Dashboard
The project dashboard remains unchanged from last quarterly report:
  - Scope **Yellow** – Scope is still evolving
  - Schedule **Green**
  - Cost **Green**

Context
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. Using RDM techniques, research data will be ingested, curated, preserved, discovered, shared, and transported. 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. This platform addresses a longstanding gap in Canada’s infrastructure for RDM.

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.

The current project is a software development and integration project, started January 2016, to build a scalable software framework for this federated research data repository (FRDR). The project proposes having this framework ready for a production service, capable of accepting research data for long-term retention and discovery, by the end of 2017. To meet that deadline, the project sees a pilot test of a production-ready service running in fall of 2017, and beta testing period in the late Spring.

The software development project has demonstrated much of the key functionality in depositing datasets, processing datasets for preservation, and permitting discovery and access to datasets. This past quarter the project has continued to prepare for alpha testing by test researchers and test curators before the end of 2016.

Progress in this quarter
This third quarter of the project covers the period from July 1 to September 30, 2016. Highlights from this quarter include:

---

1 Previous reports referred to the service as the National Research Data Repository (NRDR).
**UBC discovery interface code selected**

The Project selected and adapted the discovery interface code developed and used at the University of British Columbia in their Open Collections service. This interface is highly functional and was readily adaptable to the Globus Search service utilized in FRDR. The Open Collections discovery interface also largely conforms with the principles and recommendations for data discovery in the white paper *Research Data Discovery and the Scholarly Ecosystem in Canada* developed by the Portage Network’s Discovery Working Group on behalf of the Canadian Association of Research Libraries (CARL).

We would like to acknowledge the University of British Columbia and their developers for their work and their contribution of this code to the FRDR project. Functionality that the FRDR project adds to the discovery code, such as language internationalization, will be contributed back to UBC and the open sourced Open Collections project.

**Additional metadata harvesting**

The project has added capability to harvest metadata from additional types of data repositories. Metadata from the Government of Canada’s Open Data repository is now being added to the datasets discoverable through FRDR. Similarly, FRDR is now harvesting the metadata for datasets in the University of Alberta Libraries’ Dataverse network. It will be easy later to harvest metadata from additional repositories of these same types that expose their metadata contents using the same protocols.

**Archivematica performance improvements**

Since FRDR is being designed to handle large files and large datasets (filling a gap in repository capability) the project is also investing in improvements to Archivematica, the software for reformatting and normalizing data for long-term preservation. Initial improvements made are changing the way Archivematica 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 Artefactual Systems Inc. for inclusion in the Archivematica code base.

---

2 Open Collections provides a unified platform that delivers UBC-created and maintained digital resources. See [https://open.library.ubc.ca](https://open.library.ubc.ca)

3 For the white paper see [https://portagenetwork.ca/news/discovery-white-paper/](https://portagenetwork.ca/news/discovery-white-paper/)


5 University of Alberta Libraries’ Dataverse network “to help researchers publish, analyze, distribute and preserve their data and datasets.” See [https://dataverse.library.ualberta.ca/dvn/](https://dataverse.library.ualberta.ca/dvn/)

6 Archivematica is a web- and standards-based, open-source application which allows institutions to preserve long-term access to trustworthy, authentic and reliable digital content. See [https://www.archivematica.org](https://www.archivematica.org)

7 Artefactual Systems Inc. is the lead developer for Archivematica software. See [https://www.artefactual.com/](https://www.artefactual.com/)
Repository Developments

The FRDR repository system is based on Globus Publication\(^8\), a customization of DSpace\(^9\). The Globus customizations include the integration of Globus Connect\(^{10}\) functionality, leveraging a widely used service for efficiently handling very large data transfers.

There were several significant developments in this quarter. Compute Canada signed an agreement with DataCite Canada in preparation for issuing DOIs (digital object identifiers) for submitted datasets. DOIs for datasets are important to be able to persistently cross-reference a dataset with its related published papers.

Additional access controls were added to support specifying an embargo period for data and to support withdrawing public access to data. A method was designed to import metadata from a file, to facilitate applying similar metadata to a number of dataset submissions without manual entry of metadata into a webform. A plan was made for integrating FRDR with ORCID\(^{11}\) identifiers for researchers.

Objectives for a third development phase (through November 2016) were proposed and approved by the Steering Committee.

The full set of objectives for this past development phase are listed in Appendix A.

Communications continues to be an issue

- It is clear that there needs to be additional communications regarding the project and the proposed service and that stakeholders need to be better informed and consulted regarding the project. To that end the Steering Committee is making available resources from the communications personnel at both CC and CARL. Additionally, the experience and contacts of the CARL Portage Expert Groups (in discovery, preservation, and curation) will be pursued.

Phase 3 Goals

The plan for the next phase (October-November 2016) is to get ready for alpha user testing with enough functionality for them to provide feedback on utility. The repository software for alpha testing:

- Will still be running on test infrastructure;
- The UBC code-based discovery interface will remain pointed at a test version of Globus Search Platform;

---

\(^8\) Globus Publication: Metadata is stored in the cloud, while published data is stored on campus, institutional, and group resources that are managed and operated by external administrators. See [https://www.globus.org/data-publication](https://www.globus.org/data-publication)

\(^9\) DSpace is an open source software package typically used for creating open access repositories for scholarly and/or published digital content. See [http://www.dspace.org/](http://www.dspace.org/)

\(^{10}\) Globus Connect is an efficient and user-friendly file transfer system based on GridFTP. See [https://www.globus.org/globus-connect](https://www.globus.org/globus-connect)

\(^{11}\) ORCID (Open Researcher and Contributor ID) is a code to uniquely identify academic authors and contributors. See [http://orcid.org/](http://orcid.org/)
- We will not commit to retaining or backing up any data deposited into the repository;
- We will not issue permanent DOIs;
- We will continue with interim "FRDR" branding

**Phase 3 (October-November 2016) Goals**

<table>
<thead>
<tr>
<th>Feature / Function / Improvement</th>
<th>Details / Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Expand the scope of harvesting</td>
<td>Harvest metadata from 8 repositories; current 5 plus:</td>
</tr>
<tr>
<td></td>
<td>• 1 DDI repo (U of A)</td>
</tr>
<tr>
<td></td>
<td>• 1 French repository (i.e., monolingual French metadata)</td>
</tr>
<tr>
<td></td>
<td>• 1 repository with geospatial data (preferably a domain-oriented repository)</td>
</tr>
<tr>
<td>2. Design and build a simulated Bring Your Own Storage (BYOS) system</td>
<td>Demonstrate the planned ability to incorporate additional data storage servers into FRDR—an additional Globus Endpoint that is part of the FRDR. The preservation aspects complicate this.</td>
</tr>
<tr>
<td></td>
<td>The proposed system includes the ability to include in the repository datasets that remain housed on systems at a particular institution. This is the second of the following list of ways data gets “into” FRDR:</td>
</tr>
<tr>
<td></td>
<td>1. <strong>Deposited into FRDR directly</strong> on CC provided hardware.</td>
</tr>
<tr>
<td></td>
<td>2. Institutional storage is affiliated with FRDR such that a dataset deposited into FRDR remains on institutional hardware (a “partner” installation.)</td>
</tr>
<tr>
<td></td>
<td>3. <strong>In-place ingestion.</strong> Metadata are entered into FRDR, but the data is never transferred. E.g., very large datasets that are already housed elsewhere in CC.</td>
</tr>
<tr>
<td></td>
<td>4. <strong>Indexed only.</strong> Metadata for datasets that are exposed by other repositories are harvested by FRDR.</td>
</tr>
<tr>
<td></td>
<td>5. Globus Search is intended to also provide search results for <strong>files or content available otherwise to Globus users</strong> (e.g., working files on a local server, or files on the user’s PC.)</td>
</tr>
<tr>
<td>3. Perform feature testing of deposit interfaces</td>
<td>Demonstrate one or more collections having custom metadata into which datasets can be deposited and which can be searched using specified values of that custom metadata.</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>4.</td>
<td><strong>Investigate where Canadian research data exists</strong></td>
</tr>
<tr>
<td>5.</td>
<td><strong>Write the documentation and test plans needed for alpha testers</strong></td>
</tr>
<tr>
<td>6.</td>
<td><strong>Find appropriate people for alpha testing of curation</strong></td>
</tr>
<tr>
<td>7.</td>
<td><strong>Find appropriate people for alpha testing of discovery and preservation</strong></td>
</tr>
<tr>
<td>8.</td>
<td><strong>Improvements and additions to Discovery</strong></td>
</tr>
<tr>
<td>9.</td>
<td><strong>Improvements and additions to Repository</strong></td>
</tr>
</tbody>
</table>
10. Improvements and additions to Preservation

- Document and publish a summary of changes and improvements to Archivematica to date (including graphs/metrics/analysis)
- Extract metadata from submitted datasets and ask the submitted to vet or edit that extracted metadata. The intention is to gather additional useful metadata to aid in discovery.
- Add feature: normalization of geospatial data

11. Policy work

- Create a draft Terms of Service (for researchers who are submitting items)
- Draft the license(s) that researchers can select to apply to a dataset submission

12. Infrastructure planning

- Find out from Compute Canada when its new hardware for high-availability services is going to be available, how it fits with the FRDR project plans
- Determine how planned CC Storage Services will support the FRDR testing, pilot and production services.
- Examine the appropriate ISO standard for levels of data security and evaluate how our implementation sizes up against the ISO standard

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 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 http://www.computecanada.ca/RDM
## Phase 2 (July-September 2016) Goals and Outcomes

<table>
<thead>
<tr>
<th>Goals for Oct 2016</th>
<th>Outcomes</th>
</tr>
</thead>
</table>
| 1) Implement ACL handling for data set files | ● We made the changes necessary to support embargoed and/or withdrawn items.  
● We have considered and implemented security checks in the workflow.  
● We have implemented an embargo checker that routinely ensures compliance. |
| 2) Communication and consultation, governance. Work on the project website. | ● CC newsletter article was published in July.  
● Web site computecanada.ca/RDM exists, but doesn’t do a good job of selling the project.  
● Response to published invitation to join a stakeholder group were poor. (Google Group in CC space may not be best; invitations need to be targeted, but maybe should come from CARL?)  
● We are awaiting Steering Committee direction on involvement of CC, CARL Communications staff. [They are now engaged as of September 2016]  
● While the Portage Discovery Expert Group may be good advisors, we have not firmed up the process for consultation. We will also need advisory groups for other areas: preservation, researchers, curation |
| 3) Archivematica performance improvements | ● There are three microservices targeted for speed improvement; we have sent patches to Artefactual:  
● “removeUnneededFiles” has been changed to run once per submission rather than once per file  
● “generateFileUUID” was merged with “updateFileAndChecksum” to halve the number of micro-service calls for generating basic file information.  
● pbzip2 was replaced by the faster lbzip2 parallel bzip implementation in “compressAIP” while still retaining the same compression ratio.  
● These speed improvements should decrease the ingestion time for a submission with a large number of files (~ 10⁴) up to 18%. |
| 4) Determine Canadian repos to be harvested and add them to the index | ● As of today the repositories from which we are harvesting are: UBC Circle, SFU Radar, Open Data Canada, Scholars Portal, FRDR itself  
● We have been adding repositories in a targeted fashion, selecting different repository types, metadata schemas, and harvesting protocols, while improving the harvesting process as we go.  
● We have identified candidates for addition in the next phase. |
<table>
<thead>
<tr>
<th>5) Analysis: ORCID integration</th>
<th>● We have quantified the necessary work for integrating ORCID with Globus Publication and the benefits for the different types of integration.</th>
</tr>
</thead>
</table>
| 6) Design a way to import metadata on item submission | ● Analysis and design is done.  
● Implementation is scheduled for next phase (Oct 1 - Nov 30, 2016). |
| 7) Analysis: Generation of DOIs through Datacite Canada | ● We have contacted DataCite Canada and received the necessary documentation  
● CC has signed an agreement with them and we received DataCite account as of 2016-08-29.  
● Creating DOIs will remain free until March of 2017 which will provide us time for testing  
● Implementation is scheduled for next phase (Oct 1 - Nov 30, 2016) |
| 8) Investigate features of other repositories that we may want | ● We polled the steering committee regarding their view of the importance of a number of features  
● We reviewed the report on the features of repositories done a couple of years ago: *Research Data Repositories: Review of Current Features, Gap Analysis, and Recommendations for Minimum Requirements* by Claire C. Austin, Susan Brown, Nancy Fong, Chuck Humphrey, Amber Leahey, Peter Webster. IASSIST Quarterly (2015). |
| 9) Implement an alpha version of a discovery front end | ● We investigated ANDS, UBC and Globus discovery code.  
Recommended UBC code to the Steering Committee.  
● Steering Committee approved and finalized decision on UBC Open Collections code.  
● Alpha version is running and works against Globus Search Platform.  
● Further work scheduled to be done next phase. |
| 10) Adopt use cases from Dspace and customize for this project | ● A selected set of use cases has not yet been assembled, proposed, customized, nor approved.  
● Hundreds of use cases are available (from JISC, from DSpace) and need to be pruned.  
● We posited a list of ‘provocative’ use cases to the Steering Committee to stimulate discussion. |
| 11) Analysis: what would need to be done in order to alpha test at the end of 2016? | ● Analysis was done, we should be ready for alpha testing Dec 2016  
● Preparation for alpha testing is scheduled for next phase (Oct 1 - Nov 30, 2016)  
● The steering committee and team agree on terminology (alpha, beta, pilot) |