Dataset Discovery and Metadata

What's in it for researchers?

Alicia Urquidi Díaz, MLIS, PhD (ella/elle/she/sie) • aurquidi@oceannetworks.ca
WDS-ITO • Portage Webinar Series • May 19, 2021
Vancouver is on the traditional, ancestral and unceded territories of the ʷməθkwəy̓əm (Musqueam), Skwxwú7mesh (Squamish), and Selílwitulh (Tsleil-Waututh) Nations.

James Bay (Victoria, BC) is on the traditional territory of the Lekwungen-speaking People who are known today as the Songhees.

To this day, the Songhees, Esquimalt and WSÁNEĆ nations maintain relationships with the land occupied by the city of Victoria and the University of Victoria, just as the Musqueam, Squamish and Tsleil-Waututh do with the city of Vancouver.

For information on these and other Indigenous territories, please visit: Native Land Digital Map
Today’s Webinar

1. Metadata & Discovery
   Researchers, metadata, and discovery infrastructures.

2. Bringing Data Together
   An insider look into FRDR’s Discovery Service
   Kelly Stathis, Discovery & Metadata Coordinator, Portage Network

3. Tips & tools from WDS-ITO
Data discovery

The ability to derive (or discover) new information and knowledge from existing data sources.


Metadata

The information we create, store, and share to describe things, [which] allows us to interact with these things to obtain the knowledge we need.

Understanding Metadata (Riley & Niso: 2017)
Descriptive metadata

Information to facilitate the discovery (via search or browse) of resources, or provide contextual information useful in the understanding or interpretation of a resource.

 Understanding Metadata (Riley & Niso: 2017)

Dublin Core Elements

<table>
<thead>
<tr>
<th>Rights</th>
<th>Contributor</th>
<th>Creator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>Coverage</td>
<td>Title</td>
</tr>
<tr>
<td>Publisher</td>
<td>Identifier</td>
<td>Description</td>
</tr>
<tr>
<td>Type</td>
<td>Date</td>
<td>Source</td>
</tr>
<tr>
<td>Relation</td>
<td>Format</td>
<td>Language</td>
</tr>
</tbody>
</table>

Core elements of the Dublin Core metadata standard. (Image source)
The potential for **reusing research data** is inextricably tied to how discoverable these data are to other researchers.

Research Data Discovery and the Scholarly Ecosystem in Canada (E. Barsky, J. Brosz & A. Leahey: 2016)
Also you

Discovery portal visitors (potential users of your research data)

Aggregation & Discovery Services

Data repository

You
Bringing Data Together

An insider look into FRDR’s Discovery Service
Kelly Stathis, Discovery & Metadata Coordinator, Portage Network
Bringing Canadian Research Data Together: The FRDR Discovery Service

Kelly Stathis (they/elle)
Discovery & Metadata Coordinator, Portage Network
kelly.stathis@engagedri.ca
May 19, 2021
What is the FRDR Discovery Service?

- Federated Research Data Repository (FRDR) / Dépôt fédéré de données de recherche (DFDR)
- Collaboration between the New Digital Research Infrastructure Organization’s Portage Network (NDRIO Portage) and Compute Canada
- FRDR Discovery Service:
  - One of FRDR’s three components (discovery, deposit, and preservation)
  - A national discovery layer indexing Canadian research data repositories

https://www.frdr-dfdr.ca/
What is the FRDR Discovery Service?

Metadata Harvester
(Python)

harvest database
(postgres)

CKAN
Dataverse
Socrata
OpenData Soft
OAI-PMH
Other types

Find and Share Canadian Research Data

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

Deposit Data
Any researcher affiliated with a Canadian institution can deposit data into FRDR. The platform can efficiently ingest datasets of any size, and preservation processing is done automatically. Data professionals from the Portage Network and institutions across Canada work with researchers to curate and approve deposited items.

https://www.frdr-dfdr.ca/
What are the benefits?

- Improve discovery of Canadian research (meta)data
- Break down repository siloes
- Drive traffic to existing repository sites
- Create interoperability between Canadian and international platforms

Diagram:
- Metadata Harvester (Python)
- FRDR Discovery Service
- OAI-PMH for all harvested records
- Data Citation Index
- ProQuest Central Discovery Index (CDI)
Collaborating repositories

- Over 80 discoverable research data repositories including:
  - University repositories, including Scholars Portal Dataverses
  - Government repositories at the federal, provincial, and local levels
  - Domain-specific repositories
  - Datasets deposited in FRDR

https://www.frdr-dfdr.ca/discover/html/repository-list.html
UBC Dataverse (Scholars Portal)

Metadata Harvester (Python)

harvest database (postgres)

https://doi.org/10.5683/SP2/TYN44T
The Impacts of Burn Severity on Forest Resiliency in British Columbia

Understanding how resilient forests are after wildfire events is important to forest management practices. The objective of this study was to use Landsat-8 data to understand how the burn severity of the Little Buntall Lake wildfire has impacted forest regrowth several years later. This was done by deriving different vegetation indices to see how the changes in vegetation health were impacted by burn severity. Additionally, landscape pattern metrics were used to understand the changes in the spatial patterns of the burn severity and vegetation health over time. The results showed that the higher the burn severity, the greater the impact on vegetation health immediately after the wildfire, as well as a slower return to pre-fire conditions. When comparing to the pre-fire values, the post-fire Normalized Difference Vegetation Index and the Tasseled Cap Greenness values showed an initial drop in the NDVI and a rise in value four years after the wildfire. This differed from the Tasseled Cap Wetness values as the NDVI showed a continued drop in value. When looking at the changes in the spatial patterns for vegetation health, the Sparse Vegetation class showed the greatest change for Core Area with a 3.7 to 3.2 hectares increase every year. When looking at burn severity, the High Severity class showed the greatest decrease in core area with a loss of 2.6 to 3.2 hectares every year. These results show that monitoring vegetation regrowth can be done using Landsat-8 derived vegetation health indices as well as with spatial pattern analysis.

https://doi.org/10.5683/SP2/TYN44T
UBC Dataverse (Scholars Portal)

Metadata Harvester (Python)

harvest database (postgres)

https://github.com/frdr-dfdr/frdr_harvest
**UBC Dataverse (Scholars Portal)**

- Dataverse
- CKAN
- Socrata
- OpenDataSoft
- OAI-PMH
- Other types

**Metadata Harvester (Python)**

**FRDR Discovery Service**

- Harvest database (postgres)

https://www.frdr-dfdr.ca/
Metadata Harvester

UBC Dataverse (Scholars Portal)

Dataverse

CKAN

Socrata

OpenData Soft

OAI-PMH

Other types

harvest database (postgres)

FRDR Discovery Service

https://www.frdr-dfdr.ca/
 Metadata Harvester (Python)

Discourse

harvest database (postgres)

UBC Dataverse (Scholars Portal)

CKAN

Dataverse

OAI-PMH

Socrata

OpenData Soft

Other types

FRDR

https://www.frdr-dfdr.ca/
Metadata Harvester (Python)

Geodisy

harvest database (postgres)

Important: Geodisy provides map search functionality to supplement the FRDR discovery service. Currently in beta, the map search includes datasets from repositories indexed by FRDR with bounding box metadata. Dataverse repository datasets with location metadata and/or geospatial files are also included. To perform a text only search, return to frdr-dfdr.ca.

Explore Canadian research data

Search by location

To use the map search:
1. Move the map to display your area of interest. You can also hold the shift key and click to draw a box for the map to zoom toward
2. Click "Search Here" to see the results

https://geo.frdr-dfdr.ca/
Metadata Harvester (Python)

Geodisy

harvest database (postgres)

https://geo.frdr-dfdr.ca/
Harvest database (postgres)

UBC Dataverse (Scholars Portal)

CKAN

Dataverse

Socrata

OAI-PMH

OpenData Soft

Other types

Metadata Harvester (Python)

Geodisy

The Impacts of Burn Severity on Forest Resiliency in British Columbia (2 of 3)

Author(s): Murray, Brant

Description: Understanding how resilient forests are after wildfire events is important to forest management practices. The objective of this study was to use Landsat-8 data to understand how the burn severity of the Little Bobtail Lake wildfire has impacted forest regrowth several years later. This was done by deriving different vegetation indices to see how the changes in vegetation health were impacted by burn severity. Additionally, landscape pattern metrics were used to understand the changes in the spatial pattern...

Collection: The Impacts of Burn Severity on Forest Resiliency in British Columbia

Location: Canada, British Columbia, Little Bobtail Lake, Canada, British Columbia, Prince George

Institution: UBC Dataverse

Data Source: https://doi.org/10.5281/zenodo.3565405

https://geo.frdr-dfdr.ca/
OAI-PMH for all harvested records

ProQuest Central Discovery Index (CDI)

Metadata Harvester (Python)

UBC Dataverse (Scholars Portal)

Dataverse

CKAN

Socrata

OpenData Soft

Other types

OAI-PMH

harvest database (postgres)

FRDR Discovery Service

OAI-PMH for all harvested records

OpenAIRE EXPLORER

Data Citation Index

ProQuest Central Discovery Index (CDI)

(Open Archives Initiative Protocol for Metadata Harvesting)
OAI-PMH for all harvested records

ProQuest Central Discovery Index (CDI)

Data Citation Index

ProQuest Central Discovery Index (CDI)

https://explore.openaire.eu/
OAI-PMH for all harvested records

ProQuest Central Discovery Index (CDI)

Metadata Harvester (Python)

UBC Dataverse (Scholars Portal)

Data Citation Index

FRDR Discovery Service

OpenAIRE Explode

ProQuest Central Discovery Index (CDI)

harvest database (postgres)

Other types

OAI-PMH

OpenData Soft

CKAN

Dataverse

Socrata

The Impacts of Burn Severity on Forest Resilience in British Columbia

UBC Dataverse (Scholars Portal)
Discovery & Metadata Tips
Discovery Activity

1. Go to one of the platforms listed
2. Do a keyword search: *earthquake monitoring vancouver island*
3. Find the dataset that has 2 co-authors named Martin

Finding and Being Found through Discovery Services

Discovery Platforms

- OpenAIRE Explore
- WorldWideScience.org
- Google Dataset Search
- DataCite Commons
- DCI Web of Science*  
  *if you have access to it
Original dataset

Römer, Miriam; Riedel, Michael; Scherwath, Martin; Heesemann, Martin; Spence, George D (2016): (Table S1) Earthquakes detected in RSAM and BPR record with date and time, magnitude and location with distance to sonar-site. PANGAEA, https://doi.org/10.1594/PANGAEA.868921.

Full metadata: https://doi.pangaea.de/10.1594/PANGAEA.868921?format=metadata_jsonld
(Meta)data Deposit

PANGAEA

DOI creation

Aggregation & Discovery

5 researchers
1 dataset
1 subject repository
1 DOI minting agency
6+ different international search platforms

Römer, Miriam; Riedel, Michael; Scherwath, Martin; Heesemann, Martin; Spence, George D (2016): (Table S1) Earthquakes detected in RSAM and BPR record with date and time, magnitude and location with distance to sonar-site. PANGAEA, https://doi.org/10.1594/PANGAEA.868921,
Many sources, but only one dataset → DOIs ftw!

Tip #1

DOIs
Unique, persistent identifiers for digital objects that allow your dataset to be found, referred to and linked unambiguously.

Source: https://datasetsearch.research.google.com/
Source: https://explore.openaire.eu/search/find

OpenAIRE and Google Dataset Search
Tip #2
ORCiDs
Free, unique identifiers that researchers create and manage themselves.
Tip #3
Include machine-readable rights metadata

Information a human or machine needs to provide appropriate access to a resource, provide appropriate notification and compensation to rights holders, and to inform end users of any use restrictions that may exist.

Riley (2011): Seeing Standards

Many discovery platforms index datasets by access mode and kind of license, which powers the filters you can apply in your search (Source: https://explore.openaire.eu/search/find, https://commons.datacite.org/)
Tip #3b

Indigenous Data Sovereignty

“The right of Indigenous peoples to control data from and about their communities and lands, articulating both individual and collective rights to data access and to privacy.

IDS also raises overarching, “fundamental questions about assumptions of ownership, representation, and control in open data communities”
Tip #4
Add spatial metadata and view your dataset on a map

Google Dataset Search, FRDR’s Geodisy, and other discovery platforms have map search functionalities, which rely on geolocation fields the metadata record.

### Spatial metadata

**Coverage:**
- Median Latitude: 49.192688
- Median Longitude: -131.117628
- South-bound Latitude: -10.800000
- West-bound Longitude: 102.890000
- North-bound Latitude: 56.298000
- East-bound Longitude: -102.180000

**Date/Time Start:** 2012-06-19T13:40:00
**Date/Time End:** 2013-06-21 05:18:00

**Minimum Elevation:** -1250.0 m
**Maximum Elevation:** -1250.0 m

**Event(s):**
- **NEPTUNE**
  - Latitude Start: 24.670000
  - Longitude Start: -110.170000
  - Latitude End: 40.180000
  - Longitude End: -121.030000
  - Elevation Start: -1250.0 m
  - Elevation End: -1250.0 m
  - **Location:** off west coast of Vancouver Island, British Columbia
  - **Method/Device:** Underway cruise track measurements (CT)
  - **Comment:** Ocean Networks Canada’s North-East Pacific Time Series Underwater Networked Experiments (NEPTUNE) cabled ocean observatory.

Source: [https://doi.pangaea.de/10.1594/PANGAEA.868921](https://doi.pangaea.de/10.1594/PANGAEA.868921)
Tip #5

Standard terminology

Use standardized terms and (wherever possible) controlled vocabularies, and link to and use specific terminologies in your research field.

And → Metadata is for humans too! Use descriptive documents and free-text fields to include human-readable summaries.

Describing your dataset

Earthquakes detected in RSAM and BPR record with date and time, magnitude and location (Latitude and Longitude) with distance to sonar-site. The azimuth is the angle between sonar and earthquake location. Time delays for P- and S-wave were calculated using standard Earth reference velocity model.
Tip #5

Standard terminology (cont)

The inclination to each other of two intersecting lines, measured by the arc of a circle between the two lines forming the angle, the center of the circle being the point of intersection. An angle may be classified as following:

- less than 90°; a right angle 90°; an obtuse angle, more than 90° but less than 180°; a straight angle, 180°; a reflex angle, more than 180° but less than 360°; a perigon, 360°. Any angle not a multiple of 90° is an oblique angle. If the sum of two angles is 90°, they are complementary angles; if 180°, supplementary angles; if 360°, explementary angles. Two adjacent angles have a common vertex and a common side. A dihedral angle is the angle between two intersecting planes. A spherical angle between two intersecting great circles.
Finding recommended repositories

Tip #6
CoreTrustSeal
An international, community based, non-governmental, and non-profit organization that promotes sustainable and trustworthy data infrastructures.

- Find CTS certified repositories
- Canadian 2021 CTS certification cohort

R13. Data discovery and identification
- Inclusion in disciplinary or generic resource registries
- Presence of search facilities with
  - A searchable metadata catalogue using international standards
  - Persistent identifier systems
  - Enabled machine harvesting of the metadata
  - Recommended data citations

R14. Data reuse
- Metadata provided with dataset download
  - In formats used by target community
  - Ensure continued data understandability
  - Account for the possible evolution of formats

Source: DOI:10.5281/zenodo.3638211
Tip #7

**Metadiscovery Tools**

You can find more repositories, as well as other discovery & aggregation platforms in these service registries:

- [re3data](#)
- [EOSC Marketplace](#)
- [Scientific Data’s recommended repository list](#)
- [WDS-ITO’s Searchable Index of Metadata Aggregators](#) (will be available soon under [doi.org/10.5281/zenodo.4589050](https://doi.org/10.5281/zenodo.4589050))

---

**Finding research data services**

the “meta-meta catalogues”
Tip #7

Schema.org

A few repositories and discovery services are now offering automatic semantic enrichment to make your metadata understandable to search engines such as Google, and other semantic digital tools.

WDS-ITO has some interesting resources on Schema.org and semantic markup:

- **Portage Webinar: Schema.org for Research Data Managers** (by Chantelle Verhey)
- **Schema.org for Research Data Managers: A Primer** (Verhey & Payne, 2020, will be published in the near future).
- **Schema Crosswalks Visualization** (in collaboration with the Research Data Alliance’s Research Metadata Schemas Working Group)

Semantic Metadata

WDS-ITO • Webinar: Metadata for Discovery • Alicia Urquidi Díaz, aurquidi@oceannetworks.ca • May 19, 2021
Call for Applications: NDRIO–Portage COVID-19 Data Curation Funding

NDRIO and Portage are pleased to announce the availability of funding to support the curation of Canadian research data related to the COVID-19 pandemic. The purpose of this funding is to bring COVID-19 related data into timely compliance with the FAIR Guiding Principles (Findable, Accessible, Interoperable, and Reusable) for the management of research data, algorithms, tools and workflows. Ultimately, increasing the FAIRness of datasets makes them discoverable and re-usable by both humans and machines, enabling downstream research. As a secondary objective, this initiative aims to build capacity for research data management (RDM) in Canada by supporting the training of highly-qualified personnel (HQP) in data stewardship and curation.

21 May: Deadline for applications to NDRIO-Portage COVID-19 Data Curation Funding (for Canadian researchers and research programs)
Thank you!