Portage Data Discovery Expert Group - Collections Development Working Group

Phase One Report

Prepared by the Portage Network, Collections Development Working Group of the Data Discovery Expert Group on behalf of the Canadian Association of Research Libraries (CARL)

Berenica Vejvoda (McGill University, Chair)
Alison Ambi (Memorial University of Newfoundland)
Eugene Barsky (University of British Columbia)
Kevin Lindstrom (University of British Columbia)
Heather MacDonald (Carleton University)
Kathleen Matthews (University of Victoria)
Michael Moosberger (Dalhousie University)
Lisa O'Hara (University of Manitoba)
Susan Powelson (University of Calgary)
Kimberly Silk (Canadian Research Knowledge Network)
Allison Sivak (University of Alberta)
Kristi Thompson (University of Windsor)

JULY 2017

Portage Network
Canadian Association of Research Libraries
portage@carl-abrc.ca

www.portagenetwork.ca

portage
SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA

CARL ABRC
CANADIAN ASSOCIATION OF RESEARCH LIBRARIES
ASSOCIATION DES BIBLIOTHÈQUES DE RECHERCHE DU CANADA

Table of Contents

Background

This working group is one of two working groups created to support the Portage Data Discovery Expert Group. The goal of the Collections Development Working Group is to ensure that Canadian research data is comprehensively included and indexed in the Federated Research Data Repository (FRDR) and other search tools to support its discovery and reuse[1].

Research Data Definition

The working group developed the following definition of research data for use in this project.

> *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[2].*

This definition is based on the definition of research data from the CASRAI Dictionary[3]**.** The CASRAI definition was modified to make it less broad and to eliminate the inclusion of digital and non-digital content that merely had "the potential of becoming research data."

[1] Terms of Reference: https://portagenetwork.ca/wp-content/uploads/2017/04/DDEG-CollectionsWG-TOR-EN.pdf

[2] CASRAI Dictionary. "Research Data". http://dictionary.casrai.org/Research_data

[3] Ibid.

Types of Research Data Repositories

In addition to defining research data, a typology[4] of research data repositories was created in order to help identify potential repositories for inclusion in the FRDR. The following list is ranked by curatorial, temporal, and archival commitment, progressing from minimal to maximum commitment.

1. **Staging repositories** used before transferring data to another repository and often supporting an active research project, such as Globus (e.g. using DSpace as a staging repository until there's a better alternative);
2. **Research data deposits** made with an institutional repository that is intended primarily for publications (e.g. research datasets from institutional repositories, such as DSpace, that are clearly marked as data - dc.type:Dataset -[5]);
3. **Research data repositories that primarily provide access** to research data without necessarily archiving the data;
4. **Generic digital repositories** that are owned by publishers, such as Dryad or Figshare, or that hold specialized digital collections, such as Canadiana;
5. **Publication-related dataset repositories** that support the replication of research findings in publications (e.g. Mendeley Data);
6. **Domain repositories** that hold research data shared within a specific domain (e.g. Canadian Astronomy Data Centre);
7. **Research data repositories providing access to, and the archiving of, research data** (e.g. Scholars Portal Dataverse or UBC Abacus Dataverse Network).

[4] Typology of research data repositories provided by Chuck Humphrey, Director of Portage
[5] http://dublincore.org/documents/2008/01/14/dcmi-type-vocabulary/

Repository Collection and Criteria Development

The Re3.data registry[6] was used as the initial source of research data repositories, as it provides access to one of the most extensive listings. DataCite Canada registered repositories[7] were also included. A focused search for humanities research data repositories was done to supplement the list. Based on feedback from the Portage Data Discovery Expert Working Group, the list was narrowed down to research data repositories hosted in Canada. This resulted in an initial list of roughly 170 research data repositories.

A list of criteria was developed based on the Re3.data filters and narrowed down by group discussion and consensus. The list of criteria includes:

- Research data repository type;
- Whether the repository is run by government, a university or other body;
- Whether contact information for technical support is readily presented;
- Whether the research data repository implemented a discernable and standardized metadata schema;
- Whether an API for harvesting metadata is discernable;
- Whether the repository utilized persistent identifiers such as a Handle or DOI;
- Whether a filter for limiting to data was available for those repositories that included other material.

[6] http://www.re3data.org
[7] https://search.datacite.org/

Identification of Pilot Research Data Repositories

Through the application of these criteria to the list of research data repositories, two criteria emerged as especially relevant: research data repository type and metadata schema. Some of the criteria were challenging to assess from the front-facing website - e.g. visible API and available data filter. Criteria that were not determined a priori, but which emerged organically and also guided the analysis, included whether the data repository was still active, the recency of updates to the repository website, as well as whether or not datasets were actually available for download. Data repositories that appeared to require a password for access were excluded.

The following 10 candidates were identified.

Research Data Repository	Contact Information
Canadian Opinion Research Archive (http://www.queensu.ca/cora)	Canadian Opinion Research Archive, School of Policy Studies Queen's University, Kingston, Ontario K7L 3N6 Email: cora@queensu.ca
Ocean Networks Canada (http://www.oceannetworks.ca)	Ocean Networks Canada University of Victoria PO BOX 1700 STN CSC Victoria, BC V8W 2Y2 Phone: 250.472.5400 Email: info@oceannetworks.ca
Polar Data Catalogue (https://www.polardata.ca)	Data Manager Gabrielle Alix Phone: 519-888-4567 x 37572 Email: galix@uwaterloo.ca
All Canadian University Dataverses	Dalhousie University (in development; no link yet) Ontario Council of University Libraries Email: dataverse@scholarsportal.info

	University of Alberta Libraries' Dataverse Network https://dataverse.library.ualberta.ca University of Manitoba (in development; no link yet) University of British Columbia, Simon Fraser University, University of Victoria and University of Northern British Columbia http://dvn.library.ubc.ca/dvn Contact: eugene.barsky@ubc.ca
Mouse Atlas of Gene Expression (http://www.mouseatlas.org/mouseatlas_index_html)	BC Cancer Agency / Michael Smith Genome Sciences Centre Phone: 604 707-5900
World Ozone and Ultraviolet Radiation Data Centre (http://woudc.org)	Meteorological Service of Canada Environment and Climate Change Canada 4905 Dufferin Street Toronto, ON M3H 5T4 Canada Email form: http://woudc.org/contact.php?lang=en
Ocean Tracking Network (OTN) (http://oceantrackingnetwork.org)	Lenore Bajona Director of Data Management Phone: (902) 494-7893 Email: lenore.bajona@dal.ca
Hakai Institute (https://www.hakai.org)	Staff listing: https://www.hakai.org/people/hakai-staff
BC Conservation Data Centre (http://www2.gov.bc.ca/gov/content/environment/plants-animals-ecosystems/conservation-data-centre)	Government of British Columbia Phone: 250-356-0928 Email: cdcdata@gov.bc.ca

Limitations and Issues Encountered

- Initially it was a challenge to clearly define and understand the scope and expectations for this working group;
- Many portals do not host research data: many registered with Re3data.org seemed to be research portals hosted by researchers, but are not data repositories per se;
- Different types of data in one repository with different metadata schema;
- Difficult to isolate datasets from non-datasets (i.e. due to absence of a separate metadata field that would allow filtering by research dataset type);
- Struggle to find humanities-based research data repositories (vs. repositories with humanities publications); this may be due to the fact that humanities data are text-based, as opposed to numerical, which may require a different approach;
- Collections with no explicitly stated standardized metadata schema;
- Various methods of restricted access were encountered.

Recommendations

The Collections Development Working Group recommends that:

- The following list of criteria be used to identify research data repositories for inclusion in FRDR:
    - Research data repository type;
    - Whether the repository is run by government, a university or other body;
    - Whether contact information for technical support is readily presented;
    - Whether the research data repository implemented a discernable and standardized metadata schema;
    - Whether an API for harvesting metadata is discernable;
    - Whether the repository utilized persistent identifiers such as a Handle or DOI;

    - Whether a filter for limiting to data was available for those repositories that included other material;
    - Subsequent work should address data duplication, multiple DOIs, and metadata duplication.
- In accordance with the research data definition:
    - data which has the "potential to become" research data be excluded since almost anything has the potential to become data.

Next Steps

The Collections Development Working Group recommends the following next steps:

- To further examine research data repositories that are hosted outside Canada but have significant Canadian content (e.g. Dryad and PANGAEA);
- To revisit the inclusion of Canadian government-hosted research data repositories which were excluded in the first Phase in order to manage scope;
- To make contact with the selected repositories to further flesh out criteria and verify criteria gleaned from front-facing websites;
- Make contact with the selected repositories to explicitly determine and confirm the ability to harvest the metadata and limit the harvest to research data content;
- Before identifying additional repositories, fine tune and evolve our criteria based on outcomes of contact with the pilot repositories.