

Opening doors to data

Grant Gibson

Canadian Research Data Centre Network

Date: 2025-06-15



Introduction

- Thanks to SSHRC for supporting this “Opening doors to data” project via the Connections Grant program: RDM stream
- Thanks to CRDCN and its funders: CFI, CIHR, SSHRC, Provincial Partners and Collaborating Institutions.

Please access the project folder for this workshop:
www.github.com/CRDCN/RAD_training

Introduction

- Evolution of data discovery & open science
- Journal pressures
- Council pressures
- University pressures

Containing this

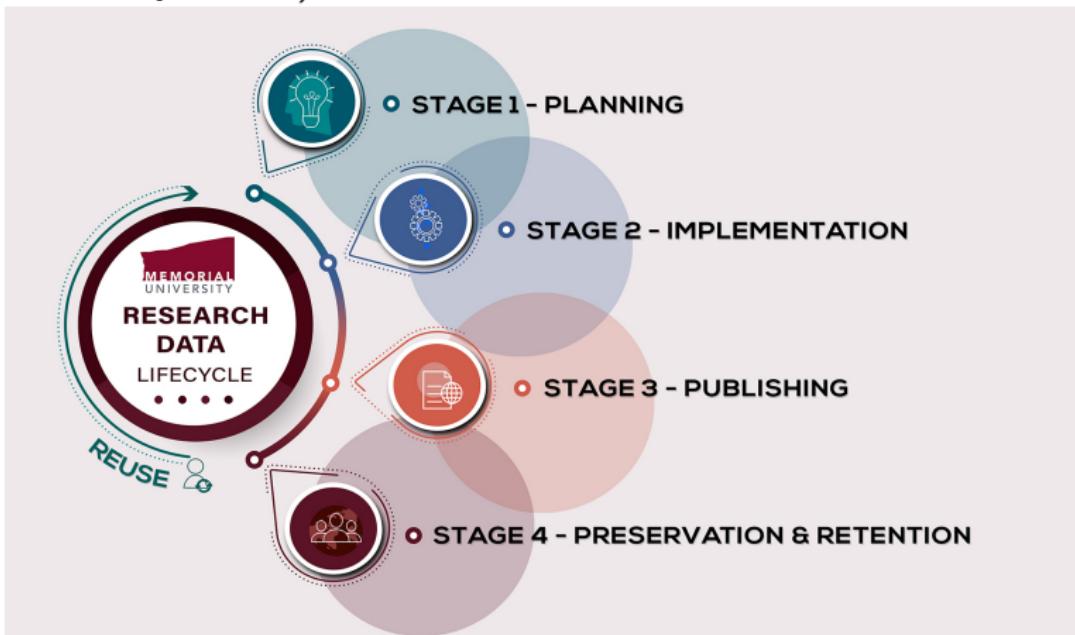
- Today we're talking about restricted access data
- Not all sensitive data, and not open data, but data that have an access process or mechanism

Targets

- FAIR data
- “As open as possible, as closed as necessary”
- This is not a binary, but a continuum

Data lifecycle

- Image of data lifecycle & description (credit: Memorial University OVPR)



Researcher support for discoverability

F - *Findable*

Researcher support for discoverability

F - *Findable*

Can anyone discover that your data even exist?

Can others even figure out where the data you used are?

Concepts:

- Persistent Identifiers
- Indexing (including metadata)

Researcher support for discoverability

A - *Accessible*

Researcher support for discoverability

A - *Accessible*

Can others access the data you used?

Can they figure out HOW to do so?

Concepts:

- Data accessibility statement
- Access metadata
- Transparent process

Researcher support for discoverability

I - *Interoperable*

Researcher support for discoverability

I - *Interoperable*

Concepts:

- Open source
- Machine-readable
- Metadata
- Control vocabularies

Researcher support for discoverability

R - *Reusable*

Researcher support for discoverability

R - *Reusable*

Concepts:

- Provenance
- Licensing
- Archiving

Case study: CBS

- Canadian Blood Services offers secondary use research data (that is, databases about donors that you can request)
 - Suppose you were a researcher interested in research on data about donors
- ① Think about a research question and evaluate whether you could answer it
 - ② Outline the process you would follow to get the data
- We'll take up these questions and discuss how "FAIR" we think the data are.

Case study: CBS

- Canadian Blood Services offers secondary use research data (that is databases about donors that you can request)
 - Suppose you were a researcher interested in research on data about donors
- ① Think about a research question and evaluate whether you could answer it
 - ② Outline the process you would follow to get the data



Takeup (GA)

FAIR Restricted data

- Findability doesn't **need** to be affected, but often is
- Accessibility might mean something different
- Interoperability can sometimes be tricky
- Reusability can be better, but this requires effort

Findability

- Organizations providing data as a non-priority activity
- Resources to make it findable may not even be considered
- Knowledge of how to make data discoverable might not exist
- Even where a core dataset from an academic source exists there can be
 - A research team primarily using the data and making it available is secondary (see item 1)
 - Because it isn't open, it's not posted, and making it findable is a separate activity (where with open data it's often put into a service that manages both curation and discoverability)

Accessibility

- Is it even considered?
 - I would argue yes, very seriously, but not with a lens of FAIR
 - Some of this is foundational see *Read et al. 2024*

Interoperability

- This goes hand-in-hand with findability and suffers the same resourcing issue
- Metadata can have the potential to disclose individuals
- Restricted datasets *don't generally have good metadata*

Reusability

- Data often belong to an organization and so don't suffer the same risks that data held by individuals do
- Similar to Interoperability/Findability issues though, improper data management makes data less reusable

What would success look like? (GA)

Data management planning

- *DMP template tool*
 - Consider what anyone following on from you will be starting from. Everything that got you from that step to another point becomes part of your research data
 - Not every part of the project data will therefore be restricted and you need to plan for that

Data Accessibility Statements

- “Access available on request” NOT sufficient
- Encourage data source to template this language! If no, DIY with review

Data Accessibility Statement Contents

Data Accessibility Statement Contents

- Who can access
 - Rank?
 - Research themes?
 - Citizenship?
- Terms of access
 - Consultation?
 - Ethics?
 - Proposal?
 - Citation?
 - TRE?
- Licensing/Agreements
 - Ex. CC-BY vs. CC-BY-NC
- Costs

Data Accessibility Statements

- Thorough description of the access process, eligibility requirements links to info, financing considerations, and licensing info/terms-of-use

Data discovery efforts

- Any datasource can now be indexed in Lunaris relatively easily.
- Metadata only deposits where there is some info to provide to potential users.
- Metadata only deposits also create a PID (persistent identifier) which makes the data a lot easier to find for someone who wants to use it in the future because *YOU CAN CITE IT*

Preservation efforts

- If a datasource overwrites my data with a new copy and doesn't say anything then someone trying to recreate my work will be very confused at best
- Ideally new versions get new PIDs and older versions point to the newest, or everything is gathered under one PID with versioning.
- Frequency of versioning will depend on frequency of access and preferences of organizations.

How do I talk to my data source?

- Raise the issue
- Highlight the benefits
- Point to resources

Case study; indexing in Lunaris

- Lunaris is a service of the Digital Research Alliance of Canada
- It collates metadata from a variety of sources
- Data sources can work with Lunaris to have their datasets indexed

Lunaris page

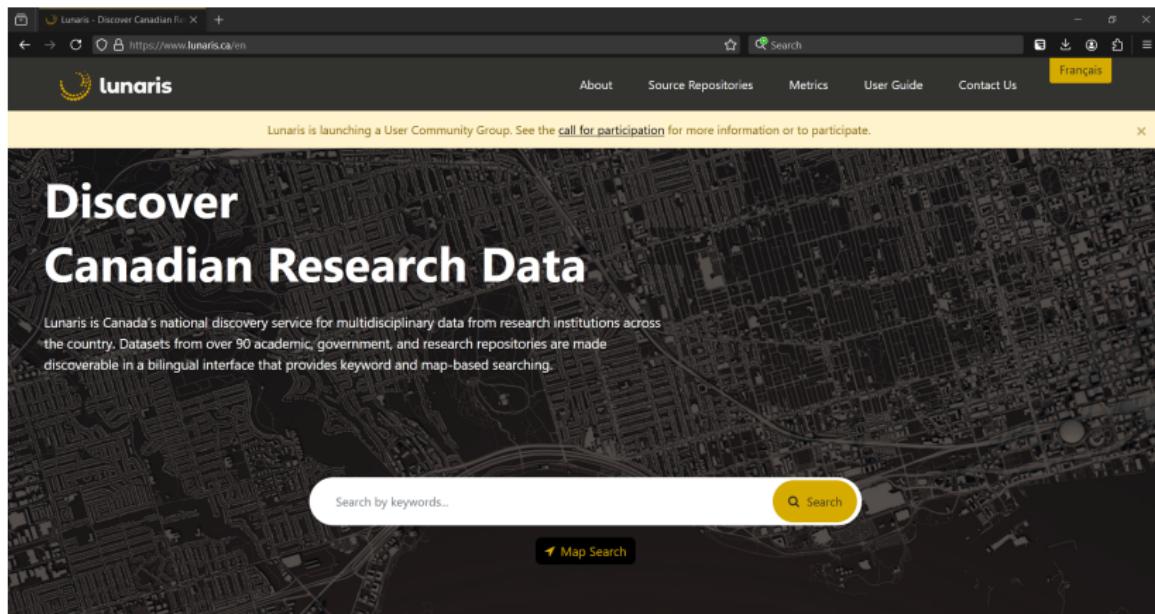


Figure 2: Lunaris webpage screenshot

System needs

- Lunaris has a metadata schema (list of information it can process and display)
- Any data source can provide a database of this information about their datasets to be “harvested”
- Work with Lunaris team to determine how the data should be formatted (samples from our case study in the workshop folder)

Case study

The screenshot shows a webpage for the Canadian Research Data Centre Network (CRDCN). The top navigation bar includes links for Home, About, vRDC, Programs, Initiatives, News, Data, Publications (which is highlighted in yellow), and Français. Below the navigation is a breadcrumb trail: Home > Data > Datasets > Farm Management Survey. The main content area features a large title "Farm Management Survey" with the subtitle "2017 - 2021". To the right of the title is a "Subjects" section with four categories: Agriculture and food, Environment, Environmental protection, and Land use. Below the title are three tabs: Overview (which is selected and highlighted in dark blue), Related Publications, and Related Data. The "Overview" tab contains sections for "Overview" and "Summary". The "Overview" section provides a brief description of the survey, stating it is a collaborative project between Statistics Canada and Agriculture and Agri-Food Canada, and that it contributes to agriculture and agri-food research by measuring selected management practices in the agriculture sector. It also mentions that data were collected using a sample of 18,000 farms. The "Summary" section includes a table titled "Available Cycles" showing the years 2017-2021 and the name "Farm Management Survey". To the right of the "Overview" section is a "Documentation" section featuring a "STATSCAN" button with a help icon. Further down is a "Publication Note" section with text about publication requirements and a "See a sample" button.

Figure 3: CRDCN data webpage screenshot

Case study

The screenshot shows a webpage for the Canadian Research Data Centre Network (CRDCN). The top navigation bar includes links for Home, About, vRDC, Programs, Initiatives, News, Data, Publications (which is highlighted in yellow), and Français. Below the navigation is a breadcrumb trail: Home > Data > Datasets > Farm Management Survey. The main content area features a yellow header "Farm Management Survey" with the subtitle "2017 - 2021". To the right of the title is a "Subjects" section with four categories: Agriculture and food, Environment, Environmental protection, and Land use. Below the title are three tabs: Overview (selected), Related Publications, and Related Data. The Overview tab contains sections for "Overview" and "Summary". The "Overview" section provides a detailed description of the Farm Management Survey (FMS), stating it is a collaborative project between Statistics Canada and Agriculture and Agri-Food Canada. It explains that the survey collects data from a sample of 18,000 farms to inform federal and provincial policy decisions. The "Summary" section highlights key findings, such as the importance of dairy, beef, poultry, pork, field crops, horticulture, and fruit, vegetable, berry, and nut crops. The Available Cycles section shows a table with one row for 2017-2021, labeled "Farm Management Survey". To the right of the main content is a "Documentation" section featuring a STATSCAN button and a "Publication Note" section. The "Publication Note" section notes that publications based on data from the FMS must include an acknowledgement of the support provided by Statistics Canada, Agriculture and Agri-Food Canada, and the University of Guelph. A "See a sample" button is also present.

Figure 4: CRDCN webpage data screenshot

Case study

ID	Format	File	Database	Protocol	Data Type	Dataset	Version	Summary	Dataset	DOI	URL	DOI	Subjects	Actions	Archiving	Content	Date
58603	Text	FarmMgmtSurvey_PMS_2017-2021_Cross-Sect	2017	2021	PM	The Farm Management Survey (PMS), conducted every five years, is a collaborative project between Statistics Canada and Agriculture and Agri-Food Canada.		Summary notes The Farm Management Survey (PMS), conducted every five years, is a collaborative project between Statistics Canada and Agriculture and Agri-Food Canada.	Dataset	DOI:10.3138/fms-2017-2021	URL:https://doi.org/10.3138/fms-2017-2021	DOI	Subjects	Actions	Archiving	Content	2021-08-24
					PM	PM data were collected using a sample of 12,000 farm households in all provinces and territories of Canada. Information on every farm, land, ownership, production crops, storage crops, and hot, vegetables, dairy and non-dairy, small farms were also nationally excluded, and large areas were considered "out of scope" because of their minimal influence on the economy.		PM data were collected using a sample of 12,000 farm households in all provinces and territories of Canada. Information on every farm, land, ownership, production crops, storage crops, and hot, vegetables, dairy and non-dairy, small farms were also nationally excluded, and large areas were considered "out of scope" because of their minimal influence on the economy.	Dataset	DOI:10.3138/fms-2017-2021	URL:https://doi.org/10.3138/fms-2017-2021	DOI	Subjects	Actions	Archiving	Content	2021-08-24

Figure 5: Screenshot of CRDCN dataset database

Step 1: Metadata crosswalk

- Sounds fancier than it is
- Map fields from my database to the schema from Lunaris
- In some cases these will be dynamic and obvious (ex. “Database name”)
 - Some will be dynamic and non-obvious
- In some cases these will be static and need to be generated (ex. “Rights”)

Crosswalk examples

- Subjects + Static fields -> Keywords
- Static access data -> Rights
- Permalink (on crdcn.ca) -> URL
- Catalogue ID -> Identifier

Step 2: Translation code

- Example in the resources (.py and .R) for our translations
- Dynamic fields mapped against the current database
- Static fields added at the end
- If you have EN/FR metadata both can be included in one file
- Write output as .json file

Step 3: Publish

- Post the .json file somewhere (anywhere) accessible on the web and share this with the Lunaris team
- Updates made to the file will be automatically reflected in Lunaris (“Harvester” checks back on occasion)

Result

- Restricted dataset is much more discoverable to those seeking information on a given topic
- Updates to metadata including access protocols, dataset info, even new data available are machine readable

Result

Lunaris example page

Option 2 - Metadata only deposits

- More detailed and of greater use to others
- Create similar metadata profile to previous example
- BUT, can include documentation

Introducing: Borealis

- National collection of Dataverse instances (one from *most* Canadian universities)
- Your dataverse ingests data, Borealis rolls it up into a national service
- Borealis info is collected by other search tools (including Lunaris).
- *Example*

Borealis & restricted data

- Can't deposit personally identifiable info into Borealis
- CAN create metadata records including deposit of supporting documents
- With the right supporting documents, Borealis can be a very effective tool for data discovery
 - Anonymized version of dataset
 - Questionnaires
 - Structural files

Borealis & restricted data

- Metadata only < Metadata w summary docs < Metadata with variable-level information
- Resource document and explainer available

Why choose Borealis or Lunaris

Lunaris Pros:

- Dynamic updates to content 'automatic'
- Minimal lift to create a series of entries from existing data

Borealis Pros:

- Permits a greater level of detail about the resource
- Allows document upload
- Creates a permanent landing page for the resource
- Mints a DOI for the resource which is 'versioned'

Why choose Borealis or Lunaris

Lunaris Cons:

- Limited and inflexible info about data source
- Needs a landing page to point to
- No “permanence” to the record, nothing is created to cite or point to
- System needs some data skill to populate automatically/navigate

Borealis Cons:

- Process to create/update records more involved
- Many fields to populate (though most are optional)
- Records cannot be deleted (this is also a pro!)
- Accessibility (for upload) limited to consortia members

When to use Lunaris

- Pre-existing and established set of landing pages
- No additional documentation to upload
- Frequent changes/additions/updates *that don't reflect changes to the underlying data*
- Data provider is not a consortium member

When to use Borealis

- Want to create a PID for the dataset
- Have additional contextual info to provide (ex. PDF uploads, more thorough metadata)

Thank you for your attention

- Questions?
- grant.gibson@crdcn.ca