Evaluating the problem of finding and accessing restricted data in Canada

Where do we go from here?

Kevin Read, MLIS, MAS
Associate Librarian
University of Saskatchewan
kevin.read@usask.ca
Land acknowledgement

I live and work on Treaty 6 Territory and the Homeland of the Métis. I pay my respects to the First Nations and Métis ancestors of this place and commit to being an ally to and partner with those who came before me.
Restricted data?

Data that are not immediately accessible immediately accessible because they are restricted or only available upon request.

Examples:

- Licensed data (e.g. commercial, organizational)
- Restricted data (e.g. health, legal)
Example

https://www.clsa-elcv.ca/access
The Discovery/Access Problem
Where's the data?

Research says:

- The ability to search for restricted data through individual data sources/websites is difficult or non-existent
- Researchers struggle to discern whether or not restricted data can be used for research purposes
- Researchers are often unaware that restricted data exists at all
Where's the data?

Issue is exacerbated when researchers report on their use of restricted data:

"None of the publications that required an application included metadata sufficiently outlining the requirements for access and approval."

Accessing the data...

“the re-use of this [restricted] data requires a set of complex approvals from multiple governing entities which are often opaque, difficult to navigate and obtain, and so pose risks to population based research”

- Uncertainty about whether they are eligible to access the data
- Difficulties both in understanding and navigating the application request process
- Lack of standardization of how to submit a data request across sources
- Amount of time it takes to apply for and acquire restricted data is prohibitive and an impediment to research
- Many data sources do not provide adequate support to help someone navigate the process

Understanding the data...

Restricted data is often of poor quality (e.g., not maintained, poorly organized, lacks standardization, data quality is low)

Documentation necessary to facilitate reuse is often absent or insufficient
These barriers have consequences

Researchers limit their research questions to data they can easily find and obtain.

Researchers may invest a substantial amount of resources into acquiring data that cannot be easily acquired and/or used.

Academic and non-academic research is limited when restricted data does not provide means to make data easily discoverable and/or accessible.
Why does this matter?

The Canadian data sharing landscape does not currently support the discovery and access of restricted data in the same way it does open data.

Restricted data can have tremendous value (e.g. health outcomes) but is not optimized for discovery, accessibility, and reuse.
Addressing the Problem
Forming a national working group

Access Limited Data Discovery Working Group operating within the Digital Research Alliance of Canada

Access-Limited Data definition broader than “restricted data”

Goals:

1. Scope the landscape of Canadian access-limited data locations, platforms and/or tools;
2. Identify the challenges associated with increasing the discovery of access-limited data;
3. Make recommendations for improving the discovery of access-limited data based on challenges identified.
Working group membership

- Kevin Read, University of Saskatchewan (Chair)
- Grant Gibson, Canadian Research Data Centre Network
- Amber Leahey, Scholars Portal
- Lynn Peterson, National Research Council
- Sarah Rutley, University of Saskatchewan
- Julie Shi, University of Toronto iSchool (Graduate Student Assistant)
- Victoria Smith, Digital Research Alliance of Canada
- Kelly Stathis, DataCite
- Jeremy Geelen, Canadian Institutes of Health Research (Special Advisor)
Research questions

**RQ1:** What types of Canadian access-limited data sources exist that include datasets that could be used for research purposes?

**RQ2:** How well do a sample of Canadian restricted health data sources identified in RQ1 make their data discoverable and accessible?

**RQ3:** What are the challenges associated with discovering and accessing restricted data from the sample of Canadian health data sources reviewed in RQ2?
Our Approach
Step 1: Scoping Canadian data sources

Identify access-limited Canadian data sources

Reviewed:

- Canadian data sources (e.g. CRDCN, Scholars Portal)
- University/college and academic partnership websites
- Government websites

Requested submissions from Alliance RDM Expert Groups and CANLIBDATA listserv

Info captured:

Geographic Region
Sector
Disciplinary Focus
# Canadian access-limited data source inventory

<table>
<thead>
<tr>
<th>Data Source Name</th>
<th>Region</th>
<th>Data Source Description</th>
<th>URL</th>
<th>Date URL Last Accessed</th>
<th>Associated Institution(s)</th>
<th>Sector(s)</th>
<th>Discipline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbvie Pharmaceutical Research &amp; Development</td>
<td>National; International</td>
<td>There are important health benefits in making clinical trial data and information available to health care providers, researchers, patients and the general public. Thus, we have adopted national and international principles and standards regarding the sharing and publication of clinical trials data and information. In addition, Abbvie conducts audits of compliance to ensure we are meeting those principles and standards.</td>
<td>Link</td>
<td>2022-02-18</td>
<td>Abbvie (private company, operates globally but headquartered in the US)</td>
<td>Private</td>
<td>Medical, health and life sciences</td>
</tr>
<tr>
<td>Agriculture and Resource Development, Government of Manitoba</td>
<td>Manitoba</td>
<td>The Agriculture and Resource Development Field Land Sample and Core Library is a secured library and viewing facility of Manitoba cores and samples collected from wells drilled under The Oil and Gas Act, the Manitoba Stratigraphic Core Hole Program, and select mineral exploration drillcore collected through the Mines Act. There is no charge to view the cores in the facility. Informational Notice 18-02 states our policies on examining, sampling, analyzing and removing these core and samples that were collected through the Oil and Gas Act.</td>
<td>Link</td>
<td>2022-02-18</td>
<td>Government of Manitoba</td>
<td>Government</td>
<td>Engineering and technology</td>
</tr>
<tr>
<td>Alberta Health Services</td>
<td>Alberta</td>
<td>Not provided</td>
<td>Link</td>
<td>2022-02-18</td>
<td>Alberta Health Services</td>
<td>Government</td>
<td>Medical, health and life sciences</td>
</tr>
<tr>
<td>atlantis</td>
<td>Alberta</td>
<td>We are committed to the continued updating, maintenance, storage, distribution and licensing of Alberta’s primary spatial datasets and our trusted partner’s imagery and LIDAR datasets.</td>
<td>Link</td>
<td>2022-02-18</td>
<td>Various (multi-partner initiative)</td>
<td>Non-profit; Private</td>
<td>Other</td>
</tr>
<tr>
<td>Association of Faculties of Medicine of Canada data</td>
<td>National</td>
<td>40 years of data on Canada’s medical education system</td>
<td>Link</td>
<td>2022-02-18</td>
<td>AFMC</td>
<td>Other</td>
<td>Medical, health and life sciences</td>
</tr>
<tr>
<td>Atlantic Canada Conservation Data Centre</td>
<td>Atlantic</td>
<td>AC CDC maintains location data for species of conservation concern (S-ranks between S1 and S354, in addition to species considered extirpated (S6) and historic (S7)) a Geographic Information System (GIS) database.</td>
<td>Link</td>
<td>2022-02-18</td>
<td>Atlantic Canada Conservation Data Centre</td>
<td>Non-profit</td>
<td>Natural sciences</td>
</tr>
<tr>
<td>Atlantic PATH</td>
<td>Atlantic</td>
<td>Atlantic PATH has recruited over 34,000 participants from all four Atlantic Provinces. The samples and information that participants have given will help researchers find out why some people develop certain chronic diseases and others don’t. This information will help to find new ways of preventing chronic diseases and to diagnose these diseases earlier, when they can be easier to treat.</td>
<td>Link</td>
<td>2022-02-18</td>
<td>Dalhousie University &amp; Canadian Partnership for Tomorrow's Health</td>
<td>Academic</td>
<td>Medical, health and life sciences</td>
</tr>
</tbody>
</table>

n=137

Step 2: Grading Canadian health data sources

Identified **55 health data sources** in our inventory

48 were eligible for review

- n=4 only permitted patient data requests
- n=3 became inaccessible during review

Each source underwent qualitative review to identify discovery/access attributes

<table>
<thead>
<tr>
<th><strong>Attributes Identified:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Discovery Attributes</strong></td>
</tr>
<tr>
<td>Data description</td>
</tr>
<tr>
<td>Data documentation</td>
</tr>
<tr>
<td>Searchability / browsability</td>
</tr>
<tr>
<td>Use of metadata standards</td>
</tr>
<tr>
<td>Attribute</td>
</tr>
<tr>
<td>---------------------------------</td>
</tr>
<tr>
<td><strong>Data description</strong></td>
</tr>
<tr>
<td><strong>Data documentation</strong></td>
</tr>
<tr>
<td><strong>Searchability/Browsability</strong></td>
</tr>
<tr>
<td><strong>Use of metadata standards</strong></td>
</tr>
<tr>
<td><strong>Data request processes</strong></td>
</tr>
<tr>
<td><strong>Data restrictions</strong></td>
</tr>
<tr>
<td><strong>Pricing information</strong></td>
</tr>
<tr>
<td><strong>Contact information</strong></td>
</tr>
</tbody>
</table>
Grading 48 health data sources
Grading 48 health data sources

- 2 authors reviewed each data source
- Tiebreaker for discrepancies
Grading goals

Evaluate the state of discovery and access in Canadian health data sources

Identify gaps and areas for improvement

Prioritize data sources for future indexing (e.g., FRDR, Dataverse)

Inform final recommendations

Grade A

Grade B

Grade C
Discovery grading criteria

**Grade A**
1. Provides description of the data that supports understanding for a broad audience
2. Describes who is responsible for creating the data
3. Describes the data’s intended use or purpose

**Grade B**
1. Description of the data requires more info to facilitate understanding and selection for reuse
2. Description of who is responsible may not be entirely clear
3. Description of data’s intended use or purpose may not facilitate understanding, selection, or reuse

**Grade C**
1. Little or no detail about the data
2. Little or no detail about who is responsible for the data
3. Little or no detail about the data’s intended use or purpose

Discovery grading criteria

**Data Documentation**

**Grade A**
1. Datasets are accompanied by multiple pieces of detailed, structured information about the data itself, as well as additional contextualizing information. Interpretation and reuse of the data is possible with the materials provided.
   - e.g. data dictionary, codebook, user guide, code, data collection materials or instruments

**Grade B**
1. Datasets are accompanied by some detailed, structured information about the data itself.
   - Data dictionary, codebook, user guide, or robust data collection instrument is provided, but more may be required to facilitate interpretation and reuse

**Grade C**
1. Datasets are not accompanied by information about their content and structure. Interpretation and reuse of the data is not possible with the available materials.

Discovery grading criteria

Searchability / Browsability

**Grade A**
1. Datasets are searchable within the data source
2. Basic keyword search is available
3. There is a mechanism to browse datasets by one or more relevant facets or variables (e.g., topic, population)
4. There is an advanced interface allowing boolean search and/or searching of specific metadata fields

**Grade B**
1. Datasets are searchable within the data source
2. Basic keyword search is available

**Grade C**
1. There is no mechanism to search or browse the data source

*Note: If a data source only had one dataset, this category was graded N/A*
Discovery grading criteria

**Grade A**
1. Structured metadata is available and structured using one or more recognizable standards (e.g. Dublin Core, ISO 19115)
2. Metadata elements are employed consistently. There is a set of core elements, including metadata needed to cite the dataset
3. Metadata is clearly presented within the web interface

**Grade B**
1. Structured metadata is available, but may not adhere to a recognizable standard
2. Metadata elements are employed consistently; there is a set of core elements that is usually present (e.g. title, description)
   a. Metadata may not be presented in a conventional interface for this tier. For example, a text document with structured fields would meet criteria.

**Grade C**
1. There is no metadata, or
2. Metadata is present but it is limited and unstructured—e.g. a title and description only
Access grading criteria

**Grade A**
1. Source **clearly explains all aspects of the request process** including application forms, instructions, timelines, review process, and submission process
2. Support for submitting data access requests is available and clearly indicated

**Grade B**
1. Source **clearly explains most aspects of the request process** including application forms, restrictions, timeline, and submission process

**Grade C**
1. Source provides **no clear instructions on how to submit a request** or what to include, or
2. Data request form is not available

Access grading criteria

Grade A
1. Source includes a section devoted to describing data restrictions
2. Specific populations or projects eligible to access data are well described and include examples
   a. e.g. “Only oncology researchers affiliated with [institution] may access these data”
   b. e.g. “Only research studies concerning [topic] may use these data”

Grade B
1. Restrictions on who can use the data are mentioned but not described

Grade C
1. No information is provided, or
2. Information is not clear enough to determine if restrictions exist
Access grading criteria

Pricing

Grade A
1. Fees are explained clearly, with enough information to estimate the cost of a specific request
2. If a quote process is in place, a researcher can:
   a. Submit a basic research plan to receive a quote;
   b. Contact someone for help estimating fees; or
   c. View a sample project with fees applied

Grade B
1. It is clear that there are fees for accessing data
2. Estimated costs are provided
3. If a quote process is in place, a full research proposal is required

Grade C
1. No information on fees is provided

Access grading criteria

Grade A
1. There is a clear, easy-to-find contact person and email address devoted to data inquiries and applications
2. Contact info is displayed within the area of the website related to data

Grade B
1. Contact information related to the application process is provided
2. Contact info is displayed within the area of the website related to data

Grade C
1. No contact information is provided, or
2. Only a generic “Contact Us” tool is provided

Results
Canadian access-limited data landscape

Region (n=137)
Canadian access-limited data landscape

Sector

- Private: 1.9%
- Non-profit: 9.3%
- Other: 13.1%
- Academic: 16.7%
- Government: 57.0%

(n=137)

Discipline

- Medical, health, and life sci: 41.7%
- Social sciences: 8.1%
- Other: 15.9%
- Natural sciences: 17.4%
- General: 17.4%
- Engineering and technology: 1.9%
Discovery grading of health data sources

Access grading of health data sources

Key takeaways

- 42% (n=20) did not receive an “A” grade in any category
- 52% (n=25) did not provide any information about data restrictions
- 56% (n=27) did not provide any data documentation to support interoperability and/or reuse
- 79% (n=38) received a “C” grade for metadata standards
- 0% (n=0) received an “A” grade for metadata standards
It’s not all bad! Potential data sources for indexing

<table>
<thead>
<tr>
<th>Rank</th>
<th>Data Source</th>
<th># of A Grades</th>
<th># of B Grades</th>
<th># of C Grades</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Canadian longitudinal study on aging</td>
<td>6</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Manitoba Population Research Data Repository</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>CanPath</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Population Data BC</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>BORN ontario</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>Canadian Institute for Health Information</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>ICES</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>Cancer Care Ontario</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>Health Data Nova Scotia</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>BC Cancer Registry</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>CHILD cohort study</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Key Barriers & Recommendations
Insufficient Infrastructure

Barrier 1
Infrastructure barriers

Navigation, workflows, and linkage to related content were challenging
Dataset information and access request information were often separate
Time investment in seeking and learning about datasets was very high
Lack of standardization / high variability in each data source
Some data sources vanished during our study – preservation concern
Infrastructure: Recommendations

1. Establish a community of practice for stewards of Canadian restricted data sources to establish commonly accepted guidelines and standards;

2. Offer funding opportunities for restricted data sources to adopt data discovery and access standards;

3. Align data request procedures as much as possible across jurisdictions to improve workflows for the acquisition of restricted data (e.g., Sensitive Data Pilot); and

4. Explore a common infrastructure model that could be adopted by all restricted data sources in Canada.
Where’s the metadata?

Barrier 2

Dataset Description: ?
Population studied: ?
Timeframe: ?
Access procedure: ?
Dataset restrictions: ?
Cost of data: ?
Metadata barriers

Grading identified reasonably good descriptions of data and access procedures, but no metadata to disseminate them;

Non-existent structured metadata for the data access request process;

Dataset metadata schemas provide little specificity in this area; and

Without metadata, these data sources will remain hard to find and their data will not be accessed.
Metadata: Recommendations

1. Develop a metadata schema (or extension) that addresses access-specific data requirements to improve discovery and access; and

2. Engage with data sources who were identified through our grading exercise to improve the discovery of their data in national aggregators (e.g., FRDR).
Metadata: Recommendations

1. Develop a metadata schema (or extension) that addresses access-specific data requirements to improve discovery and access; and

2. Engage with data sources who were identified through our grading exercise to improve the discovery of their data in national aggregators (e.g., FRDR).
Patient Centered Measurement: The Acute Inpatient Survey 2016/17

Data sources
78 Acute Care Hospitals and 2 freestanding rehabilitation hospitals in 6 health authorities (Fraser Health, Interior Health, Island Health, Vancouver Coastal Health, Northern Health, Provincial Health Services Authority) and Providence Health Care.

Date range
September 1 2016 to March 1st 2017

Description
The Acute Inpatient 2016/17 Survey asked patients about their health-related quality of life and their experiences with the quality of care and services received as an inpatient in one of 78 acute care hospitals and two freestanding rehabilitation hospitals in British Columbia.

The survey was coordinated by the BC Office of Patient-Centred Measurement (PCM) on behalf of the BC PCM Working Group, a group that includes representation from the BC Ministry of Health and the seven Health Authorities.

The Survey included items from the following Patient Experience Reported Measures (PREMs) and Patient Outcome Reported Measures (PROMs).

View a video presentation on these data below:

Access criteria:
● Project proposal required
● REB ethics approval required
● Data transfer agreement necessary

Dataset restrictions:
Only researchers with active Tri-agency grant funding are permitted to access this data.

Conditions of use:
Researchers are only permitted to access data on secure systems that have no internet access.

Timeline to access:
~1-2 months processing

Cost of data:
● $250: Access
● $40/hour: Support
Lack of Documentation

Barrier 3
Documentation barriers

We identified very little data documentation across health data sources; Mirrors challenges identified in existing research (e.g., time investment); and Many data sources do not have the personnel or expertise to develop robust documentation (no devoted data stewardship support)
Documentation: Recommendations

1. Develop targeted guidance and training for restricted data sources to articulate the value, importance, and utility of including data documentation for restricted data; and

2. Engage large organizations with robust administrative staff who have well-documented data to provide guidance for those with less funding/capacity
Concluding Thoughts
Our next steps

Publish our findings from this research

Progressing on phase 2 of our work to explore discovery and access metadata for restricted datasets:

- Extract metadata from 48 health data sources
- Identify commonalities in metadata elements
- Develop a minimal metadata standard/extension for both:
  - Restricted datasets
  - Restricted data access procedures
Going forward

Canada has room for improvement with respect to the discovery and access of restricted data

High value restricted datasets that can be used for research are often hidden, inaccessible, and/or unusable

Metadata for restricted data and their access procedures can improve discovery and reuse

Only scratched the surface with health data sources (82+ data sources remain)

Support from national data initiatives is crucial for future success in this area
References


Questions?

kevin.read@usask.ca