CANARIE 2018 Research Data Management (RDM) Community Consultation: A Summary

July 2018

Background

In January 2018 CANARIE launched a Consultation to determine community views on key gaps in the digital research infrastructure (DRI) ecosystem supporting research data management. CANARIE's Research Data Management Advisory Committee (RDM AC) reviewed the submissions and made recommendations to CANARIE for its first Research Data Management (RDM) Funding Call, which launched in May, 2018.

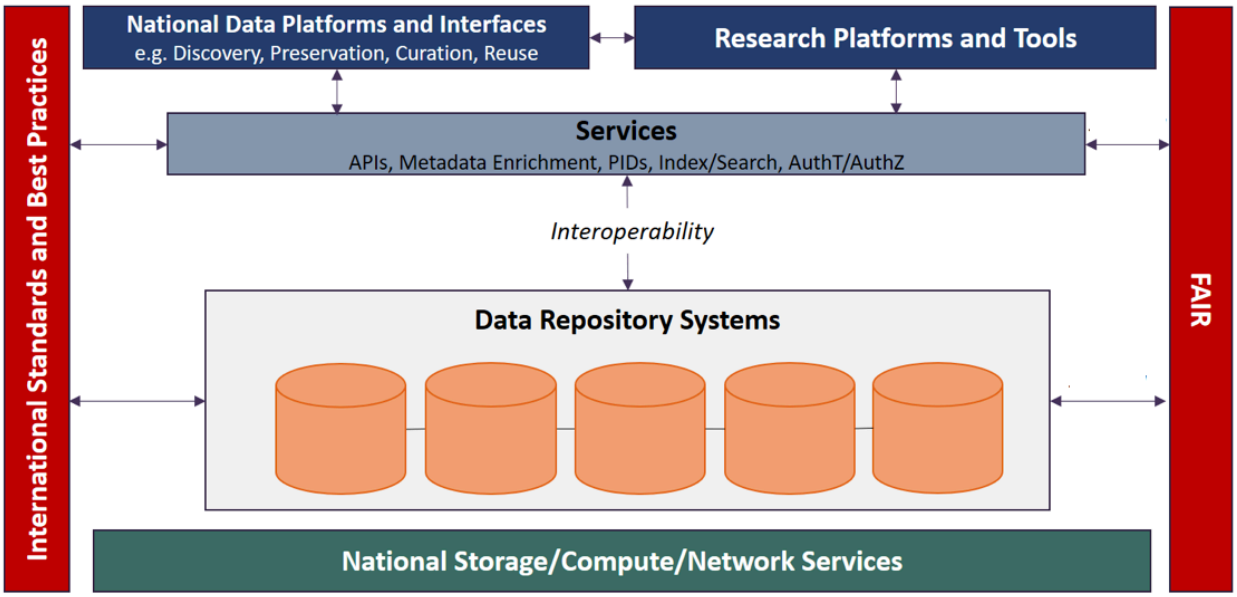
The 2018 Call is CANARIE's first call for innovation in the development of DRI that supports data management. Key goals of the Call are to: support the ability to find, access, and reuse data that has been generated by others; enable the adoption of best practices and compliance with funding policy; and align with national and international standards to facilitate broad data sharing across research institutions.

The Consultation also laid the foundation for the Call by asking respondents to articulate how their highlighted gap or opportunity intersects with existing and emerging national data services framework (NDSF) and the FAIR Principles (which states that data should be **F**indable, **A**ccessible, **I**nteroperable and **U**sable).

The NDSF is being used to frame development efforts in a way that facilitates interoperability and integration with DRI at the national and international levels. While Canada is in the early stages of efforts to build a NDSF, development of DRI that adheres to standards and protocols complements creation of a complete NDSF. Implementation of tools within the NDSF could encompass:

- development of DRI that is accessible to Canadians, where accessibility is determined through evolving policy frameworks;
- development of national data software services that are accessible to domain and generalized software platforms via appropriate interfaces;
- development and/or enhancement of existing domain and generalized software frameworks that improves or elevates their capacity to operate as platforms and/or services in a NDSF;
- collaborative teams that effect integration of multiple existing DRI platforms and software components into a cohesive NDSF.

The NDSF concept is illustrated below, including key themes that emerged from the Consultation.



Consideration of the FAIR Principles in a software development project can also take multiple forms. The FAIR Principles were designed to promote best practices that could be reflected in both human and machine services, so are a good fit for the CANARIE Call with its emphasis on software development.

Themes

After a review of the Community Consultation submissions, the RDM AC and CANARIE identified eight themes under which projects submissions would be considered. These themes reflect the feedback of Canada's RDM community, and serve to focus efforts on key DRI gaps.

In the listing below, themes are in order of the number of Community Consultation submissions that were slotted into that theme, number one representing the highest number of submissions that intersected with the theme, and eight the lowest. In the descriptions below, (meta)data is used as a term that applies to both metadata (data about data) and the data itself. The descriptions below have been extracted from the submissions to prevent the disclosure of any information that submitters may consider confidential, while at the same time providing the broader community with a good sense of the identified gaps and opportunities. Some submissions have been posted publicly in their entirety by the organizations that submitted them.

1) Enriching (meta)data and Discovery

As the diversity of research outputs increases, the challenge of finding specific researchers and resources also increases, especially at the national level. There is an opportunity to aggregate existing (meta)data from all systems and stages of the research lifecycle, as well as innovative approaches to enriching metadata (e.g. text mining, linking, entity extraction), that will improve the discovery of Canadian research outputs via an NDSF. Specific gaps or opportunities that were identified in the submissions include:

1. A national platform that aggregates metadata across the research ecosystem and leverages standards like ORCID and DOI as a core component.
2. Tools that use common APIs and ontologies to facilitate the harvesting of metadata from different systems in the research ecosystem,
3. A national platform that harvests/links (meta)data from domain-specific research platforms to facilitate discovery of data and increase the sharing of data between disciplines.
4. Tools that facilitate the identification of datasets that should be shared, and make it easy to add additional metadata to aid discovery.
5. Tools that facilitate the collection and enhancement of metadata during the active research stage.
6. Tools that facilitate the linking/discovery of data containing personal identifying information while preserving the privacy of individuals.
7. Tools that facilitate the mark-up extraction of enhanced metadata from textual outputs of research, thereby enhancing discovery and analysis.

2) Federated Repositories / Interoperability

There are in the order of 200 Canadian "data repositories", and likely many times this number of databases that contain useful research (meta)data. Each repository tends to present a unique architecture and (meta)data profile, and few are accessible via well-defined application program interfaces (APIs). There is an opportunity to use international standards and best practices to better document and make accessible Canada's (meta)data repositories, and to develop smarter approaches to harvesting and linking these sources at the national level, thereby facilitating inter-disciplinary research, discovery, and re-use. Specific gaps or opportunities that were identified in the submissions include:

1. A national data repository service and platform that ensures that all researchers, regardless of discipline, have a place to deposit their data.
2. Tools that increase interoperability between domain-specific research platforms thereby facilitating connections between researchers and datasets.
3. Adoption of common metadata, interoperability, and API standards that facilitate the federation of disparate data repositories.
4. Tools and services at the horizontal level that are accessible to any research platform or data repository, and that provide common functionality in support of RDM.
5. A national source of administrative and technical information about Canadian datasets and repositories.
6. Tools and ontologies that facilitate the harmonization of data across datasets/repositories in the same domain, and across domains.

3) Domain-Specific Repositories

Domain-specific repositories provide researchers with sources of (meta)data deposit and discovery, facilitating a response to funder/publisher mandates, and support domain-specific features that make (meta)data more useful. There is an opportunity to enhance Canada's domain repositories to ensure that they are interoperable with each other at the national level, and that these systems use internationally recognized best practices. Specific gaps or opportunities that were identified in the submissions include:

1. Repositories in disciplines not currently represented by a feature-rich repository.
2. Tools that provide more effective access to data in repositories that otherwise assume a high level of access, or knowledge, to effectively use the data.

4) Data Deposit and Curation

Publishers and funders are adopting policies that require researchers to make data from their published outputs more accessible and reusable, but the options available to the researcher can be confusing and complex. There is an opportunity to create software services to guide and document the best approach to data management for each researcher, in a standards-based context with a minimum of effort (e.g. with the assistance of machine-mediated services), and according to practices that make their data FAIR. Specific gaps or opportunities that were identified in the submissions include:

1. Tools that use machine-mediated services to facilitate stages in the data curation process, and apply standards to the creation of rich metadata.
2. Tools that facilitate more active and accurate deposit into data repositories.
3. Tools that facilitate the deposit of data in domain-specific repositories by contributors not necessarily familiar with the details of data in that discipline, or not currently contributing data.
4. Tools and services that facilitate a national coordinated approach to research data curation.

5) Preservation

While it can be a challenge to determine what research (meta)data should be preserved and for how long, it is clear that (meta)data is not accessible, and reusable over the long term, unless it is preserved. There is an opportunity to build IT tools that facilitate this decision-making process in such a way that leverages appropriate best practices that ensure research outputs are accessible in a usable way for a reasonable period. Specific gaps or opportunities that were identified in the submissions include:

1. Tools to help assess the long-term value/retention period for data after the mandatory retention period.
2. A software service/registry accessible to any software platform, that can provide canonical information about research file formats and standards, and associated actions to facilitate research workflows and preservation.
3. Integration of preservation tools into a federated system of national data repositories and platforms.

6) Persistent IDs / Citability

A key requirement for making research outputs accessible is the adoption of accepted best practices for unique IDs, whether for the researcher and their team, the variety of outputs, or the research equipment and services used. There is an opportunity to integrate Canadian research platforms with best practice standards for persistent IDs that make it easier to link the various outputs throughout the lifecycle, and attribute outputs to all research participants. Specific gaps or opportunities that were identified in the submissions include:

1. Development of support for dynamic data citations in research/data platforms.
2. Development of support for uniquely identifying physical samples in research/data platforms.
3. Tools to support the integration of persistent identifiers (such as DOI and ORCID) into existing and emerging research platforms and data repositories.

7) Data Access and Analytics

Researchers have an increasing diversity of approaches to analyzing and transforming their data, whether in local systems or in the context of large national and international high-performance computing frameworks. There is an opportunity to build services that bring disparate sources of data to the researcher to facilitate data analysis at all stages of the research lifecycle. Specific gaps or opportunities that were identified in the submissions include:

1. Tools that create advanced data products that make it easier to access and analyze large and/or complex datasets.
2. APIs that facilitate more sophisticated interactions with raw data research datasets.

8) Data Privacy and Security

In some disciplines (e.g. human health, biodiversity, social sciences), (meta)data cannot be easily shared without an extensive and lengthy process to anonymize/protect (meta)data. There is an opportunity to develop software services that facilitate this effort through semi-automated anonymization, as well as approaches that link datasets about individuals without unauthorized disclosure of personal information. Specific gaps or opportunities that were identified in the submissions include:

1. Tools that facilitate the sharing of sensitive data, including the integration of data access agreement/sharing templates into other systems.

CANARIE's goal is to facilitate the development of DRI based initially on the gaps and opportunities submitted by the broader stakeholder community as part of this consultation. As the projects from Call 1 proceed, CANARIE and Research Data Canada (which is funded by CANARIE) will work with the community to facilitate a cohesive and interoperable network of DRI accessible to all Canadians. CANARIE is also actively seeking feedback from the community on the renewal of its mandate for 2020-2025, and there is an option within the proposal to include additional funding to continue the efforts of the RDM Program.