December 14th, 2020

Digital preservation and NDRIO: a white paper

From The Portage Preservation Expert Group *and* The Canadian Association of Research Libraries Digital Preservation Working Group

> *There will be more infectious disease outbreaks in our future. We need to be able to study and learn from this one. Taking the appropriate steps now to preserve the record of ... this strange and terrifying moment would be a valuable legacy. Because remembering is a form of honouring.*

Esyllt Jones, Ian Milligan, and Shelley Sweeney,
"Will Covid-19 become the 21st century's forgotten pandemic?"
The Globe and Mail, October 5, 2020.

This proposal is endorsed by the Board of the Canadian Association of Research Libraries (CARL) and the CARL Portage Network.

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

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

Signatories

- Dale Askey, Vice-Provost (Library & Museums) & Chief Librarian, University of Alberta
- Jonathan Bengtson, University Librarian, University of Victoria & President of Canadian Association of Research Libraries (CARL)
- Carrie Breton, Repository Specialist, University of Guelph
- Corey Davis, University of Victoria & Canadian Association of Research Libraries (CARL) Visiting Program Officer
- Rebecca Dickson, Digital Preservation Coordinator, Council of Pacific and Prairie University Libraries (COPPUL)
- Jonathan Dorey, Research Officer - RDM, INRS, ENAP, TÉLUQ
- Evan Echols, Digital Collections Archivist, University of New Brunswick
- Alex Garnett, Research Data Management & Systems Librarian, Simon Fraser University
- Kenton Good, Head, Digital Preservation Services, University of Alberta
- Susan Haigh, Executive Director, Canadian Association of Research Libraries (CARL)
- Grant Hurley, Digital Preservation Librarian, Scholars Portal, Ontario Council of University Libraries (OCUL)
- Beth Knazook, Preservation Coordinator, Portage
- Steve Marks, Digital Preservation Librarian, University of Toronto
- Keltie McPhail, Digital Initiatives Librarian, University of Prince Edward Island
- Michael Moosberger, Associate University Librarian for Research & Scholarly
- Mireille Nappert, Archiviste numérique, HEC Montréal
- Rebecca Ross, Director, Marketing and Stakeholder Engagement, Canadian Research Knowledge Network (CRKN)
- Roslynn Ross, Director, Digital Preservation and Migration, Digital Operations and Preservation Branch, Library & Archives Canada
- Jean-François Ruest, Spécialiste en ressources documentaires, Université Laval
- Lara Wilson, Director, Special Collections & University Archivist, University of Victoria
- Lee Wilson, Service Manager, Portage
- Mike Winter, Compute Canada
- Jess Whyte, Digital Assets Librarian, University of Toronto

For more information, please contact:

Corey Davis
coreyd@uvic.ca
(778) 677-5746

Respectfully submitted from the traditional territories of the Lekwungen peoples

Introduction

Researchers have sounded the alarm over data gaps that are "hindering Canada's response to the COVID-19 pandemic."[1] These gaps include not only the *absence* of good data -- such as those related to COVID-19 cases among Indigenous peoples[2] -- but also an inability of researchers to access *existing* data when needed. And if accessing data today is a challenge, how much harder will it be to ensure access to these data in the future?

NDRIO must do its part to support researchers who are working diligently to save lives and livelihoods today. It may seem premature to consider future data needs in the midst of COVID-19's second wave, but, as researchers are making increasingly clear, the actions we take today will directly impact our ability to fight the pandemics of tomorrow:

> Almost as soon as COVID-19 grabbed hold of our consciousness last spring, people began turning to historical disease outbreaks to help us make sense of our current experiences. Past epidemics have become indispensable reference points.[3]

As librarians and archivists from Canadian research institutions and allied organizations, we care deeply about the long-term accessibility of research data; and we are compelled to ask: *Are the right investments being made now to ensure that our researchers are ready to face the next pandemic?*

The answer is, unfortunately, no. Research libraries, which lead Canada's research data management (RDM) efforts, are concerned about financial and other barriers to preserving data over time, especially as the volume of data continues to grow.[4] Researchers already experience significant challenges when trying to ensure the long-term accessibility of their data, especially after the research funding for a specific project ends.[5]

But there is good news too. Thanks to the hard work of research libraries, regional consortia, and other organizations over the last several decades, many of the pieces we need to build a robust Canadian digital preservation infrastructure are at hand.

---

[1] Andrew-Gee, E., & Grant, T. (2020, April 23). How Canada's crucial data gaps are hindering the coronavirus pandemic response. *The Globe and Mail.* https://www.theglobeandmail.com/canada/article-how-canadas-crucial-data-gaps-are-hindering-the-coronavirus-pandemic/

[2] For example, see Hamilton, W. (2020, December 5). Advocates call for B.C. to collect COVID-19 data specific to urban Indigenous people. *CBC News British Columbia.* https://www.cbc.ca/news/canada/british-columbia/advocates-call-for-b-c-to-collect-covid-19-data-specific-to-urban-indigenous-people-1.5802006

[3] Jones, E., Milligan, I., & Sweeney, S. (2020, October 5). Will Covid-19 Become the 21st Century's Forgotten Pandemic? *The Globe and Mail.* https://www.theglobeandmail.com/canada/article-will-covid-19-become-the-21st-centurys-forgotten-pandemic/

[4] Hurley, G. & Shearer, K. (2019). *Final report of the survey on digital preservation capacity and needs at Canadian memory institutions, 2017-18*. Canadian Association of Research Libraries. https://hdl.handle.net/1807/985535. This report highlighted low use of tools for preservation processing and low uses of preservation-friendly storage infrastructures largely due to resource issues.

[5] Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., ... & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current biology, 24*(1), 94-97. https://doi.org/10.1016/j.cub.2013.11.014

What we need now is a focused, sustained, and coordinated effort by NDRIO to ensure existing services are enhanced, connected, and scaled to meet the needs of 21st century, data-intensive research in Canada.

In the opinion of this paper's signatories, the three main components of DRI in need of support to better ensure long-term access to vital datasets are:

Trustworthy data repositories	The research enterprise is becoming "increasingly dependent on both digital data and repositories that provide access to and enable the use of such resources. Repositories must earn the trust of the communities they intend to serve and demonstrate that they are reliable and capable of appropriately managing the data they hold."[6]
Preservation processing systems	These systems integrate with repositories to create and manage archival information packages (AIPs). AIPs are meant to be intelligible independent of repository platforms, and contain a wealth of metadata that help describe what they contain and where they came from. This helps ensure datasets remain usable as the technologies that create, read, and present digital information evolve.
Future-friendly storage services	Preservation storage services are not defined by the underlying storage technologies they rely on (e.g. SSD, tape, etc.), but rather by the policies and workflows that enable multiple, independently-administered, geographically distributed, authentic and trustworthy copies of data to be managed over time.

Specific recommendations on what NDRIO might consider doing in each of these areas are provided below. But first, it may be useful to step back and get a better sense of what we mean when we talk about digital preservation more broadly.

What is digital preservation?

In the context of this paper, digital preservation is about sustaining access to data for as long as necessary. It involves actively monitoring, planning, administering, and managing data, as well as related systems and workflows -- all of which ensure persistence over time.[7] Good digital preservation practices also ensure the authenticity and reliability, and thereby the trustworthiness, of the resources preserved.

Researchers rely on a complex and interdependent mix of software and hardware environments to make data intelligible. Digital preservation focuses on ensuring data remain robust and resilient in the face of unrelenting technological change. Established best practices are deployed to 'future-proof' data so that it remains meaningful long after the original research project has ended and the contexts for data collection, analysis, and other activities are no longer available or easily understood.

---

[6] Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., ... & Khodiyar, V. (2020). The TRUST Principles for digital repositories. *Scientific Data*, *7*(1), 1-5. https://www.nature.com/articles/s41597-020-0486-7

[7] Digital Preservation Coalition. (2020). *What is digital preservation?* https://www.dpconline.org/digipres/dpeg-what-is-dp

Digital preservation activities ensure access over time by addressing risks such as:

- Hardware and/or software obsolescence rendering data unreadable
- Loss of data that is irreplaceable or unable to be reproduced
- Loss of context, identifiers, and/or documentation necessary to effectively interpret data[8]

As such, digital preservation encompasses a wide range of activities to mitigate these risks, including:

- Planning and developing strategies, policies, and workflows
- Liaising with data creators, data users, policy makers, and many others
- Ensuring repositories, preservation processing systems, and distributed preservation storage networks work together

Digital preservation is not simply storing data safely for a long time. It is a whole system of technology, people, and best practices that are intentionally deployed to increase the chances that data will survive and be accessible to researchers in the future. *These efforts are by nature ongoing, regardless of the underlying technologies that support them, and they require sustained investments across multiple institutions in order to be successful.*

Investing in Canada's digital preservation infrastructure

Significant investments in research data preservation have already been made in Canada. Research libraries have developed expertise and infrastructure over the last several decades, often in partnership with organizations like Compute Canada and CANARIE. Examples include:

Trustworthy data repositories	The Canadian Association of Research Libraries (CARL), through Portage, partnered with Compute Canada to develop the Federated Research Data Repository (FRDR). In addition, Scholars Portal, a service of the Ontario Council of University Libraries (OCUL) at the University of Toronto, with funding support from CANARIE, has launched Dataverse Canada, a national data repository service. Many institutions have also created repository infrastructure in support of research data preservation, such as the CAUL Atlantic Islandora Repository Network (CAIRN).[9]
Preservation processing systems	Regional academic library consortia are providing preservation processing through such services as the Council of Prairie and Pacific University Libraries' (COPPUL) Archivematica-as-a-Service, and OCUL's Permafrost service. Additionally, FRDR has created an Archivematica pipeline that will enable it to connect to preservation storage systems (see below), and Scholars Portal has funded the development of a Dataverse-Archivematica integration to produce standards-compliant AIPs from Dataverse. Additionally,

[8] Digital Preservation Coalition. (2020). *Risks: Failing to preserve digital materials renders them unusable.* https://www.dpconline.org/digipres/implement-digipres/dpeg-home/dpeg-risks

[9] CAUL - CBUA Atlantic Islandora Repository Network. (*n.d.*). *About CAIRN*. https://www.cairnrepo.org/

	the Canadian Research Knowledge Network (CRKN) maintains a Trusted Digital Repository (TDR) for heritage materials.[10]
Future-friendly storage services	COPPUL's WestVault preservation storage service is a high-redundancy peer-to-peer storage network built across five campus data centres in Western Canada. OCUL's Ontario Library Research Cloud (OLRC) is a high capacity, geographically distributed cloud storage provider built on Swift OpenStack, with infrastructure at 5 universities in Ontario. CRKN also has significant storage capacity as part of its TDR.

Even with all of this in place, Canada still falls short of other advanced jurisdictions. For example, Jisc in the UK supports the full lifecycle of research data management, including preservation processing and storage through systems like Archivematica.[11] The European Open Science Cloud (EOSC) ARCHIVER project is building services for data preservation, as well as providing trustworthy repository services.[12] Some US jurisdiction, such as Texas and California, have deployed State-wide data repository infrastructures through academic library consortia that interoperate with robust preservation services provided by organizations like The Chronopolis Network,[13] Lyrasis,[14] and LOCKSS.[15]

What's next for digital preservation infrastructure in Canada?

Investing in capacity and coordination

We see NDRIO continuing the work of Portage in coordinating key functional elements across multiple partner organizations to create cohesive, sustainable, and country-spanning preservation infrastructure that leverages expertise and resources from across the research community. These efforts would include:

- Coordinating technology developments between repositories, preservation processing systems, and future-friendly storage to ensure interoperability across platforms and regions
- Strengthening standards-based certification efforts for data repositories and preservation infrastructure[16]
- Negotiating agreements to ensure a trustworthy chain of custody as data moves from repository infrastructure through preservation processing into future-friendly storage
- Monitoring external developments and identifying collaborative opportunities within Canada and internationally

---

[10] Canadian Research Knowledge Network. (2020). *Trustworthy digital repository*. https://www.crkn-rcdr.ca/en/trustworthy-digital-repository

[11] Jisc. (2020). *Preservation*. https://www.jisc.ac.uk/preservation

[12] Archiver. (2020). *ARCHIVER & EOSC, the European Open Science Cloud*. https://www.archiver-project.eu/archiver-eosc

[13] The Chronopolis Network. (2020). *About*. https://libraries.ucsd.edu/chronopolis/about/index.html

[14] LYRASIS. (2020). https://www.lyrasis.org/Pages/Main.aspx

[15] LOCKSS. (*n.d.*). https://www.lockss.org/

[16] "Standards such as the CoreTrustSeal, DIN31644/NESTOR, and ISO163638 focus on four major assessment areas: organization, digital object management, technical infrastructure, and security risk management." *From* Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., ... & Khodiyar, V. (2020). The TRUST Principles for digital repositories. *Scientific Data*, *7*(1), 1-5. https://www.nature.com/articles/s41597-020-0486-7

In order to carry out this work, NDRIO must continue to fund positions dedicated to the curation and preservation of research data through Coordinator positions, with the understanding that as the organization matures and research-intensive activities grow, more people will be needed to ensure long-term access to Canada's research data resources.

Investments in infrastructure
We need NDRIO's support to improve preservation infrastructure in Canada, either through direct efforts and provisioning of infrastructure, or through targeted funding calls to enable existing providers to enhance their services to meet national preservation demands.

The highest priorities for infrastructure investments as identified by the signatories of the paper include:

Trustworthy data repositories	1. Ensure adequate repository storage resources are available for FRDR and Dataverse Canada. This should include 2-3 copies of all data in separate geographic locations. 2. Support both these repositories in developing robust, transparent policy infrastructure that clearly articulates sustainability considerations, digital object management, technical infrastructure, security risk management, and succession planning. 3. Support the development of preservation-friendly repository functionality, including fixity checking, format identification, robust and standard metadata, and enhanced integration with preservation processing and/or future-friendly preservation storage services.
Preservation processing systems	1. Enhance the Dataverse-Archivematica integration to enable more automated preservation processing of curated datasets. 2. Generalize the FRDR Archivematica pipeline for use in a variety of repository environments. 3. Create bilingual interfaces and service points and enhance processing capacity for COPPUL's Archivematica-as-a-Service and OCUL's Permafrost service.
Future-friendly storage services	1. Scale COPPUL's WestVault to meet high-redundancy storage requirements, for those data served by a higher level of preservation.[17] 2. Enhance the OLRC to integrate with Dataverse Canada, and expand the distributed Swift OpenStack infrastructure in jurisdictions outside of Ontario. 3. Refine DuraCloud Canada to support preservation service providers with connected workflows and infrastructure for national preservation and data management services.

Conclusion
Currently, Compute Canada has no mandate to provide preservation-supporting services for research data beyond the life of a specific, funded research project. This approach has proven

[17] National Digital Stewardship Alliance. (2019). *Levels of preservation.* https://ndsa.org/publications/levels-of-digital-preservation/

outdated, short-sighted, and inimical to the requirements of 21st century data-intensive research.

Digital preservation infrastructure is a critical component of research data management and DRI more broadly. As such, we encourage NDRIO to find ways of addressing this gap. If NDRIO is going to sustain ARC infrastructure beyond any one federal funding cycle, and renew and grow ARC hardware and software assets over time in support of changing researcher needs and technological innovations, there is no reason why the same commitment should not be made for preservation infrastructure as described in this report. A robust suite of repositories, processing services and storage infrastructure for depositing, discovering, and preserving research data already exists. *What is needed is coherent and strategic coordination of these resources alongside additional investment to ensure they are sustainable*.

Because of continuing population growth and human encroachment into natural habitats, not to mention the resumption of routine international travel, it is unlikely the next pandemic will wait another 100 years to strike.[18] The pace of technological change is accelerating. The complex, interdependent, hardware and software environments that enable data-intensive research today will not be the same in 10 years, much less 100.

There is a growing recognition in the research community that enabling future-friendly best-practices, workflows, and technologies in support of long-term access to research data is critical. If NDRIO is committed to supporting RDM, it must also be willing to cultivate and sustain digital preservation infrastructure in the present, for the future.

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, “the most innovative nations in the future will be those that best manage their research data today.”[19] These words have never rung more true.

[18] Hunter, D. (2020, April 5). We are fighting a 21st-century disease with 20th-century weapons. *The Guardian*. https://www.theguardian.com/world/2020/apr/05/we-are-fighting-a-21st-century-disease-with-20th-century-weapons

[19] Humphrey, C. (2012). *Research Data Management Infrastructure*. https://preservingresearchdataincanada.net/category/rdmi-1/