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

Eugene Barsky, UBC

Lee Wilson, ACENET/Portage

Image by https://www.flickr.com/photos/40032755@N06/

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

Image - https://www.flickr.com/photos/kenfagerdotcom/

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/

Source - https://www.forbes.com/sites/bernardmarr/2015/03/25/big-data-the-winning-formula-in-sports/#2a9791e234de

Defining research data

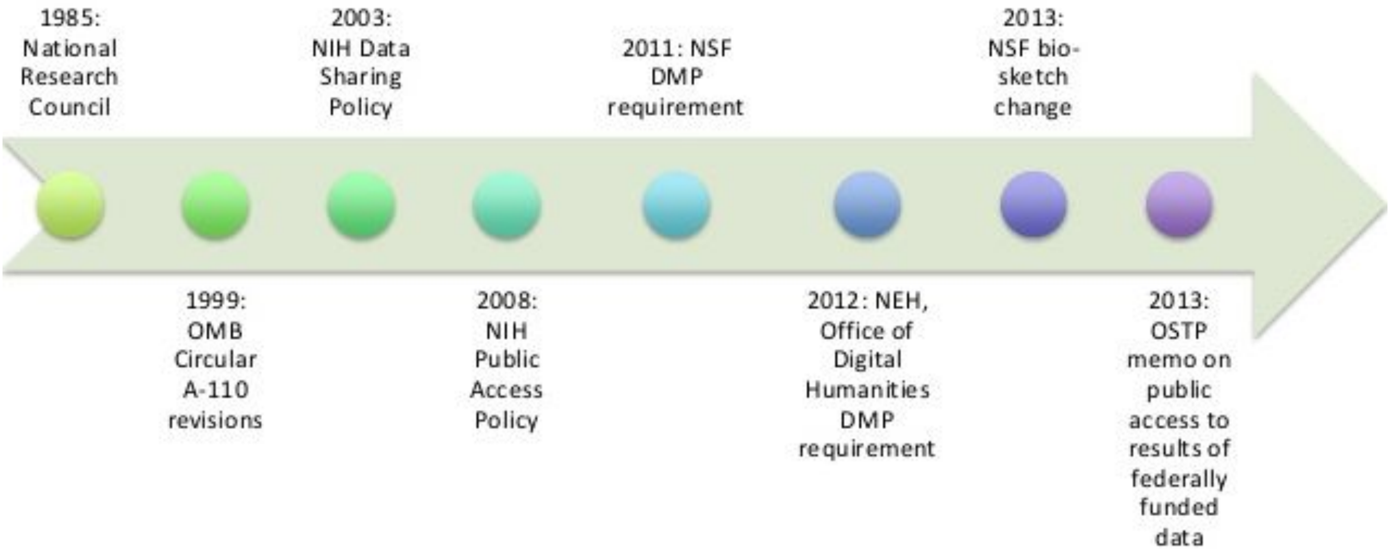
Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results.

Source - CASRAI Glossary - http://dictionary.casrai.org/Research_data

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

Why data management

- In the USA



1985: National Research Council

1999: OMB Circular A-110 revisions

2003: NIH Data Sharing Policy

2008: NIH Public Access Policy

2011: NSF DMP requirement

2012: NEH, Office of Digital Humanities DMP requirement

2013: NSF bio-sketch change

2013: OSTP memo on public access to results of federally funded data

* 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

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

NSERC
CRSNG

SSHRC CRSH

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.

2
Smith-Synchronome
Master Frequency Control Clock
STANDARD
SMITHS ENGLISH CLOCKS LTD
179 PORTLAND ST
LONDON W.I
TIME
Patent Nos 362848. 2634/31
MADE IN ENGLAND

* Image - https://www.flickr.com/photos/pamilne/

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

CIHR IRSC
Canadian Institutes of Health Research Instituts de recherche en santé du Canada

NSERC CRSNG

SSHRC CRSH

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)

image -https://www.flickr.com/photos/hms831/

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/

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/

Portage Network of Experts

Expert Groups:

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

Working Groups:

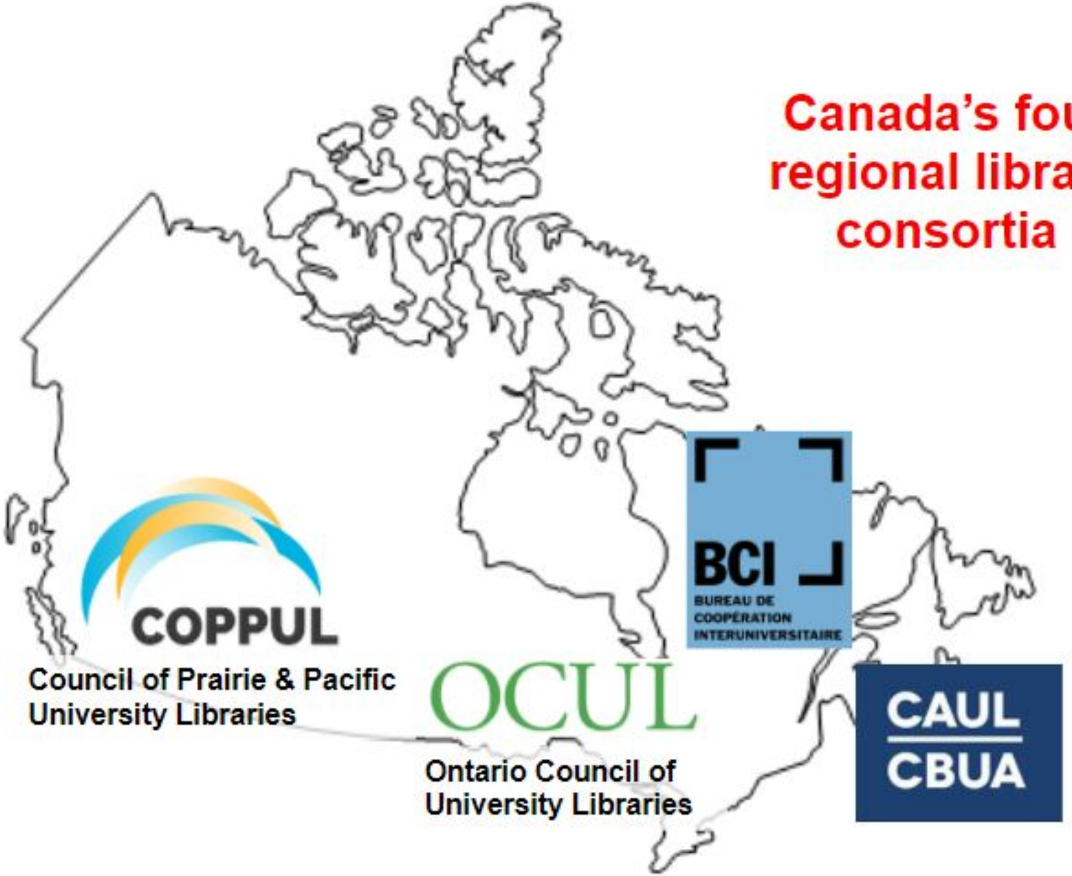
- Dataverse North
- FRDR Service Model
- Institutional Strategies
- Ethical Treatment of Sensitive Data

110+
Experts

40+
Organizations

Regional Stakeholders

Canada's four regional library consortia



COPPUL

Council of Prairie & Pacific University Libraries

OCUL

Ontario Council of University Libraries

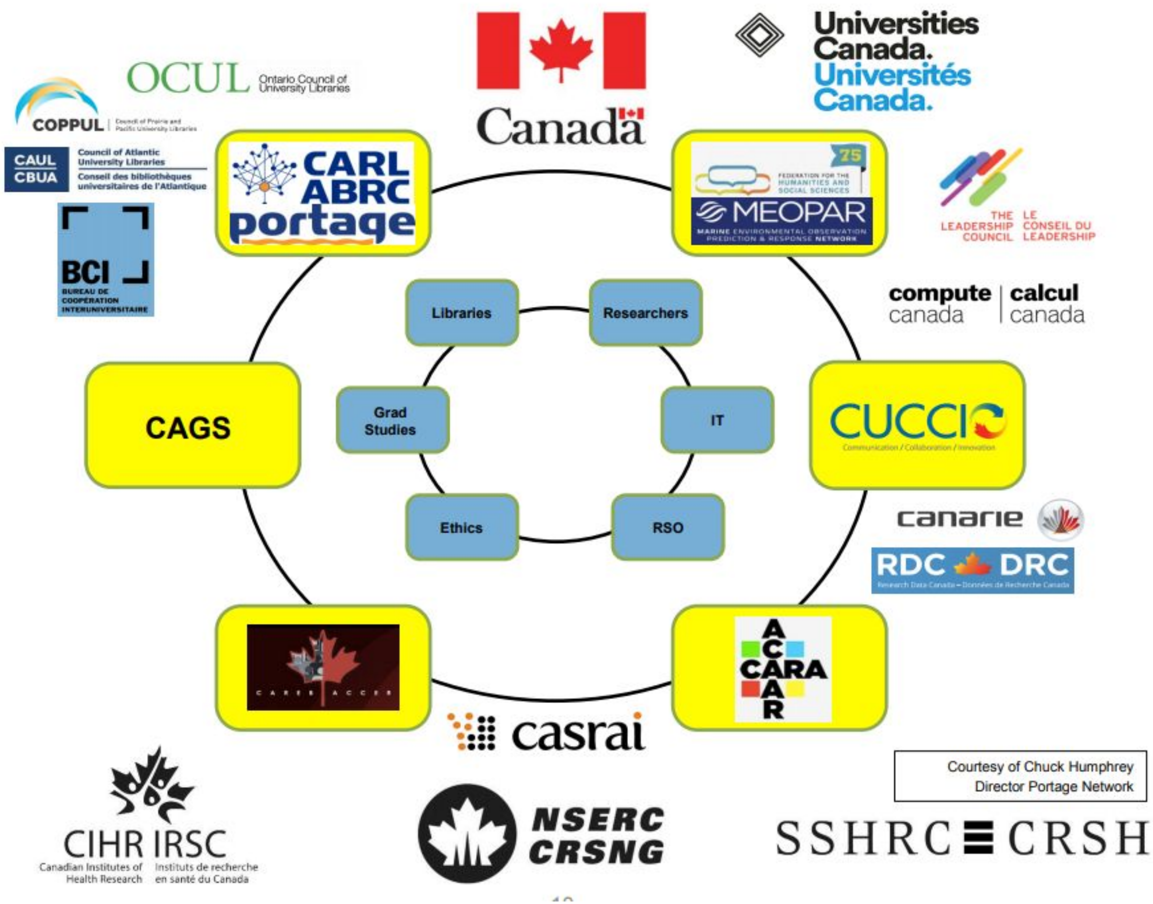
BCI

BUREAU DE COOPÉRATION INTERUNIVERSITAIRE

CAUL

CBUA

Part of a Larger RDM Ecosystem



Libraries

Researchers

IT

RSO

Ethics

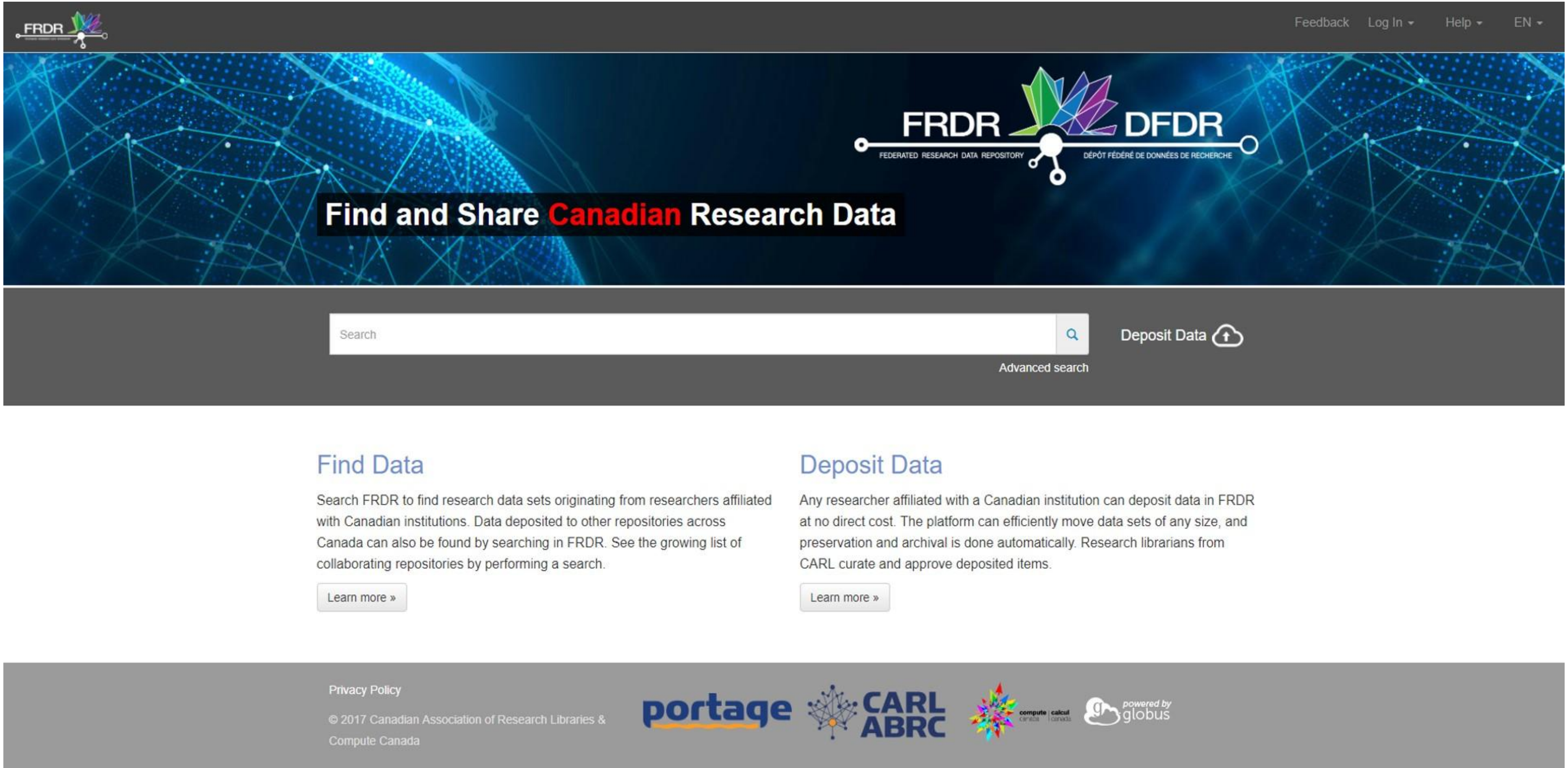
Grad Studies

CAGS

Courtesy of Chuck Humphrey
Director Portage Network

Canada
Universities Canada. Universités Canada.
OCUL Ontario Council of University Libraries
COPPUL Council of Prairie and Pacific University Libraries
CAUL CBUA Council of Atlantic University Libraries Conseil des bibliothèques universitaires de l'Atlantique
BCI BUREAU DE COOPÉRATION INTERUNIVERSITAIRE
CARL ABRC portage
75 FEDERATION FOR THE HUMANITIES AND SOCIAL SCIENCES
MEOPAR MARINE ENVIRONMENTAL OBSERVATION PREDICTION & RESPONSE NETWORK
THE LEADERSHIP COUNCIL LE CONSEIL DU LEADERSHIP
compute canada | calcul canada
CUCCIO Communication / Collaboration / Innovation
canarie
RDC DRC Research Data Canada – Données de Recherche Canada
CARES ACCES
ACCARAAR
casrai
CIHR IRSC Canadian Institutes of Health Research Instituts de recherche en santé du Canada
NSERC CRSNG
SSHRC CRSH

Focus on Data Discovery



FRDR
Feedback
Log In
Help
EN
FRDR
DFDR
FEDERATED RESEARCH DATA REPOSITORY
DÉPÔT FÉDÉRÉ DE DONNÉES DE RECHERCHE
Find and Share Canadian Research Data
Search
Deposit Data
Advanced search

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 »

Privacy Policy

© 2017 Canadian Association of Research Libraries & Compute Canada

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/

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

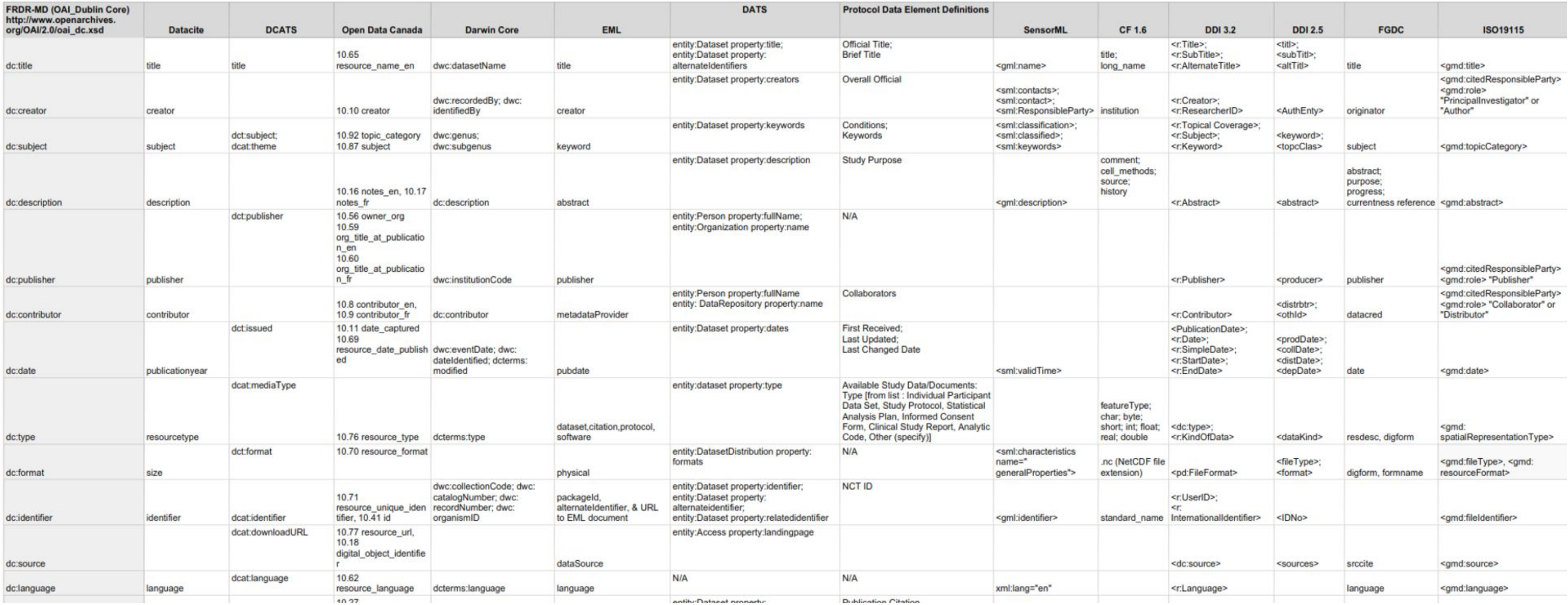
Goals:

- 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)

FRDR Discovery



FRDR-MD (OAI_Dublin Core) http://www.openarchives.org/OAI/2.0/oai_dc.xsd	Datacite	DCATS	Open Data Canada	Darwin Core	EML	DATS	Protocol Data Element Definitions	SensorML	CF 1.6	DDI 3.2	DDI 2.5	FGDC	ISO19115
dc:title	title	title	10.65 resource_name_en	dwc:datasetName	title	entity:Dataset property:title; entity:Dataset property: alternateIdentifiers	Official Title; Brief Title	<gml:name>	title; long_name	<r:Title>; <r:SubTitle>; <r:AlternateTitle>	<titl>; <subTitl>; <altTitl>	title	<gmd:title>
dc:creator	creator		10.10 creator	dwc:recordedBy; dwc: identifiedBy	creator	entity:Dataset property:creators	Overall Official	<sml:contacts>; <sml:contact>; <sml:ResponsibleParty>	institution	<r:Creator>; <r:ResearcherID>	<AuthEnty>	originator	<gmd:citedResponsibleParty> <gmd:role> "PrincipalInvestigator" or "Author"
dc:subject	subject	dct:subject; dcat:theme	10.92 topic_category 10.87 subject	dwc:genus; dwc:subgenus	keyword	entity:Dataset property:keywords	Conditions; Keywords	<sml:classification>; <sml:classified>; <sml:keywords>		<r:Topical Coverage>; <r:Subject>; <r:Keyword>	<keyword>; <topcClas>	subject	<gmd:topicCategory>
dc:description	description		10.16 notes_en, 10.17 notes_fr	dc:description	abstract	entity:Dataset property:description	Study Purpose	<gml:description>	comment; cell_methods; source; history	<r:Abstract>	<abstract>	abstract; purpose; progress; currentness reference	<gmd:abstract>
dc:publisher	publisher	dct:publisher	10.56 owner_org 10.59 org_title_at_publication_en 10.60 org_title_at_publication_fr	dwc:institutionCode	publisher	entity:Person property:fullName; entity:Organization property:name	N/A			<r:Publisher>	<producer>	publisher	<gmd:citedResponsibleParty> <gmd:role> "Publisher"
dc:contributor	contributor		10.8 contributor_en, 10.9 contributor_fr	dc:contributor	metadataProvider	entity:Person property:fullName entity: DataRepository property:name	Collaborators			<r:Contributor>	<distrbtr>; <othId>	datacred	<gmd:citedResponsibleParty> <gmd:role> "Collaborator" or "Distributor"
dc:date	publicationyear	dct:issued	10.11 date_captured 10.69 resource_date_published	dwc:eventDate; dwc: dateIdentified; dcterms: modified	pubdate	entity:Dataset property:dates	First Received; Last Updated; Last Changed Date	<sml:validTime>		<PublicationDate>; <r:Date>; <r:SimpleDate>; <r:StartDate>; <r:EndDate>	<prodDate>; <collDate>; <distDate>; <depDate>	date	<gmd:date>
dc:type	resourcetype	dcat:mediaType	10.76 resource_type	dcterms:type	dataset,citation,protocol, software	entity:dataset property:type	Available Study Data/Documents: Type [from list : Individual Participant Data Set, Study Protocol, Statistical Analysis Plan, Informed Consent Form, Clinical Study Report, Analytic Code, Other (specify)]		featureType; char; byte; short; int; float; real; double	<dc:type>; <r:KindOfData>	<dataKind>	resdesc, digform	<gmd: spatialRepresentationType>
dc:format	size	dct:format	10.70 resource_format		physical	entity:DatasetDistribution property: formats	N/A	<sml:characteristics name=" generalProperties">	.nc (NetCDF file extension)	<pd:FileFormat>	<fileType>; <format>	digform, formname	<gmd:fileType>, <gmd: resourceFormat>
dc:identifier	identifier	dcat:identifier	10.71 resource_unique_identifier, 10.41 id	dwc:collectionCode; dwc: catalogNumber; dwc: recordNumber; dwc: organismID	packageId, alternateIdentifier, & URL to EML document	entity:Dataset property:identifier; entity:Dataset property: alternateidentifier; entity:Dataset property:relatedidentifier	NCT ID	<gml:identifier>	standard_name	<r:UserID>; <r: InternationalIdentifier>	<IDNo>		<gmd:fileIdentifier>
dc:source		dcat:downloadURL	10.77 resource_url, 10.18 digital_object_identifier		dataSource	entity:Access property:landingpage				<dc:source>	<sources>	srccite	<gmd:source>
dc:language	language	dcat:language	10.62 resource_language	dcterms:language	language	N/A	N/A	xml:lang="en"		<r:Language>		language	<gmd:language>
			10.27			entity:Dataset property:	Publication Citation						

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/

Image - https://www.flickr.com/photos/debord/

Thanks! Questions?

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