RDM Workshop - Jan 23, 2019
Discussion Document

DOI: 10.5281/zenodo.2584179

One of the most recent efforts to identify gaps and opportunities in the DM landscape was the RDM Community Consultation conducted by CANARIE in June 2018, as a way to identify funding priorities for their first RDM funding call. The Consultation resulted in the development of 8 themes representing current gaps in the Canadian DM ecosystem.[1]

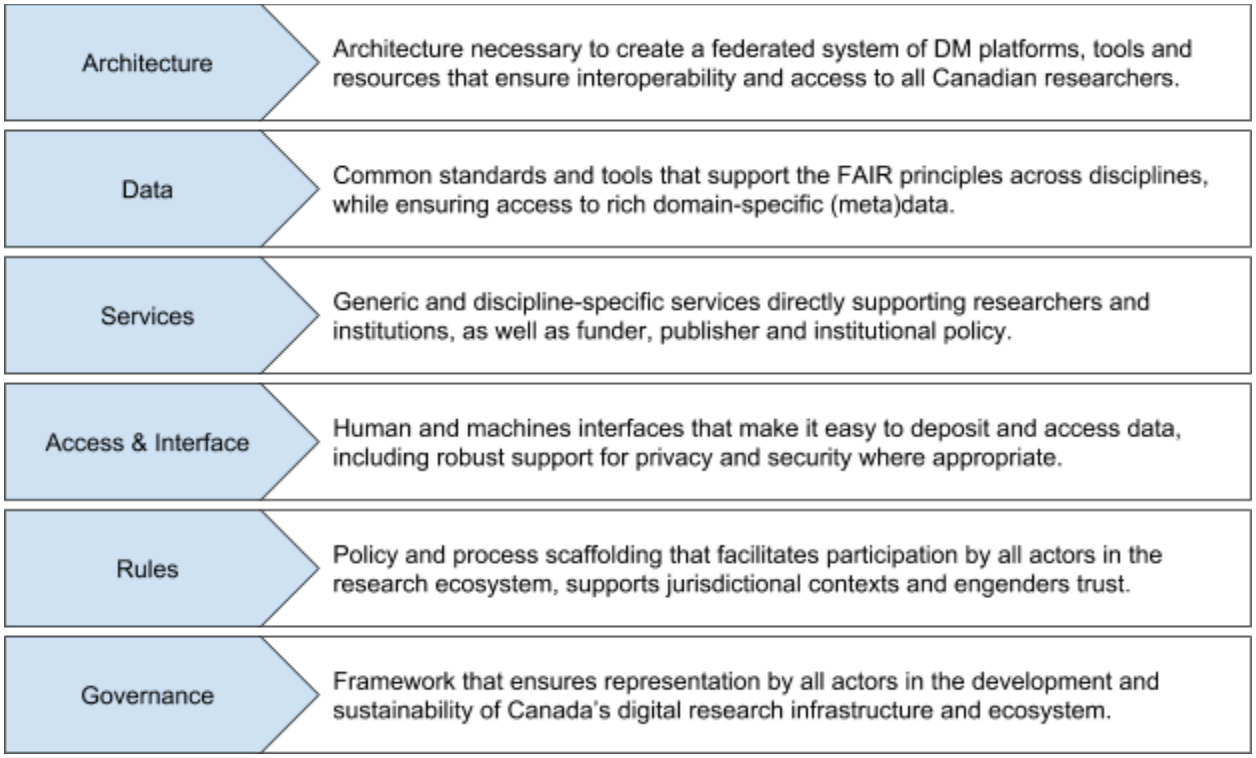
A total of nine projects were selected, each addressing one or more of the above themes, as well as the FAIR principles (Findable, Accessible, Interoperable and Reusable), and the NDSF: *a service that provides one or more data-related functions to applicable stakeholders and disciplines in a specific national context* [2]. These projects were announced in November 2018, and include national DRI for generic use (i.e. accessible to any researcher in Canada) as well as domain-specific use (i.e. intended for use by any researchers in a specific community of practice).

The 2019 RDM Workshop is intended to:

1. introduce funded recipients and project details to each other, and tools to facilitate communication between participants;
2. highlight alignment on approaches and components to a National Data Services Framework, including participating in the 1.5-day 2019 NDSF Summit;
3. learn from each other and provide feedback on how we might shape future calls, and otherwise inform future efforts.

The model below (based on the European Open Science Cloud (EOSC) model[3]) also provides the scaffolding for the discussion at the 2019 National Data Services Framework Summit[4], which immediately follows the RDM Workshop. For the RDM Workshop the goal is to define commonalities between funded projects that can lead to a more interoperable and cohesive set of digital research infrastructures (DRIs). Linking a suite of otherwise stand-alone infrastructures can be attempted at all six levels in the model below: the workshop offers an opportunity for a conversation in the context of the nine funded projects.

[1] https://www.canarie.ca/?wpdmdl=14726
[2] https://forum.casrai.org/t/national-data-service/939
[3] https://ec.europa.eu/research/openscience/pdf/swd_2018_83_f1_staff_working_paper_en.pdf
[4] https://www.rdc-drc.ca/activities/ndsf/



The discussion elements below are offered as the basis of an ongoing conversation regarding ways to achieve greater interoperability and cohesion of services in a NDSF context. Most of these were mentioned in the successful Project proposals and have been extracted and generalized here to facilitate discussion at the Workshop. The Workshop participants should feel free to add additional ideas/components under the "Other" categories in the list below.

1. **Architecture**
    a. *Federation*: given the Canadian landscape, as well as the desire to integrate as effectively as possible with international systems, a federated approach to architecture is preferred. We can think of a federated approach at two levels.
        i. Basic: systems provide access to data via an interface allowing harvesting by external systems.
        ii. Enhanced: systems provide interfaces that allow bi-directional sharing of (meta)data[5].
    b. *Storage*: one common characteristic that runs across most of the RDM projects, is the use of Compute Canada storage. There may an opportunity for RDM projects to share common storage at all three levels.
        i. Active: working copies, short term storage for the duration of the project.
        ii. Repository: dissemination copy, medium term beyond the duration of the project.
        iii. Preservation: Preservation copies, long term.

[5] (meta)data is a convenience used to refer to both metadata and data at the same time.

c. Other

2. **Data**
    a. *Metadata*: a NDSF assumes some level of agreement on common metadata standards, or at least the ability to map (meta)data from disparate platforms.
        i. Dublin Core (DC)/Datacite/Schema.org
            1. All platforms propose one or more descriptive metadata formats, and a majority propose support for common schemas, with DC being the most common, followed by Datacite.
        ii. ORCID
            1. Support for a local researcher identifier is also substantial, with ORCID being common to a number of projects.
        iii. Other
3. **Services**
    a. Tri-Council DM Policy
        i. While not available yet in a final form, the Tri-Council draft policy[6] does provide a basic level of support researchers should be able to expect from the platforms they use. Platforms that intersect with the components if a DMP, and meet the needs of data availability statements and data deposit are examples.
    b. DMP Assistant
        i. The DMP Assistant is one example of a DMP tool, and can be used as a guide to the types of services and resources researchers might expect from a platform.
    c. Other
4. **Access & Interface**
    a. *Interfaces*: there are a number of standards, protocols and interfaces that are unique to the funded projects, typically reflecting support for specific domains. There are also a number of these that are common among many of the funded projects, providing opportunities for collaborations and interoperability.
        i. Globus API
        ii. Open Archives Initiative (OAI)
        iii. ORCID API
        iv. Datacite API
        v. Other
    b. While not mentioned in the proposals, the SmartAPI[7] is worth considering as a way to describe services and resources according to the FAIR Principles.
5. **Rules**
    a. National Access

[6] https://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html
[7] https://smart-api.info/

i. Federated Identity Management (FIM): most of the funded projects propose support for a form of FIM. The most common are CAF and LDAP.

b. CoreTrustSeal

i. One of the funded projects mentions CoreTrustSeal as a framework that can guide the provision of sustainable data repository services, as well as an achievable certification standard that helps repositories provide a robust services, and researchers to make decisions about where to deposit their data.

c. Other

6. Governance

a. National Data Services Framework

i. Canadian NDSF Committee(s): the NDSF could provide a governance context for DRI platforms via membership and engagement on committees.

b. RDC-DRC Best Practices Designation

i. The recent RDC-DRC Best Practices Designation[8] is intended to highlight Canadian services and resources that define a best practices approach. There are two levels to the Designation:

1. Basic Designation: a service or resource that provides an example of a best practice in one of the 4 areas: Infrastructure; Outreach and Education; Policy; Standards and Protocols.
2. NDSF Designation: a service or resource that meets the Basic Designation, and also meets the requirements for functioning at a national level.

c. Other

The nine Projects are described below, and the 2018 RDM Project map provides additional detail.

1. **Canadian Health Omics Repository, Distributed (CanDIG CHORD)** – *Led by Dr. Guillaume Bourque, McGill University*
    a. CanDIG is a national project that allows collaborative analysis of human health genomics data distributed across the country, enabling stewards of this data complete, auditable control over data access. The CHORD project will create a federated Canadian national data service for privacy-sensitive genomic and related health data. It will also broaden the Canadian health research community's access to the technologies and services being built by CanDIG and its international partners in the Global Alliance for Genomics and Health.
2. **Dataverse for the Canadian Research Community** – *Led by Kate Davis, University of Toronto*

[8] https://docs.google.com/document/d/1sqbX_F0BllX3vVMjhcwUJBR_O6rOqGPc5oh9wG7uy6Y

    a. Dataverse is an open-source research data repository platform, developed by Harvard University's Institute for Quantitative Social Science with adopters and contributors from Canada, the US, and Europe. Originally architected to serve the needs of social science researchers with small to medium size data files, this project will adapt Dataverse's software architecture to address the needs of a broad range of researchers in Canada through improved scalability, support for large data files, curation workflows, and integration with Canadian storage and authentication providers.
3. **DuraCloud – Linking Data Repositories to Preservation Storage** – *Co-led by Corey Davis, Council of Prairie and Pacific Research Libraries and Stephen Marks and Kate Davis, University of Toronto*
    a. Canadian researchers have access to many storage services suitable for the long-term preservation of digital content, including research data. The DuraCloud project will connect several Canadian preservation storage services via this software, which is maintained by the DuraSpace Foundation. As a result, Canadian researchers will be able to seamlessly access different storage services through a single interface.
4. **FAIR Repository for Annotations, Corpora and Schemas (FRACS)** – *Led by André Lapointe, CRIM*
    a. Artificial intelligence-based applications require access to massive quantities of data. To enable Canada's academic researchers to scale their AI-based projects such that they are competitive with private sector applications, large volumes of data must be coupled with detailed annotations. Annotated data sets allow models to be effectively trained and validated by machine learning algorithms. The FRACS project will simplify the management of large scale datasets by facilitating the creation, storage, search, manipulation and sharing of their annotations.
5. **Federated Geospatial Data Discovery for Canada** – *Co-led by Eugene Barsky, Evan Thornberry, and Paul Lesack, University of British Columbia Library*
    a. Traditionally, research data repositories have relied on text-based searching. However, there is increasing demand for geographic components in research, examples of which include migration paths, the distribution of agricultural yields, infrared satellite imagery, the distribution of artifacts in an archaeological site, and the flow routes of water. The goal of this project is to create an extensible, open-source software method to search and discover Canadian geospatial research data using an interface specifically designed for maps, enabling users to discover geospatial resources in a more spatially-intuitive way.
6. **Making Identifiers Necessary to Track Evolving Data (MINTED)** – *Led by Reyna Jenkyns, Ocean Networks Canada (ONC), University of Victoria*
    a. ONC operates world-leading ocean observatories and dynamic data repository services. While there has been a growing recognition of the benefits and need for data citations made evident by the introduction of the FAIR Principles, existing platforms and tools are currently only able to serve the needs of static or

non-frequently updated datasets. The MINTED project will apply best practices for dynamic dataset citation, Digital Object Identifiers (DOIs), and researcher ORCIDs into ONC's Oceans 2.0 digital infrastructure.

7. **Radiam: Management Software for Active Research Data** – *Led by Dr. Kevin Schneider, University of Saskatchewan*
    a. Research data, which may have value beyond the research for which it was collected, is often distributed across multiple storage devices, tools, and platforms. Simply knowing that a dataset exists, let alone finding it, presents a significant challenge. Radiam will provide a project-level metadata index of research data, regardless of where or how it is stored. Radiam will improve researchers' ability to find and cite existing datasets by not only storing the location of the data, but also the standard and custom metadata records associated with it.
8. **Managing the Research Data Lifecycle using Islandora** – *Co-led by Donald Moses and Rosemary Le Faive, University of Prince Edward Island (UPEI)*
    a. In collaboration with Simon Fraser University and the Islandora Foundation, UPEI will build research data management capacity and integrations using the latest version of Islandora, also known as CLAW. Islandora is an open-source software framework designed to help organizations collaboratively manage, discover, and share digital assets using a best-practices, standards-based approach. The project will develop integrations with identifier, metadata, authentication, storage, and dissemination systems, supporting the FAIR principles and the research data lifecycle.
9. **Research Portal for Secure Data Discovery, Access and Collaboration** – *Co-led by Dr. Elizabeth Theriault, Ontario Brain Institute and Moyez Dharsee, Indoc Research*
    a. The Ontario Brain Institute (OBI) and Indoc Research have developed Brain-CODE, an extensible neuroinformatics platform designed to manage the collection, curation, analysis and sharing of different data types across several brain disorders. To address the RDM needs of researchers studying disorders of the brain and other disease areas, this project will develop data portal software that will enable research teams to securely and seamlessly capture, query, and visualize patient data; collaborate and share datasets; and access support and training resources. The project will serve the needs of teams using Brain-CODE as well as those from collaborating institutions and the broader medical research community.