Research Data Management

As their digital datasets grow, researchers across all fields of inquiry are struggling to manage those datasets. For an individual researcher, there is a strong motivation to be able to find, access, and analyze their own data once it is produced. There is a need to share datasets with colleagues, to preserve datasets for later reuse, or to combine their dataset with others from within or beyond their disciplinary area. For the research community, the reproducibility of a scientific result drives a need for open, managed, accessible datasets that allow results to be independently validated. For policy makers, the data output of government funded research is seen as a valuable asset to be preserved and shared for general good.

The recent Tri-Agency Statement of Principles on Digital Data Management (June 15, 2016) sets the expectation: “Research data resulting from agency funding should normally be preserved in a publicly accessible, secure and curated repository or other platform for discovery and reuse by others.” The Statement also includes expectations for researchers to create data management plans, and to provide the necessary metadata to facilitate understanding and reuse.

Using Research Data Management (RDM) techniques, data can be ingested, curated, preserved, discovered, shared and transported. Researchers need the technical tools, skills and support to enable effective RDM.

CC and CARL Partnership for a national platform for Research Data Management

Compute Canada (CC) and the Canadian Association of Research Libraries (CARL) have agreed to collaborate to build a scalable national platform for research data management and discovery.

The proposed pan-Canadian platform will provide tools and services to support the curation, access, discoverability, and preservation of research data, allowing researchers across Canada in a range of disciplines to have improved access and control of large amounts of data. This addresses a longstanding gap in Canada’s infrastructure for digital research data management.

The Portage Network of CARL will assist with the requirements for and design of the national platform service, providing metadata and data workflow solutions and testing the platform.

Compute Canada will provide project management and software development expertise and necessary computational power.

This RDM service is not intended to serve as a monolithic solution for all of Canada’s research data needs. Rather, it is meant to provide a framework that allows existing and future data repositories to be federated within a coherent system. At the same time, it will provide a flexible repository and preservation system for Canadian researchers and institutions who do not have a solution already in place.
Proposed Core Features of the Service

- **Federated storage model**: Individual institutions or organizations can deploy storage locally and can federate their local repository into the national system.
- **Federated support model**: On-campus support for the researchers generating the data to manage those data.
- **Nationally integrated**: While the storage and support are distributed, a coherent national service is provided to researchers regardless of their location or field.
- **Scalable model**: The system can scale as adoption by researchers and stored data grows.
- **National data discovery**: While different data collections can be hosted in different locations, with different access controls and different metadata, the various data collections are discoverable through a web-based, federated search tool.
- **Data preservation**: Researchers and institutions can choose to preserve data in multiple locations in long-term preservation formats. A long-term institutional commitment is required for any preserved datasets.
- **Suitable for a broad range of data types**: The favoured solution is suitable for managing diverse datasets from a broad spectrum of disciplines, typically referred to as the long tail of data.
- **Bulk data and metadata ingestion**: The system is able to ingest and index existing data and metadata from Canadian researchers.
- **Access control mechanisms**: The solution allows fine-grained control of who can discover and download each dataset, and supports embargo.

The technology platform as envisioned

Most data that is indexed and made available for national discovery is expected to be housed in curated collections. The role of the curator is to ensure the quality of both the data and the associated metadata. Curation could be performed at project level (e.g. by a member of the research team), at an institutional level, or at a national level through granting collection curation privileges to specific people. A self-serve option for small collections without curation is also envisioned, primarily for active research datasets.

Once data is ingested by the system, any researcher can then discover and access the collection, if allowed by the data’s access policies (which are highly granular). If the researcher can access the data, she or he can use Globus Connect to transfer the data to a Compute Canada processing facility, to any other Globus-enabled facility in the world, or to their laptop for further analysis. Each dataset has an associated (unique) identifier (such as a DOI) assigned, so that the dataset can be cited, the data owner can be credited, and the work can be reproduced by others.

It will also be possible to federate existing data repositories into a national service. The national service can harvest metadata from existing repositories. This allows search over all repositories from a single web interface. This basic level of federation requires only the support of agreed-upon data exchange standards and protocols. A deeper federation would allow Globus transfers of any discovered dataset from the federated repository to any Globus enabled facility, as with data deposited directly into the national service. This requires repository-by-repository software development.
Building the Infrastructure

Leveraging the work of a national federated pilot project convened by Research Data Canada (RDC) in 2014-2015, CARL and Compute Canada launched a 2 year joint project in January 2016 to build a national-scale research data repository tool and preservation suite, capable of providing the technical foundation for a national RDM service. This document provides only a high-level overview of this project.

The project technology leverages three existing products: Globus Data Publication (a data repository service), Globus Connect (large file transfer service) and Archivematica (a Canadian open-source data preservation package). The technology also includes custom-built software to integrate these software packages into a coherent solution.

Compute Canada has also partnered with Globus (www.globus.org) to assist with the integrations as well as accelerate vendor development of features in Globus software that researchers need.

Curation, Training, Support, and Infrastructure

Clearly, the technology development is only one piece of what would be necessary to establish and operate a national research repository platform, and to ensure it is well-used and meets researcher needs. Other pieces would include: the engagement of researchers and of librarians and curators working out of their institutions across the country; training and support of researchers throughout the country in RDM practices and the technology used; and the required IT infrastructure and its operation to underpin the national repository and federated discovery.

Project Governance

The Steering Committee for the development project comprises representation from Compute Canada and the Canadian Association of Research Libraries:

- Dugan O'Neil dungan.oneil@computecanada.ca
- Chuck Humphrey chuck.humphrey@ualberta.ca
- Steve Marks steve.marks@utoronto.ca
- Jason Hlady jason.hlady@usask.ca

Stakeholder Group:

A broad stakeholders group is being set up to keep interested parties informed about progress in the project and the service. Anyone can request to be added to an email list to become part of this Stakeholder Group and to receive updates and comment about the evolving National RDM Service. This list is run as a Google Group at rdm@computecanada.ca

To ask to be added to the Stakeholder Group, send email to jrsouza@computecanada.ca

Contact the Technology Project:

- Project Sponsor jason.hlady@usask.ca
- Lead Developer todd.trann@computecanada.ca
- Project Manager keith.jeffrey@computecanada.ca

Web site www.computecanada.ca/RDM
Progress to June 2016

The technology development was started in January 2016. Six months in to development there are a number of successes to report. All work is early development only so far; there are no production services and no versions ready for end-user testing.

Currently, datasets (including datasets with large files or large numbers of files) can be uploaded into the development data repository running on Compute Canada hardware. Data transfers, which could take hours to complete for very large datasets, are performed asynchronously. Datasets normally automatically receive some degree of standards-based preservation processing to help ensure the dataset will be stored in usable formats for future use. Metadata (information that describes the data) is collected from the repository and indexed for discovery, together with metadata collected about the datasets held in other data repositories (institutional, regional, and domain-specific) in Canada. Indexed data can be searched, restricting results by filtering by facets such as data type, date, subject and source repository. The dataset can be accessed or copied for reuse from whichever repository in which it is held, depending on the access restrictions that may exist for that dataset at the source.
Specific technical achievements of the development project January-June 2016:

- Through a formalized two-year project partnership agreement between Globus and Compute Canada, the development team obtained access to the Globus Publication software code and created development instances of Globus Publication on Compute Canada hardware.
- Developed reproducible methods for automated deployment of Globus Publication and Archivematica.
- Deployed development instances of Archivematica on Compute Canada hardware.
- Developed a robust integration between Globus Publication and Archivematica, such that datasets submitted to the repository technology can be automatically processed by Archivematica and the resulting Dissemination Information Package (DIP) is passed back to Globus Publication. Data submitters can monitor the process.
- Designed and built a metadata harvester to collect metadata from selected Canadian research data repositories using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Any and all metadata exposed by repositories through OAI-PMH is harvested and made searchable through discovery interfaces.
• Globus has developed a new Globus Search Platform, running on Amazon Web Services, with an API that will be used for developing a new discovery user interface for the repository. The discovery interface will enable users to find and access datasets in the national repository or other repositories from which metadata has been harvested. A prototype of an interface has been integrated with the development repository. The API supports faceted search and supports searching on custom metadata in repository collections.

• Improved the Globus Publication interface to support internationalization supporting both official languages of Canada, and completed project branding for the development stage of the proposed service.

• Completed the first stage of performance profiling for the execution of Archivematica in preparation for future development to improve performance, scalability, and implement new features.

• Shared improvements and contributed bug fixes back to Archivematica and DSpace open source projects.

• Delivered presentations on the technology and project:
  o Lightning talk at Globus World, Chicago, April 2016
  o Presentation at 5th National Data Service Consortium Workshop, Chapel Hill, North Carolina, April 2016
  o Presentation at CANHEIT/HPCS, Edmonton, June 2016