Submitted on: 15-08-05

Portage: Supporting Canadian innovation through shared expertise and stewardship of research data

Kathleen Shearer
Research Associate, Canadian Association of Research Libraries, Ottawa, Canada.
E-mail address: kathleen.shearer@carl-abrc.ca

Susan Haigh
Executive Director, Canadian Association of Research Libraries, Ottawa, Canada.
E-mail address: susan.haigh@carl-abrc.ca

Martha Whitehead
Vice-Provost and University Librarian, Queen's University, Kingston, Canada
E-mail address: martha.whitehead@queensu.ca

Abstract:

The billions of dollars that are invested every year in research generate vast and diverse amounts of research data. Sound research data management (RDM) practices, with due respect for confidentiality and intellectual property, accelerates scientific progress by allowing researchers to access and re-use others' data for their own scientific purposes, thereby adding value to those data and speeding up the rate of new discoveries.

The paper will describe Canada's Portage project to develop a coordinated, library-based national research data network. Led by the Canadian Association of Research Libraries (CARL), in collaboration with regional academic library associations and other important national infrastructure organizations, Portage has two major components:

- *A library-based network of expertise for research data management; and*
- *A project that connects the various infrastructure and service components into a national preservation and discovery network for research data that will evolve and expand over time.*

Plans for preservation services include a technical infrastructure consisting of software that supports the entire research data lifecycle (ingest, preservation, discovery, access, repurposing), data

replication services, and networked data storage. This infrastructure will be highly distributed with local, regional and central nodes and will also be based on standards to ensure interoperability. Close collaboration with other partners and stakeholders, including organizations providing high-capacity compute and storage resources and the national high-speed network, will be essential for the development and ongoing maintenance of this infrastructure.

Keywords: Research data management, libraries, infrastructure, sustainability, collaboration

Introduction

Billions of dollars are invested every year in research, an investment that generates vast and diverse amounts of data. If properly managed, these data have virtually limitless potential to be re-used in innovative ways. Sound research data management, with respect for confidentiality and intellectual property, accelerates scientific progress by allowing researchers to access and re-use others' data for their own scientific purposes, thereby adding value to those data and speeding up the rate of new discoveries. It also leads to greater efficiencies in research by preventing duplication in data creation; and enables greater transparency and verification of research findings.

Many countries including Australia, Netherlands, United Kingdom, and United States are investing in national policies, infrastructure and services to support more comprehensive research data management (RDM). In Canada, the federal government has recently published Canada's Action Plan on Open Government 2014-16[1], which contains a section on Open Science. The Plan includes deliverables aimed at improving access to publications and data resulting from federally funded scientific activities. It also calls for the development and adoption of policies, guidelines and tools to support effective stewardship of scientific data. Meanwhile Canada's three federal granting agencies have been consulting with the research community about RDM policy implementation that would improve the way research data is managed in Canada.

In order to derive maximum benefits from research data, it must be managed appropriately across the data lifecycle. This means we need services and infrastructure for researchers throughout and beyond the lifespan of a project. Without this holistic kind of support, the majority of research data will be lost or inaccessible to others. In Canada there are gaps in the current landscape. A recent environmental scan published by the tri-councils asserts, "Canada still lacks infrastructure, services and funding mechanisms to support widespread RDM. Infrastructure funding remains focused on domain-based solutions that support research excellence, rather than data sharing and preservation after the lifespan of the project."[2]

[1] http://open.canada.ca/en/content/canadas-action-plan-open-government-2014-16
[2] Shearer, K. http://www.science.gc.ca/1E116DB8-E7F3-4B6F-BB44-83342BAAA030/Comprehensive%20Brief%20on%20Research%20Data%20Management%20Policies.pdf

The Broader Context

Much of the focus for improving the accessibility and preservation of research data has been targeted at the massive datasets produced through large science projects. However, there are thousands of research projects every year that produce data which rest on the hard drives of researchers and are neither catalogued nor accessible. A 2011 survey of 1700 researchers across disciplines undertaken by the journal, *Science*, found that 48.3% of respondents were working with datasets that were less than 1GB in size and over half of those polled store their data only in their laboratories.[3] The same situation exists in Canada: a 2013 survey of over 300 Canadian researchers undertaken by Susan Mowers et.al found that only 4% of respondents shared their data through a "curated digital data repository", and another 14% used an institution repository or a "public domain archive" The vast majority, 81% of respondents, indicated that they stored data on their local hard drives.[4]

Many researchers are not yet comfortable with sharing their data. A survey of more than 2,250 responses from researchers across a wide array of disciplines and countries found that only about half of respondents (52%) indicated they had made their data publicly available, although this varies significantly across disciplines.[5] The major reason cited for not sharing data are IP and confidentiality concerns. Furthermore, many researchers simply do not have the current "know how" for data sharing. A survey in the US of researchers at five different institutions found, "None of the scholars interviewed during this study expressed satisfaction with their level of expertise in data management, and few had access to individuals who could provide knowledgeable guidance. On the contrary, most participants reported feeling adrift when establishing protocols for managing their data and added that they lacked the resources to determine best practices, let alone to implement them."[6]

Libraries and Research Data Management

In this burgeoning era of 'data intensive scholarship', roles and responsibilities are still being established amongst the many stakeholders in this landscape. Libraries, in many ways, are well positioned to develop valuable services to support research data stewardship. Librarians are already recognized on campus for expertise in preserving and providing access to other types of research content and they have links with the disciplinary communities. However, there are a variety of challenges. For groups not familiar with current library services, libraries may not be obvious partners and stakeholders in terms of developing services for research data. As well, managing this kind of data presents special challenges. Research data are very diverse in format, metadata, and the analytical tools applied. In addition, 'data collections,' which have use beyond the original research project for which the data were created, are community resources. The value of sharing across institutional boundaries is as important as demonstrating that library data services are directly benefiting the researchers at a given institution.

Libraries can play a wide range of roles in the area of research data management. An introductory course for research librarians developed by the Canadian Association of

[3] Ferguson, DOI: 10.1126/science.331.6018.692
[4] Mowers et.al.: http://gsg.uottawa.ca/data/open/aa-interim-survey-report/20130801-en.pdf
[5] Ferguson, DOI: 10.1126/science.331.6018.692
[6] Jahnke et.al.: http://www.clir.org/pubs/reports/pub154/problem-of-data

Research Libraries categorized library services into 4 categories: collections, users, access, and preservation. Within each category, numerous service areas were identified, including providing information and support services, managing data repositories, long-term preservation services, and reference services. (CARL, 2013)

In 2012, LIBER, the European Association of Research Libraries, published, *Ten recommendations for libraries to get started with research data management* Recommendations include activities such as:[7]

- Offer research data management support, including data management plans for grant applications, intellectual property rights advice and information materials. Assist faculty with data management plans and the integration of data management into the curriculum, and
- Offer or mediate secure storage for dynamic and static research data in co-operation with institutional IT units and/or seek exploitation of appropriate cloud services.

Institutional versus Collective Action

The Canadian Association of Research Libraries (CARL) has for some years focused on data management as an issue through a Data Management Subcommittee (DMSC) and several reports were prepared for distribution within CARL institutions and more broadly.[8] The DMSC also supported an education program for research librarians that was very well received and indicated a thirst for further collective action. CARL has also advocated in its submissions to federal budget consultations and elsewhere the importance of national support for research data sustainability. In addition, many Canadian research libraries have provided data services through a research data library as far back as the 1990s, focusing on purchased or licensed Statistics Canada data, financial data, and consortial collections, such as the ICPSR, in many instances through collaborative development and support of storage and retrieval tools. They were also involved in the ultimately unsuccessful National Consultation on Access to Scientific Research Data[9], which was intended to work towards a national research data strategy.

Some Canadian research libraries are already involved in the development of regional networks, but there are others whose research data management needs are not yet being addressed. Although the current focus of the regional networks must be their regional constituents, the individuals involved have the knowledge, experience, and vision needed to develop a national network; and they are willing to collaborate.

While some RDM services are most appropriately delivered locally by the individual institution, many other services can be undertaken collectively by library consortia, especially given the resources and expertise involved with RDM. In addition, by its nature, research data management is a highly collaborative effort. Not only does it involve research libraries, but also the participation of research services and ethics offices reporting to Vice-Presidents Research, information technology service units, the national high performance computing

[7] LIBER, 2012

[8] See the CARL website, Research Data Management section: http://www.carl-abrc.ca/en/scholarly-communications/data-management-sub-committee.html

[9] Humphrey, C 2014From National institution to National Infrastructure (http://preservingresearchdataincanada.net/category/national-data-infrastructure/)

community (Compute Canada), and the national high-speed research network (CANARIE). To that end, CARL envisioned the collaborative development of national, shared services that provide some level of support for all Canadian research libraries, through the development of a national, library-based research data management network.

The Portage Network

In March 2014, CARL launched ARC, a one-year project to lay the foundation for implementing the national library-based research data management network. This project involved participation from a wide variety of stakeholders in the community, including all four regional academic library associations, as well as other important stakeholders representing some of Canada's foremost research data management experts. The project adopted the following vision and principles for the network[10]:

> The availability of digital data is dramatically redefining the nature and scope of the research endeavour across all domains in the 21st century. We envision a future in which Canada capitalizes on the trend towards data intensive research and is a world leader in research and innovation. This future is achievable with comprehensive support for research data management at a national scale.
>
> [The Network] aims to lay the foundation for a library-based research data management network that will improve our national capacity for the management, preservation, and re-use of research data.
>
> The work of [The Network] is guided by the following underlying principles:
>
> - Data are a public good
> - Intelligent access: openness, with respect for privacy
> - Collaborative approaches: cost savings and sharing expertise
> - Inclusiveness: aim to serve all researchers and create a more level playing field
> - Commitment to standards and interoperability
> - International relationships: liaise internationally and ensure our work is in keeping with international practices
> - Respect for differences: flexibility to meet the needs of different regions, institutions, and disciplines
> - Open source: Tools will be contributed back to the community

[10] Canadian Association of Research Libraries, 2014

- Stewardship: a sense of responsibility for managing research data over the long term

As of March 31, 2015 the network was officially launched as “Portage” which will become a full-fledged service for research data management and build on the previous years’ work. It will have two components:

1. Network of Expertise

The Portage Network of Expertise will provide access to a comprehensive set of resources that direct users to the most up-to-date, relevant, and trusted sources about research data management. In addition, Portage is developing a national bilingual data management planning tool that will be launched in September 2015. The tool will be available to all researchers in Canada and provide support for planning, organizing, and managing their research data. Consulting services will also be provided, drawing on expertise from across the country. This service will target the following areas: privacy, security, and confidentiality; skills and training; data management plans; data access and dissemination; data discovery; data preservation; and data curation.

Portage DMP Builder

One immediate objective of the Portage Network of Expertise is to develop a national, bilingual data management planning service that will assist researchers in preparing data management plans (DMPs). For research data to be shared and re-used, they must be actively managed throughout their lifecycle, beginning at the time they are first envisioned. A DMP helps researchers decide how their data will be managed throughout the research cycle. In short, data management planning is a critical aspect of good research practice.

The DMP service allows the creation of national templates to meet specific requirements from funding bodies or customized templates for individual institutional use. This online web service will be available to all researchers in Canada and will guide researchers through a series of questions in composing their plan. It will be hosted by the University of Alberta Libraries and is based on an implementation of the DCC DMP Online tool. The Portage default template represents a generic national data stewardship plan and is not tied to any specific research funder. The tool is divided into seven sections, representing the most important areas in research data management.

This Portage DMP template is being developed and maintained by the DMP Experts Group (DMPEG), a group of data management experts from across the country. The template is currently in beta testing, and the bilingual Portage DMP Builder tool will be available for all researchers in Canada to use in September/October 2015.

2. National Preservation and Discovery System

Portage has also been working to connect the various infrastructure and service components needed for a national preservation and discovery platform. The ultimate aim is to enable all interested universities to participate, whether or not they have their own local infrastructure, by coordinating shared repositories and services under a cost model that recognizes varying institutional investments and needs.

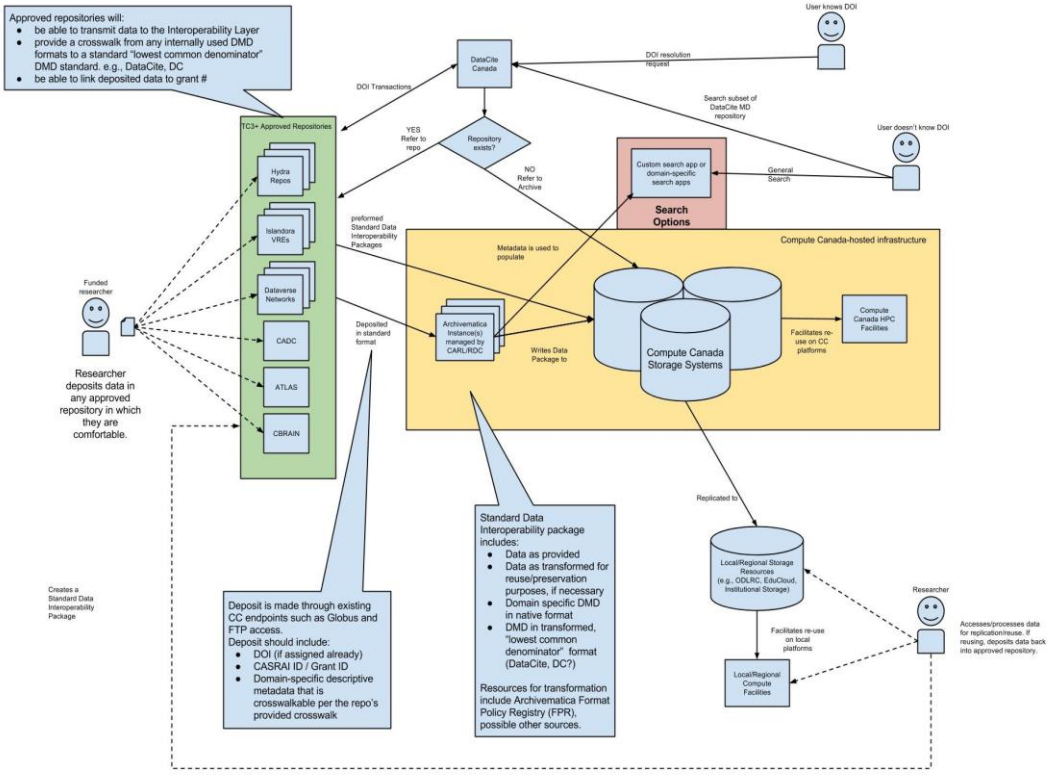
It is anticipated that there will be a three-year transition period for Portage to expand its activities and become fully operational. The ARC working group has been disbanded, but task-specific groups such as the Portage DMP Experts Group and an infrastructure working group continue to make progress on various parts of network. The intention is to continue in these directions, while establishing the governance structure and sustainability framework as Portage services are implemented over the next three years.

The proposed business model for Portage will likely be a mixed model of institutional membership fees, as well as external funding contributions. This will ensure long-term sustainability of the infrastructure and services, while also enabling the network to develop more quickly with targeted investments in priority areas. In addition to these funding sources, it is also expected that some in-kind contributions will continue to support the work of Portage during the transition to full-fledged operations. This framework will be refined in the coming months with input from key stakeholder communities.

The network will provide ingest and preservation services requiring a robust technical infrastructure consisting of software that supports the entire research data lifecycle (ingest, preservation, discovery, access, repurposing), data replication services, and networked data storage. This infrastructure will be highly distributed with local, regional and central nodes and will also be based on standards that ensure interoperability across nodes and data types.

In addition to ingest and preservation, a complementary set of services will support the discovery of data contained in data repositories across Canada. For this, metadata from repositories will be aggregated into an open registry through which discovery tools will be built to enable searching across data collections and repositories. Datasets from distributed repositories will be interoperable because they will map to common metadata standards allowing an appropriate level of integration. The following diagram represents the basic infrastructure model adopted by Portage. As pilot projects evolve, this diagram will go through a number of new iterations.

Diagram 1: Portage Infrastructure Diagram



Close collaboration with other partners and stakeholders is essential for the development and ongoing maintenance of this infrastructure. Both cash and in-kind resources will be needed to support one-time developments and initial implementation, as well as ongoing operations.

Operations

- Provide ingest and repository services for datasets. The network will ingest research datasets and the appropriate metadata to ensure that datasets are interoperable, preserved, and discoverable. In addition, it will support ingest processes from external data repositories into the preservation environment.

- Provide preservation services that maintain the data over the long-term, ensuring that they remain understandable and reusable into the future. Services will include: the addition of descriptive and administrative metadata; standardizing data formats so they can be managed as a "collection"; migrating data to new formats when their current formats become obsolete. Data replication in at least three separate locations across the network is also an important principle of the Portage model of preservation.

- Develop, implement, and maintain an aggregated discovery tool that provides access to the Canadian research data in collaboration with other partners.

- Develop metadata guidelines and procedures. Research data management must produce metadata that both make the data discoverable and make the data independently understandable (an OAIS principle). The latter type of metadata is much more detailed than the former and may include workflow documentation.

Service Model

- One-time development, upgrading, and/or expansion of the technical infrastructure will be funded through a variety of financial mechanisms in collaboration with the various constellations of participants in the network. Portage will seek to obtain support from sources such as government and granting agencies and also rely on participant fees for specific initiatives.

- The ongoing technical infrastructure will be maintained in a distributed but collective manner that appropriately reflects the local, regional and national configuration of the preservation network. This is seen as a way of achieving long-term sustainability. Local infrastructure and services will primarily be the responsibility of the local institution. Resource and cost sharing mechanisms, that also include major stakeholders and partners, will be developed to support multi-institutional, regional and/or national activities and various service nodes.

- Institutions lacking local resources and mechanisms/repositories for ingesting data into the network will have access to such services via other participants acting as host sites for others. Appropriate financial mechanisms will be developed as the needs and requirements of these institutions become clearer.

Conclusion

The network will gradually develop over the next several years and the governance structures and sustainability framework will evolve as the network matures. A Portage Steering Committee will be appointed by the CARL Board to represent the stakeholders involved in the Portage Network. It is envisaged that this committee will have majority representation from the major institutions and organizations contributing to the network, as well as of the broader stakeholder community (e.g. non-CARL members, researchers, research administrators, and funding agencies).

As CARL begins to set up the governance and operational structures for the Portage network, there is already significant momentum in several areas. A Portage Director has been seconded and will begin work in September 2015. Several task-specific groups such as the Portage DMP Experts Group and the preservation and discovery working group are continuing to make progress on various components of the network. The intention is to continue with these directions, while establishing the governance structure and business model.

Acknowledgments

Numerous people have contributed to the development and will remain engaged with the Portage Network, including members of the Project ARC Working Group and members of the DMP Experts and Technical Groups. Special thanks to Alan Darnell, Chuck Humphrey, Mark Leggott, Steve Marks, Dugan O'Neil, Brian Owen and Leanne Trimble for the detailed development of many of the network concepts.

References

Canadian Association of Research Libraries. 2014. *Project ARC Vision and Principles*, 2014. Available at: http://data-carl-abrc.ca/project-arc/project-arc-vision-and-principles/

Ferguson, Liz. 2014. “How and why researchers share data (and why they don’t)”. *Exchanges*, November 3, 2014. Available at: http://exchanges.wiley.com/blog/2014/11/03/how-and-why-researchers-share-data-and-why-they-dont/

Humphrey, C. 2012. **“**From National institution to National Infrastructure”. *Preserving Research Data in Canada: The Long Tale of Data***.** Available at: http://preservingresearchdataincanada.net/category/national-data-infrastructure/

Jahnke, Lori and Andrew Asher. 2012. “The Problem of Data”. *CLIR Reports*, pub 154. August 2012. Available at: http://www.clir.org/pubs/reports/pub154/problem-of-data

Christensen-Dalsgaard, Birte. 2012. *Ten recommendations for libraries to get started with research data management*. July 4, 2012. Available at: http://libereurope.eu/wp-content/uploads/The%20research%20data%20group%202012%20v7%20final.pdf

Mowers, Susan, Chuck Humphrey, and Carol Perry. 2013. *Summary Report: Survey of Researchers Needs and Practices Regarding Research Data Management in Canada*. August 1, 2013. Available at: http://gsg.uottawa.ca/data/open/aa-interim-survey-report/20130801-en.pdf

Science 11 February 2011: Vol. 331 no. 6018 pp. 692-693. Available at: DOI: 1126/science.331.6018.692

Shearer, Kathleen. 2015. *Comprehensive Brief on Research Data Management Policies.* Science.gc.ca, April 2015. Available at: http://www.science.gc.ca/default.asp?lang=En&n=1E116DB8-1