Understanding the challenges associated with finding and accessing restricted data in Canada: a mixed methods study

Read, Kevin B.; Gibson, Grant; Leahey, Amber; Peterson, Lynn; Rutley, Sarah; Shi, Julie; Smith, Victoria; Stathis, Kelly

Publisher's version / Version de l'éditeur:
https://doi.org/10.1139/facets-2023-0102
FACETS, 9, pp. 1-9, 2024-08-01

FACETS

OPEN ACCESS | Research Article

Understanding the challenges associated with finding and accessing restricted data in Canada: a mixed methods study

Kevin B. Read[a], **Grant Gibson**[b], **Amber Leahey**[c], **Lynn Peterson**[d], **Sarah Rutley**[a], **Julie Shi**[e], **Victoria Smith**[f], **and Kelly Stathis**[g]

[a]University of Saskatchewan, University Library, Saskatoon, SK, Canada; [b]Canadian Research Data Centre Network and McMaster University, Hamilton, ON, Canada; [c]Scholars Portal, Toronto, ON, Canada; [d]National Research Council of Canada, Ottawa, ON, Canada; [e]University of Toronto, University Libraries, Toronto, ON, Canada; [f]Digital Research Alliance of Canada, Ottawa, ON, Canada; [g]DataCite, Hannover, Germany

Corresponding author: **Kevin B. Read** (email: kevin.read@usask.ca)

Abstract

Data that are restricted are historically challenging for researchers to find and even more difficult to access. While efforts to support open data have expanded in Canada, the same cannot be said for restricted data. To better understand the landscape of restricted data in Canada, this study aimed to accomplish two primary goals: (1) identify data sources where data were restricted and (2) assess a subset of health sciences data sources to determine how well they make their data discoverable and accessible. Our study identified 137 Canadian data sources, where 48 health sciences sources were evaluated for discoverability/accessibility. Data sources received poor grades with respect to data discovery due to a lack of metadata standards (38/48, 79%), an inability to find datasets through searching and browsing (32/46, 70%), and a lack of data documentation to support reuse (27/48, 56%). The absence of pricing information (31/48, 65%) and opaque dataset restrictions (25/48, 52%) were identified as key barriers to the data access request process. This study highlights significant room for improvement with respect to improving the discovery of and access to restricted data in Canada and makes recommendations for how to better support restricted data sources on a national scale.

Key words: restricted data discovery, data sharing, research data management, FAIR principles, open metadata, data access processes

Introduction

Over the past 5 years, there has been a significant investment in developing digital research infrastructure, data management support, and establishing data sharing best practices to promote accessibility and usability of research data in Canada. For example, Canada's Tri-agency, which is composed of the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council, and Social Sciences and Humanities Research Council, released a data management policy (Government of Canada 2021), which requires prospective applicants to submit a data management plan and indicate where they will deposit their data. In addition, the Digital Research Alliance of Canada (Alliance) was formed to support digital research infrastructure including data repositories as well as research data management (RDM) software and support for Canadian researchers. This organization has established and supports national data sharing platforms, including the Federated Research Data Repository, Lunaris, a national Dataverse repository, and a discovery service to index, store, and provide access to Canadian research data from across disciplines and sectors (Digital Research Alliance of Canada 2022*b*, 2022*c*). Finally, the federal government released a Roadmap for Open Science in 2020, calling for all government-funded research outputs to be made openly available by default (Government of Canada 2020*a*).

Finding and accessing data that is restricted or sensitive remains a common challenge for researchers despite these widespread efforts to improve data sharing. This challenge is enhanced by the fact that restricted data are by their very nature more difficult to access. For the purposes of this study, we define restricted data as data that are not immediately accessible because they are restricted due to commercial, ethical, or legal reasons, or they are only available upon request.

Recent studies have shown that author-reporting on the availability and accessibility of restricted data in publications is poor, particularly with respect to providing clear language and information in journal data availability statements (Read et al. 2021; Gabelica et al. 2022; Page et al. 2022). A previous study completed by the corresponding author found

that when CIHR-funded authors reported that data could be shared or an application was required, none provided details about how to access the data (Read et al. 2021). Gabelica et al. (2022) encountered similar issues when trying to access data from authors who stated in their data availability statements that their data were available upon request; only 6.8% of authors provided the requested data.

Beyond the insufficient author-reporting of restricted data, existing research highlights a wealth of reasons why researchers encounter barriers when trying to locate and access restricted data. Researchers are not always adept at discerning whether a particular restricted access dataset might be available to them (van Schaik et al. 2014; Ho et al. 2018; Bekemeier et al. 2019; Mpango and Nabukenya 2019; Clayton et al. 2021; Hanna et al. 2021; Pongiglione et al. 2021; Knosp et al. 2022). Even when they have found a source for the restricted data in question, they often face uncertainty about their eligibility to access the data and are consequently hesitant to invest time in finding out as inquiries can sometimes take months or years (Boland et al. 2017; Ho et al. 2018; Lugg-Widger et al. 2018; Rahimzadeh et al. 2018; Knosp et al. 2022).

Another set of challenges arise from poor infrastructure for searching for, locating, and requesting restricted access data, specifically poor search interfaces (or lack thereof), difficult-to-use software, a lack of standardization for how to submit data requests, and a lack of support from restricted data stewards (Nancarrow 2013; Choudhury et al. 2014; Sarwate et al. 2014; van Schaik et al. 2014; Lugg-Widger et al. 2018; Rahimzadeh et al. 2018; Garrison et al. 2019; Bonomi et al. 2020). More generally, researchers have reported difficulties in both understanding and navigating the data application process (Sydes et al. 2015; Siu et al. 2016; Ho et al. 2018; Prince et al. 2018; Rahimzadeh et al. 2018; Bekemeier et al. 2019; Saulnier et al. 2019; Knosp et al. 2022). For the purposes of this study, we define data stewards as individuals or organizations responsible for dataset documentation, quality, storage, preservation, and access in any given data source; the responsibilities of a data steward may also include managing the technical infrastructure used to house those datasets (CODATA no date).

Lugg-Widger et al. (2018) highlight that the difficulties in acquiring restricted data for reuse can impede health research, specifically "the re-use of this 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". Finally, several studies highlight that when a researcher is finally able to obtain restricted data, the data are often of poor quality and lack the documentation necessary to enable effective reuse (Nancarrow 2013; van Schaik et al. 2014; Boland et al. 2017; Prince et al. 2018; Bekemeier et al. 2019; Byrd et al. 2020; Pongiglione et al. 2021).

These barriers to discovering, accessing, and using restricted data have resulted in several consequences: researchers have reported that they may limit their research questions to data that can be more easily accessed (van Schaik et al. 2014); researchers may invest a substantial amount of resources (e.g., time and money) into acquiring data that cannot be easily found and/or used (Siminski et al. 2021); and researchers have highlighted that academic and non-academic research (particularly in the health sector) is strongly limited when restricted data lack sufficient means to facilitate discovery and access (Lugg-Widger et al. 2018; Rahimzadeh et al. 2018; Pongiglione et al. 2021).

While many of the studies above have examined the high-level ethical and legal barriers to restricted data sharing, few (Leahey 2014) have explored the barriers that data providers across sectors face when trying to make their datasets discoverable and accessible. To help address these barriers, a Canadian national working group was formed to identify and evaluate the challenges related to discovering and accessing restricted data in Canada and to inform the ongoing development of Canadian infrastructure for research data. To that end, this study was undertaken to answer three primary research questions:

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

RQ2: Based on a sample of Canadian restricted health data sources identified in RQ1, how well do these sources 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 identified in RQ1?

The answers to these questions provide an opportunity to address barriers associated with finding, accessing, and using restricted data while examining it through a national lens. Ultimately, the findings identified in this study can inform how Canadian restricted data sources and the Canadian RDM landscape can work toward improving the discovery, access, and use of these data. While this study focuses on the Canadian RDM landscape and restricted data sources, it can be used to inform progress toward improving the discovery, access, and use of both Canadian and international data.

Materials and methods

Design and setting

To identify and evaluate the ability for restricted data sources to make their data discoverable and accessible, this study took a two-step mixed methods approach that included scoping the landscape of Canadian access-limited data sources and grading Canadian health sciences data sources based on their ability to make datasets easily discoverable and accessible. This study began in January 2021 and was completed in November 2021.

Step 1: scoping the landscape of Canadian access-limited data

The first stage involved an environmental scan to identify the number and characteristics of Canadian data sources that contained access-limited data. In this first stage, we broadened our search beyond restricted data to identify any Canadian data sources that were considered "access-limited", meaning they had significant barriers to discovery, access, or use, even if the data may have been publicly available.

To achieve this, we developed a search strategy to identify federal and provincial government resources, academic institutions, private sectors, and not-for-profit sectors that contained access-limited data (see Supplemental File 2). We also looked within individual university, college, and government websites to capture data sources we may have missed with our original strategy. Finally, we consulted with members of the Alliance RDM Network of Experts (Digital Research Alliance of Canada 2022*a*) in meetings and through email and the broader Canadian research data community via listservs to fill in gaps that may have been missed in our search efforts.

Data sources were included if they indicated that some or all data were available by request or if data could be located but little to no information was provided about the datasets themselves—even if they were publicly downloadable. When a data source was marked for inclusion, the following information was captured: data source name, region, description, URL, host institution, sector, and discipline.

Data analysis for scoping exercise

For every data source identified, we examined the counts for the region from which the data source originated, the sector, and the disciplinary focus of the data source. For the latter, we used the six top-level disciplinary divisions provided by the Canadian Research and Development Classification for Field of Research (Government of Canada 2020*b*). The data dictionary is available on the Open Science Framework (Read et al. 2022).

Step 2: grading discovery and access attributes in health data sources

Grading a subset of health sciences data sources

Our scoping exercise identified health sciences data sources as the most common discipline ($n = 55$) in our sample. We chose to use these data sources as a subset for our grading analysis because they represented the largest disciplinary sample and provided us with the ability to consistently compare discovery and access characteristics within a single discipline. These data sources contained but were not limited to administrative, clinical trial, medical education, and registry, as well as patient and population health research data. Prior research discussed in our introduction suggests that these types of health data in particular are difficult to discover and access (van Schaik et al. 2014; Lugg-Widger et al. 2018; Rahimzadeh et al. 2018; Pongiglione et al. 2021), despite the recognition of their value for research purposes, thus adding further justification to focusing on their discovery and access challenges. Using this subset, we identified common attributes related to discovery and access that we could use to grade each data source.

Grading these data sources served to reveal key challenges and areas for improvement with respect to restricted data discovery and access in Canada.

Identifying data discovery and data access attributes

Using the health sciences data sources subset, we identified common attributes that represented the positive and negative characteristics associated with data discovery and access within them. Authors were assigned to review a sample of 7–10 health data sources and write narrative summaries about how well each data source provided discovery and access information by answering the following questions:

- How well does the data source describe/list the datasets that are available?
- How thoroughly are individual datasets described?
- If applicable, do datasets have metadata and how detailed are they?
- Do datasets use an established metadata schema? (e.g., Dublin Core, DCAT)
- Can the datasets be searched within the data source (e.g., database, catalogue, etc.), if applicable?
- How thoroughly are dataset access requirements described?
- How detailed and/or transparent are the instructions for obtaining access to the data?
- If the data access has associated costs, how transparently are the costs described?
- Is a contact person/steward/governing body for the dataset(s) clearly indicated?

Narrative summaries were then collated and summarized into specific attributes related to the broader categories of data discovery and data access. For the data discovery category, we identified dataset description, dataset documentation, searchability and/or browsability of datasets, and metadata standards as the common attributes. For the data access category, data access request processes, data restrictions, quality of pricing information, and contact information were identified as common attributes. These attributes are fully defined in Table 1.

Developing and applying a data discovery and access grading rubric

Using the attributes in Table 1, we developed grading criteria to evaluate the 55 health data sources based on their data discovery and access attributes using a letter grade system of A through C (see Supplemental File S3). Using the finalized criteria, each health data source was reviewed and graded by two reviewers. If a conflict was present between the assigned scores for a particular source from the two reviewers, a consistent, third reviewer was assigned to make a final decision.

Data analysis of grading exercise

The final grading results were analyzed by identifying the frequency of scores across all health data sources in our sample to identify areas of strength and weakness related to data discovery and access. We then examined the degree to which there was consistency between data discovery and access scores within individual data sources using the Kendall rank correlation coefficient, which quantifies the similarity

Table 1. Data discovery and access attributes.

Attribute	Definition
Data discovery	
Dataset description	Description of the data itself, including summary describing the purpose, nature, and scope of the data collection, special characteristics of its contents, major subject areas covered, and major research questions.
Dataset documentation	Detailed, structured information about the data itself that supports its interpretation and use.
Searchability/browsability	If a data source has more than one dataset, there is the presence of mechanisms for searching and browsing them (e.g., full search interface, table, list, etc.)
Metadata standards	The presence of metadata standards applied within the data source. Metadata standards establish a common way of structuring data about the dataset. For data sources that do not have individual datasets within them, this section can be applied to the data source as a whole.
Data access	
Data request process	The presence, completeness, and clarity of information and/or content (e.g., application forms) required to successfully submit a data access request and to understand the data access request process.
Dataset restrictions	The criteria of persons/organizations/projects who are eligible to access the data. Data restrictions should inform a user whether or not they would be eligible to acquire the data and should indicate what parameters are necessary to meet eligibility.
Quality of pricing information	The transparency of the costs associated with acquiring and using dataset(s).
Contact information	The presence of information about a contact person who can support a data requester with understanding the data or completing the application process.

of two different rankings of a set of objects (Kendall and Maurice 1990). To make this calculation, we first sorted the data sources based on the highest number of "A" grades and then the highest number of "B" grades among our data discovery attributes. We then assigned ranks based on this sorting (A = 2, B = 1, C = 0), allowing for ties among the data sources. We then repeated this same sorting and ranking for the data access attributes in our sample of data sources. The goal of this analysis was to discern whether data sources that scored highly in data discovery received equivalent scores when describing their data access request processes.

Results

Canadian landscape of access-limited data sources

Our scoping exercise was successful in identifying 137 access-limited data sources. Of these data sources, 49.6% (n = 68) were provincial in scope, 48.9% (n = 67) were classified as national, 2.0% (n = 3) were from the territories, and 1.5% (n = 2) were classified as international with Canadian datasets hosted within them. Each data source may have had more than one geographic category assigned to it.

With respect to the sectors these data sources represent, 61.3% (n = 84) were from the government, 26.3% (n = 36) originated from academic institutions, 15.3% (n = 21) were from the non-profit sector, and 4.4% (n = 6) were hosted in the private sector. 10.9% (n = 15) of data sources were classified as "other", as they did not fit within the other predefined categories. Categories assigned to each data source were not mutually exclusive.

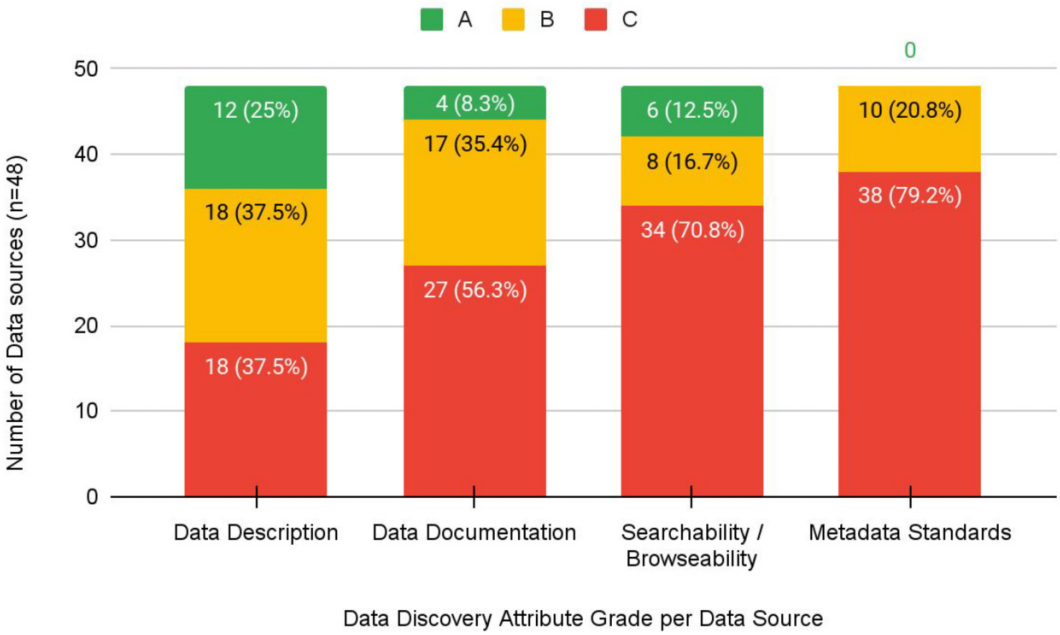
In terms of the disciplinary focus of the data sources identified, 40.8% (n = 56) were categorized as being related to medical, health, and life sciences, 16.8% (n = 23) were considered general purpose, 16.8% (n = 23) were from the natural sciences, 5.8% (n = 8) were from the social sciences, 2.2% (n = 3) were from agricultural and veterinary sciences, and 2.2% (n = 3) were from engineering and technology. The 16.8% (n = 23) of data sources that were labelled as general purpose represented data provided by provincial government ministries and agencies and included a variety of categories within our predefined list, including but not limited to government administrative records, the census, and parliamentary debate transcripts. 15.3% (n = 21) were classified as other as their focus was outside the scope of the predefined categories used in our analysis. For example, many data sources drawn from Canada's National Research Council focused on property management, purchasing and requisition files, and business opportunity data that did not fit within either of the general purpose or discipline-specific categories. The complete list of data sources identified is available in Supplemental File S4.

Health data source scoring results

From the health data sources identified, 48 of the 55 data sources were included in the final analysis. Seven data sources were excluded because they either only accepted requests from individuals for their own personal health record data (n = 4) or the data source became inaccessible during our study due to broken URL links (n = 3).

The grading exercise identified that 42% (n = 20) of data sources did not receive an "A" grade in any category; however, 44% (n = 21) received an "A" grade in two or more categories. The majority of datasets received a "C" grade for a lack of metadata standards (38/48, 79%), an inability to explore and discover datasets through searching and browsing (34/48, 71%), or lack of data documentation to support interpretability and reuse (27/48, 56%) (Fig. 1). Descriptions of datasets themselves fared better in that 25% (n = 12) received an "A" grade, and 37.5% (n = 18) received a "B" grade.

Fig. 1. Grading results examining data discovery attributes in Canadian restricted health data sources.



The absence of or lack of clarity with respect to pricing information (31/48, 65%) and vague or non-existent information related to dataset restrictions (25/48, 52%) were identified as key barriers to the data access request process (i.e., received a "C" grade). It is likely that many of the 31 sources without information about pricing are accessible for research purposes at no cost; however, providing this information to potential users is an important part in streamlining access. The actual description of the data request process, however, identified that 70.8% of data sources received an "A" (13/48, 27%) or "B" (21/48, 43.8%) grade in this category. Similarly, contact information was generally well described, as only 22.9% ($n = 11$) received a "C" grade (Fig. 2).

The degree to which individual data sources were consistent across discovery and access attributes was considered "fair" according to Kendall's tau ranked test (tau-$b = 0.31$, $p = 0.0059$), indicating that data sources with higher discovery grades also had higher access grades on average relative to the other data sources in our sample.

Discussion and conclusions

This study provides preliminary insight into the discovery and access characteristics of restricted health data sources in Canada. In particular, this study highlights three key barriers associated with data of this kind that both mirror findings in previous research and introduce new complexities to the restricted data sharing landscape. First, this study calls attention to the challenges with discovery and accessing data due to a lack of sufficient infrastructure. Second, not a single data source received an "A" grade for metadata; for the majority of data sources, metadata were either sparse or non-existent. Finally, the availability of documentation for facilitating the reuse of datasets was lacking, resulting in an inability to understand or interpret data in the sources we identified.

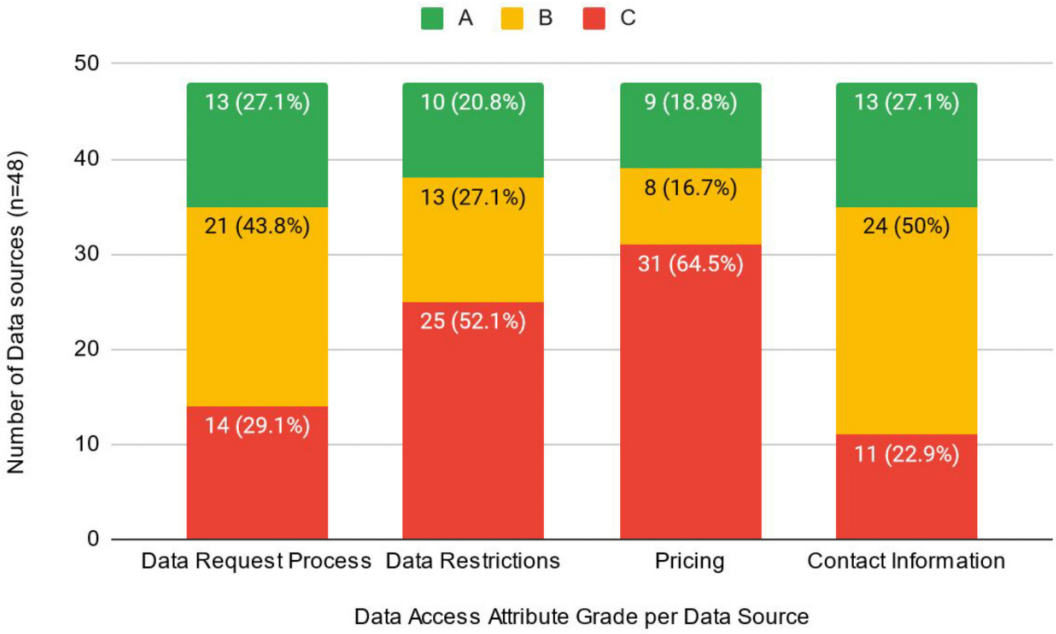
To forge a path toward removing barriers in these three areas, we elaborate below on the challenges associated with these areas and suggest how to improve and support the discovery and access of restricted data in relation to them. In addition, we suggest approaches for examining data sources beyond the health sector to provide a more holistic view of the Canadian restricted data sharing landscape.

Improving restricted data infrastructure for discovery and access

Infrastructure was observed across the sources at varying levels ranging from individual research projects, institutional websites, large scale research organizations with support for data access, regional health data centres, and government data access programs. This infrastructure diversity reflects variations across research disciplines, institutions, and jurisdictions that may govern restricted data sharing, and highlights the difficulties of standardizing discovery and access to these data. The limited availability and high variability in workflows across data sources can lead to challenges around standardization, equitable access, and sustainability at the human, project, and technological levels. Without consistency and standardization across restricted data sources, datasets will remain difficult to find, access, and use. In a worst-case scenario, data sources that are under-resourced may result in data loss—an issue we encountered during our study as three sources became inaccessible during our analysis.

This study highlights that data sharing infrastructure in Canada does not adequately support making restricted data FAIR (findable, accessible, interoperable, reusable) (Wilkinson et al. 2016). The ideal state of restricted data dis-

Fig. 2. Grading results examining data access attributes in Canadian restricted health data sources.



covery and access infrastructure has not been fully defined or envisioned yet and requires a variety of stakeholders and experts to come together to define relationships and roles across institutions and jurisdictions. Researchers need access to reliable, secure, institutionally approved infrastructure and workflows for storing, sharing, and preserving research data to comply with funder and/or journal policies, and to facilitate reproducibility and reuse of restricted data across sectors. Without sufficient infrastructure, these data will remain hidden, difficult to access, and may even be lost due to lack of support. For an infrastructure model that supports restricted data discovery and access to be successful on a national scale, we recommend that government (specifically provincial/territorial health bodies), academic institutions, and national data management initiatives (e.g., the Alliance) work together to develop a standardized model that supports restricted data discovery and establishes standard workflows for accessing restricted data for research purposes.

Developing and adopting robust discovery and access metadata standards

We found no consistent metadata describing restricted datasets or their access-specific requirements among data sources we assessed. While some sources included non-standardized metadata elements specific to restricted data, the use of specific metadata schemas was absent. Our grading results demonstrate that data sources provide reasonably good description of their data and the request process to access them; however, a lack of metadata with which to structure or disseminate this information impedes their ability to be discovered by the research community.

Restricted data sources could adopt existing metadata schemas to positively impact discovery, use, and standardization. The inclusion of structured metadata would help researchers by reducing time and resources expended on identifying valuable datasets. The findings from our grading exercise emphasize this, as over half of the data sources did not provide any information about who is eligible to access their datasets. Structured metadata would also improve discovery by increasing the potential for restricted data to be harnessed by national aggregators. National aggregators play an important role in making data discoverable, accessible, and therefore, reused; however, the extent to which they can perform this function is limited by the lack of standardized, openly available metadata. Those interested in using restricted data can use metadata harvested by aggregators to filter and sort data sources and datasets, so that they can identify data that suit their needs.

Presently, existing metadata standards may not be sufficient to adequately describe restricted data, specifically with respect to the data access request process. Although generalist metadata schemas include generic elements to describe access restrictions, guidance for their application is often left open-ended; for example, Dublin Core's "accessRights" element does not provide sufficient structure to describe access procedures (Dublin Core Metadata Innovation 2020). In practice, the values of these elements are often free text and do not have consistent values. Moreover, vocabularies to describe access—such as the Confederation for Open Access Repositories Access Rights vocabulary (2021)—describe the access status of a resource (such as "open access" or "restricted access") but do not describe the *conditions* of access restrictions (such as who can access the resource or at what cost).

Data sources would benefit from standard metadata elements to describe data access procedures, as expenditures of resources deployed to make data discoverable, to create data governance and access frameworks, and to ensure com-

pliance may be reduced by addressing current ambiguities in best practices. Researchers would also benefit from access procedures being clearly defined in standardized metadata. For example, the cost to access restricted data varies greatly among data sources. If cost were to be included as a metadata element by multiple data sources, a researcher could search within a national aggregator to identify datasets that may be accessed at no cost. This would greatly improve the ability of researchers to identify data appropriate for their particular study in relation to its funding.

One way for restricted data sources to adopt existing metadata schemas and improve discoverability is to connect with global research infrastructure. Digital object identifiers (DOIs) are a type of persistent identifier commonly used for research outputs—such as data and publications—with accompanying metadata. If restricted data sources were to begin registering DOIs for their datasets, they would be providing standardized metadata—for example, according to the DataCite Metadata Schema (2021)—and this metadata would be publicly searchable and usable by aggregators.

Data are valuable insofar as researchers are aware they exist and know they may access them. The gaps in existing metadata practices we found among data sources show the importance of *how* information about data sources and data access is communicated, not merely *what* is communicated. Existing metadata schemas are necessary but need to improve their ability to improve discovery of restricted data; they currently do not provide sufficient structured metadata to document restricted data access procedures.

Based on the lack of metadata identified in the 137 data sources from this study, we believe a logical first step to improving their discoverability would be to have them develop metadata that align with a schema used by national aggregators like Lunaris (Digital Research Alliance of Canada 2023)—Canada's research data discovery platform—so that they can be harvested by those systems. Creating metadata for these data sources would in turn make them harvestable in other systems such as Google Dataset Search.

Another recommendation resulting from this study is that existing metadata standards bodies like DataCite (DataCite 2021), the Data Documentation Initiative (2023), and the W3C Data Catalog Vocabulary (World Wide Web Consortium 2023) must begin to develop more robust metadata to account for the restricted data access request process. Without metadata to sufficiently describe that process, it may be findable, but the access challenges researchers face determining whether they are eligible to access a dataset will persist.

Support and training for data documentation best practices

From our grading exercise, only four data sources received an "A" grade for the "Data Documentation" attribute. This component of discoverability is critical for researchers as, without the ability to view and/or interact with the data, a researcher cannot know whether the data are suitable to answer the research question unless they are sufficiently well documented.

Careful and deliberate documentation of data can be onerous and time consuming but provides the minimum benefit for the researcher/organization who collected the data and a maximum benefit for secondary users of the data. These incentives are clearly reflected in the sub-optimal documentation of the health data sources we investigated. The organizations that fared well in our grading exercise were evidence of this as large organizations with many administrative staff were shown to have well-documented data, compared to data sources from small research groups run mostly by academics who had little to no documentation.

While this study has focused primarily on the issues of discovery and access of restricted data, the lack of documentation found in the health data sources indicate that much of the data we identified may not be usable. Better training and support for data stewards to develop strong data documentation are needed to facilitate secondary use of restricted data, including but not limited to user guides, data dictionaries, documented code, and readme files. Funding bodies implementing data management or sharing requirements, national organizations like the Alliance who promote data management best practices, and academic institutions who host or support restricted data sources should invest in the development of training programs for the data stewards responsible for these sources to ensure that they have the skills necessary to develop sufficient documentation to enable data repurposing or reuse.

Exploring the national landscape of access-limited data sources

Returning to the first step of this study, our preliminary scoping of access-limited data in Canada identified data sources from across sectors and disciplinary areas, capturing a diverse landscape of data related to the sciences, infrastructure, environment, demography, agriculture, and others. While our analysis was restricted to health-related data sources within our sample, we see value in applying (or adapting) our methodological lens, materials, and findings to an exploration of the remaining data sources that were not our focus. Questions worth investigating include the following: Do discovery and access trends among restricted data sources vary widely by field/topic and sector? How well do non-health sources perform against our grading rubric? Is a standardized framework for making disparate restricted data more discoverable and accessible a desirable or possible outcome? At minimum, a thorough comparison of the discovery and access characteristics of non-health data sources in our inventory will provide insight into the challenges users and curators face when interacting with data that cannot be shared openly.

This study provides a preliminary evaluation of Canadian restricted health data sources' ability to make their data discoverable and accessible and highlights key gaps that need to be addressed by the RDM community to improve their use. As the Canadian government continues to release policies that call for data to be FAIR, data sharing infrastructure must accommodate the complexity of needs required for restricted data to be found, acquired, and used. Without suitable digital infrastructure, better metadata, standardized workflows for accessing data, and robust documentation to facilitate sec-

ondary use, these valuable datasets will remain hidden, insufficiently supported, and underutilized.

Limitations

We acknowledge that the 137 data sources identified in our study do not represent all Canadian access-limited data sources, and some data sources may have been omitted. That said, our team of experts in data discovery and access searched as comprehensively as possible and reached out to experts in the field to identify sources we may have missed. One area that may be less represented than others is in the province of Quebec, where we relied on the expertise of others to provide us with French data sources.

With respect to the grading exercise, our scores were limited to what was available from the public version of the data source only. It may be possible that additional description, metadata, and documentation are available once a data request is made, and data are acquired. We did not want to burden data sources with false requests, which is why we omitted this step. Furthermore, our approach was meant to mimic someone who was interested in the data and browsing for it in the data source as if they were looking for data for the first time.

Article information

Editor
Carolyn Emery

History dates
Received: 14 June 2023
Accepted: 18 March 2024
Version of record online: 1 August 2024

Copyright

Data availability
The grading rubric, access-limited data inventory, data dictionary, and de-identified grading results of restricted health data sources are freely available via the Open Science Framework at https://osf.io/ubzn2/. The grading results have been de-identified because our study's focus was to evaluate the current state of discovery and access in Canadian restricted data sources, not shame restricted data sources that may have received low grades. Identified grading results are available from the corresponding author upon request. If making a request, please provide the corresponding author with information about why and how you will use the data.

Author information

Author ORCIDs
Kevin B. Read https://orcid.org/0000-0002-7511-9036
Grant Gibson https://orcid.org/0000-0003-0750-3850
Amber Leahey https://orcid.org/0000-0001-6476-223X
Lynn Peterson https://orcid.org/0000-0002-3574-0346
Sarah Rutley https://orcid.org/0000-0001-5645-5803
Kelly Stathis https://orcid.org/0000-0001-6133-4045

Author contributions
Conceptualization: KBR, GG, AL, LP, SR, VS, KS
Data curation: KBR, GG, AL, LP, SR, VS, KS
Formal analysis: KBR, GG, AL, LP, SR, JS, VS, KS
Investigation: KBR, GG, AL, LP, SR, JS, VS, KS
Methodology: KBR, GG, AL, LP, VS, KS
Project administration: KBR
Resources: KBR
Supervision: KBR
Validation: KBR, GG, AL, LP, SR, JS, VS, KS
Visualization: KBR
Writing – original draft: KBR, GG, AL, SR, VS, KS
Writing – review & editing: KBR, GG, AL, LP, SR, JS, VS, KS

Competing interests
The authors declare there are no competing interests.

Supplementary material

Supplementary data are available with the article at https://doi.org/10.1139/facets-2023-0102.

References

Bekemeier, B., Park, S., Backonja, U., Ornelas, I., and Turner, A.M. 2019. Data, capacity-building, and training needs to address rural health inequities in the Northwest United States: a qualitative study. Journal of the American Medical Informatics Association, **26**(8–9): 825–834. doi:10.1093/jamia/ocz037.

Boland, M.R., Karczewski, K.J., and Tatonetti, N.P. 2017. Ten simple rules to enable multi-site collaborations through data sharing. PloS Computational Biology, **13**(1): e1005278. doi:10.1371/journal.pcbi.1005278. PMID: 28103227.

Bonomi, L., Huang, Y., and Ohno-Machado, L. 2020. Privacy challenges and research opportunities for genomic data sharing. Nature Genetics, **52**(7): 646–654. doi:10.1038/s41588-020-0651-0. PMID: 32601475.

Byrd, J.B., Greene, A.C., Prasad, D.V., Jiang, X., and Greene, C.S. 2020. Responsible, practical genomic data sharing that accelerates research. Nature Reviews Genetics, **21**(10): 615–629. doi:10.1038/s41576-020-0257-5. PMID: 32694666.

Choudhury, S., Fishman, J.R., Mcgowan, M.L., and Juengst, E.T. 2014. Big data, open science and the brain: lessons learned from genomics. Frontiers in Human Neuroscience, **8**: 239. doi:10.3389/fnhum.2014.00239. PMID: 24904347.

Clayton, G.L., Elliott, D., Higgins, J.P.T., and Jones, H.E. 2021. Use of external evidence for design and Bayesian analysis of clinical trials: a qualitative study of trialists' views. Trials, **22**(1): 789. doi:10.1186/s13063-021-05759-8. PMID: 34749778.

Committee on Data for Science and Technology (CODATA). no date. Data steward definition[Website]. Available from https://codata.org/rdm-terminology/data-steward/ [accessed November 2023].

Confederation of Open Access Repositories. 2021. Controlled vocabularies for repositories: access rights. Available from https://vocabularies.coar-repositories.org/access_rights/ [accessed June 2022].

DataCite. 2021. DataCite Metadata Schema[Website]. DataCite Schema. Available from https://schema.datacite.org/ [accessed June 2022].

Data Documentation Initiative. 2023. Metadata Schemas[Website]. Available from https://ddialliance.org/products/overview-of-current-products [accessed December 2023].

Digital Research Alliance of Canada. 2022a. Network of Experts. Available from https://alliancecan.ca/en/services/research-data-management/network-experts [accessed October 2023].

Digital Research Alliance of Canada. 2022b. Research data management. Available from https://alliancecan.ca/en/services/research-data-management [accessed June 2022].

Digital Research Alliance of Canada. 2022c. Canadian digital research infrastructure needs assessment. Available from https://alliancecan.ca/en/initiatives/canadian-digital-research-infrastructure-needs-assessment [accessed June 2022].

Digital Research Alliance of Canada. 2023. Lunaris. Available fromhttps://www.lunaris.ca/en [accessed June 2022].

Dublin Core Metadata Innovation. 2020. DCMI metadata terms: access rights. Available from https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ [accessed April 2022].

Gabelica, M., Bojčić, R., and Puljak, L. 2022. Many researchers were not compliant with their published data sharing statement: mixed-methods study. Journal of Clinical Epidemiology, **150**, 33. doi:10.1016/j.jclinepi.2022.05.019.

Garrison, N.A., Barton, K.S., Porter, K.M., Mai, T., Burke, W., and Carroll, S.R. 2019. Access and management: indigenous perspectives on genomic data sharing. Ethnicity & Disease, **29**(Suppl 3): 659–668. doi:10.18865/ed.29.S3.659.

Government of Canada. 2020a. Roadmap for Open Science. Available from https://www.science.gc.ca/eic/site/063.nsf/eng/h_97992.html [accessed June 2022].

Government of Canada. 2021. Tri-Agency Research Data Management Policy. Government of Canada Policies and Guidelines. Available from http://science.gc.ca/eic/site/063.nsf/eng/h_97610.html [accessed April 2021].

Government of Canada, S.C. 2020b. Canadian Research and Development Classification (CRDC) 2020 Version 1.0—Field of Research (FOR). Available from https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=1278187 [accessed June 2022].

Hanna, C., Lemmon, E., Ennis, H., Jones, R., Hay, J., Halliday, R., et al. 2021. Creation of the first national linked colorectal cancer dataset in Scotland: prospects for future research and a reflection on lessons learned. International Journal of Population Data Science, **6**(1): 1654. doi:10.23889/ijpds.v6i1.1654.

Ho, H.K.K., Görges, M., and Portales-Casamar, E. 2018. Data access and usage practices across a cohort of researchers at a large tertiary pediatric hospital: qualitative survey study. JMIR Medical Informatics, **6**(2): e32. doi:10.2196/medinform.8724.

Kendall, M.G., and Maurice, G., 1907-1983. 1990. Rank correlation methods. 5th ed. E. Arnold, London.

Knosp, B.M., Craven, C.K., Dorr, D.A., Bernstam, E.V., and Campion, T.R. 2022. Understanding enterprise data warehouses to support clinical and translational research: enterprise information technology relationships, data governance, workforce, and cloud computing. Journal of the American Medical Informatics Association, **29**(4): 671–676. doi:10.1093/jamia/ocab256.

Leahey, A. 2014. Building a web based health data search tool using DDI. Available from http://summit.sfu.ca/item/13952 [accessed June 2022].

Lugg-Widger, F.V., Angel, L., Cannings-John, R., Hood, K., Hughes, K., Moody, G., and Robling, M. 2018. Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: managing the morass. International Journal of Population Data Science, **3**(3): 432. doi:10.23889/ijpds.v3i3.432.

Mpango, J., and Nabukenya, J. 2019. A qualitative study to examine approaches used to manage data about health facilities and their challenges: a case of Uganda. AMIA - Annual Symposium Proceedings, **2019**(101209213): 1157–1166.

Nancarrow, S. 2013. Barriers to the routine collection of health outcome data in an Australian community care organization. Journal of Multidisciplinary Healthcare, **6**(101512691): 1–16. doi:10.2147/JMDH.S37727.

Page, M.J., Nguyen, P.-Y., Hamilton, D.G., Haddaway, N.R., Kanukula, R., Moher, D., and Mckenzie, J.E. 2022. Data and code availability statements in systematic reviews of interventions were often missing or inaccurate: a content analysis. Journal of Clinical Epidemiology, **147**(22): 1–10. doi:10.1016/j.jclinepi.2022.03.003.

Pongiglione, B., Torbica, A., Blommestein, H., De Groot, S., Ciani, O., Walker, S., et al. 2021. Do existing real-world data sources generate suitable evidence for the HTA of medical devices in Europe? Mapping and critical appraisal. International Journal of Technology Assessment in Health Care, **37**(1): e62. doi:10.1017/S0266462321000301.

Prince, K., Jones, M., Blackwell, A., Simpson, A., Meakins, S., and Vuylsteke, A. 2018. Barriers to the secondary use of data in critical care. Journal of the Intensive Care Society, **19**(2): 127–131. doi:10.1177/1751143717741082.

Rahimzadeh, V., Schickhardt, C., Knoppers, B.M., Sénécal, K., Vears, D.F., Fernandez, C.V., et al. 2018. Key implications of data sharing in pediatric genomics. JAMA Pediatrics, **172**(5): 476–481. doi:10.1001/jamapediatrics.2017.5500.

Read, K.B., Ganshorn, H., Rutley, S., and Scott, D.R. 2021. Data-sharing practices in publications funded by the Canadian Institutes of Health Research: a descriptive analysis. CMAJ Open, **9**(4): E980–E987. doi:10.9778/cmajo.20200303.

Read, K.B., Gibson, G.A., Leahey, A., Peterson, L., Rutley, S., Smith, V., et al. 2022. Canadian data source identification and evaluation datasets. Available from https://osf.io/ubzn2/ [accessed June 2022].

Sarwate, A.D., Plis, S.M., Turner, J.A., Arbabshirani, M.R., and Calhoun, V.D. 2014. Sharing privacy-sensitive access to neuroimaging and genetics data: a review and preliminary validation. Frontiers in Neuroinformatics, **8**, 35. doi:10.3389/fninf.2014.00035.

Saulnier, K.M., Bujold, D., Dyke, S.O.M., Dupras, C., Beck, S., Bourque, G., and Joly, Y. 2019. Benefits and barriers in the design of harmonized access agreements for international data sharing. Scientific Data, **6**(1): 1–6. doi:10.1038/s41597-019-0310-4.

Siminski, S., Kim, S., Ahmed, A., Currie, J., Benns, A., Javanbakht, M., et al. 2021. A virtual data repository stimulates data sharing in a consortium. Online Journal of Public Health Informatics, **13**(3): e19. doi:10.5210/ojphi.v13i3.10878.

Siu, L.L., Lawler, M., Haussler, D., Knoppers, B.M., Lewin, J., Vis, D.J., et al. 2016. Facilitating a culture of responsible and effective sharing of cancer genome data. Nature Medicine, **22**(5): 464–471. doi:10.1038/nm.4089.

Sydes, M.R., Johnson, A.L., Meredith, S.K., Rauchenberger, M., South, A., and Parmar, M.K.B. 2015. Sharing data from clinical trials: the rationale for a controlled access approach. Trials, **16**(1): 1–6. doi:10.1186/s13063-015-0604-6.

van Schaik, T.A., Kovalevskaya, N.V., Protopapas, E., Wahid, H., and Nielsen, F.G. 2014. The need to redefine genomic data sharing: a focus on data accessibility. Applied & Translational Genomics, **3**(4): 100–104. doi:10.1016/j.atg.2014.09.013.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al. 2016. The FAIR guiding principles for scientific data management and stewardship. Scientific Data, **3**(1): 160018. doi:10.1038/sdata.2016.18.

World Wide Web Consortium. 2023. W3C Data Catalog Vocabulary[Website]. Available from https://www.w3.org/TR/vocab-dcat-3/ [accessed November 2023].