Improving the Discovery of Access-Limited Data

White Paper for Canada’s National Digital Research Infrastructure Organization (NDRIO)

2020-12-14

Primary Contact:
Kevin Read (kevin.read@usask.ca), University of Saskatchewan

Co-Contributors:
Amber Leahey, Scholars Portal; Sarah Rutley, University of Saskatchewan; Victoria Smith, Portage Network; Kelly Stathis, Portage Network

Background

As increasing emphasis is placed on open government and research reproducibility in the landscape of Canadian data, more questions arise about the discovery of and access to all types of data for use in research. Many types of data that are valuable for research and teaching do not currently meet the standards for open discovery and accessibility that we have come to expect from a national digital research infrastructure (DRI).

The Tri-Agency Statement of Principles on Digital Data Management and Canada’s Roadmap for Open Science set expectations that research data underlying publications should be shared for discovery and reuse by others (1,2). However, many types of research data cannot be ethically and/or legally shared openly; sensitive health data, Indigenous data, commercial data, embargoed data, and licensed and proprietary data, among others, have conditions or restrictions placed on access. In addition, some data suffer from limitations to discovery and access due to data size and/or complexity, organizational factors such as resourcing, as well as the technological and historical conditions associated with data that impact their accessibility and discoverability. While the sensitive, restricted, and/or access-limited nature of these data present barriers to integration with current modes of discovery, improvements can be made to better incorporate these data into national DRI for broad discovery and potential reuse.

This white paper addresses the current challenges associated with making access-limited data discoverable, and provides recommendations for how NDRIO can improve the discoverability and reusability of access-limited data in the transition to Canada’s new DRI.

What is access-limited data?

Access-limited data can be defined as data that is not immediately accessible or for which access or discovery is limited. Such data includes:

A. data that can be made available under certain conditions, or
B. data that lacks the appropriate infrastructure to be made accessible or discoverable more broadly.

While concerted efforts have been made to make research data open and accessible via repositories like Scholars Portal and the national Federated Research Data Repository (FRDR), infrastructure limitations currently exist that make access-limited research data difficult to discover. Similarly, there is a wealth of undiscovered or undiscoverable data in Canada with potential value for research—government data, institutionally licensed data (e.g., geospatial data, health data access programs and registries), and nationally licensed consortial data (e.g. Statistics Canada microdata and Census data)—but to date these collections have not been the subject of efforts to enhance discoverability, accessibility, and usability in a systematic way.

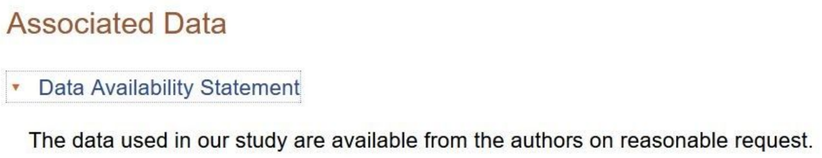
Access-limited data possess great value for research purposes, but lack the infrastructure, policy, and standards to make them more discoverable. Below we describe the challenges associated with the discoverability of access-limited data, and make recommendations for NDRIO to consider as they transform Canada's DRI.

Challenges Associated with the Discovery of Access-Limited Data

Discovery of access-limited research data underlying published articles

Access-limited research data is often referenced in a publication, described on a research lab or consortial website, or hosted by an institutional body or third-party. Because there are few mechanisms for making access-limited research data discoverable, public-facing information about these datasets is often non-existent, and specific procedures for accessing data of this kind are seldom available. Here we provide an example of a data availability statement from a CIHR-funded publication (Fig 1) to highlight a common issue concerning the discoverability and availability of access-limited research data; the statement provides no detail about the research data itself and does not describe the request process—

Fig 1. A typical data availability statement from a publication with access-limited data



Associated Data

▾ Data Availability Statement

The data used in our study are available from the authors on reasonable request.

Because statements like these are often hidden inside the body of a publication, their discoverability is extremely limited. This is concerning because access-limited research data have substantial value, often due to the nature of how it is collected. For example, access-limited data is frequently generated by health sciences disciplines, where human subjects participate in potentially high risk research studies. Improving the discoverability of access-limited research data can help avoid research redundancy and improve collaboration between researchers in specific fields to yield better health outcomes. Research on SARS-CoV-2 outcomes for example, requires access to patient health records and other private health information. At present, access to such data is dependent on specific jurisdictional requirements which do not promote or encourage data-sharing or inter-jurisdictional research.

Beyond the health sciences, the same principles can be applied to other kinds of access-limited data such as commercial, licensed, or embargoed research data; by increasing the discoverability of their data, stakeholders within these disciplines can garner interest from suitable research communities while maintaining the necessary access procedures to ensure their data is used responsibly. Without adequate metadata, clear access instructions, and sufficient infrastructure to make access-limited data discoverable, this valuable data will remain hidden and unusable.

Discovery of access-limited data that can be used for research

In addition to access-limited research data that underlies published findings, there is the challenge of identifying data that could be discoverable and valuable for research purposes, but does not currently have the support or infrastructure necessary to enable its discovery and/or use. Addressing this category of access-limited data is challenging because it requires identifying and assigning value to collections that are unexposed or untouched. Administrative, academic, government and institutionally licensed data such as health registry, water security, property, and industry data all have great potential to be used for research; however a commitment is needed to locate these types of data, establish relationships with the stewards of these data, and make decisions about which data are valuable enough to receive resources to enhance discoverability and reusability.

Key Recommendations to Improve Access-Limited Data Discovery

Develop Suitable Data Infrastructure

For a variety of reasons, access-limited data suffer from a lack of reusable infrastructure to support discovery by researchers and data managers. Due to data sensitivities and/or data licensing restrictions that preclude these kinds of data from being shared broadly, infrastructure may not readily exist or be available to support the reuse of access-limited data. Complexities surrounding the type of data, the size or volume, and or the availability of metadata can also pose barriers to making access-limited data discoverable in the current DRI. Some infrastructure does exist to support workflows for sharing restricted data, including built-in permissions for restricting access in data repositories such as Dataverse (3). However, these generic workflows do not support sensitive human health data, nor do they support data licensed to institutions that require researchers to navigate data agreements, licensing terms, and access conditions that are often required (but not intuitive to use or available to them openly).

Access-limited data can also suffer from a lack of generalizability for broad audiences and can have certain conditions that do not make it suitable or at times possible to be disseminated in the same ways as other research data. For example, the access conditions required for certain access-limited data is not currently supported in generic workflows for deposit and sharing in most data repositories. These data often require metadata and curation not easily connected to or managed in current national DRI tools. For sensitive data, curators must assess the privacy and ethics requirements for certain data sets, review consent documentation in relation to privacy legislation and ethics policy, develop and support data transfer and storage access workflows, and assess potential risk to participants related not the potential use and access to the dataset (which is distinct from the risks associated with participating in the research).

Recommendations

- Support the development of bilingual national infrastructure for federated discovery of access-limited data, including:
    - Enhancements to the FRDR Discovery Service to align with Canadian access-limited data resources (e.g., repositories, registries, catalogues, research portals);

    - Support existing access-limited data repositories to improve interoperability, harvestability, and sustainability to be discoverable in a federated system;
    - Engage with communities that currently provide access and or support access to licensed, restricted and sensitive data for use in research.
- Develop data repository infrastructure to accommodate sensitive data, including mechanisms to control access and secure data transfer, including:
    - Support for the initiative to add zero-knowledge encryption to FRDR datasets, making it possible for FRDR to securely accept sensitive data submissions (e.g. human subjects research data) (4);
    - Similarly, for Dataverse, this can include enhancements to make the connection to Canadian trusted secure remote storage locations where sensitive data are currently being stored and managed (5), and, improved mechanisms for tagging sensitive data for end-user understanding about sensitive data access conditions (6).
- Develop data repository infrastructure to accommodate licensed and restricted data, including mechanisms to enable customized terms of use and data use agreements;
- Establish and scale up existing communities to develop best practices and coordinate the management and stewardship of access-limited data across institutions and jurisdictions;
- Ensure that bilingual training programs are developed for researchers, research administrators, data managers, and other institutional supports across the research ecosystem as access-limited infrastructure is created.

Develop Metadata and Reporting Standards for Access-Limited Data

An existing barrier to making access-limited research data discoverable is that few suitable metadata standards exist or are adopted to support the discovery, access, and use of access-limited data. For example, while many researchers make their data available by request or application, there are no formal standards for indicating these requirements or describing them for requestors/applicants. Furthermore, there is a lack of reporting standards for access-limited research data from funders and publishers, which has resulted in inconsistent data sharing practices from Tri-agency funded researchers who collect sensitive or restricted data. Without reporting standards, researchers will not be held accountable to provide sufficient descriptive information to find, access, and share their access-limited research data, and therefore this valuable data will remain undiscoverable and unusable.

Recommendations

- Develop metadata standards to support the discovery and reuse of access-limited research data, including elements that describe data application/request processes;
- Weave access-limited metadata standards into DRI infrastructure throughout the data lifecycle, with attention to both researcher and end-user needs;
- Establish reporting standards in collaboration with funders and publishers to increase the support and expectation for making access-limited data discoverable.

Establish Guidance and Policy to Improve the Discovery of Access-Limited Data

Along with the need for stronger infrastructure to support access-limited data, efforts should be made to support access-limited data discovery through improved guidance and policy. Presently, there are no existing recommendations on what, how, and where researchers should make access-limited data discoverable. Without strong guidance, valuable data will remain hidden and without adequate stewardship. Policy innovations to support the discovery of access-limited data must address research guidelines governing data creation and collection, along with guidelines governing data management and sharing.

For access-limited data that can be used for research, no sustainable initiatives exist to identify, evaluate, and expose for discovery high value collections across sectors. To address this, close communication with diverse stakeholders responsible for this data will be key. By devoting resources to this effort, NDRIO can establish stronger relationships with Canadian data community partners, develop guidelines for identifying and assessing the value of hidden data collections, establish a community of experts to aid in the decision making process of making access-limited data discoverable, and take the necessary steps to ensure chosen data collections are continually supported as they are released to the public.

Recommendations

- Engage stakeholders from research ethics, institutional research offices, funding bodies, repositories, publishers, and others to ensure access-limited data requirements are embedded in their policies and aligned to support inter-jurisdictional discovery and access;
- Create opportunities for the abovementioned stakeholders to coordinate their policy innovations to create an integrated policy framework;
- Identify and facilitate the discovery of access-limited data that may be hidden or unsupported;
- Convene an expert advisory board to review and select access-limited data for discovery;
- Incentivize owners and stewards of select access-limited data to make their data discoverable;
- Foster collaborative partnerships with academic researchers to improve access-limited data discovery.

Concluding Remarks

As Canadian research continues its trajectory towards open science (2), there is an opportunity for a paradigm shift in how research data are created, such that the capacity to retain, make discoverable, and potentially share access-limited data constitute the new default model. This model will require new standards and best practices to protect restricted, sensitive, or licensed data while ensuring that data can still be discovered, interpreted, and reused by researchers when possible. Identifying and selecting under resourced, hidden data that have potential value for future research must also be woven into this model. The success of an access-limited data model will depend on integrated support across jurisdictions and reflect the entirety of the research lifecycle, from the funding application to the creation and collection of data, through to the preservation and sharing of published datasets. NDRIO is singularly positioned to develop and coordinate efforts for improving the discovery and usability of access-limited research data, and uncovering dormant access-limited data that can be utilized in future research efforts.

This proposal is endorsed by the Canadian Association of Research Libraries (CARL) and the CARL Portage Network.

References

1. Government of Canada, “Tri-Agency Statement of Principles on Digital Data Management” (Innovation, Science and Economic Development Canada, December 21, 2016), https://www.ic.gc.ca/eic/site/063.nsf/eng/h_83F7624E.html.

2. Government of Canada, “Roadmap for Open Science” (Office of the Chief Science Advisor of Canada, February 2020), https://www.ic.gc.ca/eic/site/063.nsf/eng/h_97992.html.

3. Dataverse Project, “Dataset + File Management” (Dataverse.org, November 9, 2020), https://guides.dataverse.org/en/latest/user/dataset-management.html.

4. Portage Network, “SFU Working Towards Zero Knowledge Encryption of Sensitive Data in FRDR,” June 1, 2020, https://portagenetwork.ca/news/sfu-working-towards-zero-knowledge-encryption-of-sensitive-data-in-frdr/

5. Dataverse Trusted Remote Storage Agents (TRSA) Project https://cyberimpact.us/dataverse-trusted-remote-storage-agent-update/

6. Dataverse DataTags Project https://privacytools.seas.harvard.edu/datatags-dataverse