Research Data Preservation in Canada
A White Paper

Prepared by the Portage Network, Preservation Expert Group (PEG) on behalf of the Canadian Association of Research Libraries (CARL)

Umar Qasim, PEG Chair (University of Alberta)
Corey Davis (University of Victoria and Council of Prairie and Pacific University Libraries)
Alex Garnett (Simon Fraser University)
Steve Marks (University of Toronto)
Michael Moosberger (Dalhousie University)

APRIL 2018

Portage Network
Canadian Association of Research Libraries
portage@carl-abrc.ca

www.carl-abrc.ca

portage
SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA

CARL ABRC
CANADIAN ASSOCIATION OF RESEARCH LIBRARIES
ASSOCIATION DES BIBLIOTHÈQUES DE RECHERCHE DU CANADA

Table of Contents

Executive Summary

This paper provides a framework within which digital preservation is defined, and provides a set of guiding principles that reflect the values and commitments of organizations and communities currently involved in this area of work. Based on this foundation, a federated approach to research data preservation in Canada is proposed, which would build on Portage's ongoing efforts to build networks of expertise and communities of practice across the country.

Using the Open Archival Information System (OAIS) Reference Model as a guide, it is recommended that Portage support the development of a distributed coordinated network of archives, collectively referred to as "Preservation Service Providers" (PSPs), in order to meet the data management needs of Canadian researchers. The model proposed would disaggregate six OAIS functions into three areas:

Repository Services, which would handle the *Ingest*, *Access* and *Data Management* functions of the OAIS model. This could include the two repository technologies currently supported by Portage, the Federated Research Data Repository (FRDR) and Dataverse, as well as any number of domain and institutional repositories based in Canada.

Preservation Services, which would handle the *Archival Storage* functions within the OAIS model and which could be undertaken as through a coordinated network of digital preservation service providers (PSPs).

Planning and Monitoring, in which Portage would be responsible for *Administrative* functions, and where *Preservation Planning* responsibilities would be shared between Portage and the PSPs.

Next steps proposed for the Portage Network to achieve this vision of a federated digital preservation network include:

- Building a common understanding of basic digital preservation requirements in order to determine the core attributes of a sustainable and distributed digital preservation infrastructure for research data in Canada.
- Cultivating and nourishing partnerships with national and regional stakeholders to align existing and emerging services, coordinate communications, outreach and advocacy, and explore collaborative funding opportunities.
- Continuing to articulate a clear and unified message on all issues related to the development of sustainable national research data preservation infrastructure.
- Defining core competencies in support of training for those responsible for research data and digital preservation activities in their respective institutions and organizations.

Introduction

The Preservation Expert Group (PEG) was created to advise Portage on developing research data management (RDM) infrastructure and best practices for preserving research data and metadata in Canada.[1] The members of PEG have written this White Paper as a foundation document to describe the current digital preservation landscape, highlighting some of the digital preservation work already being undertaken in Canada, and to identify challenges that need to be addressed by Portage and other stakeholders to develop and improve RDM capacity and infrastructure across the country. This White Paper is intended to stimulate discussion of digital preservation initiatives and directions in Canada, and how Portage, in collaboration with other stakeholders, can play a role in helping to define and advance these initiatives.

Defining Digital Preservation

"Digital Preservation is the active management of digital content over time to ensure ongoing access."[2] Digital preservation involves a series of activities, such as selecting content for preservation, preparing and maintaining it in an environment that enables access, and having strategies in place to ensure that this content can be made available over time.

Digital preservation best practices are constantly being developed and refined as technological, economic, political, and broad social contexts change. One of the more mature frameworks that has effectively guided many organizations and initiatives in their digital preservation efforts is the Open Archival Information System (OAIS) Reference Model, originally developed by the space data community and now widely adopted as a best-practices framework. The OAIS model describes "an Archive, consisting of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community."[3] This model not only provides a framework for digital preservation practice, it also provides the digital preservation community with a common vocabulary to talk about their activities and initiatives.

---

[1] https://portagenetwork.ca/working-with-portage/network-of-expertise/portage-preservation-expert-group/about-us/
[2] http://www.digitalpreservation.gov/about/
[3] https://www.iso.org/standard/57284.html

Guiding Principles for Portage's Digital Preservation Efforts

Guiding principles reflect the values that an organization or community holds and the level of commitment they have for preserving digital content under their stewardship. These fundamentals are an essential part of any digital preservation effort and may vary from organization to organization, or from community to community. They serve as a foundation for developing policies that guide actions to address the current and future challenges of digital preservation.

The following guiding principles underlie the analysis and recommendations contained in this White Paper. The following list is not intended to be prescriptive or comprehensive. Rather, it reflects the experience of PEG members in developing and implementing systems to support digital preservation, the current state of the field, and the goals of Portage more broadly.

Principle 1: Processes and tools should be community-based, transparent, and open

Digital preservation presents a series of challenges that are not solvable within a short period of time, or through simple solutions. Long-term preservation encompasses a range of complex technical, legal, and ethical dimensions. As such, there will always need to be a wide variety of voices informing the process, and this can only be done in an environment where decision-making is open and subject to the scrutiny and direction of stakeholder communities. A willingness to collaborate and engage a wide variety of stakeholders, as well as a commitment to transparency and openness, are key to the ongoing development and sustainability of digital preservation efforts in Canada.

Principle 2: Access is a primary goal of digital preservation

Digital preservation is not an end unto itself. The objective of preserving any content is to enable appropriate access to that content in the future. Digital content should be provided in formats that meet the needs of designated user communities and that are usable and understandable at an appropriate level.

Principle 3: Not all data can or should be preserved

Preserving everything would be prohibitively expensive and time-consuming. Reasonable and defensible criteria for selecting what data are preserved over the long-term are an important aspect of preservation planning and critical to the success of any efforts in this area.

Principle 4: Digital preservation is an ongoing risk management exercise

Digital preservation is not a prescribed series of static actions to be carried out, after which we can say that something is "preserved." It is an ongoing process that involves the careful evaluation of risks posed to a specific body of digital content. Risks can originate from the nature of the data itself, its formats or the media it is stored on, the platform(s) it resides in, and any number of larger economic and political factors. As such, constant monitoring of the preserved data, technological changes, user community needs, and the broader societal context, is critical.

Principle 5: Metadata is crucial

Metadata is an essential component of digital preservation strategies.[4] Organizations must facilitate the creation, and ensure the maintenance of, robust metadata to ensure the reliability, authenticity, and usability of the digital objects entrusted to their care, in accordance with standards and best practices.[5] Metadata that demonstrates chains of custody and authenticity is of particular importance to those research data required to validate findings and other published conclusions.

---

[4] https://en.wikipedia.org/wiki/Preservation_metadata
[5] For e.g., see: http://www.dpconline.org/docs/technology-watch-reports/894-dpctw13-03/

Challenges Associated with Research Data Preservation

Economic risks

Perhaps the greatest single threat to the long-term accessibility of digital information is economic. Digital preservation is expensive, and the increasing size and complexity of research data creates formidable challenges in terms of costs. According to a study conducted in 2014, global research output doubles every nine years.[6] Who pays for the data ingest, storage, and access infrastructure, as well as the necessary expertise, to support this explosion of research data output?

Heterogeneity in research data

The highly variable structure and content of research data also creates unique challenges. Detailed, context-rich metadata must be preserved along with research data in order to make them independently understandable and reusable by future researchers. This creates a series of challenges for data stewards, and it means that metadata expertise is best deployed near the beginning of the research lifecycle. This is different for other types of research outputs, like electronic journal articles and e-books, which tend to become part of a preservation strategy near the end of the research lifecycle.

Changes in technology

Changes in technology also introduce a range of challenges, such as a lack of backward compatibility and the dependencies on specific, often proprietary hardware and software environments. These issues are compounded by the many highly specialized hardware and software environments associated with multiple fields of study. Whereas a large corpus of electronic journal articles or other published outputs representing dozens of disciplines might be stored in only a handful of file formats, such is not the case with research data, where a large number of data types and file formats need to be managed over time. Finding hardware- and software-agnostic ways of producing and preserving archival copies of complex datasets will be key to enabling long-term access.

---

[6] http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html

The Canadian Context for Research Data Preservation

> "*To increase our capacity to preserve digital information, we need a framework to strengthen, coordinate and better communicate our collective efforts*".[7]

There is an increasing recognition in Canada that a coordinated national strategy for digital preservation is needed.[8] This is recognized across a broad range of disciplines, from the digital humanities[9] to the sciences, especially where 'big data' research projects necessitate a coordinated approach to the long-term storage, management, and preservation of large datasets[10].

And while there is increasing awareness of the importance of a coordinated approach to digital preservation in Canada:

> "*...[t]o date, Canada has lacked a 'master plan' to guide its scientific, cultural, and education communities, businesses, and civil society in the production, use, sharing and preservation of its vast and growing body of digital information.*"[11]

There are many Canadian organizations active at the intersections of research data management and digital preservation. Overall, but often in relative isolation, they have made significant investments in developing expertise, building organizational capacity, and deploying digital research infrastructure to ensure the long-term viability of important research data and other digital materials. To date, however, there has been no national mechanism to enable and implement preservation infrastructure and expertise to meet the needs of the research community in a coordinated and sustainable way.

---

[7] http://www.lac-bac.gc.ca/obj/012033/f2/012033-1000-e.pdf (p. 7)
[8] http://www.lac-bac.gc.ca/obj/012033/f2/012033-1000-e.pdf (p. 4)
[9] http://www.cwrc.ca/cwrc_news/lasting-change-sustaining-digital-scholarship-and-culture-in-canada/
[10] http://www.scienceadvice.ca/uploads/eng/assessments%20and%20publications%20and%20news%20releases/memory/CofCA_14-377_MemoryInstitutions_WEB_E.PDF (p. 34)
[11] http://www.lac-bac.gc.ca/obj/012033/f2/012033-1000-e.pdf (p. 7)

There are a number of key organizations and stakeholders in Canada focused on the provision of digital research infrastructure for long-term preservation. These organizations represent significant expertise and capacity in digital preservation and their engagement will be critical to the success of Portage's efforts.

National Organizations	Regional Organizations
● Canadian Association of Research Libraries ● Canadian Research Knowledge Network ● Compute Canada ● CANARIE and Research Data Canada ● Library and Archives Canada ● Leadership Council for Digital Research Infrastructure	● Regional academic library consortia (COPPUL, OCUL, BCI, CAUL-CBUA) ● OCUL Scholars Portal **Institutions** ● A number of research libraries in Canada provide high-capacity and scalable digital preservation services to their campus communities.

Model for a National Distributed Digital Preservation Service

With the understanding that many organizations in Canada are currently involved in digital preservation activities, the PEG recommends that Portage spearhead efforts to establish a distributed, OAIS-type preservation network to provide Canadian researchers with an easy, reliable way to deposit, find, share, and preserve research data. Background information on the OAIS Reference Model is provided in order to effectively frame this recommendation.

The Open Archival Information System (OAIS) Reference Model

The OAIS Reference Model identifies and describes a core set of functions an organization uses to meet its primary mission of preserving information over the long-term for its community.[12]

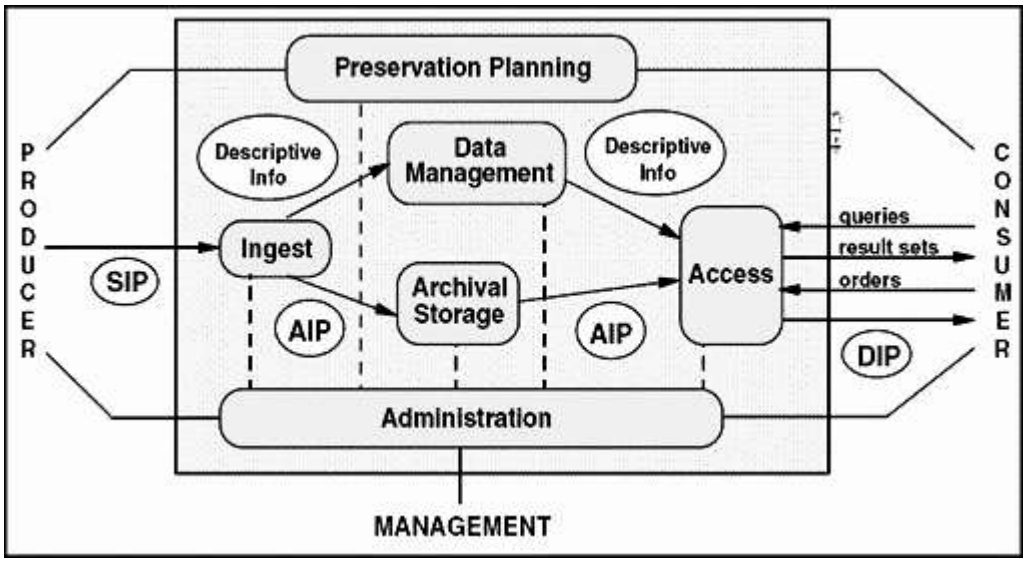

Figure 1: The OAIS Reference Model[13]

[12] "Designated Community: 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." https://public.ccsds.org/pubs/650x0m2.pdf (p. 21)
[13] https://commons.wikimedia.org/wiki/File:OAIS-.gif

1. The **Ingest** function accepts information from content producers, validates it and prepares it for inclusion in the archive. Creation of descriptive metadata is also part of this function. The Ingest function results in the output of Archival Information Packages (AIPs)[14], which are moved to an archival store for long-term retention.
2. The **Data Management** function maintains descriptive metadata, manages the administrative data, and supports search and discovery of the archived content.
3. The **Archival Storage** function manages the long-term storage and maintenance of the digital content in the archive.
4. The **Access function** disseminates the content in the archive.
5. The **Preservation Planning** function monitors continuously the condition of the archive and employs appropriate preservation strategies to keep the content accessible.
6. The **Administration** function manages operations of the archive and coordinates the activities of the other five functions.[15]

Disaggregating the OAIS model

Typically, a single organization is responsible for the people and infrastructure needed to support an OAIS-compliant system. However, it is possible to take a disaggregated approach to implementing various OAIS functions. With this approach, even a single OAIS function can be implemented across multiple organizations, with each organization providing a unique service addressing different segments of the functional model. Through a disaggregated service model, the core functions of an OAIS-based archival system can be allocated to different organizations within a network, especially under conditions where a single organization is unable to perform all the core functions in isolation because of resource constraints and/or other limitations.

Portage's Role in Coordinating a Disaggregated Approach to Research Data Preservation in Canada

The PEG recommends that Portage lead efforts to establish a distributed, OAIS-type preservation network for research data in Canada. Some network functions are already in place and others are being developed in partnership with several organizations. In this model, Portage would primarily play a coordination role, helping to harmonize the activities of various organizations to ensure that all the necessary functions articulated

---

[14] An AIP as defined by the OAIS reference model, is an information package that is used to transmit archival objects into a digital archival system, store the objects within the system, and transmit objects from the system. An AIP contains both metadata that describes structure and content and the actual content itself.

[15] http://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file

in the OAIS model are in place to support the long-term preservation of research data in Canada.

OAIS-based functions of a distributed network would be organized into three main areas:

1. **Preservation services**, especially archival storage
2. **Repository services** for ingest, access, and data management
3. **Preservation planning and administration**

Preservation Services

Dedicated preservation services within the distributed service model would support the archival storage function of the system. Providers of archival storage and related preservation services (Digital Preservation Service Providers, or PSPs) would work collaboratively with Portage to determine the scope of the preservation planning function overall, with Portage serving a coordinating function by helping to bring multiple PSPs into the preservation planning process overall.

PSPs would be organizations within Canada with existing digital preservation capacity. They may include, for example, research libraries that already provide digital preservation services for their campus communities, especially in terms of data ingest and persistent archival storage, or domain-based repositories serving a specific academic or research community. They may include regional academic library consortia, such as OCUL/Scholars Portal and COPPUL, both of whom offer a wide range of preservation-related services to their members. They may also include organizations centered around ultra-high-speed research networks and high-performance computing facilities, such as Compute Canada, which, in partnership with Portage, offers preservation processing as part of the FRDR platform.

Archival Storage

Archival storage is the function of the network which manages the long-term storage and maintenance of content. Regular maintenance activities such as format migration, media refreshment, error checking, and disaster recovery planning are an important part of this service. Archival storage doesn't provide any direct access to end users; rather, it interacts with other functions in the preservation network.

In the proposed preservation network service model, the archival storage function would be filled by organizations with existing capabilities and the associated expertise and experience to perform this role effectively. Partnering with an existing and diverse set of PSPs allows the network to mitigate against the risks of working with a single and specific technical platform. At the same time, Portage would work with PSPs to develop a set of shared requirements for archival storage and related services, while the

operation and maintenance of infrastructure associated with individual PSPs would be left to their host organizations.

Repository Services

Data repositories, in general, provide ingest, access, and data management functions for deposited content. In a distributed model, any Canadian data archiving service provider able to meet some basic community-established terms, conditions and technical requirements, could be included as part of the preservation network.

As part of its vision for a common national data services framework in Canada, Portage currently supports two repository technologies: the Federated Research Data Repository (FRDR) and Dataverse. FRDR, created in partnership with Compute Canada, is designed to fill known gaps in the repository landscape; namely, the ability to handle large datasets ("big data"), the provision of a national discovery layer, and the development of a preservation pipeline through Archivematica integration. [16] Meanwhile, Portage's Dataverse North Working Group is a community-led collaboration among university libraries in Canada with the goal of building capacity and support for Dataverse. [17] Dataverse North is exploring options for the creation of a national Dataverse instance that would be available to all Canadian researchers. We envision both of these repository technologies being well-suited to fulfill the ingest, access, and data management functions of the OAIS model.

Preservation Planning and Administration

Portage would be responsible for the overall administration function of the preservation network, the scope of which would be determined in consultation with PSPs, and would share the responsibility for preservation planning with those same organizations and initiatives.

Preservation planning

Preservation planning involves documenting preservation strategies and appropriate practices to maintain accessibility of digital content over time. As part of preservation planning, an organization must keep fully abreast of changes to the external environment in order to implement new strategies for risk mitigation when necessary.

In the proposed preservation network service model, Portage would work with PSPs to perform this function. As PSPs are preserving significant amounts of data in their own right, they are natural partners to engage in the preservation planning function of the network. PSPs would work with Portage to ensure that the preservation requirements

[16] https://portagenetwork.ca/frdr-dfdr/
[17] https://dataverse.org/

and risk mitigation strategies being employed locally are consistent with those of the overall preservation network.

Based on evolving community needs and technology environments, Portage would recommend strategies and practices to PSPs in order to safeguard preserved content from becoming inaccessible. Likewise, PSPs may identify new practices or risks that require action on the part of Portage or other PSPs. Portage would also work with PSPs to determine specifications for creating archival information packages (AIPs). Furthermore, Portage would maintain a registry of Portage-related preserved content within each PSP. A central registry of file formats recommended for long-term preservation would also need to be developed by Portage, in collaboration with other national and international preservation initiatives.

Administration

The administration function is responsible for day-to-day operations of the preservation network, including coordinating activities between various network functions and between participating organizations. In the proposed model, Portage would be responsible for maintenance of the network from a logistical and policy perspective. Policy work would likely entail negotiating service agreements and reciprocal arrangements with PSPs, and overseeing the transition of materials from one PSP to another, should the need arise. Policy work might also include developing guidelines and processes for certification of PSPs as trustworthy repositories, appraisal criteria for ingested content, and terms and conditions for deposit licenses. More generally, Portage would also help ensure that all network participants remain up-to-date on current practices in digital preservation, and that overall preservation planning is consistent and compatible between PSPs.

Next Steps

Canada needs a coordinated preservation program with core activities that support research data preservation nationally. Creating a distributed OAIS-compliant preservation network in partnership with PSPs is crucial to the long-term preservation of research data in Canada. As Google's Richard Whitt states, "the chief…challenge is not to supplant what activity is taking place today, but to help coordinate and expand and deepen those efforts."[18]

In consideration of the current landscape related to RDM, digital infrastructure, and digital preservation in Canada, and with special consideration of past efforts, the PEG recommends that Portage, in the next two years, focus on the following areas in order to advance a distributed preservation network for research data in Canada.

- **Build a Common Understanding of Basic Digital Preservation Requirements:** Strike a working group under the auspices of the Portage Network, with representation from a broad range of stakeholders, to determine the core attributes of a sustainable and distributed digital preservation network for research data in Canada, including options for funding long--term archival storage.

- **Focus on Partnerships:** Establish agreements with national and regional stakeholders in order to align existing and emerging digital preservation infrastructure, coordinate stakeholder communications, outreach, and advocacy, and explore collaborative funding opportunities.

- **Unify Messaging:** Align advocacy and outreach efforts across stakeholder groups around the need for a coordinated approach to the deployment and management of research data preservation infrastructure nationally. Advocate for a rationalization of mandates across several organizations to enable the establishment of significant persistent storage resources in conjunction with high-performance computing or ultra-high-speed research networking infrastructure and facilities.

- **Articulate Core Competencies:** Articulate core competencies in Canada needed for research data preservation activities. This may include providing training opportunities for individuals interested in joining this field, and facilitating discussions and knowledge sharing among experts across the country.

---

[18] http://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=1609&context=chtlj (p. 79)

Conclusion

PEG envisions a future in which Canada becomes a world leader in research and innovation by effectively managing data-intensive research assets. A key to this vision is the establishment of an effective national strategy to tackle the myriad challenges associated with enabling meaningful access to research data over the long-term. By spearheading the development of a distributed digital preservation network for research data and metadata in Canada, and with an accompanying commitment to collaboration, transparency, and openness, we believe that Portage can make a real difference to the lives of Canadians. When data are effectively preserved, they can be more readily shared and reused, allowing Canadian researchers to build upon the work of others, stimulating new discoveries and leading to more transparency and accountability within the research enterprise. In the words of former Portage Director Chuck Humphrey, "[t]he most innovative nations in the future will be those that best manage their research data today."[19]

[19] https://preservingresearchdataincanada.net/category/rdmi-1/