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 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

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

DOI: 10.1145/nnnnnnn.nnnnnnn

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.

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, 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.
- 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. https://doi.org/10.6084/m9.figshare.7795829.v1

[2] K. Dohe, D. Durden, "The cost of keeping it: toward effective cost-modeling for digital preservation," presentation, open science framework, 2018. https://doi.org/10.17605/OSF.IO/HVD5F

[3] N. Tallman, L. Work, "Approaching appraisal: framing criteria for selecting digital content for preservation," presentation, Open Science Framework, 2018. https://doi.org/10.17605/OSF.IO/8Y6DC

[4] S. Marks, *Becoming a Trusted Digital Repository*, Chicago, IL: Society of American Archivists, 2015, pp. 46-49.

[5] S. Marks, *Becoming a Trusted Digital Repository*, Chicago, IL: Society of American Archivists, 2015, pp. 50.

[6] K. Green, K. Niven, G. Field, "Migrating 2 and 3D datasets: preserving AutoCAD at the Archeology Data Service," *ISPRS International Journal of Geo-Informatio*n, vol. 5, no. 4, pp. 44-56, April 2016.

[7] D. Rosenthal, "Emulation & virtualization as preservation strategies: a report commissioned by The Andrew W. Mellon Foundation," New York, 2015. https://mellon.org/Rosenthal-Emulation-2015

[8] 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.

[9] D. Minor, "The Digital Preservation Network," presentation, 2014. http://web.stanford.edu/group/dlss/pasig/PASIG_September2014/20140918_Presentations/20140918_04_DigitalPreservationNetwork_DavidMinor.pdf

[10] J. Kunze, J. Littman, E. Madden, J. Scancella, C. Adams, "The BagIt File Packaging Format (V1.0)." https://tools.ietf.org/html/rfc8493

[11] The Texas Digital Library. https://www.tdl.org/about-tdl

[12] 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. doi:10.1061/(ASCE)NH.1527-6996.0000246.

[13] 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. http://dcpapers.dublincore.org/pubs/article/view/3846

[14] File Information Tool Set (FITS). https://projects.iq.harvard.edu/fits/home

[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

[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. https://www.researchgate.net/publication/45456621_Risk_management_foundations_for_DLs_DRAMBORA_digital_repository_audit_method_based_on_risk_assessment

[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

[18] Integrated Rule-Oriented Data System (iRODS). https://irods.org/