Data Management Roadmap – 2019-2024

For Submission to Innovation, Science and Economic Development (ISED)

Authored by:

David Castle (University of Victoria/RDC), Laura Gerlitz (RDC), Susan Haigh (CARL), Mark Leggott (RDC/CANARIE), Lori MacMullen (CUCCIO/LCDRI), Jeff Moon (CARL/Portage), Benoit Pirenne (University of Victoria-ONC/WDS-ITO), Robbin Tourangeau (Compute Canada), Bo Wandschneider (University of Toronto/CUCCIO)

March 29, 2019

NOTE: This document was originally submitted to ISED to facilitate the process of creating a new national DRI organization. This version has been created as background material for the 2020 NDSF Summit. It includes the following sections from the original document: Vision, Key Functions, and Activities.

Vision

This document provides recommendations for the framework for a 5-year roadmap for funding priorities in the Canadian data management (DM) ecosystem. A great deal of work has preceded this[1], and a number of efforts continue to refine the discussion, so we will not duplicate that material here. We will, however, repeat the Vision statement from the LCDRI submission to ISED:

> *An innovative and coordinated research data management community, providing responsive services and resources that support Canadian researchers in advancing the research that is critical to building and sustaining Canada's economic and social prosperity.*[2]

This Vision was supported by three primary Goals[3]:

> *A. Build innovative services and resources that are distributed across research institutions, nationally coordinated, internationally recognized, and sustainable, while being responsive across the full spectrum of researcher needs and disciplines. These services and resources should respect researcher-focused, discipline-specific, national, and institutional data stewardship policies, and be based on best-practice standards and protocols.*
>
> *B. Advance and adopt DM processes and procedures that are informed by researcher, institution, and discipline-specific needs, to improve the overall quality of research data and to advance best practices. This will require flexible and adaptive tools and platforms supporting data planning, creation, curation, deposit, access, discovery, and reuse.*
>
> *C. Establish a community of practice that is supported by a distributed network of specialists who can provide expert advice, support, and training in DM best practices to researchers.*

These Goals are also closely aligned with the Kanata Declaration from the 2019 National Data Services Framework (NDSF) Summit, which reflects a broad stakeholder perspective. In addition, **the elements of this Roadmap directly support the soon to be finalized Tri-Agency Data Management Policy**, which together with other provincial, publisher, and national and international funder mandates, will have a substantial impact on the prioritization and provision of DM supports for researchers.

Lastly, these goals have been reflected in the ongoing and foundational work of university libraries, researchers in their respective disciplines, and other partners. Recognizing the critical importance of their collective leadership in furthering DM, university libraries have worked collaboratively to establish the CARL Portage Network. Through this Network, they have tested and demonstrated the essential role that libraries play in supporting and encouraging researcher uptake of DM platforms and services. In addition, researchers have made critical contributions to improving DM practices through targeted work in their own discipline-specific areas.

This proposal provides a roadmap for how funds can be most strategically invested in addressing these critical goals on behalf of Canada's research community.

---

[1] LCDRI Data Management Position Paper for ISED, Aug 2017; Consolidated Response to Questions from ISED on the LCDRI DM Position Paper, Fall 2017; LCDRI Coordination Position Paper for ISED, Jan 2018; LCDRI Backgrounder, Jan 2018;

[2] LCDRI Data Management Position Paper for ISED, Aug 2017, p. 26

[3] LCDRI Data Management Position Paper for ISED, Aug 2017, p. 27

Key Functions

The Roadmap uses three core Activities to reflect priorities and approaches to act on the Vision and Goals. It is important to highlight that these Activities intersect at all levels and with all of the key DM functions that are listed below, and described in detail in the LCDRI DM Document. In some cases all 5 Functions are addressed in the context of a single activity, while others are more focused.

The Functions required for successful DM must be implemented and owned within and across many communities and types of organizations (including institutions and funders but also software suppliers and publishers) both domestically and internationally. Relying solely on domain and/or regional approaches to DM leaves gaps that will prevent achievement of an effective and sustainable ecosystem. The facilitation and coordination of so many diverse and independent stakeholders is most effectively accomplished through a federally funded, neutrally-governed, and collaborative not-for-profit organization with designated mandates for accomplishing this goal.

National coordination is essential to enabling coherent, efficient, and effective RDM in Canada, and to prevent barriers to data sharing. Specifically it:

- ensures that all researchers and administrators across Canada have access to RDM services and platforms, regardless of their discipline, geographical location, or the size of their institution;
- enables collaborative and efficient development of the policies, standards, protocols, processes, and procedures essential to ensuring that researchers can find, access, and reuse research data generated in Canada and elsewhere in the world;
- ensures that tools and platforms, such as data repositories and archival storage, are interoperable across Canada and internationally, facilitating access and sharing, rather than creating barriers;
- facilitates consistency of practice and approach to RDM across Canada by supporting strong communities of practice, access to networks of experts, and shared training;
- leverages expertise and shared investment across research organizations, government, and other funders, increasing quality, impact, and financial efficiency; and
- builds a collaborative RDM culture that engages researchers and research administrators across Canada, creating awareness of and support for RDM and ensuring that rather than stepping on each other's feet, they are standing on each other's shoulders to advance RDM in Canada.

The statements below highlight a need for and the importance of this federal role in the context of each Function. The descriptions that follow in Activities suggest levels of federal support that would be appropriate.

1. ***Policies***
   Development of policy is relevant at all levels of the ecosystem, and the regional/institutional activity here is largely in response to policy developed at the federal level: coordination of policy development amongst the various stakeholders needs to receive federal support to roll out to the institutional level and be effective.
2. ***Standards and Protocols***
   Development of, and adherence to, standards and protocols is particularly relevant at the international and domain-specific levels: coordination of activities at all levels in response to best practices, and how they can be supported with sustainable funding, can only be facilitated with federal support.
3. ***Processes and Procedures***
   Development of, and adherence to, processes and procedures is particularly relevant at the institutional and researcher level: ensuring consistency in the deployment of research infrastructure, and the accessibility to research outputs, is a key responsibility of a coordinated federal response.
4. ***Leadership, Advice, Support and Training***
   Provision of leadership in DM, the deployment of effective training to all researchers and supporting units, and the delivery of appropriate services is relevant at all levels of the ecosystem: a federally

supported and coordinated response is the most effective and efficient way to direct funds through a national network, and ultimately to the researcher.

5. ***Tools, Platforms, and Storage***
   Coordinated and sustainable research infrastructure is required to ensure that research activity is supported at all stages, and for a sustainable period of time: federal funding is key to the provision of the "glue" that ensures an interoperable and cost-effective national framework.

Activities

1. Coordination

The ISED DRI Discussion Paper referred to a future-state organizational entity, referred to here as the "DRI Organization" (henceforth "DRIO"), that would facilitate the development of details for efforts related to Advanced Research Computing (ARC/storage), DM, and research software (RS). We will use that terminology here, recognizing that this is subject to change, particularly as the organizational structures evolve over the next 1-2 years.

We also recognize that while DRIO is in the initial stages (i.e. 2019-2021), identified funding priorities may be deployed and managed by existing organizations, but these services would eventually merge into the DM program under DRIO.

Coordinated oversight for DM, RS, and ARC/storage will realize significant new opportunities for synergy among all these programs, with the goal of delivering a simpler and more cohesive service to researchers. This is alluded to below in the context of storage, but can equally apply to the synergies between DM and RS, which is an opportunity already being implemented in the current CANARIE programs. This Roadmap assumes close collaboration between DM, ARC and RS, but does not consider the details of RS funding, activity, or oversight, which is assumed to be a separate program under DRIO, albeit with close ties to DM activities.

There are also opportunities for greater coordination of programs and services among CANARIE, DRIO, CFI and the Tri-Agencies, to name a few. The Capacity component has clear synergies with the priorities of the Tri-Agencies, and Infrastructure Development and Delivery with CFI and the NREN, for example. Similarly, various regional organizations are developing services that support DM, RS and ARC/storage, and exploring collaborative efforts in these areas should be a key priority of DRIO. Providing funding at the national level to create and support these synergies is key to realizing the potential.

Focus: Enhancement and expansion of national DM coordination efforts, including merging Portage and RDC facilitation efforts, directing resources to priorities in the first year, defining details for subsequent years, and driving better outcomes.

Outcomes: A single national DM unit within DRIO that facilitates community agreement around best practices, and the effective distribution of DRI funds to best meet the needs of researchers.

Federal Remit: In order to effectively intersect with, and deliver on the promise of better research, as well as the federal priorities of open government (and especially open science), there needs to be a strong national voice that transcends national and international borders, with a focus on: leadership, national vision and coordination; standards and best practices; program oversight; and international engagement. It is recommended that Coordination activities be funded with 100% federal contribution.

Transitional Considerations

- With the expectation that RDC will continue to be funded past the current Contribution Agreement framework with CANARIE, and until DRIO assumes this role, coordination with other DM stakeholders will continue during this period.
- The Portage Network, like RDC, has a national role for DM, as reflected in its current suite of researcher-focused service offerings. The specific funding priorities identified by Portage for FY 2019-20 would be met via the activities identified below, especially 2.1.1-2.1.5, 2.2, 2.3, and 3.1.1-3.1.5, and need to be considered in the transition period.
- In order to ensure the merging of appropriate elements of these two organizations into a single national entity over the next two years, the two organizations would start with strengthening collaboration on their national facilitation efforts, especially those defined by the work of stakeholders (e.g. Advisory/Standing Committees and Working Groups).

2. Infrastructure: Tools and Platforms

There is a clear need and solid return on investment to develop and deliver national Infrastructure supporting DM. Various efforts over the last few years have positioned the DM community to act quickly on the needs of researchers: the priorities below represent programs that would support these needs. However, we acknowledge that DRIO, once established, will be responsible for monitoring and addressing these needs through a variety of programs and partnerships on an ongoing basis.

Focus: Provide support for the development, sustainability, and use of national DM infrastructure and services that meet the needs of all researchers.

Outcomes: Through effective support for researchers across Canada, there is an increase in access to, and preservation of, FAIR[4] data in all disciplines.

Federal Remit: Federal funding will address the development and ongoing operation of a cohesive and interoperable suite of national tools, platforms and repositories, that are available to all Canadian researchers. The level of federal funding will vary based on the role of the components in the national platform.

2.1. Support for National Tools and Platforms

There are a range of systems, platforms, tools and services that are integral to effective data management at scale, such as data platforms, repositories, and middleware. There are a number of examples that can be highlighted as essential in the national context.

2.1.1. *A bilingual, customizable data management planning tool* - One of the 3 pillars of the Tri-Agency Data Management Policy is the recommendation (and in some cases requirement) for researchers to create data management plans. The Portage DMP Assistant is the primary DMP application in use in Canada today, and as a fully bilingual platform, provides all Canadian researchers with a tool to create and maintain DMPs.

2.1.2. *A researcher dashboard* - This is essentially a “one-stop-shop” for researchers to discover DM services in their discipline, and to help them respond to journal and funder requirements.

2.1.3. *National, multi-disciplinary repository options* - In order to ensure that all researchers have options for data deposit and preservation, the provision of one or more national

[4] Findable, Accessible, Interoperable, Reusable.

multi-disciplinary repositories that can address the long-tail of research activity is critical. Current national multidisciplinary repositories include the Dataverse North Network and the Federated Research Data Repository (FRDR). These platforms would also be expected to achieve an international, community-endorsed level of certification (e.g. CoreTrustSeal).

2.1.4. *Nationally Coordinated Preservation Services* - The community needs to develop, and implement a nationally coordinated, sustainable network of preservation service providers, taking into account the roles of storage and compute, curatorial oversight, and the need for appropriate funding.

2.1.5. *A bilingual national discovery layer* - Leveraging metadata from the diverse array of research data repositories in Canada, a national, bilingual, data discovery service is essential to facilitate discovery of and access to research data. The discovery layer of the Portage/Compute Canada Federated Research Data Repository (FRDR) has started to fulfill this role.

2.1.6. *An endorsed researcher ID system* - The ORCID researcher ID system, and the associated ORCID CA Consortium provide a similar service to Digital Object Identifiers (DOIs), but for researchers. The ORCID Persistent IDentifier (PID) ecosystem also intersects with other PID frameworks, some of which are well established (e.g. DOIs) and others which are emerging (e.g. Research Activity IDs, RaIDs). Given the evolving nature of the PID ecosystem, it is critical to complement institutional investment through the ORCID-CA consortium with federally-supported services that ensure the accessibility of key standards, and also the middleware and services that may be developed to facilitate adoption.

2.1.7. *An endorsed dataset ID system* - DataCite Canada, and related middleware and services, is currently supported by the National Research Council (NRC), and provides a robust ID system for data. Virtually every data repository needs to have support for minting data IDs, and Datacite DOIs are the most common. The NRC and other national players (including the DRIO), need to determine what is needed for the long-term sustainability of this infrastructure, and how it can be enhanced to facilitate broader adoption.

2.2. Support for a Sustainable Network of Data Repositories

In addition to the national multi-disciplinary repository options (as noted in 2.1.3), there are domain data repositories in the Canadian ecosystem that meet the needs of researchers in the national context, typically for all researchers in a specific domain. Together, these form a Canadian Network of Data Repositories. The priority in the next 2 years is to identify these domain repositories, and facilitate their adherence to a minimum standard of certification as developed and endorsed by the community and DRIO. Canada is the current host of the World Data System - International Technology Office (WDS-ITO) which provides support for repository certification, WDS membership and overall DM training, and would therefore be a natural partner in this context.

Some of the domain repositories in this network have, or are developing, business models that bring in revenue, while others have institutional, regional, or project-based funding. Given that some repositories are sustained through a combination of revenue models, the goal with the DRI programs outlined here is to ensure these repositories that meet minimal criteria have baseline sustainable funding to continue operations. Sustainable funding for domain repositories is critical to ensuring that researchers can respond to funder mandates and the open science commitment, and that Canadian research data can be stewarded for the long-term. These repositories, as well as others, would be harvested in a national discovery layer to facilitate access to Canadian research outputs.

The level of federal funding needed will vary, and will be determined based on criteria to be more clearly defined as part of DRIO program execution, but should include: endorsement by the community of practice (ie. researchers and publishers); accessible to all Canadian researchers in that domain; adherence

to an international repository certification; support for best practice domain standards and file types; a robust governance model; and integration with the NDSF. As the NDSF is more clearly defined, including the availability of shared repository storage and compute services, it will be feasible to support specific operational aspects of the domain repositories. An example would be the repository and preservation storage needs, which could be met by the DRIO, provided they meet the minimum criteria. These efforts would not only improve stewardship and access to research data, but would also achieve economies of scale that would help build a sustainable network of data repositories.

Once domain repositories have been identified and have begun the process of certification, additional funding should be available to allow these repositories to update and enhance their software infrastructure to ensure they meet the certification standards and the ability to participate in the network of data repositories. This would typically take the form of Highly Qualified Personnel that have expertise in software development, and especially key international standards. Where appropriate, domain repositories would make use of the tools and platforms noted above in section 2.1 to facilitate interoperability at the national and international levels. It is important to highlight the potential opportunity for synergies with RS programs (which is not addressed in this document) to include funding for enhancements in this area.

2.3. Provision of Storage

From the perspective of data management, there are three forms of storage required. They differ by their life expectancy and the level of curation required. Active storage addresses needs during the research process itself, when data are being collected, modified, or otherwise analyzed -- active storage needs vary depending on the length of the research project and the amount of data being created. Curatorial decisions are then made regarding what data to deposit into Repository storage which supports future discovery and appropriate access, in the medium term (1-2 decades). Further curatorial decisions will be made to migrate selected data into long-term Preservation storage that will guarantee access and readability essentially forever. Like the availability of high-speed networking, the provision of appropriate storage (Active, Repository, Preservation) is key to a functional Canadian DM ecosystem. Support and development is needed for each component of this storage continuum to ensure researchers have access to appropriate storage along the entire research life cycle. Active storage will require ongoing discussions with the HPC/compute/library community. Repository storage needs, both domain-specific and multidisciplinary (see 2.1.3), are of immediate concern to ensure researchers can respond to Tri-Agency and journal data deposit requirements. Preservation storage, and the role of institutional, regional, and national long-term storage solutions, will need to be developed and deployed, again requiring ongoing discussions with the HPC/compute/library community.

3. Capacity: Fostering a Network of Experts and Community of Practice

The need to effect cultural change to realize a broader adoption and use of good data management practices is a common thread in the DM community, as is the importance of HQP in that culture change. Within the research community, and within the library and other communities that support the research DM ecosystem, there is a need to strengthen capacity.

Focus: Increasing DM capacity in the research community through a national network of DM expertise.

Outcomes: Widespread adoption of DM best practices, an increase in the available training opportunities for researchers in all disciplines, participation by all publicly funded research organizations in training initiatives, and an increase in HQP in the DRI community.

Federal Remit: Fostering a Network of Experts and Community of Practice to ensure that the federal investment in Coordination and Tools and Platforms is reflected by a change in the culture of research,

that researchers receive coordinated support throughout the lifecycle, and that domain best practices are fostered in all research disciplines.

3.1. Development and Support of a Network of Experts

The Network of Experts includes a core team of full-time staff that facilitate all aspects of the DM ecosystem, and builds on the proof of concept work of the Portage Network of Experts, and similar networks in the domains.

This would include working with stakeholders in the form of committees and working groups, but also in overseeing the development and deployment of relevant programs under the DRIO umbrella. Key areas of focus would be: repositories, curation, preservation, and training. This team would provide support to institutional partners who provide the front-line interface with researchers. The priority for this team in the next 1-2 years is to work with the community to further define needs, and identify intersections with funded programs.

To be most effective, the Network of Experts would include support for a distributed model and regular meetings of the community, and would also have access to capacity-building funds to support efforts at institutional, regional, national, and international levels.

3.2. Development and Provision of Training and Capacity Building

With support from DRIO, the Network of Experts would facilitate the programs identified above, as well as the capacity building efforts listed below. As with other core costs, the costs for the coordination of training and appropriate national meetings would be reflected in the Coordination mandate. This would include working with the community to define a National Training Framework that could be deployed at any research institution in the country. The assumption is that the program would have a train-the-trainer approach. The Network of Experts would be the interface for local support leads, who would be the primary point of contact with researchers.

In order for Canada to align its DM practices with accepted international standards and best practices, support for membership, sponsorship, and participation in international DM organizations (e.g. Research Data Alliance (RDA), RDA North America, World Data Systems, International Technology Office (WDS-ITO), CODATA, CASRAI) would be included throughout the 5-year mandate.

3.3. Community Leadership (Governance)

Key to a successful strategy over the full mandate, is ensuring that the voice of researchers and DM service providers is an integral part of the conversation. It is recognized that the very nature of DM services depends on a strong, distributed, collaborative network, including academic libraries, domain communities and others. The success of the DM program will depend on a governance structure that appropriately reflects stakeholders' contributions and interests.

We recommend two key stakeholder bodies for the DM program of DRIO: an Advisory Committee with a researcher focus, and a Steering Committee. There would be additional representation from stakeholder groups on appropriate sub-committees and working groups.

The key function of the Advisory Committee is to ensure that the needs of researchers are being met, and that the evolution of DM supports and programming is in tune with emerging trends in the domain communities of practice.

The key role of the Steering Committee is to provide: direction and overall vision and strategy for priorities for development needs, service offerings, and expenditures to the DRIO program on behalf of the stakeholder community; effective and regular communication to stakeholders; promotion and support of its activities; a forum for the exchange of information, knowledge and experience of DM; effective leadership and participation in DM activities internationally; advice and strategic support to stakeholders.

We also envisage strong representation of DM on the DRIO Board, reflecting the diversity of the community, and recognizing the leadership role of the library community, and especially its role as a professional DM community of practice and provider of infrastructure and services.