People, Platforms, and Content: The Importance of Libraries in the Emerging National DRI
Larry Alford, University Chief Librarian, University of Toronto Libraries
December 14, 2020

The University of Toronto Libraries (UTL) are pleased to submit the following brief to NDRIO on the current state of Research Data Infrastructure in Canada. Our views represent the perspective of a large academic library serving a diverse community of researchers with a wide range of needs in terms of data management, data publication, data preservation, and data discovery and reuse. Beyond the scale of these needs at UTL, however, our experiences, we hope, will resonate with libraries at other research-intensive universities across Canada, with which UTL works through a variety of regional and national collaborations such as the Health Sciences Information Consortium of Toronto, CARL Portage, DataCite Canada, ORCID Canada, the Ontario Council of University Libraries (OCUL), the Canadian Research Knowledge Network (CRKN), and most recently, Dataverse Canada, a new collaboration of 57 university libraries across Canada using a shared data management platform.

With significant technological infrastructure of its own, and professional and technical staff devoted to building collaborative services, UTL supports not only its local community but acts as a service provider to libraries across Canada for a number of data-related platforms, including the new Dataverse Canada platform. Our recommendations to NDRIO, then, will focus on the opportunities and challenges of leveraging library-based services and platforms to enhance Canada's emerging digital research ecosystem.

What are the tools, services and resources currently used?

People

Like many large and mid-sized academic libraries, UTL provides direct support to researchers through a network of expert data librarians and technicians with specialized skills in data discovery, data management, metadata standards, licensing, curation and preservation. They utilize their expertise to assist researchers with managing their own data for the purposes of research, sharing, and scholarly communications. They assist in research planning throughout the research lifecycle, including design, data management, data standards, and identifying storage and publication options for datasets. They also provide advice on appropriate metadata standards to support discovery, reuse and preservation.

These information professionals also work in a national context, through Portage Network Expert Groups and regionally through various OCUL communities. While UTL believes that effective and responsive local support for researchers is critical to the future success of a national DRI, we also strongly support the continued development of a national network of support, as envisioned by Portage, building capacity across Canada.

Platforms

Libraries in Ontario have a long history of collaboration in building discovery and management systems for our shared content. These include publication-focussed platforms, such as Scholars Portal Journals and Scholars Portal Books, as well as data-focussed platforms such as ODESI and the Scholars GeoPortal. ODESI gives researchers access to thousands of numeric datasets from providers such as Statistics Canada and many of Canada's major polling firms. The GeoPortal provides access to subscription-based geospatial data from DMTI Spatial, now licensed nationally through CRKN, and to major collections of imagery data through an arrangement with the Ontario Ministry of Natural Resources and Forestry. The GeoPortal is also used to deliver municipal geospatial resources and historical topographical maps to researchers across the province and, where content is available under Open Access licenses, to researchers across Canada. All of these platforms are hosted and managed on behalf of OCUL by Scholars Portal, a team of twenty-five librarians and technical professionals located at UTL and other libraries in Ontario. All of these platforms, moreover, rely on a distributed storage service called the

Ontario Library Research Cloud (OLRC). Funded by the Ontario Ministry of Colleges and Universities through its Productivity Improvement Fund (PIF), and supported now through member subscriptions, the OLRC is a petabyte-scale replicated storage service managed by Scholars Portal staff with nodes located at university data centers across the province.

Aggregating content on shared platforms addresses a number of critical issues that will need to be resolved in constructing a new national DRI. First, it aids researchers by simplifying the number of interfaces that they need to deal with when searching for data. Second, it promotes the use of high-quality metadata, which is critical for providing context and meaning to data resources, facilitating reuse, and allowing the data to be indexed in national and international registries. Third, it encourages the use of standard identifiers for citing data and identifying authorship, such as DataCite and ORCID, thereby integrating data publications into the larger body of published scholarly literature. Fourth, it makes preservation at scale possible. And fifth, it provides a framework for pooling scarce financial resources to build high-quality, sustainable data services while allowing partners to shape these services through participation in their shared governance.

With these same advantages in mind, the four academic library consortia in Canada – the Sous-comité des bibliothèques of the Bureau de coopération interuniversitaire (BCI), the Council of Atlantic University Libraries (CAUL-CBUA), the Council of Prairie and Pacific Libraries (COPPUL), and OCUL – have agreed this year to pool resources to develop a shared platform for managing, publishing, and ultimately preserving research data for their member libraries. Leveraging external funding made available through CANARIE, CARL Portage, and NDRIO (via CARL Portage), the Dataverse service hosted at UTL has been scaled to meet the needs of Canadian libraries and their researchers. In a short time, Dataverse Canada has grown to support more than 2,700 researchers and research groups across 57 academic institutions in Canada. These users have created 3,000 unique datasets comprising over 60,000 individual files. The collective investment of these libraries in Dataverse Canada represents an annual commitment of about $360,000. Ongoing funding from NDRIO, available in the current year in the amount of $250,000, would contribute not only towards sustainability of the service but would support its growth in key strategic areas: enhanced accessibility for researchers with visual disabilities; full bilingual support, including training, technical support and documentation; repository preservation and certification as a trusted digital repository; and closer integration with virtual research environments for support computation.

Content

One of the characteristics of current data-driven research is the adoption of new computational methods from the field of artificial intelligence (AI) as well as semantic technologies such as linked data. These tools in turn have changed our understanding what data is. UTL has amassed large collections of digital content in a variety of formats – text, video, audio, imagery – that can be understood not just as individual digital objects but as bodies of data that can be usefully analyzed as a corpus through the application of new computational approaches such as machine learning. Use of library collections for these purposes, though still not widespread, is becoming more common, and so the need to better integrate library content – to repurpose it as data resources – is both one of the great opportunities and great challenges in developing a national DRI. Projects developed in other jurisdictions, such as the Haith Trust Research Center or CADRE (from the Big Ten Academic Alliance), are being adopted at UTL and many other Canadian academic libraries to provide library-based computational support to researchers. These projects can serve as useful models for developing similar tools as part of a Canadian digital research ecosystem.

What challenges do researchers face when using our tools?

Academic libraries in Canada have made great progress in building new services to support data intensive research at their universities. But the challenges for researchers are still formidable. These can be considered in a few broad categories: challenges related to discovering suitable data resources; challenges related to

accessibility of those resources; and challenges related to data reuse, specifically, the integration of data into computational environments.

Discovery Issues

The landscape for data in Canada is fragmented. Research data, government produced data, and published commercial data are all housed in different repositories. Despite the collaborative work underway by CARL Portage and Compute Canada to build a Federated Research Data Repository (FRDR) for Canadian data, there is still no single place that a researcher can look to (1) discover what is available across all data sources or (2) assess the suitability of available data for a particular project. Success in this area requires researchers to know their destinations before they begin their searches. But one of the most common questions data librarians continue to hear is, "*where do I start?*" Another aspect of this fragmented landscape is the lack of connection between various data silos. It is a non-trivial task for a researcher who has discovered a dataset of interest to expand that search to include, for instance, related published datasets or published articles that may have referenced the dataset.

Accessibility Issues

This is a broad area of concern, touching not only on technical issues related to reuse, such as understanding licensing restrictions or privacy requirements, but also on a range of equity issues from complying with accessibility legislation to ensuring that DRI tools can be used fully by researchers working in either of Canada's official languages. UTL and the Université de Montréal partnered recently to develop a French-language version of Dataverse as part of the Dataverse Canada collaboration. To achieve this at a high level required a significant investment in time and resources from both parties. But full support for bilingualism is fundamentally important, not only in the Canadian context but also in light of international research collaborations that Canadian scholars are increasingly engaged in.

Integration Issues

Data is often tightly bound to the code or software that was used to create it. The decision to use even a well-documented collection of data can be difficult if the accompanying code cannot be analyzed or re-executed to verify the reproducibility of the original research results. This is even more challenging in a long-term perspective because of rapid changes in technology, especially in software platforms. Active research in digital preservation includes identifying platform-independent approaches to the preservation of code and data. In this way, digital preservation can be thought of not as external to the research process but as a critical foundation for data reusability and so an important element of any DRI. As described in detail in another white paper submitted to NDRIO by the Portage Preservation Expert Group (PEG), active preservation policies and strategies must be developed to ensure the integrity and accessibility of data into the future.

What would a cohesive DRI ecosystem look like?

In a few points below, we sketch out what we think a cohesive ecosystem might include and how NDRIO might foster those elements, confining our comments to aspects of the DRI ecosystem related to data discovery and data management and the support of researchers in these tasks.

First, we believe that it is critical for the success of any new DRI to take advantage of the work that is already being done in academic libraries to support researchers. Building on existing relationships between librarians and researchers rather than trying to recreate that kind of support at a national level will lower barriers to the adoption of DRI tools and services and respond better to the unique needs of each researcher.

Second, we believe it is also important for NDRIO to continue to support the development of networks of expertise, as exemplified by the work of CARL Portage, to establish best practices in RDM services across Canada. The work of Portage has been critical in advancing the state of data management in libraries over the last few years.

Third, we believe that the collaborative model on which Dataverse Canada is built is an exemplary model for developing other kinds of shared services in the new DRI. We hope that NDRIO will continue to fund services like Dataverse Canada that leverage community investments in sustainability and allow for effective aggregation and scaling while still providing for local involvement in governance as well as local branding and integration with local RDM support services.

Fourth, we believe that there are significant advantages to consolidating the number of individual data repositories across the country and that NDRIO should encourage that process in future funding opportunities. A less fragmented ecosystem will allow each repository to scale more effectively and allow the whole community to adopt practices that will encourage interoperability.

Fifth, we support initiatives such as DataCite Canada and ORCID Canada and encourage NDRIO to continue to support both as a way to ensure that data resources are fully integrated into the broader scholarly publishing ecosystem.

Sixth, with our colleagues in BCI, we support adoption as a first principle the commitment to bilingual support for all aspects of the new DRI, including not just interface design but also training, support, and the ability to participate fully in governance. All systems funded by NDRIO should meet this basic standard.

Seventh, we also encourage NDRIO to target its funding to support projects and services that conform to the highest standards for accessibility, regardless of individual provincial mandates. This is a critical equity issue and one that will allow us to reflect our values internationally through a national DRI.

Eighth, we encourage NDRIO to take a broad view when thinking about what data is and what data-intensive research encompasses. The challenges of managing large volumes of observational data, such as those generated by scientific instruments, are difficult but well understood. We would encourage NDRIO, however, not to overlook the value of Canada's digital library collections, including web archives, as important data sources for the DRI. We would further encourage NDRIO to support research and pilot projects to better understand how these resources can become integral parts of the DRI, through building seamless connections between this kind of data and computational services of the DRI.

Ninth, we encourage NDRIO to consider published data sources as important aspects of the DRI. These include, as we have noted, important collections of numeric and geospatial data published by various agencies and corporations. Over twenty years ago, the Canadian Foundation for Innovation (CFI) supported academic libraries in a funding request to lower the barriers to access to electronic journals through the establishment of a national site licensing program, giving rise to the Canadian National Site Licensing Programme (CNSLP) and later CRKN. We encourage NDRIO to consider supporting a similar initiative to license important data resources at a national level to help level the playing field for researchers across the country in terms of access to core research data and to reduce costs for the system as a whole.

Tenth, and last, we encourage NDRIO to build on the work of Portage FRDR and explore the development of a national data catalogue and data archive, including a common platform for indexing, storing and preserving research data resources for the long term, as a way to reduce barriers to access and reuse, as described elsewhere in this document.

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