

Capitalizing on Big Data

CARL's response to the consultation document *Capitalizing on Big Data: Toward a Policy Framework for Advancing Digital Scholarship in Canada*

December 12th 2013



**Canadian Association
of Research Libraries**

**Association des bibliothèques
de recherche du Canada**

Who we are

The Canadian Association of Research Libraries (CARL) represents the country's twenty-nine largest university libraries. Enhancing research and higher education are at the heart of our mission. CARL promotes effective and sustainable scholarly communication, and public policy that enables broad access to scholarly information.

Overview

CARL thanks the Social Sciences and Humanities Research Council (SSHRC), the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council (NSERC) and the Canada Foundation for Innovation (CFI), in collaboration with Genome Canada —the “TC3+” — for consulting the broader stakeholder community on a potential policy around the stewardship of research data through their paper *Capitalizing on Big Data: Toward a Policy Framework for Advancing Digital Scholarship in Canada*.

CARL members are deeply engaged in the issues of stewardship of research data in our own association, our institutions and regional library consortia, and through very active roles in Research Data Canada (RDC) and the Leadership Council for Digital Infrastructure (LCDI). CARL has had the opportunity to contribute to RDC and LCDI responses to the consultation document, and we are aware that other related organizations are providing submissions as well. In this submission from CARL we address the issues through the particular lens of research libraries in the national context.

CARL members approach these matters from the perspective of the researchers we work with on a daily basis. Those researchers want to acquire and use data in new computational ways, they want to plan for the effective management of data generated in new research projects, and they want to present and preserve their data for future use. Through this researcher lens, CARL is pleased to provide the following recommendations, which are explained in more detail below, as a response to the consultation paper:

1. Work with key stakeholder communities to raise awareness of the importance of research data management
2. Provide guidelines for specific elements in data stewardship plans
3. Leverage the stewardship strengths of research libraries to develop research data management services
4. Establish preservation certification processes to operationalize suitable data repositories
5. Facilitate coordination of and build on existing initiatives
6. Develop appropriate funding mechanisms that support the management of research data across the data lifecycle
7. Invest in training and skills development across the research data management lifecycle

CARL commends the TC3+ on their intention to bring a strong focus to the stewardship of Canada's research data. As you know, research data represent a valuable publicly-funded asset that, if managed and shared appropriately, can be further exploited and lead to new discoveries and greater innovation. In addition, research results cannot be validated or reproduced unless the research data on which they are based are accessible and understandable to others. The gaps in Canada's capacity to steward research data have already been well documented. That being said, Canada also has some strengths that can be brought to this endeavour, including various discipline-based research data management activities, as well as the collaborative activities of research libraries in managing a range of essential research products. Further actions by TC3+ will help to bring Canada into line with other developed countries that have already begun to put the infrastructure and policies into place to support the wide-spread sharing and preservation of research data.

Recommendations

Establishing a Culture of Stewardship

Recommendation 1: Work with key stakeholder communities to raise awareness of the importance of research data management

Research data management (RDM) in the university context involves a range of stakeholders including administrators, researchers, the libraries and IT departments, all of which have important roles to play. Vice-Presidents Research, assisted by their research support offices, will provide campus leadership along with University Librarians and Chief Information Officers and it is important for the TC3+ to engage with these groups. In addition, it will be critical to raise awareness of the issues in the research community by involving scientific and scholarly societies and faculty associations. The T3+ plan for developing capacity should include a program of support for on-campus advocacy and education that will assist institutions in building awareness and skills. For example, such a program could include teams of experts available to advise institutions in various elements of RDM governance and support.

Recommendation 2: Provide guidelines for specific elements in data stewardship plans

Requiring data stewardship plans as part of funding applications is a good first step that will provide an incentive for researchers to consider how the data they produce will be managed. Given the varying levels of awareness of the value of long-term RDM among researchers and the fragmented state of support across the country, initial requirements for data stewardship plans may need to be relatively broad and flexible. However, requirements should be strengthened over time as both community understanding and acceptance increases and as capacities and infrastructure become more fully developed in Canada. The experiences of funders elsewhere have shown that any policy requirements for RDM plans should identify the specific elements to be included in the plan. The research library community has already begun to work with researchers on RDM plans and CARL, with Research Data Canada, would be pleased to provide specific recommendations on desirable RDM plan elements.

Recommendation 3: Leverage the stewardship strengths of research libraries to develop research data management services

Research libraries have a long tradition of stewardship. Our core values include preservation, access to information, and protection of privacy. Our librarians are skilled not only in the

practices of sound information management, but also in teaching and research. Some research libraries, especially those of the larger universities, already have librarians with data management expertise who can offer assistance to researchers on data management planning. CARL itself has provided some RDM training to library staff, offering a four-day institute in 2013, and will likely offer a second, more specific or advanced institute in 2014. However, such programming could be expanded if greater funding for this were made available. In addition, there still remain significant gaps in services and the TC3+ should work together with research libraries to build capacity in this area. (See also Recommendation 7.)

Recommendation 4: Establish preservation certification processes to operationalize suitable data repositories

One of the elements of the data stewardship plan proposed by the TC3+ includes “suitable data repositories.” It will be important to establish a common understanding of data repositories, to foster an understanding that these are more than simply storage space for computational analysis or other short-term storage. In particular, digital preservation and data curation must be integral to data stewardship. Research libraries have considerable expertise and experience to offer in this regard (see Recommendation 5). TC3+ and CARL, with Research Data Canada, should work together to develop appropriate criteria and a certification process for repositories. The Research Data Alliance (RDA) has launched a group to develop global data repository certification criteria that will bring together various regional certification initiatives. This work could provide a suitable framework for determining such parameters here in Canada.

Coordination of Stakeholder Engagement

Recommendation 5: Facilitate coordination of and build on existing initiatives

The major components of a RDM infrastructure already exist to some degree, but coordination is necessary and resources will be needed in order to “knit together” the extant components and to address remaining gaps. While we concur that coordination will always be a good thing, we would suggest that at this point, duplication or overlapping interests among the providers of digital research infrastructure (DRI) services is not the issue; rather, the well-known infrastructure gaps (at least in terms of RDM) are a much greater concern.

In particular, TC3+ should engage with both the Leadership Council for Digital Infrastructure (LCDI) and Research Data Canada (RDC) as they provide a breadth and level of representation with numerous stakeholders across the country. RDC has a particular focus and strength in the RDM aspect of DRI. In terms of long-term data storage, Compute Canada may be able to devote several petabytes of disk space, which could potentially be structured to accommodate an active preservation regime.

Preservation involves not only securing appropriate storage space but also the long-term management of data. The research library community has been working at the forefront of digital preservation in Canada. The Ontario Council of University Libraries’ (OCUL) Scholars Portal, hosted at the University of Toronto, developed the first certified Trustworthy Digital Repository in Canada. This certification was issued under the ISO 16363 standard in February 2012 following an audit of the Scholars Portal Journals repository. This repository has been in place for a decade and contains more than 37 million journal articles. OCUL is now developing the Ontario Digital Library Research Cloud, a 1.2 PB storage cloud with nodes distributed across the

province that will integrate with repository applications. Members of the Council of Prairie and Pacific University Libraries have experience with the LOCKSS Program, based at Stanford University Libraries, which provides a digital preservation environment for digital content. A LOCKSS network for Canadian government information was initiated by a group of eleven academic libraries in the fall of 2012. As well, several university libraries (Alberta, British Columbia and Simon Fraser) in Western Canada are collaborating in the development of a digital preservation environment building on the open source, standards-based system Archivematica.

We should note as well that there are a number of other library-based initiatives that are of value for RDM, which include DataCite, the Canadian node of which is provided by the National Research Council's Canada Institute for Scientific and Technical Information (NRC-CISTI), which provides unique DOIs (digital object identifiers) for archived datasets, and the Canadian Polar Data Network (CPDN), which has developed a successful regime for distributed data storage and active data preservation. In addition, several individual university libraries are archiving datasets in local digital repositories as well as offering services that address other important aspects of RDM: education, training, advice and assistance to researchers.

As a means of coordinating the various library-based initiatives and broadening support for RDM at institutions and across regions, CARL is sponsoring the development of a "network" of such initiatives, the precise form and goals of which are currently in discussion. The role of the regional university library consortia and of the Canadian Research Knowledge Network (CRKN) will be an important consideration in this approach since their members collectively represent all of the public university libraries across Canada. CRKN, which has a successful track record of creating a national platform, may be an appropriate vehicle for administrative coordination and distribution of funds to regional nodes of a network.

While these library-based initiatives will contribute to meeting some RDM needs for preservation storage, training and services across Canada, it is not anticipated that they will address more than a small part of the whole range of needs; the university libraries, directly or by way of CARL, will be pleased to collaborate and partner with other organizations that can contribute to the national RDM effort.

Developing Capacity and Future Funding Parameters

Recommendation 6: Develop appropriate funding mechanisms that support the management of research data across the data lifecycle

In order to achieve a "sustained operation" of "world-class centres specializing in data management", the parameters for the funding of national-scale DRI will need to be re-assessed. The TC3+'s convening power will be an important incentive for the various stakeholders to work together including engaging the provinces in a pan-Canadian approach to digital infrastructure. With regard to funding national research platforms, CFI has been successful in the past in funding pan-Canadian infrastructure with contributions from the provinces. A good example is the funding of the Canadian National Site-Licensing Program (CNSLP), which became the Canadian Research Knowledge Network (CRKN). A similar approach to funding should be considered for research data management.

Recommendation 7: Invest in training and skills development across the research data management lifecycle

A key capacity challenge is the development of skills and knowledge across stakeholder communities. RDM is complex and requires skilled personnel for all parts of the research lifecycle, which will include researchers, data scientists and librarians. The TC3+ may want to consider investing in training programs for RDM across stakeholder communities. In this context, the Digital Curation Centre (DCC) in the UK could be an interesting model for Canada. The DCC is a centre of expertise in digital information curation with a focus on building capacity, capability and skills for RDM across the UK's higher education research community. Such a model would serve to support the kinds of training program CARL has initiated and those being developed in some institutions, and it would address gaps being considered by Research Data Canada's Education and Training Subcommittee.

Conclusion

In conclusion, CARL would be pleased to participate in discussions on these various proposals as they may find stakeholder acceptance and be developed in greater detail. In general, we are very much encouraged by the proposals put forth in the consultation paper as they indicate a resolve on the part of the TC3+ to make progress on both the coordination of the development and funding of DRI and building and sustaining of a Canadian approach to the stewardship of research data. Again, we thank the TC3+ for their consulting of the community and we look forward to moving forward as a nation towards greater and more effective RDM.

Canadian Association of Research Libraries

Tel. 613.482.9344 203-309 Cooper St.
info@carl-abrc.ca Ottawa, ON
www.carl-abrc.ca K2P 0G5

