or community preservation solutions do not fully meet the specialized needs, requiring rethinking or expanding approaches to issues like migration or storage.

III. Reliable Robust and Resilient Digital Infrastructure for Nuclear Decommissioning

Jenny Mitcham of the DPC will present on a collaborative project with the Nuclear Decommissioning Agency (NDA) in the UK [2]. The NDA has been charged with the complicated task of decommissioning and cleaning the seventeen principal nuclear energy plants in the UK, a task described as the largest and most important environmental restoration programme in Europe. Amongst its many challenges, the NDA is by default required to become a trusted leader for information management and digital preservation, thus the project draws on the preservation understanding of the DPC and its wider membership.

Jenny will report on this ongoing project and share some of the challenges and themes that have emerged thus far. These include the high number of legacy data systems to manage, new data collection methodologies and technologies constantly evolving, compliance within a highly regulated environment and managing semi-current records for long periods.

IV. Digitizing Films Of Nuclear Tests

Jim Moye has over 40 years experience in the motion picture post-production industry and as an expert in film preservation. For the past five years he has been under contract by Lawrence Livermore National Laboratories, to digitize the scientific films from the above-ground nuclear tests.

There were 210 atmospheric nuclear tests and another nine nuclear cratering tests performed by the United States during the 1940s, 1950s, and early 1960s. Every test was extensively recorded with high quality motion picture films and photographs, providing a unique historic record of circa 10,000 films/photos. These newly digitized films/photos are now being re-analyzed using modern image processing techniques to obtain more accurate and precise measurements of the nuclear yield of each test and other scientific data pertaining to nuclear weapon effects.

Jim will discuss the challenges with preserving this historic scientific record, including the unique technical requirements of digitizing the films to allow accurate analysis, the software for automated analysis, the requirements for secure storage and the management of more than 10 million scanned images.

V. IPRES 2019: Opening A Dialogue

The discussion following the presentations, moderated by William Kilbride of the Digital Preservation Coalition, will highlight the ways we learn from each other and how information and preservation professionals in the nuclear field can contribute to and gain from the digital preservation community. We will address topics such as: Is digital preservation in the nuclear industry different? Why or how? What are impediments to collaboration with the wider digital preservation community or even each other? How can we overcome them? Does the length of time material needs to be accessible and the length of time material needs to remain classified have an effect on the preservation approaches? If yes, how is this addressed? What records

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

management and appraisal approaches are taken? How do we balance the importance of making the right decisions and the importance of not keeping too much? These questions and input from the audience will further open the dialogue and enable a discourse on how we can make progress together.

References

[1] Nuclear Knowledge Management Challenges and Approaches, Vienna: International Atomic Energy Agency, 2018, pp. 26-27.

[2] Reliable, Robust and Resilient: DPC supports Digital Infrastructure Project for Nuclear Decommissioning Authority, Digital Preservation Coalition. https://www.dpconline.org/news/dpc-supports-digital-infrastructure-project-for-nda

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Stronger together; collaboration as a matter of course

Publishers Round Table

Eefke Smit
the International Association of STM Publishers
The Netherlands
smit@stm-assoc.org

Abstract – A discussion panel will bring together the views of content providers, DP-service providers and user experts to investigate how cross stakeholder collaboration can ensure better digital preservation.

Keywords – cross stakeholder collaboration, size of research data, user perspective, publishers.

Conference Topics – Collaboration: a Necessity, an Opportunity or a Luxury?; Building Capacity, Capability and Community.

I. Contents

A discussion panel will bring together the views of content providers, DP-service providers and user experts to investigate how cross stakeholder collaboration can ensure better digital preservation.

Evidence will be gathered from examples around CLOCKSS, PORTICO and the KB e-depot in The Netherlands. Forefront players at the side of content providers will explain their stake in the game and will provide an outlook of next challenges to cope with. How will the sheer size of research data be catered for ? An expert on user-research will bring in the perspective of The User of Digital Preservation facilities; what need are we trying to fulfil.

The panel will consist of 2 publishers (Elsevier, SpringerNature), 2 Digital Preservation service providers (CLOCKSS, Portico) and a User expert

The panel will be moderated by Eefke Smit, Director for Standards and Technology of the International Association of STM Publishers.

Panel members confirmed so far: Craig van Dyke (CEO CLOCKSS), Wouter Haak (VP Research Data, Elsevier), Maurits van der Graaf (MD Pleiade Market Research), Kate Wittenberg (CEO Portico), more names TBA soon.

The strength of the panel lies in its composition:

- bringing together representatives of two global service providers for Digital preservation, both frontier pioneers at the time this all started: Portico and CLOCKSS, each of them working along very different philosophies;
- representatives of two large content providers who also have played a frontrunners' role in the establishment of e-depots and digital archives;
- a researcher who has investigated the use that people make of such archives; much more than was anticipated and in his view proves how important this work is.

The message that the panel aims to bring across is that multi-stakeholder collaboration generates many benefits and is a necessity to accelerate getting the right shared infrastructures in place. With this in mind, the questions will focus around the following topics:

1. (Portico and Clockss) Explain what your organisation does and how it developed itself to this position -- why are their approaches so different ? Is there a benefit to doing things differently ?
2. (Nature, Elsevier) Why is it so important for content providers to fund and support such infrastructure initiatives, what goals does it serve? How do they benefit ?
3. (researcher) How does the user benefit ? How can we grow the number of users and usage ? (ALL + audience)
4. What does the DP-community need for the next 5 to 10 years ? (ALL + audience) What will pose the new challenge ? (ALL + audience)

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

5. How important is multi-stakeholder collaboration for that ? (ALL + audience) How reliable is the support from commercial parties -- what would be the ideal funding model ? (ALL + audience)

The discussion will take on a workshop-like format, to enable the audience to participate actively and intensely. Especially on questions 3, 4 and 5, the moderator will ensure a lively interaction with the audience so that their view is included in the discussion. Conflicting views and debate will not be avoided, with the aim to explain better how collaboration is the overall goal.

The panel is scheduled to last 90 minutes.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

PIDs & Preservation

Incorporating persistent identifiers in a preservation strategy

Helena Cousijn
DataCite, Netherlands
helena.cousijn@datacite.org
https://orcid.org/0000-0001-6660-6214

Ginny Hendricks
Crossref, UK
ghendricks@crossref.org
https://orcid.org/0000-0002-0353-2702

Alice Meadows
ORCID, USA
a.meadows@orcid.org
https://orcid.org/0000-0003-2161-3781

A persistent identifier (PID) provides a long-lasting reference to an entity. PIDs should be open and unique and should resolve to a digital representation of the entity. Used in this way, PIDs can serve as an important component in digital preservation strategies for academic resources.

In this panel discussion, representatives from leading PID organizations will explore roles that PIDs can play in digital preservation strategies. We invite the participants to engage in a conversation on how PID providers can work with the preservation community, and what preservation strategies they should be deploying with PID metadata.

Keywords – Persistent identifiers, metadata, infrastructure, preservation

Conference Topics – Building Capacity, Capability, and Community; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

Crossref, DataCite, and ORCID are well-established providers of persistent identifiers; Crossref and DataCite provide DOIs for research outputs and ORCID provides ORCID identifiers (iDs) for researchers. Ensuring that these identifiers are truly persistent—even beyond the lifetime of the research output or research—is critical for continued access to the entities and/or their metadata. As such, the three organizations have committed to persistence in their organizational practices. All are open, not-for-profit, community-governed, and community-led and all are committed to a collaborative, community-driven approach to ensuring the long-term preservation of research is part of a trusted research information infrastructure.

PID organizations have a goal complementary to that of preservation organizations. Where preservation organizations focus on ensuring that the object remains digitally available over time, PID providers ensure the identifier and identifying metadata for connected objects remain discoverable over time. We work with the community to connect identifiers for the underlying person-place-object entities and make these relationships and associated metadata openly available [1]. By ensuring persistence of digital representations of entities and associated identifiers, preservation and PID organizations contribute to making research outputs FAIR [2] and enabling reuse of research outputs over time.

II. Different approaches to PID persistence

Whilst all three organizations have much in common, each one is approaching the overall question of preservation differently.

Crossref has direct agreements with many archival organizations such as Portico, CLOCKSS, and the Internet Archive, with whom its entire metadata corpus is routinely and systematically backed up, which ensures that the identifiers remain persistent.

However, the existence of a DOI does not alone ensure preservation. The more than 100 million metadata records require commitment from the asserter (e.g., publisher, funder) to maintain the

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

records for the long-term. For this reason, we call Crossref DOIs persist-able rather than persistent. At the most basic level, when members stop publishing or go out of business (and when they tell Crossref) there is a hosted defunct DOI page that asks for information about other locations. In 2018, Crossref introduced the obligation to archive content in their membership terms [3]. In addition, Crossref also has individual agreements with some national libraries and other archiving organizations to redirect DOIs when needed, using our 'multiple resolution' service when multiple archives are involved.

DataCite similarly focuses on the persistence of the identifier and associated entity metadata rather than the research output itself. DataCite members register DOIs with DataCite and take responsibility for maintenance and preservation of the entities for which DOIs are registered. In cases where individual organizations are no longer able to maintain the DOI record, larger organizations (e.g., national libraries) often take on this task.

While preservation and access are often treated as separate functions, DataCite treats them as complementary: preservation aimed at providing access over time, while access depends upon preservation at a point in time [4]. This is particularly critical when working with non-traditional scholarly outputs such as datasets. Providing access to datasets, related data, versions, software and other outputs allows the provenance of the materials to be accessible and usable over time. To support this, DataCite systematically and regularly checks the health of its DOIs by checking for successful resolution.

In the worst-case scenario, when entities disappear from the digital domain and the connection with their DOI breaks and return a "404 error", the DOI Foundation offers a "DOI not found" form for people to report lost DOIs. The DOI Foundation also alerts the appropriate DOI Registration Agency via a daily email which is manually actioned.

Most digital preservation work has focused on digital representations of research outputs. However, the organizations where research is performed or funded, and the people that carry out research are just as important to the integrity of the research process. Digital representation of organizations (such as ROR IDs) [5] and persons are at a much earlier stage of community understanding and technological capability. ORCID provides a persistent identifier and landing page for researchers, open APIs, and annual public metadata files, all components of a persistence strategy. However, many of the connections to a person's identifier are to entities that have fleeting if any digital representation (e.g., student or faculty webpage, peer review activity). ORCID has therefore started conversations with preservation specialists and is in the early stages of defining a preservation strategy.

III. Challenges remain

With the proliferation of PIDs and a growing number of low-barrier providers, PIDs are becoming commoditized. At the same time, more mandates are coming into play - whether to "get a DOI" or "get an ORCID". Without a community understanding of the needs and requirements for long-term commitment to stewarding digital content, PIDs cannot be useful for preservation

IV. Panel discussion - an education-based approach

While there are clear opportunities and intentions for PID infrastructures to support the preservation of research outputs, it takes time for these services to be fully supported, and adoption is variable. Ongoing efforts need to emerge from cross-community understanding and a collective commitment to digital representations and preservation. This panel discussion will be led by Craig Van Dyck, Executive Director of CLOCKSS. To start this discussion, we will first ask the audience some important questions: 1) where should the responsibility for the preservation of the content underlying PIDs lie? 2) How should PID providers preserve their metadata? and 3) how can the different communities work together to meet preservation challenges? Based on the responses from the audience, the facilitator will lead a panel discussion on how to develop a cohesive approach to preservation, with persistent identifiers and metadata as core connecting components.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

References

[1] https://www.project-freya.eu/en

[2] Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. https://doi.org/10.1038/sdata.2016.18

[3] https://www.crossref.org/membership/terms/

[4] Abrams, Cruse and Kunze (2009). Preservation is not a Place. https://doi.org/10.2218/ijdc.v4i1.72

[5] https://ror.org

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

CLOUD ATLAS

Navigating the Cloud for Digital Preservation

Andrea Goethals
National Library of New Zealand
New Zealand
andrea.goethals@dia.govt.nz
orcid: 0000-0002-5254-9818

Jefferson Bailey
Internet Archive
USA
jefferson@archive.org
orcid: 0000-0002-0830-6325

Roslynn Ross
Library and Archives Canada
Canada
roslynn.ross@canada.ca
orcid: 0000-0001-6006-6034

Nicholas Taylor
Stanford Libraries
USA
ntay@stanford.edu
orcid: 0000-0002-9504-0640

Abstract – Cloud services have become the increasingly dominant paradigm for many other types of IT-based services; why not for digital preservation? Cloud services are by now a familiar, though not mature, part of the digital preservation landscape. It is worth considering further how commercial cloud services are currently or might prospectively be used for digital preservation, as well as the implications for memory institutions, individually and collectively, if digital preservation shifts primarily in that direction. This panel will offer contrasting institutional perspectives on the potential or the perils of the cloud for digital preservation, featuring case studies on how memory institutions can leverage the cloud in deliberate and mission-supporting ways, and how some are working to build alternative, community-based infrastructures.

Keywords – cloud, community, digital preservation, infrastructure, storage

Conference Topics – Exploring New Horizons; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

The hallmarks of cloud-hosted services are unparalleled flexibility and scalability, features that have proven sufficiently compelling as to bring about a sea-change in the way that organizations think about their IT infrastructures. The cultural heritage community is no exception; both consideration and application of cloud services for digital preservation has accelerated, as memory institutions negotiate the evolution of their parent organizations' IT strategy, if not the shift in the macro IT environment. How to best leverage cloud services to serve digital preservation or build complementary or alternative community-based services are areas of active exploration. Application of cloud services for digital preservation should ideally consider both fitness and externalities.

Notwithstanding their dynamic impact for IT and society, the large commercial cloud service providers have lately come in for a good deal of criticism. Some in the digital preservation field have raised questions regarding the fitness of commercial cloud providers in meeting fundamental characteristics that the best practices in digital preservation necessitate. For example, David S.H. Rosenthal points out that many cloud storage services have a number of conspicuous shortcomings: continuity of storage contingent on continuity of payment, opaque data integrity assurance mechanisms, vulnerability to privileged insiders, vulnerability to operator error, attractiveness as a target of attack, steep fees for data egress, and organizational immaturity.[1]

Cultural heritage institutions would do well to mind alignment with their goals and values. The library profession has traditionally held privacy protection as a core tenet. How well do the business models of cloud services support this value, or can cloud services be leveraged in such a way that this concern is mitigated? How can memory institutions utilize cloud services to enhance their impact and

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

relevance, and not cede either capacity or commitment to serving as information stewards? Can, or how can, memory institutions enlist cloud service providers to partner in meaningful ways to support community needs, given our rich missions and often modest resources, as well as the near-monopoly power of some of the key companies?

These questions highlight areas of possible mission misalignment between memory institutions and cloud service providers. The latter typically prioritize sustainability rather than profit, openness rather than capture, diversification rather than vertical integration, long-term stewardship rather than market-contingent product commitment, expansive dissemination rather than tolled access, community rather than market predominance.

At the same time, the trends in the IT environment are both unmistakable and unignorable, and cloud services - whether those specifically provided by well-known commercial companies, or the paradigm of hosted services more generally - offer advantages that definitely make them worthy of consideration for use in digital preservation. Acknowledging that cloud services play a growing and important role in digital preservation, it is crucial that those entrusted to digital preservation approach this new environment with the ability to assess trade-offs between benefits and risks, a vision for the infrastructures to help realize as a field, and an interest in mapping what alternatives the "cloud" might afford

II. Session objectives

The panel members will offer case studies of how their respective organizations, individually and in partnership with one another, are negotiating the role of cloud services as part of or in juxtaposition to their digital preservation infrastructure portfolios. Along the way, the panelists and facilitator will engage in and stimulate audience discussion on a number of important and timely questions:

How can cloud services best be utilized to enhance and extend digital preservation practice and capabilities without compromising core mission or ceding core capacities?

How can the cultural heritage community nurture complementary, principled, and practical alternatives to commercial cloud services, and why does that matter?

Where do cloud services offer the greatest potential for impact for digital preservation? Where do they pose the greatest risks?

III. Contributors

Andrea Goethals is the Digital Preservation Manager at the National Library of New Zealand where she manages the Preservation Research & Consultancy Team. Andrea will serve as a facilitator for this session.

Jefferson Bailey is Director of Web Archiving and Data Services at Internet Archive. Jefferson will discuss Internet Archive's practically cloud-scale infrastructure from several angles, including operational considerations, effective and efficient scaling, and pilot services exploring a more explicit repositioning of Internet Archive infrastructure as a commercial cloud alternative for some use cases.

Roslynn Ross is the Director of Digital Preservation and Migration at Library and Archives Canada. Roz will discuss the challenges of managing a large digital collection and considerations when working with diverse teams to implement a technical solution leveraging cloud capabilities.

Nicholas Taylor is the Program Manager for LOCKSS and Web Archiving at Stanford Libraries. Nicholas will discuss the LOCKSS Program's long-standing support of community-based, local, distributed digital preservation infrastructure as a juxtaposition to commercial cloud services. While the LOCKSS Program has generally maintained a skeptical orientation towards cloud services for digital preservation storage, in particular, a recent, major software re-architecture offers opportunities to better leverage cloud-like infrastructures as well as strengthen the model for LOCKSS as a hosted service.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

References

[1] Rosenthal, D. (2019). Cloud For Preservation. [online] DSHR's Blog. Available at: https://blog.dshr.org/2019/02/cloud-for-preservation.html [Accessed 20 Mar. 2019].

[2] Joseph, H. (2018). Securing community-controlled infrastructure: SPARC's plan of action. College & Research Libraries News, [online] 79(8), p.426. Available at: https://crln.acrl.org/index.php/crlnews/article/view/17246/18986 [Accessed 20 Mar. 2019].

Understanding and Implementing PREMIS

A tutorial

Karin Bredenberg
National Archives of Sweden
Sweden
Karin.bredenberg@riksarkivet.se
0000-0003-1627-2361

Eld Zierau
Royal Danish Library
Denmark
elzi@kb.dk
0000-0003-3406-3555

Angela Di Iorio
Sapienza University of Rome
Library System, Italy
angela.diiorio@uniroma1.it
0000-0002-2828-096X

Abstract – This half day tutorial will provide participants with an short introduction to the PREMIS Data Dictionary [1]. There after the focus is implementation. It will give a basic overview of the standard and explore different models of implementation.

Keywords – Preservation strategies and workflows; systems, and tools; Case studies, best practices and novel challenges; Training and education

Conference Topics – 2. Designing and Delivering Sustainable Digital Preservation: 5. The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

The PREMIS Data Dictionary for Preservation Metadata is a specification that provides a key piece of infrastructure for digital preservation activities, playing a vital role in enabling the effective management, discovery, and re-usability of digital information. Preservation metadata provides provenance information, documents preservation activity, identifies technical features, and aids in verifying the authenticity of digital objects. PREMIS is a core set of metadata elements (called "semantic units") recommended for use in all preservation repositories regardless of the type of materials archived, the type of institution, and the preservation strategies employed.

II. Summary of tutorial

The PREMIS Data Dictionary was originally developed by the Preservation Metadata: Implementation Strategies (PREMIS) Working Group in 2005 and revised in 2008 and 2015. It is maintained by the PREMIS Editorial Committee and the PREMIS Maintenance Activity is managed by the Library of Congress [2].

We have seen a constant call for PREMIS to undertake tutorials, such as this, as more and more organizations come to grips with digital preservation. This tutorial provides in its first part an introduction to PREMIS and its data model and an examination of the semantic units in the Data Dictionary organized by the entities in the PREMIS data model, objects, events, agents and rights.

As the second part, it presents how the preservation community can use PREMIS metadata support tools for the implementation of software, repository systems and data management practices.

As the third part, it presents examples and case studies of PREMIS implementation, using PREMIS in XML and PREMIS in RDF, in relation to the PREMIS Ontology.

It will include examples of implementation experiences through the institutional experience of the tutors as well as invited speakers.

The tutorial aims at developing and spreading awareness and knowledge about metadata to support the long-term preservation of digital objects.

III. Content outline

The draft outline for the tutorial is outlined below.

A. Introduction to PREMIS
- Background (brief history and rationale of PREMIS)
- Benefits of implementing PREMIS
- Outline of main Entities

B. Implementation community support tools
- Data Dictionary
- Ontology

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

- Website, PIG, id.loc.gov

C. Implementation case studies

- PREMIS in METS
- PREMIS and Semantic Web Technology
- PREMIS Conformance and repository interoperability

D. Wrap up

IV. Intended audience

The tutorial will benefit individuals and institutions interested in implementing PREMIS metadata for the long-term management and preservation of their digital information but who have limited experience in implementation. Potential audience includes cultural heritage operators, researchers and technology developers, professional educators, and others involved in management and preservation of digital resources.

V. Expected learning outcomes

A. Participants will understand:

- What PREMIS is and why it exists;
- The benefits of implementing PREMIS;
- The nature of the existing PREMIS community;
- The critical role PREMIS plays in the digital preservation community.

B. In addition, participants will get insight into:

- How PREMIS may be used in conjunction with METS;
- How different organisations implement PREMIS within their own repositories;
- How PREMIS, deals with Semantic Web Technology, and how it impacts on the data interoperability of repository systems.

VI. Short biographies of organizers

Karin Bredenberg is a Senior Technical Advisor on metadata at the Swedish National Archives. She currently serves as the chair of PREMIS EC, co-chair of TS EAS, chair of the DILCIS Board as well as a member of the METS Board. Currently Bredenberg is the activity lead for specifications in the project E-ARK4ALL and the eArchiving Building block.

Eld Zierau is member of the PREMIS Editorial Committee, since 2013. She is a digital preservation researcher and specialist, with a PhD from 2011 within digital preservation. Originally, she is a computer scientist, and has worked with almost all aspects of IT in private industries for 18 years, before starting in digital preservation in 2007. She has been working with many aspects of digital preservation, and she is involved as an architect or a consultant on major initiatives such a new digital repository including data modelling of metadata for preservation.

Angela Di Iorio is a PhD in engineering computer science, and she is responsible for the Digital Library - Repository System of the Sapienza University of Rome. She was involved in digital preservation projects as an expert since 2007, and is a member of the PREMIS Editorial Committee, since 2009.

References

[1] PREMIS Editorial Committee. 2015. PREMIS Data Dictionary for Preservation Metadata. Accessed 2019 located at http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf , Web archived: archive.org,, archive time: 2017-02-10 06:23:29 UTC archived URL: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

[2] PREMIS website. Accessed 2019. Located at http://www.loc.gov/standards/premis/index.html

[3] METS website. Accessed 2019. Located at http://www.loc.gov/standards/mets/index.html

Understanding and Implementing METS

Karin Bredenberg
National Archives of Sweden
Sweden
Karin.bredenberg@riksarkivet.se
0000-0003-1627-2361

Sean Mosely
National Library of New Zealand
Te Puna Mātauranga o Aotearoa
New Zealand
Sean.Mosely@dia.govt.nz
0000-0001-6462-8970

Juha Lehtonen
CSC – IT Center for Science
Finland
juha.lehtonen@csc.fi
0000-0002-9916-5731

Abstract – This half day tutorial will provide participants with an introduction to the Metadata Encoding and Transmission Standard (METS) and the METS Primer [1]. It will give a short basic overview of the standard and thoroughly explore different models of implementation using two different use cases. The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within digital libraries as well as digital archives, expressed using the XML schema language of the World Wide Web Consortium. It is maintained by the METS Board and the METS Maintenance Activity is managed by the Library of Congress [2]..

Keywords – Metadata and information strategies and workflows; Infrastructure, systems, and tools; Case studies, best practices and novel challenges; Training and education

Conference Topics – 2. Designing and Delivering Sustainable Digital Preservation; 5. The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

The METS Primer describes the metadata and information being stored or transferred in a METS document. METS provides a key piece of infrastructure for digital transfer as well as digital preservation activities, playing a vital role in enabling the effective management, discovery, and re-usability of digital information. METS metadata provides for descriptive information, administrative information, and structural information about digital objects. By working in conjunction with other standards, METS gives information regarding documents preservation activity, identifies technical features, and aids in verifying the authenticity of digital objects. METS contains a set of metadata elements recommended for use in all transfer as well as archiving situations regardless of the type of materials being transferred or archived, the type of institution, and the transfer strategies employed.

II. Summary of tutorial

The Making of America II project (MOA2) [3] originally attempted to address the issues of digital object metadata in part by providing an encoding format for descriptive, administrative, and structural metadata for textual and image-based works. METS, originally a Digital Library Federation initiative, built upon the work of MOA2 and provided an XML document format for encoding metadata necessary for both management of digital objects within a repository and exchange of such objects between repositories (or between repositories and their users). Depending on its use, a METS document could be used in the role of Submission Information Package (SIP), Archival Information Package (AIP), or Dissemination Information Package (DIP) within the Open Archival Information System (OAIS) Reference Model.

We have seen a growing call for the METS Board to undertake tutorials, such as this, as more and more organizations come to grips with digital transfer and

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

digital preservation. This tutorial introduces METS and its elements and gives an introduction to the elements in the Primer.

In addition, and with focus of the tutorial, it presents two examples of using METS metadata and a discussion of implementation considerations made in these two use cases. The tutorial will also show how to use METS in combination with the "Preservation Metadata: Implementation Strategies" (PREMIS) [4] standard. It will with these two uses cases show examples of implementation experiences through the institutional experience of the tutors.

The tutorial aims at developing and spreading awareness and knowledge about metadata to support the transfer and long-term preservation of digital objects.

III. Content outline

The draft outline for the tutorial is outlined below.

A. Introduction to METS
- Background (brief history and rationale of METS)
- Benefits of implementing METS

B. METS in detail
- Core elements
- Simple example to build familiarity

C. Implementation
- METS Profile
- Case studies
- The case of using PREMIS in METS
- Support and the METS community
- Conformance

D. Next Steps
- Round table discussion for institutional plans

E. Wrap up

IV. Intended audience

The tutorial will benefit individuals and institutions interested in implementing METS for transfer as well as for the long-term management and preservation of their digital information but who have limited experience in implementation. Potential audience includes cultural heritage operators, researchers and technology developers, professional educators, and others involved in management and preservation of digital resources.

V. Expected learning outcomes

A. Participants will understand:
- What METS is and why it exists;
- The benefits of implementing METS;
- The nature of the existing METS community;
- The critical role METS plays for transferring digital object in the digital preservation community.

B. In addition, participants will get insight into:
- How METS may be used in conjunction with PREMIS;
- How different organizations implement METS within their own repositories;
- The nature of conformance with METS.

VI. Short biographies of organizers

Karin Bredenberg is a Senior Technical Advisor on metadata at the Swedish National Archives. She currently serves as the chair of PREMIS EC, co-chair of TS EAS, chair of the DILCIS Board as well as a member of the METS Board. Currently Bredenberg is the activity lead for specifications in the project E-ARK4ALL and the eArchiving Building block.

Juha Lehtonen is a Senior Applications Architect at CSC – IT Center for Science located in Espoo, Finland. He acts as a main designer of methods, models and specifications for the national digital preservation services, and is a coordinator between the partner organizations using the services and digital preservation developers. In 2012-2014, he attended in APARSEN – EU/FP7 Network of Excellence project and was leading the activities related to preservation policies. He has been a member of the METS Editorial Board since 2017. Before his career at CSC, he has participated in high-tech industry related projects in spectral color research and has acted as a technical coordinator of digitization services for natural history collections of Finland. Lehtonen received his Ph.D. in Computer Science from University of Joensuu in 2009.

Sean Mosely is a Digital Preservation Technical Specialist at the National Library of New Zealand Te Puna Mātauranga o Aotearoa. Over the past ten years, Sean has held various positions relating to physical and digital preservation across national institutions in Australia and New Zealand. Sean

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

has developed various tools that utilise METS and complementary metadata standards, with a particular focus on automated and scalable digital ingest solutions. Sean has been a member of the METS Editorial Board since 2017.

References

[1] METS Board, version 1.6 2010. Accessed 2019. METS Primer located at http://www.loc.gov/standards/mets/METSPrimer.pdf

[2] METS website. Accessed 2019. Located at http://www.loc.gov/standards/mets/index.html

[3] MOAC2 Website. Accessed 2019. Located at https://old.diglib.org/standards/dlfmoaii.htm

[4] PREMIS website. Accessed 2019. Located at http://www.loc.gov/standards/premis/index.html

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Preserving complex digital objects

Workshop

Caylin Smith
BritishLibrary&UKLDLs
United Kingdom
caylin.smith@bl.uk
0000-0001-6340-5708

Patricia Falcao
Tate
United Kingdom
Patricia.Falcao@tate.org.uk
0000-0003-2798-5631

Sara Day Thomson
Digital Preservation Coalition United Kingdom
sara.thomson@dpconline.org
0000-0002-3896-3414

Abstract - This workshop addresses the opportunities and challenges generated by complex digital objects - objects created using innovative technologies - and current strategies for preserving them. The workshop will present a definition of complex digital objects with examples and an overview of the preservation challenges they pose, such as deciding what to collect and problem- solving for software and hardware dependencies. Using case studies, participants will identify and analyze a set of challenges to preserving actual works that UK cultural heritage institutions are looking to acquire. This will be done through small group exercises and will draw on approaches based on work carried out by Tate, the UK Legal Deposit Libraries (LDLs), and the Digital Preservation Coalition (DPC).

Keywords – file formats, time-based media, technology watch, collaboration, capacity building

Conference Topics – The Cutting Edge: Technical Infrastructure and Implementation; Exploring New Horizons.

I. Learninggoals

- Participants will collaborate and exchange knowledge and experiences with other group members to enhance community understanding of this problem and to further build a definition of 'complex digital objects'.
- Participants will gain practical know-how to get started in planning for the preservation of complex digital objects at their home institutions.

II. Description

At the heart of preserving digital objects, particularly of non-standard or experimental content types, lies an inherent paradox that cannot be resolved by any single sequence of preservation actions, reference model, tool, or service. This paradox arises from the new and unprecedented types of content that can be generated by creators using innovative new technology.

These new and unprecedented works can be considered 'complex digital objects', which are:

- Born-digital with no print counterpart. They are defined by their native format and their intended accessenvironment
- Constituted by formats that are complex and often networked and made of components that often consist of more than one mediatype
- Comprised of non-standard format and metadata types that might never becomestandardized
- Device-dependent and often require proprietary devices and platforms to enable the intended delivery of content and a meaningful user experience
- Not typically part of existing collections. Even mature collecting institutions might not have identified the necessary capability and infrastructure to managethem
- At risk of rapid obsolescence due to the transient and rapid change of the digitalmarketplace

While the resulting innovative objects providenovel ways for a creator to realize their vision, engage in the digital marketplace, and reach wider audiences, they also pose significant challenges for the institutions and individuals tasked with ensuring their preservation and access. No matter how up-to-date, responsive, and well-resourced an institution's response to digital preservation might

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

be, the knowledge neededto manage and preserve these objects will always lag behind the growth of the technology used in their creation.

This workshop addresses this paradox head-on. The organizers will apply the research they have undertaken in this area to small group activities. This approach will help to engage members of the digital preservation community to cultivate shared knowledge and to anticipate similar challenges that their institutions will encounter. While the 2018 iPres workshop 'Preservation of a Collaborative Community-Based Virtual Reality Collection' focused on strategies for engaging with the community to tackle the challenges of virtual reality objects, this workshop will more broadly address common challenges to preserving an array of different types of complex digital objects.

III. BACKGROUND

A. *UK LDLs' Emerging Formatsproject*

The UK LDLs' Emerging Formats project was set up to investigate the digital publishing landscape in the UK for more complex works that are in scope to acquire but not currently collected under the UK's Non-Print Legal Deposit Regulations. Within the scope of the project, the LDLs decided to focus on three formats: eBooks created as mobile apps, web-based interactive narratives, and structured data. The project used a sample of publications to determine how the libraries could acquire, preserve, and provide access to works created in these formats. Based on project findings, the LDLs created a methodology to inform how to address complex publications. This research is informed by existing digital preservation practices as well as new approaches that have come out of the project

B. *Tate's Time-based MediaConservation*

Tate's Time-based Media (TiBM) conservation team is responsible for the preservation of Collection artworks using performance, film, slides, video, audio, and software. In some cases, the object of preservation is not necessarily the software or data but the experience of the artwork. To address this issue, the TiBM team has developed risk assessment and analysis processes to evaluate the vulnerability of individual artworks. The team weighs up the diverse options for preservation (from storage to migration and emulation). They then take steps to document the artwork and its technical history, while also making the work more sustainable, pre-empt future issues, and/or intervene to maintain the work's functions in the present

These strategies at the UK LDLs and Tate provide a model, and in some cases tools, documentation, and procedures - that can be adapted (or even re-used outright) by other institutions.

IV. CONTENT

This workshop will discuss definitions for complex digital objects and provide an overview of the known challenges to preserving them. The first section of the workshop will focus on three predominantchallenges:

1) *Defining the digital object and its significant properties and using this information to decide what to preserve.*
2) *Problem-solving technical dependencies, including software and hardware environments.*
3) *Strategizing for digital rights management and intellectual property rights.*

The organizers will present two to three case studies that exemplify these challenges.

Participants will then break out into small groups for activities designed to analyze and problem-solve the challenges of preserving complex digital objects. They will be asked to indicate their preferred activity from a choice of four different options while arriving at the workshop so that groups can be pre-arranged during presentations. The activity options will include:

1) *an advocacy exercise to create a press release directed at building a preservation program;*
2) *a digital preservation workflow planning exercise;*
3) *a risk management exercise aimed at evaluating vulnerabilities; and*
4) *a donor agreement exercise that addresses significant properties, rights, and licensing issues.*

The small group activities will be directed by Sara Day Thomson who co-delivers digital preservation training at multiple skill levels and across

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

multiple professional sectors. Activity structure and worksheets will be adapted from the DPC's training resources.

In the final 30 minutes of the workshop, participants will feed back the results of their small group activities and discuss common trends as well as divergent approaches. Feedback will be collected and recorded in order to document the ideas and analysis generated by participants. This feedback will be shared with participants and published in the conference proceedings. The workshop will aim to identify opportunities for collaboration in the development of new approaches.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

A Cost Model for Analysing and Controlling the Cost of Long-Term Digital Accessibility as a Strategic Decision-Making Tool

Workshop by Eye Filmmuseum and BMC

H. Uffen
The Netherlands
herman.uffen@bmc.nl

W. Swagemakers
The Netherlands
walterswagemakers@eyefilm.nl

E. van Velzen
The Netherlands
ernstvanvelzen@eyefilm.nl

I. Background: a cost model for long-term digital accessibility

For some time now, heritage organisations and other institutions have been in agreement about the need for the long-term preservation and accessibility of valuable digital resources. What is much less clear, however, is how much this would cost. In many cases, not enough consideration is given to the long-term cost of curating digital collections, often because very little is actually known about this aspect. As a result, such management costs are frequently not included in the ordinary operating expenses of the institutions in question. Moreover, many heritage institutions tend to assume that the long-term costs are very high, partly due to the exponential rise in the volume of material, whether digitised or born-digital. In many cases, incidental revenue (i.e. project income) is used to pay for long-term management costs. But is this assumption correct? And how can we control the cost of curating digital collections in the long term?

II. Experiences with the cost model as a strategic decision-making tool for digital preservation.

The second part of the workshop focuses on the practical use of the cost model and its potential as a tool for supporting strategic decisions on the design and management of a digital infrastructure. We will be looking specifically at two use cases involving Eye Filmmuseum in Amsterdam. In the first of these cases, Eye used the cost model as a tool for supporting strategic decisions of its own. The second case involved another Amsterdam-based organisation in addition to Eye. This was LIMA, a platform for media art, new technologies and digital culture. In a joint project, Eye and LIMA sought to ascertain whether storing LIMA's digital collection in Eye's e-store would generate the benefits it was intended to generate.

Not only was Eye closely involved in the development of the cost model, it also took part in the trial and was one of the first institutions to adopt the model in practice.

During the period in which Eye made use of the model, it was confronted by rising costs of digitisation and of curating and providing access to its digital heritage collection. These rising costs – a trend spanning a number of years – came over and above the cost of curating the analogue collections. Having previously outsourced the management of its storage facilities and digital assets, it now had to take a strategic decision on whether to create a digital infrastructure itself, work in partnership with another heritage institution or subcontract the activities to an external supplier. In order for an institution to reach the right decision, it needs to have a clear picture of the costs throughout the entire ingest, storage and access chain.

Part 2A: Eye use case

Using Eye Filmmuseum as a case study, this part of the workshops looks at how the cost model works in practice, what sort of information it generates, and how this information can be used as the basis for strategic decisions on how to proceed in the future. We will also be demonstrating how the cost model

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

can help arts institutions to supply their funders (which tend to be government bodies) with information on the long-term funding they need in order to develop and manage a digital infrastructure, and on the reasons for the choices they make. Eye's experiences are particularly interesting given that the results of the cost model were audited and validated by an independent research agency. It was partly on the basis of these figures that the Dutch Ministry of Education, Culture and Science decided to allocate long-term funding to Eye's digital infrastructure.

Part 2B: Eye-LIMA use case

As part of a Digital Heritage Network project, LIMA and Eye partnered up in order to ascertain whether storing LIMA's digital collection in Eye's e-store would generate the benefits it was intended to generate. The assumption at the start of the trial was that LIMA would be able to make efficient use of the economies of scale offered by Eye's e-store and that the additional storage would not place an unacceptable burden on Eye's system nor require any disproportionate adjustments to be made to Eye's infrastructure. Thanks in part to the application of the cost model, Eye and LIMA were able to make a clear decision on which strategy to follow.

III. Workshop Format

The workshop on the cost model for long-term digital accessibility consists of three parts:

1. Introduction to the cost model:
   a. How the idea came about
   b. Structure and operation
   c. Results
   d. Link with other fields: preservation policy and shared services

2. Use cases: tutorial
   a. Eye's experience with the cost model as a tool for supporting strategic decisions
   b. Experience with the cost model as a tool for exploring the possibility of a partnership between Eye and LIMA

   Discussion based on statements and questions, combined with an exercise based on the participants' own practical needs. The participants play an active role in this part of the workshop: our aim is to encourage them to reflect on practical problems in their own specific fields of work.

IV. Practical information

- Number of participants: maximum of 20-25. This is an interactive workshop. Participants will be expected to share their experiences and to think about using the cost model in their own work.
- Duration of the workshop: maximum of 2½ hours including break.
- Room layout: cabaret-style.
  The workshop is given partly in the form of tutorials and partly in the form of discussion and exercises. For the latter part of the workshop, the participants need to sit at tables in groups, so that they can work together on the questions and exercises.
- We need an LED projector and a screen.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Machine Learning For Big Text

A Tutorial On Using Predictive Coding Tools To Process Large Archival Datasets

Brent West
University of Illinois
United States
bmwest@illinois.edu
https://orcid.org/0000-0001-6961-9141

Joanne Kaczmarek
University of Illinois
United States
jkaczmar@illinois.edu
https://orcid.org/0000-0001-8081-4570

Abstract – Big datasets can be a rich source of history, yet they pose many challenges to archivists. They can be difficult to acquire and process due to the varied formats and sheer volume of files. Sensitive content must be identified in advance of making materials publicly available. These challenges inhibit access for research purposes and often dissuade archivists from acquiring big datasets. Predictive coding can alleviate these challenges by using supervised machine learning to: augment appraisal decisions, identify and prioritize sensitive content for review and redaction, and generate descriptive metadata of themes and trends. Following the authors' previous work processing Capstone email, participants will learn about innovative and effective practices to enable digital preservation of large textual datasets at scale. Hands-on experience with specific tools is provided.

Keywords – access, active learning, appraisal, automatic classification, descriptive metadata, digital archives, digital humanities, digital preservation, e-discovery, email archiving, ingest, natural language processing, PII, preserving email, privacy, redaction, restricted records, scalability, sustainability, software-as-a-service, supervised learning, technical infrastructure, technology-assisted review, text mining, unstructured data

Conference Topics – The Cutting Edge: Technical Infrastructure and Implementation; Exploring New Horizons.

I. Introduction

The Records and Information Management Services (RIMS) office of the University of Illinois, in conjunction with the University Library and the Illinois State Archives (ISA), is nearing completion on a project to acquire, process, and provide access to a collection of email messages from senior government officials of the State of Illinois [1]. The project is generously funded by a three year grant through the National Historical Publications and Records Commission (NHPRC). A unique aspect of this project is the application of commercial tools to efficiently process this large dataset. In particular, the project leverages tools developed by the legal community for electronic discovery (e-discovery) to augment the preliminary archival review and increase processing output. This tutorial provides direct, hands-on access to the tools used by the project team [2] so that participants gain practical experience.

II. Description

This tutorial provides an introduction to predictive coding, a subset of machine learning, and its potential to help archivists make appraisal decisions about large textual datasets. Facilitators will describe in detail, and demonstrate, specific tools used for a project to appraise and make available a large dataset of government email. Participants will be given access to the tools and a dataset prepared in advance for their use during the tutorial as part of a hands-on exercise. Participants may optionally bring their own dataset to use with the tools. During the exercise, participants will explore visual display features, conduct faceted searches, and actively train the tool's predictive coding model to see how the training process works. They will also learn about the limitations of predictive coding tools in this setting, and how to calculate costs of manual review versus computer-assisted review. Participants will

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

engage in an in-depth discussion, driven by their own experiences, about challenges facing archivists looking to appraise, process, and make available to the public large datasets. This tutorial is anticipated to last 3-4 hours

A. Target Audience

This tutorial is designed for Archivists, Digital Curators, and Collections Managers who currently, or may in the future, work with large textual datasets. No prior knowledge is necessary, other than a general familiarity with archival appraisal concepts and general familiarity with personal computers, as would be expected for most conference attendee. Familiarity with The Future of Email Archives [3] is beneficial.

B. Learning Goals

Participants will:

1. Gain a basic understanding of machine learning generally, and predictive coding in particular.
2. Identify challenges associated with appraising and processing large textual datasets and learn how predictive coding may help remediate those challenges.
3. Practice working with machine learning tools to prepare large textual datasets for public access.

C. Agenda

1. Introductions and project overview
2. Discussion of participants' experiences and challenges with big data
3. Predictive coding methodology overview
4. Demonstration and hands-on lab
5. Wrap-up discussion of lessons learned and potential use cases

References

[1] University of Illinois Records and Information Management Services. 2018. Processing Capstone Email Using Predictive Coding. http://go.uillinois.edu/capstone.

[2] Kaczmarek, J. and West, B. 2018. Email Preservation at Scale: Preliminary Findings Supporting the Use of Predictive Coding. In *15th International Conference on Digital Preservation*, (Boston, USA). https://osf.io/yau3c/.

[3] Council on Library and Information Resources. 2018. The Future of Email Archives. https://www.clir.org/pubs/reports/pub175/.

Achieving Criticality of Preservation Knowledge

Sustaining Digital Preservation in the Nuclear Field

Elizabeth Kata
International Atomic Energy Agency
Austria
E.Kata@iaea.org
https://orcid.org/0000-0002-9916-5854

William Kilbride
Digital Preservation Coalition
United Kingdom
william.kilbride@dpconline.org
https://orcid.org/

Jenny Mitcham
Digital Preservation Coalition
United Kingdom
jenny.mitcham@dpconline.org
https://orcid.org/0000-0003-2884-542X

Jim Moye
J&A Preservation
United States
moye2@LLNL.gov
https://orcid.org/0000-0002-4905-1445

Abstract – The concerns of the digital preservation community are shared by those working in the nuclear field - preserving information for as long as it is required, taking into account the needs of diverse stakeholders, explaining needs to IT professionals, preserving data from legacy systems and obsolete media, managing accessibility and information security. Yet the length of time knowledge needs to be accessible and understandable, the accountability for public safety, and non-proliferation concerns present further challenges in the preservation of information related to nuclear science and its applications. Preservation efforts in this field must be approached and executed with these challenges in mind while also aligning with community good practices.

Practitioners from the nuclear field will present their experiences and. the panel discussion will address collaboration and impediments, appraisal and records management in light of the length of time material must be available, and preservation approaches. The panel will demonstrate the ways in which those in the nuclear field benefit from and apply good preservation practices and highlight how practitioners in the nuclear field can contribute to the broader digital preservation community thus opening a dialogue on how we can progress together.

Keywords – digital preservation, nuclear knowledge management, decommissioning, information security, digitization.

Conference Topics – Exploring New Horizons; Building Capacity, Capability and Community.

I. Breaking Down The Silos

Knowledge management has been a topic of discussion in the nuclear field since the outset of the 21st century, yet digital preservation of nuclear knowledge may be relegated to a side note or addressed only cursorily[1]. Preservation experts, archivists, and records and information managers in the nuclear field only come into limited contact with one another or the larger digital preservation community. The sometimes classified nature of our holdings can make it more difficult to exchange information and experiences. Without making claims to cover the full range of issues in preserving nuclear knowledge, in particular research data management aspects, this panel seeks to bring representatives working on digital preservation issues in the nuclear field together to come into dialogue with one another and with a wider digital preservation public.

Practitioners working on the preservation of nuclear knowledge from the International Atomic Energy Agency (IAEA), the Digital Preservation Coalition (DPC), and J&A Preservation will each address, in 8-10 minute presentations, challenges in the field. They will discuss how they relate to digital preservation challenges in general, how the approaches in the nuclear field could further inform the efforts of the digital preservation community, and how through sustained dialogue we might achieve criticality in our preservation knowledge.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

II. Long-Term Accountability To Many Stakeholders

Elizabeth Kata will introduce some of the digital preservation challenges in the nuclear field she has encountered in her work at the IAEA. With a dual mission to promote the peaceful uses of atomic energy while preventing the proliferation of nuclear weapons, the IAEA leads knowledge sharing projects on the one hand and tightly guards certain information on the other. Stakeholders are not only the Agency itself, but rather all Member States, the public, and commercial industries, which requires a heightened awareness for building and maintaining trust across the spectrum of stakeholders.

The need to balance access to authorized users with information security over time (such as a timeframe of 100 years used for nuclear power plants) presents a specific set of problems. Due to the complex IT environment, some aspects of commercial or community preservation solutions do not fully meet the specialized needs, requiring rethinking or expanding approaches to issues like migration or storage.

III. Reliable Robust and Resilient Digital Infrastructure for Nuclear Decommissioning

Jenny Mitcham of the DPC will present on a collaborative project with the Nuclear Decommissioning Agency (NDA) in the UK [2]. The NDA has been charged with the complicated task of decommissioning and cleaning the seventeen principal nuclear energy plants in the UK, a task described as the largest and most important environmental restoration programme in Europe. Amongst its many challenges, the NDA is by default required to become a trusted leader for information management and digital preservation, thus the project draws on the preservation understanding of the DPC and its wider membership.

Jenny will report on this ongoing project and share some of the challenges and themes that have emerged thus far. These include the high number of legacy data systems to manage, new data collection methodologies and technologies constantly evolving, compliance within a highly regulated environment and managing semi-current records for long periods.

IV. Digitizing Films Of Nuclear Tests

Jim Moye has over 40 years experience in the motion picture post-production industry and as an expert in film preservation. For the past five years he has been under contract by Lawrence Livermore National Laboratories, to digitize the scientific films from the above-ground nuclear tests.

There were 210 atmospheric nuclear tests and another nine nuclear cratering tests performed by the United States during the 1940s, 1950s, and early 1960s. Every test was extensively recorded with high quality motion picture films and photographs, providing a unique historic record of circa 10,000 films/photos. These newly digitized films/photos are now being re-analyzed using modern image processing techniques to obtain more accurate and precise measurements of the nuclear yield of each test and other scientific data pertaining to nuclear weapon effects.

Jim will discuss the challenges with preserving this historic scientific record, including the unique technical requirements of digitizing the films to allow accurate analysis, the software for automated analysis, the requirements for secure storage and the management of more than 10 million scanned images.

V. IPRES 2019: Opening A Dialogue

The discussion following the presentations, moderated by William Kilbride of the Digital Preservation Coalition, will highlight the ways we learn from each other and how information and preservation professionals in the nuclear field can contribute to and gain from the digital preservation community. We will address topics such as: Is digital preservation in the nuclear industry different? Why or how? What are impediments to collaboration with the wider digital preservation community or even each other? How can we overcome them? Does the length of time material needs to be accessible and the length of time material needs to remain classified have an effect on the preservation approaches? If yes, how is this addressed? What records

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

management and appraisal approaches are taken? How do we balance the importance of making the right decisions and the importance of not keeping too much? These questions and input from the audience will further open the dialogue and enable a discourse on how we can make progress together.

References

[1] Nuclear Knowledge Management Challenges and Approaches, Vienna: International Atomic Energy Agency, 2018, pp. 26-27.

[2] Reliable, Robust and Resilient: DPC supports Digital Infrastructure Project for Nuclear Decommissioning Authority, Digital Preservation Coalition. https://www.dpconline.org/news/dpc-supports-digital-infrastructure-project-for-nda

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Preservation Metadata Workshop

A workshop

Karin Bredenberg
National Archives of Sweden
Sweden
karin.bredenberg@riksarkivet.se
0000-0003-1627-2361

Neil Jefferies
Bodleian Digital Libraries University of Oxford UK
neil.jefferies@bodleian.ox.ac.uk
0000-0003-3311-3741

Sarah Romkey
Artefactual Systems
Canada
sromkey@artefactual.com
0000-0003-3833-7648

Andreas Nef
Docuteam GmbH
Switzerland
a.nef@docuteam.ch

Abstract – Metadata is a fundamental component of all digital preservation activities. This half day workshop offers a space to explore and discuss preservation metadata.

The workshop will be based on the community standards of PREMIS, METS and OCFL and facilitated by members of the boards of all three. This is the first time that the three boards have collaborated on such a session.

The purpose of the workshop is to connect the maintainers of the standards with the digital preservation community and offer implementers of preservation metadata an opportunity to discuss issues, successes and future directions.

Keywords – Metadata, Preservation metadata, Standards, Standardization, Collaboration

Conference Topics – 1. Collaboration: a Necessity, an Opportunity or a Luxury?: 3. Exploring New Horizons

I. Introduction

This workshop offers a space to explore and discuss preservation metadata.

The workshop will be based on the community standards of PREMIS, METS and OCFL and facilitated by members of the boards of all three standards. This is the first time that the three boards have collaborated on such a session.

The purpose of the workshop is to connect the maintainers of the standards with the digital preservation community and offer implementers of preservation metadata an opportunity to discuss issues, successes and future directions.

PREMIS is a standard addressing the information you need to know to preserve digital content in a repository, and METS is a standard for transfer and storage of metadata and digital objects, where OCFL is a standard for storage of metadata and digital objects.

II. Outline of workshop content

A. Overview of the standards

1) PREMIS: The PREMIS Data Dictionary for Preservation Metadata [1] is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. This session will reflect on changes that have happened across the last year and offer a view on upcoming activities and future direction. It will in particular focus on:

- the linked data version that has been developed over the last two years;
- changes in preservation controlled vocabularies;
- reporting back on work that is being carried out to understand the reach and impact of PREMIS.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

2) METS: The METS schema [2] is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library as well as digital archives, expressed using the XML schema language of the World Wide Web Consortium. This session will report back on activities from the last year and report on planned activities. In particular, feedback will be sought on the future direction of METS.

3) OCFL: The Oxford Common File Layout [3] initiative began as a discussion among digital repository practitioners about the ideal layout and characteristics for persisted objects. It grown into an open community effort defining an application independent way of storing versioned digital objects with a focus on long term digital preservation. The OCFL represents the community's recommendations addressing five primary requirements: completeness, parsability,, robustness, versioning and storage agnosticism.

B. Community participation

The primary purpose of this workshop is to encourage community participation: continued community engagement with standards ensures their continued relevance and appropriateness for purpose.

The workshop is one of the few times that implementers can come together to show and discuss their implementations. They are crucial for not only fostering a sense of community, but also for institutions and individuals to get direct feedback on critical questions and challenges in their digital preservation programmes.

Likewise, this workshop is a rare opportunity for the Boards of the standards to get direct feedback and comment from the community. The workshop will be driven by the community and is not didactic in nature. We will encourage active participation in two ways:

- Encouraging participation of the community in the workshop through contributions regarding implementations of, questions about, and suggestions for the future of PREMIS and/or METS and/or OCFL.
- Creating a safe, participatory workshop space where all participants can engage meaningfully. The workshop will be crafted to ensure it is not didactic, but rather involving and collaborative.

III. Intended audience

The workshop is designed for those who are interested in or utilise preservation metadata. This includes digital preservation practitioners (digital librarians and archivists, digital curators, repository managers and those with a responsibility for or an interest in preservation workflows and systems) and experts of digital preservation metadata, and transfer and storage of objects.

IV. Process for soliciting contributions

Contributions will be solicited from the PREMIS Implementers' Group via its discussion list. Likewise, the METS and OCFL community will be asked for contributions. A general call will also go out through wider fora. All proposals will be reviewed and selected on relevance. Direct invitations may also be issued to organizations and people that the organizers feel will bring important and new directions to the discussion.

V. Outcomes

We believe this workshop to be key for a number of reasons, and the proposed outcomes reflect that.

1. Support for implementers from those that develop preservation metadata standards.
2. A closer connection between the community and the boards that maintain the standards.
3. A physical meeting for those who implement or have an interest in preservation metadata standards. While the communities for these standards operate mainly through virtual environments, experience tells us that there are meaningful benefits from face-to-face interactions.
4. Closer connections between the standards. This workshop should trigger a deeper level of connection between the different communities.
5. A strong indication of future directions for preservation metadata and PREMIS, METS and OCFL in particular.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

VI. Short biographies of organizers

Karin Bredenberg is a Senior Technical Advisor on metadata at the Swedish National Archives. She currently serves as the chair of PREMIS EC, co-chair of TS EAS, chair of the DILCIS Board as well as a member of the METS Board. Currently Bredenberg is the activity lead for specifications in the project E-ARK4ALL and the eArchiving Building block.

Neil Jefferies is Head of Innovation for Bodleian Digital Library Systems and Services at the University of Oxford. He is PI and Community Lead for SWORDV3, a protocol for machine-to-machine transfer of digital objects, a co-author of the Oxford Common File Layout for preservation-oriented object storage and Technical Strategist for "Cultures of Knowledge", an international collaborative project to "reconstruct the correspondence and social networks of the early modern period". He also serves on the Steering Committee for the Preservation and Archiving SIG and is a Judge for the DPC's Digital Preservation Awards. Previously, he was also a co-creator of the International Image Interoperability Framework.

Andreas Nef holds a MA in History and Computer Science from the University of Zurich. Since 2004, he has been working as archivist and project manager at Docuteam GmbH and is currently Technical Lead for the development of digital archiving solutions. The company provides information management services, and implements/manages physical and digital archives for which the Matterhorn METS Profile is one of the main package standards. Since 2014, he is member of the METS Editorial Board. A regular member of the International Council on Archives, Andreas Nef joined the Swiss Expert Pool for Civilian Peacekeeping in 2009 and is also serving as archives expert for Swisspeace. In these functions he has been allowed to support a number of activities preserving records of archives at risk.

Sarah Romkey is the Program Manager for Archivematica, the open-source digital preservation software developed by Artefactual Systems. She has a Masters in Archival Studies and a Masters in Library and Information Studies from the iSchool at University of British Columbia. She was previously an archivist for the Rare Books and Special Collections division of University of British Columbia Library. Sarah has been a member of the PREMIS Editorial Committee since January 2019.

References

[1] PREMIS Editorial Committee 2015. PREMIS Data Dictionary for Preservation Metadata, accessed 2019, located at http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

[2] Metadata Encoding and Transmission Standard, accessed 2019, http://www.loc.gov/standards/mets/

[3] Oxford Common File Layout, accessed 2019, https://ocfl.io/

Preservation Action Rules Workshop

Parcore: See one, Do one, Teach one

Justin Simpson
Artefactual Systems Inc.
Canada
jsimpson@artefactual.com
https://orcid.org/0000-0002-1281-1455

Matthew Addis
Arkivum
United Kingdom
matthew.addis@arkivum.com
https://orcid.org/0000-0002-3837-2526

Jack O'Sullivan
Preservica
United Kingdom
jack.osullivan@preservica.com
https://orcid.org/0000-0002-0306-761X

Carl Wilson
Open Preservation Foundation
United Kingdom
carl@openpreservation.org
https://orcid.org/0000-0003-1535-1770

Sarah Romkey
Artefactual Systems Inc.
Canada
sromkey@artefactual.com
https://orcid.org/0000-0003-3833-7648

Jon Tilbury
Preservica
United Kingdom
jon.tilbubry@preservica.com

Abstract – The Preservation Action Registries project (http://parcore.org) is developing a common and consistent way to describe and execute preservation policies and actions at a technical level. An initial data model has been created which defines a human and machine-readable way of describing preservation actions and the associated business rules that together make up preservation policies.

Using a skill acquisition technique borrowed from the medical profession (See One, Do One, Teach One) this workshop will lead participants through a three-step process, intended to improve the participants ability to develop and describe their own digital preservation policy in both human and machine-readable forms. The workshop will help participants to better express and share preservation policies in a concise, comprehensive and unambiguous way. Through the participants activities in the workshop, the PAR data model will be tested, validated and further improvements identified.

Participants will be introduced to the PAR project and data model, using examples of working preservation actions and business rules. In smaller group, participants will be led through the process of describing a new preservation policy using the PAR data model. Finally, groups will present their work to the other groups

Keywords – preservation rule, preservation policy, controlled vocabulary, interoperability, best current practice

Conference Topics – Designing and Delivering Sustainable Digital Preservation; Building Capacity, Capability and Community

I. Introduction

The Preservation Action Registries (PAR) project was introduced in a paper presented at iPres 2018 [1]. This paper describes a series of problems that arise in the development of digital preservation platforms such as Archivematica and Preservica. These systems lack a common and consistent way to describe and execute preservation policies and actions. Each system must implement some mechanism to define what tools and rules to use when 'doing' digital preservation. There is currently

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

no practical way to share technical information between organisations using different preservation platforms. This limitation presents a barrier to the development of best current practice in the field.

The PAR project proposes a solution in the form of a common data model describing the elements of preservation policies, implemented using machine readable data structures. This data model includes the ability to describe preservation actions that can be performed on digital content, what tools are used to perform these actions, the inputs needed, what results and outcomes are achieved, and in what context the actions should be taken.

Initially funded by Jisc, the project has been taken up by the Open Preservation Foundation. Through a collaboration between three software vendors (Arkivum, Artefactual and Preservica), the project has defined a data model using json-schema [2]. A standardized Application Programming Interface (API) has also been defined [3], using the Swagger API specification language. This allows different digital preservation platforms to implement a common method for sharing this kind of data.

II. Workshop Objectives

The PAR project has initially focussed on technical interchange of information between preservation systems. The concepts and data model provide a mechanism for organisations to describe digital preservation policies at a detailed and actionable level. In particular, PAR addresses the issue of how to describe the specifics of digital preservation actions so that a description can be taken by someone else in the community and executed either manually or automatically by their preservation system of choice. It is this 'specification' aspect of PAR that we will explored in the workshop. The workshop will ask participants to use the PAR concepts and framework to describe their own preservation policies. Through the 'teach one' part of the workshop methodology, the descriptions of preservation policies will be 'tested' to see if they are sufficiently detailed, have an unambiguous interpretation, and have enough contextual information about when to apply the policies and why. Through this approach, the participants will gain valuable skills described below while the PAR consortium will also be able to answer key questions including:

- Can the PAR model effectively describe real world preservation scenarios?
- What changes are needed to the PAR model to make it more effective, and in what preservation scenarios are they needed?
- How easy is it to use the PAR model in practice and how could it be simplified or improved?

III. Learning outcomes and benefits

The benefits of the workshop can be split into three areas: benefits to individual participants; benefits to the PAR consortium; and benefits to the wider digital preservation community.

Workshop participants will benefit through:

- Understanding of how to specify preservation policies in a concise, comprehensive and unambiguous way.
- Ability to better communicate and share preservation policies with peers and stakeholders.
- Framework for critical evaluation of their existing preservation policies, e.g. helping to identify gaps, inconsistencies and ambiguities.
- Ability to communicate preservation needs more effectively with vendors or other providers of preservation systems.
- Understanding of how to express the context for applying preservation policies, e.g. as a set of rules and priorities.
- Ability to analyze the technical aspects of preservation in their organization by comparing tools/systems/infrastructure/techniques with good practice of others.

The PAR initiative will benefit through:

- Real-world examples and use cases that will allow us to test the capabilities of the PAR model and identify gaps or problems.
- A set of requirements and priorities for further work.
- Growth of the PAR community and user base.

The digital preservation community will benefit through:

- A published set of preservation policies that cover a range of preservation actions and tools.
- A published set of preservation scenarios of how

to describe preservation contexts, e.g. business rules for selecting, comparing and applying policies based on priorities and constraints.
- PAR will become better aligned to and informed by a wide range of real-world preservation use cases.

IV. PRE-REQUISITES

The only pre-requisites for attending the workshop are:

- Participants should bring one or more preservation policies that they already use in their organization (or would like to implement). They should be prepared for these to be described, analyzed and shared in the workshop.
- Participants are expected to actively participate in the workshop, especially in the 'see one, do one, teach one' process. For example, all participants will be expected to 'teach one' in small groups or one-to-one with other participants.

While PAR supports machine-readable descriptions of preservation actions, e.g. json, we do not expect attendees to take this approach. Instead, attendees will be creating human readable descriptions using the concepts from PAR (actions, tools, rules, formats, objects etc.) and a controlled vocabulary of terms (e.g. preservation action types, file formats, and file properties).

V. Workshop Motivation

Collaborative development of best current practice in any field is an ambitious and difficult undertaking. A common method for describing and implementing digital preservation techniques should reduce the cost of developing best current practice, and reduce the learning curve for individuals and organizations attempting to improve their knowledge and expertise in the field.

The PAR project offers a starting point for developing a common method or approach. The project needs input from a wider audience of digital preservation practitioners to validate the approach, test and improve the existing data model and api, and to develop a corpus of preservation actions and business rules that represent actual best current practice in the field.

Practitioners will need to learn about the PAR data model and develop expertise in working with it. We propose to borrow a skill acquisition technique that is well known in the medical profession, commonly referred to as 'See One, Do One, Teach One' [4]. We believe this simple methodology allows the greatest chance for the concepts and techniques of the PAR project to be shared with a new audience.

VI. Workshop Structure

A. Introductions and overview of PAR

The PAR consortium will provide an overview of the PAR project, its objectives and results so far. This will include a review of the PAR data model and how this can be used to define preservation policies based on preservation actions, tools and rules. The PAR data model will provide the framework for the See one, do one, teach one part of the workshop and allow project participants to describe their preservation policies in a consistent way.

B. See one

The PAR consortium will provide the attendees with ready-worked examples of how the PAR approach can be used to describe preservation policies. These will be based on user scenarios (e.g. 'An organization is responsible for preserving audiovisual content and wishes to use EBUCore to standardize the way it describes its assets') and the specific preservation actions needed to achieve this (e.g. Use the MediaInfo tool to characterize the technical aspects of video files and store the results as EBUCore in XML format). Examples will cover several of the main PREMIS preservation event types (e.g. fixity, characterization, validation, format normalization). This will enable the attendees to 'see one' before they are asked to develop their own descriptions.

C. Do one

Attendees will be split into small groups and asked to select one or two preservation policies to work on from the candidates brought by the attendees. We will group attendees with similar policies. Each group will discuss the policies and develop a description using the PAR framework. A member of the PAR consortium will facilitate each group. Each group member will be involved in the 'Do one' activity, e.g. by focusing on different aspects of the description such as the business rules, tool specification, action

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

definition etc. GoogleDocs or similar will be used for collaborative authoring and to help the PAR consortium to capture the results of the group work.

D. Teach one

Each group member will present their group's description of a preservation policy to a member of another group (teach one). This will be done in pairs or in small groups. Everyone will get the chance to 'teach' as well as listen and comment on other people's policy descriptions. Comments will be captured in shared GoogleDocs alongside the preservation policy descriptions. We plan to identify where descriptions are incomplete, ambiguous or can be better specified. This will test whether PAR can capture all information necessary to enact a preservation action in a real-world environment. This part of the PAR workshop will focus on the 'teaching' aspect of 'see one, do one, teach one' and the use of question/response as part of teaching as a way to validate that the information being exchanged between the 'teacher' and the 'student' is clear, concise and complete

E. Feedback and next steps.

The final stage of the workshop will be used to capture general comments and feedback on the participants experience with the PAR approach. This will be done as a plenary involving all participants so everyone has chance to contribute and share their suggestions for PAR improvements.

References

[1] Digital Preservation Interoperability through Preservation Action Registries https://figshare.com/articles/Digital_Preservation_Interoperability_through_Preservation_Actions_Registries/6628418

[2] PAR Json Schemas, Github https://github.com/JiscRDSS/rdss-par/tree/master/schemas

[3] PAR REST API Specifiation, Github https://github.com/JiscRDSS/rdss-par/tree/master/api

[4] See One, Do One, Teach One – Not Just for the Medical Profession https://spin.atomicobject.com/2011/04/29/see-one-do-one-teach-one-not-just-for-the-medical-profession/

Introducing the Memento Tracer Framework for Scalable High-Quality Web Archiving

Martin Klein
Los Alamos National Laboratory USA
mklein@lanl.gov
https://orcid.org/0000-0003-0130-2097

Herbert Van de Sompel
DANS
The Netherlands
hvdsomp@gmail.com
https://orcid.org/0000-0002-0715-6126

Web archiving at scale and with high-quality is not a trivial endeavor. Given the dynamic nature of the web and the increasingly complex features incorporated into web pages, sophisticated capturing and archiving approaches are needed. This workshop introduces the novel Memento Tracer Framework and invites attendees to a hands-on experience with our framework that was designed to track, capture, and archive scholarly artifacts. Memento Tracer takes an institutional perspective and focuses on artifacts created by or relevant to individual communities. Workshop participants will learn about scalable and high-quality web archiving and leave with the satisfaction of having actively contributed to saving parts of the (scholarly) web.

Memento Tracer, Web Archiving at Scale, High-Quality Web Archiving

Designing and Delivering Sustainable Digital Preservation; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction and Background

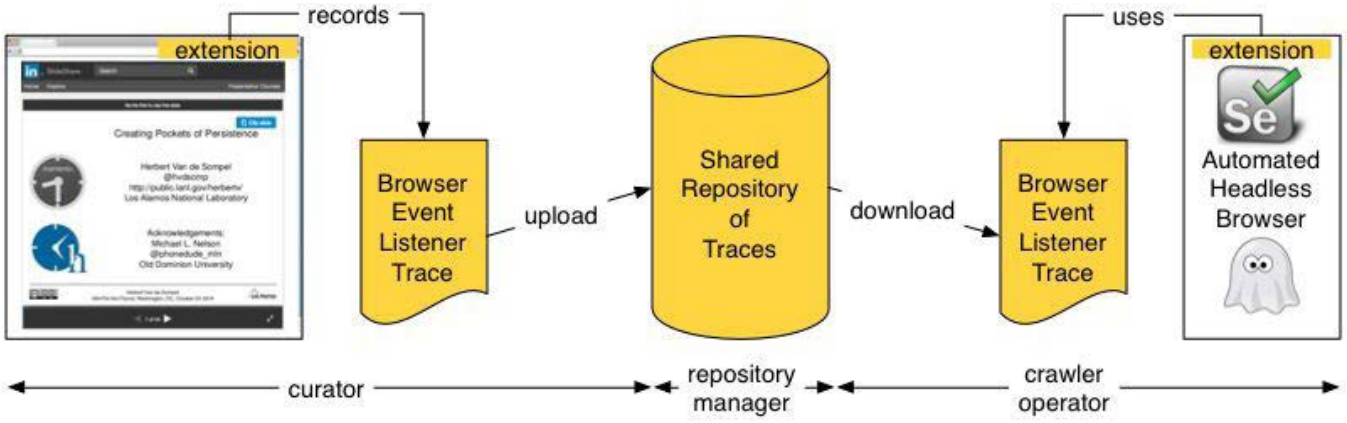
Current web archiving approaches often either excel at capturing at scale or with high quality. Despite various attempts [1], approaches that combine scale and quality remain elusive. For example, the Internet Archive's crawler is optimized for scale and hence enables an archive of more than 703 billion web resources [2]. However, the quality of the captures varies is often hindered by dynamic elements and interactive features contained in the captured resources. For instance, the CNN.com homepage has not been properly archived (and hence can not be replayed correctly) in the Internet Archive since November 2016 [3].

An example on the other end of the spectrum is Webrecorder [4]. While browsing a web page, this tool archives the page and captures all the elements the user interacts with. With this approach, Webrecorder provides high-fidelity captures but lacks the ability to archive resources at web scale as only individual user interactions with the single web resource trigger the archiving process, similar to a screen recording session

The Memento Tracer framework aims to find a balance between operating at web scale and providing high-quality archival records. The Tracer approach, visualized in the figure below, is based on a human curator interacting with a web resource to establish its essential components (boundary), and to record these interactions as a trace. This Trace will then be used to guide a browser-based capture process that automatically replays the recorded transactions and thereby delivers high-quality captures of web resources. A trace can be applied to all resources of the same class and hence enabling the framework to operate at scale. To enable transparency, foster collaboration, and avoid duplicate efforts, traces and versions of traces can be shared with a community of practice in a shared repository.

In the "Scholarly Orphans" project, we focus on archiving scholarly artifacts - web resources scholars across disciplines and throughout the research life cycle create in productivity portals such as GitHub, FigShare, Publons, and SlideShare to conduct aspects of their research and to communicate research outcomes. Hence, we designed Memento Tracer an institutional pipeline to track, capture, and archive these artifacts. The workshop will introduce the Memento Tracer framework to the participants and allow for hands-on exploration of the entire pipeline.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.



II. Workshop Topics

The workshop will begin with a high-level overview of the Memento Tracer framework and outline its capabilities by means of intuitive and easy-to-follow examples. This introduction of the novel framework is necessary in order to provide attendees with the background and knowledge to fully take advantage of the second part of the workshop. Here we will collect web archiving use cases from participants and identify scenarios where and how Memento Tracer can be applied. The workshop participants are then encouraged to try out the framework themselves while the organizers will provide guidance and (technical) support. Participants will create traces for productivity portals such as Github or SlideShare, upload their traces to a shared repository, use their own or a publicly shared trace to capture and archive artifacts, replay the archived record, and provide feedback about its quality.

At the conclusion of the workshop, participants will have gained a deep level of understanding of various practical aspects of web archiving, identified archiving use cases relevant to their home organization, utilized the Memento Tracer framework for the creation of archival records related to their use case, and done a high-level quality analysis of the created archival record.

III. Workshop Logistics

This is a hands-on workshop where the attendees are encouraged to participate, ask questions, and provide feedback. We will keep the lecture portion to a minimum and allow significant time for exploration of the Memento Tracer framework and participants' use cases.

A. *Audience and Attendees*

This workshop aims to bring together librarians, archivists, and other professionals that are passionate about addressing challenges in modern web archiving.

Since this will be the first time we organize this workshop, we have no past data to rely on but we expect 20-30 people attending this workshop.

B. *Format and Duration*

We propose a half-day workshop, ideally in the morning, that does not exceed four hours. We plan on a short break around the halfway mark and will leave plenty of time for questions and feedback from the attendees.

C. *Special Requirements*

The workshop requires a space where attendees can sit down and use their laptops. Internet connectivity as well as power supply is essential. In addition, the organizers will need a projector in the room for slides and live demonstrations.

References

[1] Squidwarc. https://github.com/N0taN3rd/Squidwarc

[2] https://twitter.com/brewster_kahle/status/1087515601717800960

[3] CNN.com has been unarchivable since November 1st 2016. https://ws-dl.blogspot.com/2017/01/2017-01-20-cnncom-has-been-unarchivable.html

[4] https://webrecorder.io/

Managing And Preserving Digital Content With Fedora

David Wilcox
DuraSpace, Canada
dwilcox@duraspace.org
0000-0001-5411-9208

Andrew Woods
DuraSpace, USA
awoods@duraspace.org
0000-0002-8318-4225

Daniel Bernstein
DuraSpace, USA
dbernstein@duraspace.org

Abstract – Fedora is a flexible, extensible, open source repository platform for managing, preserving, and providing access to digital content. For the past several years the Fedora community has prioritized alignment with linked data best practices and modern web standards. We are now shifting our attention back to Fedora's digital preservation roots with a focus on durability and the Oxford Common File Layout (OCFL). This tutorial will provide an introduction to the latest version of Fedora with a focus on digital preservation functionality and workflows.

Keywords – fedora, repository, ocfl, preservation, standards

Conference Topics – The Cutting Edge: Technical Infrastructure and Implementation.

I. Introduction

Fedora is a flexible, extensible, open source repository platform for managing, preserving, and providing access to digital content. Fedora is used in a wide variety of institutions including libraries, museums, archives, and government organizations. For the past several years the Fedora community has prioritized alignment with linked data best practices and modern web standards. We are now shifting our attention back to Fedora's digital preservation roots with a focus on durability and the Oxford Common File Layout (OCFL). This tutorial will provide an introduction to the latest version of Fedora with a focus on digital preservation functionality and workflows. Both new and existing Fedora users will be interested in learning about and experiencing Fedora features first-hand.

II. Curriculum

Attendees will be given pre-configured virtual machines that include Fedora bundled with the Solr search application and a triplestore that they can install on their laptops and continue using after the workshop. These virtual machines will be used to participate in hands-on exercises that will give attendees a chance to experience Fedora by following step-by-step instructions. The tutorial will include three modules, each of which can be delivered in 1 hour or less:

A. Introduction And Feature Tour

This module will feature an introduction to Fedora generally, with a focus on the latest version, followed by an overview of the core and extended Fedora features. It will also include a primer on data modeling in Fedora.

B. Digital Preservation Workflows

Fedora has a number of features that support digital preservation, including fixity checking, versioning, and backup/restore. Fedora also provides a robust REST-API that can be used to integrate with other applications and services in a broader digital preservation workflow. This module will demonstrate both Fedora's internal digital preservation features and API-driven integrations.

C. Oxford Common File Layout

The OCFL is an application-independent approach to the storage of digital objects in a structured, transparent, and predictable manner. It is designed to promote long-term access and management of digital objects within digital repositories. This module will provide an overview of the OCFL and present the design and prototyping work that will lead to OCFL support in the next major version of Fedora.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

III. Learning Outcomes

Tutorial attendees will:

1. Become familiar with core and extended Fedora features and functionality.
2. Learn how to exercise Fedora's digital preservation features.
3. Understand the OCFL and its potential impact on digital preservation using Fedora.

IV. Target Audience

This tutorial is intended to be an introduction to Fedora - no prior experience with the platform is required. Repository managers and librarians will get the most out of this tutorial, though developers new to Fedora would likely also be interested.

V. Conclusion

This tutorial will provide an introduction to the core and extended features of Fedora, along with an overview of Fedora's digital preservation features. Participants will have an opportunity to experience this functionality through hands-on exercises. Finally, participants will learn about the Oxford Common File Layout and how it will be implemented in Fedora to enhance support for digital preservation.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Level up on preservation:

*Updating and Mapping the next generation of the **Levels of Preservation***

Bradley J. Daigle
Academic Preservation Trust United States
bradley@aptrust.org

Aaron Collie
Federal Reserve Bank of St. Louis United States
Aaron.Collie@stls.frb.org

Corey Davis
Council of Prairie and Pacific University Libraries (COPPUL) Canada
corey@coppul.ca

Helen Tibbo
University of North Carolina Chapel Hill United States
tibbo@ils.unc.edu

Lauren Work
University of Virginia United States
lw2cd@virginia.edu

Abstract – This workshop will update the community on the work of the Levels of Preservation (LoP) Reboot Working Group and engage with the participants to gather feedback on our work so far and what form the final product(s) should take. Updating the LoP will require a more dynamic user interface and user experience. Working with the participants, we will create a list of requirements and options for the final products. In addition, the workshop will discuss a methodology for keeping the LoP updated on an ongoing basis.

Keywords – sustainable digital preservation, community practice, preservation guidelines, digital stewardship, technical implementations

Conference Topics – Designing and delivering sustainable digital preservation, building capacity, capability, and community

I. Intent and Background

The National Digital Stewardship Alliance (NDSA) created its landmark Levels of Preservation, Version 1.0 (LoP) guidelines in 2013. The original intent of the LoP was to create a tiered set of recommendations for either preservation practitioners who were just starting out or for those looking to deepen their preservation practice. Not meant as a comprehensive preservation strategy, but rather, a lightweight tool to encourage organizations to think through preservation issues, the LoP are organized into five functional areas that are at the heart of digital preservation systems: storage and geographic location, file fixity and data integrity, information security, metadata, and file formats. By design, they do not cover policy, staffing, risk, or budgetary considerations, rather they are considered to be a technical implementation of a variety of preservation decisions. Since 2013, various groups and individuals have undertaken the process of amending and extending the LoP; however, these activities have been distributed and siloed from a central body or structure (such as the NDSA) and have not been incorporated into new versions.

To this end, the NDSA launched a **Levels of Preservation Reboot Working Group** whose primary task is to provide the methodology by which this important document can be adapted more readily—taking in the broadest possible feedback in the process. Workshops are one of the critical methods of communication to both share and gather feedback.

This hands on workshop will discuss the revised LoP - new additions, clarifications, and "views" into the guidelines. We have heard that practitioners want to preserve the simplicity of the original LoP while, at the same time, adding various components that will extend them. This poses significant challenges to the expression of version 2.0. This new version will need to take advantage of newer user experience technology to be of broadest possible use. What that might look like and how we could achieve it will be a primary topic for this session. A

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

secondary component of the workshop will focus on the methodology for updating the LoP going forward so that the guidelines remain relevant and useful as the preservation landscape shifts.

II. Outcomes

Participants will be engaged with the new version of the LoP - individual updates from the subgroups and, in particular, how those disparate elements are brought together. In particular, participants will

- Understand, review, and provide feedback on the revised LoP document and how it is actively used as well as a discussion of best practices for adopting and adapting the LoP.
- Discuss how future changes and improvements could be undertaken.
- Discuss the user interaction with the new LoP and what form that could take.

By the time of iPres, the LoP Reboot Group will have combined the data gathered and refined by the subgroups. The revised document will include some possible approaches to how future revisions of the LoP might be undertaken and how successful our approach to linking the various elements together has been.

III. Audience

Any level of practitioner of digital preservation will find this workshop valuable, from novice to expert, since the LoP encompass the broad spectrum of preservation practice. Of particular value will be different approaches to crafting "views" on the document. Everything from keeping the original grid format to exploring expanded modules for the LoP. This will be the greatest challenge for the LoP Reboot--making a final product that can be of use by any level of preservation practice. Users across disciplines will also help us gauge the effectiveness of the new version and discuss possible ways the LoP can be used by administrators, curators, budget managers, archivists, and preservation experts.

IV. Process and Structure

The workshop will be designed to provide a brief overview of the LoP Revision work and then a facilitated group discussion of the currently revised state. Participants will discuss if there are any additional elements that need to be added or if current elements have the appropriate amount of clarity. After that, small groups will focus on brainstorming the necessary user experience for the new LoP. Part of this discussion may take into account the fact that additional grant funding may be required to execute this vision. Finally, the group will discuss how the LoP should be iterated over time and what that methodology might look like. Participants should leave with a clear sense of what the new LoP will look like, how we might experiment with various "views" into the content, and the means by which community feedback and updating can take place in the future.

- Review and Update [large group]
  - Overview [subgroup leads]
  - Discussion [all]
- User Experience/Interface
  [small groups - facilitated]
  - UI and final product [small]
  - UI Recap and discussion [all]
- Updating the LoP going forward
  - Discussion of future iteration [all]

Shared note taking and whiteboarding will be critical components to gathering feedback. This method was successfully employed at iPres in 2018 by the LoP team since it provides a transparent communication process to the broader community - both at the conference and remotely.

Getting ongoing feedback from the broadest range of users is critical to updating the LoP. The original survey of interest revealed a large number of non-US practitioners, so iPres is an ideal venue for this type of discussion.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

Index

Johan van der Knijff 25
Timothy Robert Hart 37
Denise de Vries 37
Carl Mooney 37
Marion Humbert 46
Stéphanie Roussel 46
Édouard Vasseur 46
Eld Zierau 53
Mark Cooper 64
Aly DesRochers 64
Jesse Johnston 64
Camille Salas 64
Neil Jefferies 72
Trevor Owens 64
Karin Bredenberg 72
Angela Dappert 72
Jack O'Sullivan 81
Richard Smith 81
Alan Gairey 81
Kevin O'Farrelly 81
Mustapha Mokrane 92
Jonas Recker 92
Rawia Awadallah 101
Iyad ALAgha 101
Tomasz Miksa, Andreas Rauber 101
Joy Davidson 101
Adnan Yahya 101
Anna Maria Tammaro, Stefano Caselli 101
Janet Anderson, David Anderson 101
Nael Salman 101
Yousef AbuZir 101
Zack Lischer-Katz 114
Ashley Blewer 126
Matt Cook 114
Nathan Hall 114
Juliet Hardesty 114
Jamie Wittenberg 114
Jennifer Johnson 114
Robert McDonald 114
Tara Carlisle 114
Sarah Romkey 126
Marjolein Steeman 134
Peter May 142
Ross Spencer 126
Maureen Pennock 142
David A. Russo 142
Jaye Weatherburn 155
Lyle Winton 155
Peter Doorn 167
Sean Turner 155
Marjan Grootveld 167
Claudia Roeck 179
Francesca Morselli 167
Euan Cochrane 191
Rafael Gieschke 179
Klaus Rechert 179
Julia Noordegraaf 179
Klaus Rechert 191
Seth Anderson 191
Jessica Meyerson 191
Ethan Gates 191
Martin Klein 201
Lyudmila Balakireva 201
Harihar Shankar 201
Maureen Pennock 212
Peter May 212
Michael Day 212
Michelle Lindlar 221
Meghan Goodchild 234
Pia Rudnik 221
Andrew Weaver 245
Grant Hurley 234
Ashley Blewer 245
David Giaretta 254
David Wilcox 250
John Garrett 254
Mark Conrad 254
Eld Zierau 254
Terry Longstreth 254
J Steven Hughes 254
Matthias Hemmje 254
Esther Plomp 260
Felix Engel 254
Yan Wang 260
J.S. Love 260
Nick Krabbenhoeft 265
Tobias Wildi 271
Alain Dubois 271
Eld Zierau 276
Sibyl Schaefer 276
Nancy Y McGovern 276
Andrea Goethals 276
Johan Kylander 290
Simon Lambert 286
Tobias Steinke 282
Heikki Helin 290
Kimmo Koivunen 290
Juha Lehtonen 290
Lozana Rossenova 297
Dr Karin de Wild 297
Dragan Espenschied 297
Katherine Boss 305
Vicky Steeves 305
Rémi Rampin 305
Fernando Chirigati 305
Brian Hoffman 305
Paul Stokes 311
John Kaye 311
Dom Fripp 311
Tamsin Burland 311
Tom Davey 311
Matt McGowen 311
Sheila Morrissey 317
Amy Kirchhoff 317
Katherine Thornton 323
Ana van Meegen Silva 335
Angela Beking 329
Kenneth Seals-Nutt 323
Melinda Haunton 340
Jo Pugh 340
Jaana Pinnick 360
James Travers 340
Jonathan Tilbury 366
Sharon McMeekin 348
Vincent Joguin 354
Remke Verdegem 366
Euan Cochrane 366
Helen Hockx-Yu 372
Donald Brower 372
Hideyuki Uesugi 377
Masayuki Uesugi 377
Ashley Adair 382
Teruhiro Tani 377
Maria Esteva 382
Anna Oates 388
Benn Chang 382
Remco van Veenendaal 396
William Schlaack 388
Frederik Holmelund Kjærskov 396
Kati Sein 396
Jack O'Sullivan 396
Anders Bo Nielsen 396
Merle Friedrich 399
Phillip Mike Tømmerholt 396
Jacob Takema 396
Sarah Middleton 401
Alec Mulinder 404
Sharon McMeekin 401
Sonia Ranade 404
David H Underdown 404
Kate Wittenberg 406
David Millman 406
Jaye Weatherburn 409
Jeffrey van der Hoeven 412
Karen Hanson 406
Craig Van Dyck 406
Susan Doerr 406
Ben O'Brien 412
João Fernandes 414
Jamie Shiers 414
Bob Jones 414
Sara Pitonnet Gaiarin 414
Wendy Hagenmaier 417
Christa Williford 417
Monique Lassere 417
Lauren Work 417
Jessica G. Benner 417
Edith Halvarsson 420
Seth Erickson 417
Sarah Mason 420
Teru Agata 423
Yosuke Miyata 423
Atsushi Ikeuchi 423
Kyryll Udod 426
Volodymyr Kushnarenko 426
Maureen Pennock 429
Stefan Wesner 426
Michael Day 429
Afelonne Doek 439
Annalise Berdini 432
Evanthia Samaras 429
John Durno 434
Zdeněk Hruška 437
Tim Gollins 439
Vicky Steeves 443
Genevieve Milliken 443
Pierre-Yves Burgi 445
Chao Wang, Zhenxin Wu, Jiancheng Zheng
Hugues Cazeaux 445
Lydie Echernier 445
Marjolein Steeman 450
Peter May 451
Yvette Hollander 450
Christopher A. Lee 453
Maureen Pennock 451
David Russo 451
Kam Woods 453
Pamela Vízner 455
Bertram Lyons 455
Paul Stokes 458
John Kaye 458
Tamsin Burland 458
Dom Fripp 458
Bert Lemmens 460
Tom Davey 458
Robert Gillesse 460
Lisa LaPlant 463
Jamie Shiers 463
Irfan Zuberi 463
Maureen Pennock 463
Trevor Owens 466
Karl-Rainer Blumenthal 469
Maureen Pennock 466
Faye Lemay 466
Tobias Steinke 466
Peggy Griesinger 469
Julia Kim 469
Shira Peltzman 469
Vicky Steeves 469
Erwin Verbruggen 469
Jeanne Kramer-Smyth 472
Montserrat Canela 472
Elizabeth Kata 475
Ineke Deserno 472
William Kilbride 475
Jenny Mitcham 475
Eefke Smit 478
Helena Cousijn 480
Jim Moye 475
Ginny Hendricks 480
Alice Meadows 480
Andrea Goethals 483
Jefferson Bailey 483
Roslynn Ross 483
Karin Bredenberg 486
Nicholas Taylor 483
Eld Zierau 486
Angela Di Iorio 486
Karin Bredenberg 488
Sean Mosely 488
Caylin Smith 491
Juha Lehtonen 488
Patricia Falcao 491
H. Uffen 494
Sara Day Thomson 491
W. Swagemakers 494
Brent West 496
E. van Velzen 494
Elizabeth Kata 498
Joanne Kaczmarek 496
William Kilbride 498
Jenny Mitcham 498
Jim Moye 498
Karin Bredenberg 501
Neil Jefferies 501
Sarah Romkey 501
Andreas Nef 501
Justin Simpson 504
Matthew Addis 504
Jack O'Sullivan 504
Carl Wilson 504
Sarah Romkey 504
Jon Tilbury 504
Martin Klein 508
David Wilcox 510
Herbert Van de Sompel 508
Andrew Woods 510
Bradley J. Daigle 512
Daniel Bernstein 510
Aaron Collie 512
Corey Davis 512
Helen Tibbo 512
Lauren Work 512