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

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. 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 was initiated 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 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 restrict access to their datasets.

II. A VERSATILE ARCHITECTURE

The DLCM solution 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

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

DOI: 10.1145/nnnnnnn.nnnnnnn

allow researchers to prepare their data for preservation, namely: to submit them with a pre-ingest step followed by ingest, to store them physically (archival-storage), to index the metadata (data-management) and to be able to access them according to specific rights (access). For uploading large datasets (Tera Bytes 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, the Genohm 's SLIMS tool has been 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 unit 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 for period of times longer than 10 years (which is the default value).

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.