Portaging Along: Developing a Collaborative National Research Data Management Network in Canada

Eugene Barsky, UBC
Lee Wilson, ACENET/Portage

Contact - eugene.barsky@ubc.ca
Spring 2018
Outline

● Background
● Tri-Agencies’ directions in Research Data Management (RDM)
● Portage’s national work
● Focus on Data Repositories and Discovery
● Federated Research Data Repository (FRDR) - a national discovery layer for research data
Data rich

Soccer clubs, like Arsenal, record on average 10 data points per second for every player on the field, or about 1.4 million data points per game.

Image - https://www.flickr.com/photos/kevlar/
Defining research data

Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results.


* Image - https://www.flickr.com/photos/34547181@N00/
Why data management

- In the USA

* From Developing data services: a tale from two Oregon universities - http://www.slideshare.net/amandawhitmire/20140618-rml-rendezvousfinal
TRI-AGENCY POLICY DEVELOPMENT
BACKGROUND

2013
Capitalizing on Big Data: Toward a Policy Framework for Advancing Digital Scholarship in Canada

2016
Tri-Agency Statement of Principles on Digital Data Management

2017-2018
Draft Tri-Agency Research Data Management Policy
Timeline

- Tri-Council to finalize RDM policy in April or May 2018.
- Public consultation for a period of two-three months.
- Six months after the policy has been publically available, institutions will be expected to enact RDM policies.
- Realistic timeline - Fall 2019 for compliance.

* Image - [https://www.flickr.com/photos/pamiine/](https://www.flickr.com/photos/pamiine/)
DRAFT
TRI-AGENCY DATA MANAGEMENT POLICY

- For consultation
- Feedback will inform final policy
- Proposed policy includes 3 possible requirements:
  1. Institutions: Institutional Strategy
  2. Researchers: Data Management Plans
  3. Researchers: Data Deposit
- Implementation: Phased, incremental
Tri-Agency expectations for RDM

Institutions:

- Institutional Data Strategy

- Provide researchers access to repositories that securely preserve, curate and provide access to research data

- Provide researchers with guidance to properly manage their data, including Data Management Plans (DMPs)
Tri-Agency expectations for RDM

Researchers:

- Incorporate RDM **best practices** (in their discipline), including **Data Deposit** for publications

- Develop **Data Management Plans** (DMPs)

- Follow institutional **policies** and standards

[Image - https://www.flickr.com/photos/jdhancock/]
Tri-Agency expectations for RDM

Funders:

- Develop **policy** and requirements that facilitate responsible data management
- Provide clear guidance for fulfill RDM requirements
- **Promote** the importance of excellent RDM
- Provide **peer-reviewers** with guidance for applications assessment

Image - [https://www.flickr.com/photos/sonson/](https://www.flickr.com/photos/sonson/)
What is the Portage Network?

● “Portage is a national, library-based research data management network that coalesces initiatives in research data management to build capacity and to coordinate activities better”

● Goals:
  ○ Build a community of practice for research data management (RDM)
  ○ Engage and advocate for research data management with stakeholder communities
  ○ Facilitate and provide leadership in the development of RDM infrastructure

● [https://portagenetwork.ca/](https://portagenetwork.ca/)
Portage Network of Experts

Expert Groups:

• Data Management Planning
• Curation
• Data Discovery
• Preservation
• Training
• Research Intelligence

Working Groups:

• Dataverse North
• FRDR Service Model
• Institutional Strategies
• Ethical Treatment of Sensitive Data
Regional Stakeholders

Canada’s four regional library consortia

Council of Prairie & Pacific University Libraries

Ontario Council of University Libraries

BCI

CAUL/CBUA
Part of a Larger RDM Ecosystem
Focus on Data Discovery

Find and Share Canadian Research Data

Find Data
Search FRDR to find research data sets originating from researchers affiliated with Canadian institutions. Data deposited to other repositories across Canada can also be found by searching in FRDR. See the growing list of collaborating repositories by performing a search.

Learn more »

Deposit Data
Any researcher affiliated with a Canadian institution can deposit data in FRDR at no direct cost. The platform can efficiently move data sets of any size, and preservation and archival is done automatically. Research librarians from CARL curate and approve deposited items.

Learn more »
FRDR Overview

- As you know, there are many research data repositories in Canada

- For instance, UBC Abacus Dataverse, Open Data Canada, Hakai Institute, and dozens more…

- We have worked to create the national research data discovery layer with **Federated Research Data Repository (FRDR)** - A scalable, federated platform for digital research data management and the discovery of Canadian research data - [https://www.frdr.ca/](https://www.frdr.ca/)
FRDR Stakeholders

- Partnership between Compute Canada (CC) and the Canadian Association of Research Libraries (CARL)
- Hosted on Compute Canada hardware and infrastructure, with CC providing development and technical support
- Service operated by Portage, including curation and data management support, with steering and input from CARL, the Network of Experts, and individual institutions
FRDR Discovery

FRDR’s harvester indexes data repositories across Canada to make research data held in many repositories discoverable from a single platform

Currently supports OAI-PMH, CKAN, CSW, Marklogic standards with plans to add more

Goals:

● supplement existing repository sites
● improve discovery
● breakdown repository siloing
● avoid being “just another repository”
FRDR Discovery

- Portage’s Data Discovery Expert Group identified and mapped 13 well-used and mature metadata standards to FRDR’s metadata model (Dublin Core/DataCite)
- Crosswalk emphasizes core elements across all standards, allowing varied discipline-specific metadata to be displayed in a single discovery interface
- Some detail/granularity lost when crosswalking to general standards (e.g., Dublin Core)
- Future work will explore more advanced ways of linking contextual metadata to FRDR (linked data approach)
<table>
<thead>
<tr>
<th>Field</th>
<th>EML</th>
<th>DCATs</th>
<th>Open Data Canada</th>
<th>Darwin Core</th>
<th>DATS</th>
<th>Protocol Data Element Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc:title</td>
<td>title</td>
<td>title</td>
<td>title</td>
<td>title</td>
<td>entity:Dataset property title</td>
<td>entity:Title</td>
</tr>
<tr>
<td>dc:creator</td>
<td>creator</td>
<td>creator</td>
<td>creator</td>
<td>creator</td>
<td>entity:Dataset property creators</td>
<td>entity:Creator</td>
</tr>
<tr>
<td>dc:subject</td>
<td>subject</td>
<td>subject</td>
<td>subject</td>
<td>subject</td>
<td>entity:Dataset property keywords</td>
<td>entity:Keywords</td>
</tr>
<tr>
<td>dc:description</td>
<td>description</td>
<td>description</td>
<td>description</td>
<td>description</td>
<td>entity:Dataset property description</td>
<td>entity:Abstract</td>
</tr>
<tr>
<td>dc:publisher</td>
<td>publisher</td>
<td>publisher</td>
<td>publisher</td>
<td>publisher</td>
<td>entity:Person property/fullName</td>
<td>entity:Publisher</td>
</tr>
<tr>
<td>dc:contributor</td>
<td>contributor</td>
<td>contributor</td>
<td>contributor</td>
<td>contributor</td>
<td>entity:Person property/fullName</td>
<td>entity:Contributor</td>
</tr>
<tr>
<td>dc:date</td>
<td>date</td>
<td>date</td>
<td>date</td>
<td>date</td>
<td>entity:Dataset property dates</td>
<td>entity:Date</td>
</tr>
<tr>
<td>dc:type</td>
<td>resource_type</td>
<td>resource_type</td>
<td>resource_type</td>
<td>resource_type</td>
<td>entity:Dataset property type</td>
<td>entity:Dataset</td>
</tr>
<tr>
<td>dc:format</td>
<td>resource_format</td>
<td>resource_format</td>
<td>resource_format</td>
<td>resource_format</td>
<td>entity:Dataset property type</td>
<td>entity:Format</td>
</tr>
<tr>
<td>dc:identifier</td>
<td>identifier</td>
<td>identifier</td>
<td>identifier</td>
<td>identifier</td>
<td>entity:Dataset property identifier</td>
<td>entity:ID</td>
</tr>
<tr>
<td>dc:language</td>
<td>language</td>
<td>language</td>
<td>language</td>
<td>language</td>
<td>entity:Dataset property language</td>
<td>entity:Language</td>
</tr>
</tbody>
</table>
FRDR Deposit

- A place for Canadian researchers to deposit large datasets
  - Big data transfer using **Globus File Transfer**
- A place to deposit datasets if researcher does not have a local or domain-specific option
- Support for custom metadata schemas
- Designed for **scalability**
- Storage may be distributed or managed centrally through infrastructure providers (e.g., Compute Canada)
FRDR Data Preservation

- Archivematica integration: **Digital preservation** processing for long-term usability of datasets
  - Converting file formats into future-friendly formats (e.g. docx-->PDF)
  - Creating Archival Information Packages (AIPs)
- Scalable, automated Archivematica processing for datasets up to 300 GB or 25,000 files (distributed over multiple VMs in CC Cloud)
FRDR - Feature List

- Direct deposit and download of datasets through Globus File Transfer
- Direct download of small datasets through HTTPS
- Automatic processing of datasets with Archivematica
- Support for custom metadata schemas
- Embargo support
- API for automated deposit
- Issuing DOIs through DataCite
- Bilingual user interface for both repository and discovery
- Indexing items from selected Canadian repositories
- Support for multiple licenses
- Faceted search in the discovery interface
- ORCID integration

Image - https://www.flickr.com/photos/danielygo/
Acknowledgements

- Steering committee: Dugan O’Neil, Jason Hlady, Jeff Moon, Umar Qasim, Lee Wilson, John Simpson, Jay Brodeur
- CARL/Portage experts: DDEG / CEG / PEG
- Portage Secretariat: Jeff Moon, Shahira Khair, Julie Morin, Lee Wilson
- CARL: Susan Haigh, Donna Bourne-Tyson, Kathleen Shearer
- UBC and the Open Collections team: Eugene Barsky, Schuyler Lindberg
- Compute Canada: Cloud East and Cloud West teams, Communications team, Translators, Support
- FRDR Development team: Alex Garnett, Keith Jeffrey, Todd Trann, Mike Winter, Adam McKenzie
- And a special thanks to the former Portage Director, Chuck Humphrey
Further Information

Production site:

https://www.frdr.ca/

Demonstration site:

https://demo.frdr.ca/

More information:

http://frdr.thedev.ca/

Thanks! Questions?

lee.wilson@ace-net.ca or eugene.barsky@ubc.ca