CARL/Portage and Compute Canada

Federated Research Data Repository

Project Update January 2017

Compute Canada (CC) and the Canadian Association of Research Libraries (CARL) are collaborating in building a scalable federated platform for digital research data management (RDM) and discovery.

CARL, through its Portage network, is developing an RDM community of practice and platforms for researchers and institutions. This includes expertise in data management planning, consultation on metadata, data curation and preservation, and support of a federated research data repository and data discovery tool (FRDR). CC is building and operating the technological platform for this repository and discovery tool.

The current effort is a software development and integration project, started January 2016, to build scalable software for this Federated Research Data Repository. The project is scheduled to have the software ready for a production service, capable of accepting research data for long-term retention and discovery, by the end of 2017.

Presently, halfway through the software development project’s timeframe, much of the capability has been demonstrated and testers are providing feedback regarding the user interfaces.

This report describes the status as of the end of calendar year 2016 and highlights some of the developments of the past year.

Background

CC and CARL are collaborating to build a scalable federated platform for digital RDM and discovery. Using best practices and techniques, research data will be transferred, ingested, curated, preserved, discovered, and shared. This partnership’s pan-Canadian platform will provide tools and services to support researchers across our country in a range of disciplines to have improved access and control of large amounts of data. Furthermore, it addresses a longstanding gap in Canada’s infrastructure for RDM. Other national jurisdictions are working on similar initiatives: JISC in the United Kingdom¹ and ANDS in Australia² for example.

The federated data repository and discovery service are 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 an existing solution.

RDM practices increase accountability for use of public funds, improve the completeness and understandability of data that is retained, improve the veracity of research findings by permitting other researchers to reproduce the results, improve the discoverability of data by other researchers, and ultimately accelerate new research outcomes.

¹ https://www.jisc.ac.uk/rd/projects/research-data-shared-service
CARL, through its Portage network, is establishing services in data management planning; consultation on metadata, data curation and preservation; and support of a federated research data repository and data discovery tool. CC is developing and operating the technological platform for these services.

In the context of supporting research data across the stages of the research lifecycle, CARL/Portage introduced its data management planning tool in October 2015. CC has contributed its licensed Globus Connect to facilitate data transfers between lifecycle stages. CARL and CC, following a pilot project in 2014-15 coordinated with Research Data Canada, began developing a data repository with scalable automatic preservation and archiving features to fill specific gaps in the lifecycle.

As the primary recommendation of this pilot and in order to accommodate the expected scale of files and datasets for a pan-Canadian service, the Globus Publication platform was selected as the foundation for the data repository and the proprietary Globus Connect software was chosen for efficient transfer of large amounts of data. CC also entered into a contractual arrangement with Globus to open-source the Publication code base and to leverage their effort on the development project.

At the same time, preserving data for the long term was also known to be critical and the Canadian standards-based preservation tool Archivematica was selected as the basis for automated processing of research data to help ensure its retention and readability for the long term. Preservation steps can involve changing file formats to standards that are expected to be supported in the long term, as well as bundling all of the data files together with as much metadata as is available into a self-contained archive format for long-term storage. The service development for FRDR will include modifications to Archivematica to increase its ability to handle larger numbers of files and larger datasets.

Presently, halfway through the software development project’s two year timeframe, much of the capability has been demonstrated and testers are providing feedback regarding the user interfaces.

Core Features of the Proposed 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 who are generating the data for which management services are needed.
- 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 to accommodate growth in adoption by researchers and in the quantity of data stored.
- 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

---

4 Globus is a not-for-profit organization based out of the University of Chicago.
long-term preservation formats.

- Suitable for a broad range of data types: Diverse datasets from a broad spectrum of disciplines, typically referred to as the long tail of data, can be properly managed.
- 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

Current Software Development
The current phase of the project (Alpha) is evaluating and adjusting the user interfaces for submitting, curating, and searching datasets. Development also continues in preparation for the next Beta testing phase (Spring 2017) and in preparing organizationally for the platform and services that will be needed during a pilot phase and, later, production.

Highlights so far in the development project to build the software framework for a scalable federated platform for digital RDM and discovery include:

Repository

- The project leverages existing products: Globus Data Publication (a data repository service) and Globus Connect (large file transfer service) packages into a coherent solution. The project has implemented a customized Globus Publication tool on CC cloud computing facilities.
- It has been demonstrated that FRDR can utilize storage facilities in multiple locations and regions. This permits institutions to retain their data locally while at the same time contributing to FRDR and making their data more widely available. Data whose location is constrained to be kept in Canada or some province can be accommodated in FRDR. (Public metadata regarding FDRD datasets is copied to Globus’ indexing and search tool currently running in the USA.)
- FRDR accommodates custom metadata. Collections can be created to include custom metadata (e.g., specific to a particular discipline). Depending on requirements, additional customizations would be needed for metadata entry forms.
- FRDR automatically issues and registers DOIs\(^5\) for datasets that have been submitted and accepted. Have arranged licensing for CC through DataCite Canada.
- Testing is underway to gain feedback from selected users regarding the interfaces for submitting datasets, searching for datasets, and curation of datasets.

Preservation

- The customized Globus Publication software integrates with Archivematica to be able to automatically perform preservation processes on submitted datasets.
- Archivematica processing has been profiled for performance characteristics.
- Archivematica performance has been improved by changing the way it launches its microservices and substituting a data compression routine. For datasets with a large number of files, these changes are expected to result in performance improvements up to 18%. These changes have been submitted to be added into the Archivematica code base.

\(^5\) A DOI is a persistent identifier intended to refer to a dataset uniquely over the long term.
- Apache Tika has been incorporated into the workflow to extract additional metadata contained within the submitted data. This additional metadata can improve the discoverability of data.
- A resource management system has been designed that will run multiple Archivematica instances in parallel to scale up throughput of preservation processing.

**Discovery**
- Searching now uses the Globus Search Platform, Globus organization’s implementation of a powerful new cloud-based discovery (search) backend.
- The University of British Columbia contributed their Open Collections discovery interface code, which has been adapted to work with the new Globus Search Platform backend.
- Multilingual capability (implemented in French and English) has been developed for the UBC Open Collections code and contributed back to UBC.
- The metadata of other Canadian data repositories are harvested and included in the index that FRDR searches, allowing researchers to discover relevant datasets in many repositories. Repositories that have been indexed include Scholars Portal, UBC Circle, UofA Dataverse, SFU Radar, and Concordia Spectrum. FRDR also indexes Open Data Canada, the Canadian government’s open data repository. So far, approximately 209,000 datasets are discoverable through the FRDR federated discovery index. Additional repositories can be harvested with modest additional effort.

**Consultation**
The partnership between CC and CARL is proving to be both effective and productive. As intended, the software development is being informed by the expertise of academic librarians, who represent different groups within the Portage Network. For example, members in the Discovery, Curation, and Preservation Expert Groups are preparing white papers documenting best practices to be followed, recommending policy choices, and contributing their time and suggestions as testers.

These expert groups will be leading developments in licensing, ongoing commitments, and terms of service within an academic library network distributed across multiple institutions. They will provide directions regarding criteria through which repositories will be incorporated into federated services and for establishing collections and a national team of curators.

**Communications**
In order to inform the various communities that will be interested in the federated data repository and related services, a number of communications and presentations have been made, including:

- A conference presentation at CANHEIT/HPCS 2016
- A presence at the September 2016 Research Data Alliance meeting
- A presentation at conference Globus World 2016
- A booth and demonstrations at Supercomputing 16 conference
- Presentation to Innovation, Science and Economic Development Canada, CARL submitted a formal proposal to ISED in December 2016.
- FRDR featured in July Compute Canada Newsletter
- Web site www.computecanada.ca/RDM with historical documents and quarterly project progress reports
## Current Development Phase:

**Alpha Phase Objectives (Dec 1, 2016 to March 31, 2017)**
approved by FRDR Steering Committee

<table>
<thead>
<tr>
<th>Feature / Function / Improvement</th>
<th>Est. Effort (days)</th>
<th>Deliverable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implement design for increased preservation throughput.</td>
<td>25</td>
<td>Demonstrate scalable implementation of increased preservation throughput</td>
</tr>
<tr>
<td>API to submit items to repository</td>
<td>15</td>
<td>Demonstrate functional API submission to Globus Publication</td>
</tr>
<tr>
<td>Implement advanced search queries in discovery interface</td>
<td>13</td>
<td>Demonstrate discovery interface that includes the UBC advanced search</td>
</tr>
<tr>
<td>Analysis and design to increase throughput for preservation (process more items at a time)</td>
<td>10</td>
<td>Analysis document and design document and plan for implementing order-of-magnitude throughput improvements in preservation processing</td>
</tr>
<tr>
<td>UI change to import metadata from uploaded file on submission</td>
<td>10</td>
<td>Demonstrate functional option for applying item metadata by specifying a file to upload. Will permit the option to subsequently add or alter metadata before publishing</td>
</tr>
<tr>
<td>Placeholder to implement changes requested by Portage Discovery Expert Group</td>
<td>5</td>
<td>Will triage feedback and will fix some of the issues reported by alpha testers</td>
</tr>
<tr>
<td>Placeholder to implement changes requested by Portage Curation Expert Group (workflow, notifications, hierarchy?)</td>
<td>5</td>
<td>Will triage feedback and will fix some of the issues reported by alpha testers</td>
</tr>
<tr>
<td>Placeholder to implement changes requested by Portage Preservation Expert Group (data validation, DIPs, AIP regeneration, data expiration?)</td>
<td>5</td>
<td>Will triage feedback and will fix some of the issues reported by alpha testers</td>
</tr>
<tr>
<td>Perform an end-to-end security analysis of all products</td>
<td>5</td>
<td>A report detailing security risk analysis.</td>
</tr>
<tr>
<td>Archivematica improvement: API interface calls replace watched folders</td>
<td>5</td>
<td>Implemented API call for Globus Publication to initiate preservation processing by Archivematica. Implemented API call for Archivematica to hand off preserved data CC object storage</td>
</tr>
<tr>
<td>List the use cases for items submitted to FRDR requiring access controls and give details on whether they are already done, in progress, to be done later, or out of scope.</td>
<td>4</td>
<td>Report with recommendation regarding handling of these use cases. CC (E.g., add contact info for submitter as a metadata field?) (e.g., add a metadata field with content to display regarding access)</td>
</tr>
<tr>
<td>Create and deploy alpha environment</td>
<td>4</td>
<td>Deployment with sufficient scale, stability, and performance to support alpha testing.</td>
</tr>
<tr>
<td>Home page design for service (see UBC OC; ANDS)</td>
<td>3</td>
<td>New home page for FRDR including new graphics</td>
</tr>
<tr>
<td>Coordinate Alpha testing to expose missing features from discovery interface</td>
<td>3</td>
<td>Sufficient discovery Alpha testing completed.</td>
</tr>
<tr>
<td>Coordinate Alpha testing to understand organization and curation issues</td>
<td>3</td>
<td>Sufficient curation Alpha testing completed.</td>
</tr>
<tr>
<td>Automate metadata extraction on item submission</td>
<td>3</td>
<td>Demonstrate integrated metadata extraction processing. (E.g., switch to enable per collection; email to</td>
</tr>
</tbody>
</table>
Remaining project timeframe

**Beta Phase**

The next phase (starting April 2017) is to conduct Beta testing of both the interface and the underlying functionality of the software. This testing will incorporate a broader range of users to increase the diversity of subjects and datasets, as well as to test the user reactions with varying degrees of familiarity with RDM and data repositories.

In this case, a user support function is required to help beta testers and to begin learning about the amount and type of support users will need.

Also during this phase, the technical and support environment for both the pilot phase and for production needs to be finalized.

**Pilot Phase**

The Pilot phase is scheduled for the fall of 2017 and will be tested with a limited number of users to gain an understanding of the technical platform and the service organization needed once production starts. The pilot phase is intended to debug the technical systems supporting the structure.

It will include a production-candidate software version, the production hardware environment and a support team.

In all cases, the Steering Committee prioritizes the work within a phase.

**Production**

The FRDR schedule plans to be ready to support CARL launching FRDR as a service to Canadian researchers at the beginning of 2018.

The pilot phase and production may be contingent upon obtaining funding for the necessary support structure for researchers and librarians and for compute capacity and technical support necessary to properly operate
and maintain the service.

**Planned Features at Release Time**

When the FRDR service launches in January 2018, the following functionality will be supported (not all of this functionality is currently demonstrable because development is ongoing):

- CARL will be responsible for the FRDR service as one of a number of RDM services that they will offer to help researchers and institutions plan, preserve, and share their research data.
- Compute Canada will provide the IT expertise, server and storage capacity under arrangement with CARL. Technical support for using FRDR and for using Globus Connect file transfer service will be available through CC.
- Repository features will include:
  - Capability to handle large files and large datasets. FRDR uses Globus Connect protocol to transfer large datasets efficiently.
  - Core repository technology will be operated on CC facilities in Canada, will be run on high-availability infrastructure and under a service management regimen. Institutions can opt to provide local storage for FRDR so datasets can be submitted to FRDR and stored at the institution.
  - Metadata for extremely large datasets that are already security stored and preserved can be added to the FRDR centralized repository and discovery engine without actually moving the data.
  - Metadata can be entered directly in a web form or uploaded from a stored metadata file. An API for submitting a large number of datasets will be available.
  - The repository will automatically register persistent identifiers (DOIs) for datasets added to the FRDR repository.
  - Researchers will receive a commitment from CARL/CC to retain their data for some period of time and keep it accessible, thereby helping researchers to comply with granting agency requirements.
  - Researchers will be able to select from an approved set of licenses to apply governing how their data should be used and cited.
  - Researchers will be able to control whether a dataset should be embargoed for a period of time and who has access to their data.
  - FRDR will have partial integration with ORCID, an organization maintaining a unique, persistent identifier for researchers (worldwide and independent of institutional affiliation.) Integration will include “person lookup using ORCID” functionality, to ensure that dataset creators are appropriately listed in the dataset’s metadata.

- Preservation features will include:
  - Datasets submitted to FRDR can undergo automatic preservation processing to convert file formats to standards expected to remain readable over longer periods of time.
  - Some collections can be queued for review by a trained curator to help ensure that datasets are complete and adequately described to maximize the reusability of the data.

---

6 [https://orcid.org/](https://orcid.org/)
Discoverability features will include:

- The Search function within FRDR will permit simple or advanced queries to find datasets deposited in FRDR.
- Faceted search capability will be supported, enabling researchers to refine their searches.
- FRDR will facilitate discovery of datasets that are deposited into many research data repositories in Canada. FRDR harvests the metadata exposed by other repositories and indexes this information. To access a dataset that is found in another repository, FRDR will redirect the researcher to the landing page of the appropriate repository.
- An API will be available for the Search functionality, permitting additional search capability to be developed by others.

While these services are planned to be available at the beginning of 2018, it is expected that development work will continue. Data repository services are still new and standards continue to evolve. Continued development will be needed to keep up with evolving standards, to add functionality to the repository, improve preservation services, make the search interface more flexible, and to add customizations to be able to harvest metadata from additional data repositories.

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

- Dugan O'Neil dugan.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 identified to keep interested parties informed about progress in the project and the service. Anyone can request to be added to the email list for this Stakeholder Group and to receive updates and comment about the evolving FRDR 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

February 23, 2017