Where is all the research data?

Opportunities for Libraries and Portage to Improve the Discovery of Canadian Research Data
Who am I?

Health Sciences Librarian at the University of Saskatchewan

Working in research data discovery for >8 years:

- NYU School of Medicine, Lead of Data Discovery
- National Library of Medicine, Associate Fellow

Current Chair of Portage’s Data Discovery Expert Group

kevin.read@usask.ca
The plan for today

Introduce a research project designed to locate Canadian research data

Identify challenges associated with improving the discovery of Canadian research data

Engage in a discussion about:

- Strategies for libraries to improve the discovery of Canadian research data
- Portage’s role in making Canadian research data more discoverable
The catalyst

http://science.gc.ca/eic/site/063.nsf/eng/h_97610.html
The catalyst

http://science.gc.ca/eic/site/063.nsf/eng/h_97610.html
The catalyst

http://science.gc.ca/eic/site/063.nsf/eng/h_97610.html
Data Management Plans

Data Deposit

"all digital research data, metadata and code that directly support research conclusions..."
“all digital research data, metadata and code that **directly support research conclusions**...”
Further motivation

Adopt an Open Science approach to federally funded scientific and research outputs.

Develop strategies and tools to implement FAIR data principles to ensure interoperability of scientific and research data.

The Data Strategy Roadmap and the Open Science Action Plan should be aligned.

http://science.gc.ca/eic/site/063.nsf/eng/h_97992.html
My questions

How have Tri-agency funded researchers shared their data in the past (if at all)?

Where can you find Tri-agency funded research data?

How well have researchers utilized RDM best practices when sharing their data?
My questions

How have Tri-agency funded researchers shared their data in the past (if at all)?
Where can you find Tri-agency funded research data?
How well have researchers utilized RDM best practices when sharing their data?

How prepared are researchers for these new policies?
The research project

Research goals:

1. Document how and where CIHR-funded researchers share their data

2. Compare Canadian researchers’ current data practices to the Tri-agency’s proposed framework for RDM and sharing

Research team:

- Kevin Read, MLIS, MAS, University of Saskatchewan
- Heather Ganshorn, MLIS, University of Calgary
- David Scott, MLIS, MA, University of Lethbridge
- Sarah Rutley, MLIS, MA, University of Saskatchewan
What we did

Extracted metadata from all CIHR-funded research articles that indicate data was shared in a repository.
What we did

Extracted metadata from all CIHR-funded research articles that indicate data was shared in a repository

Extracted metadata from all CIHR-funded research articles that had a Data Availability Statement
What we did

949 articles


4039 articles

https://www.ncbi.nlm.nih.gov/pmc
<table>
<thead>
<tr>
<th>Metadata Element</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal title</td>
<td>The journal title abbreviation, full journal title, or ISSN number</td>
</tr>
<tr>
<td>Publication date</td>
<td>The date that the article was published.</td>
</tr>
<tr>
<td>Publication Type</td>
<td>Used to identify the type of article indexed for MEDLINE</td>
</tr>
<tr>
<td>Author affiliation</td>
<td>Institutional affiliation and address (including email address, when available) of the authors of the article as it appears in the journal.</td>
</tr>
<tr>
<td>Grant number</td>
<td>Includes research grant numbers, contract numbers, or both that designate financial support by funding sources</td>
</tr>
<tr>
<td>Data Availability Statement</td>
<td>Author instructions on if/where data is available</td>
</tr>
<tr>
<td>Data Repository</td>
<td>Where data related to the article has been deposited. The complete list of databanks is available at <a href="http://www.nlm.nih.gov/bsd/medline.databank.source.html">//www.nlm.nih.gov/bsd/medline.databank.source.html</a>.</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>Includes all words in the acknowledgement section of an article (e.g., “figshare[ack]”).</td>
</tr>
<tr>
<td>MeSH Major Topic Headings</td>
<td>A MeSH term that is one of the main topics discussed in the article.</td>
</tr>
<tr>
<td>Article body - Key Terms</td>
<td>Includes all key terms in the body of an article except for the Abstract and References.</td>
</tr>
</tbody>
</table>
What we did

- CIHR-funded article
- Article metadata

How/where are the data shared (if anywhere)?

Are researchers’ data availability statements informative?

What documentation did researchers share in addition to the research data?
What we’ve found so far
Outcomes to date

Reviewed a random sample of 354 articles

- Data sharing methods are varied and inconsistent
- Data availability statements are unclear
- Included documentation is limited beyond the tables and figures commonly shared within an article
- 55% did not share data
Data sharing status

- Data not shared: 195
- Data accessible: 119
- Data available: 66
- Data sharing not applicable/possible: 40

n=354
Data sharing status

- Data not shared: 195
- Data accessible: 119
- Data available: 66
- Data sharing not applicable/possible: 40

n=354
Data sharing status

We could actually retrieve the data

- Data not shared: 195
- Data accessible: 119
- Data available: 66
- Data sharing not applicable/possible: 40

n=354
Data sharing status

We could actually retrieve the data

Authors said data was available (e.g. by request)

n=354

Data not shared | Data accessible | Data available | Data sharing not applicable/possible
---|---|---|---
195 | 119 | 66 | 40
Data sharing status

- We could actually retrieve the data: 195
- Authors said data was available (e.g. by request): 119
- Reasons provided for not sharing:
  - Data not shared: 66
  - Data sharing not applicable/possible: 40

Total: n=354
Data sharing methods

- Available via a repository: 89
- Available within the contents of the article: 84
- Data sharing is not applicable/possible: 72
- Available upon reasonable request: 40
- Available upon request via application: 36
- Available via a website: 21
- Data will be made available at a future date: 13
- Available at a future date: 12
- Not applicable: 4

n=354 [not mutually exclusive]
Data sharing method: Repositories

- Available via a repository: 89 (n=354)
- Available within the contents of the article: 94
- Data sharing is not applicable/possible: 72
- Available upon reasonable request: 40
- Available upon request: 36
- Available upon request via application: 21
- Available via a website: 13
- Available at a future date: 12
- Data will be ins: 4

n=354 [not mutually exclusive]
Data repositories used

![Bar chart showing the frequency of data repositories used, with Protein DataBank being the most frequent at 31, followed by Gene Expression Omnibus at 19, and so on. The chart includes repositories such as Gene Expression Omnibus, European Genome-Phenome Archive, GitHub, ISRCTN, BioProject, Dryad, Zenodo, and Uniprokd. The total number of repositories used is n=89.](image)
Data sharing method: Available in the article...

n=354 [not mutually exclusive]
Data sharing method: Available in the article...

- Available via a repository: 89
- Available within the contents of the article: 94
- Data sharing is not applicable/possible: 72
- Available upon reasonable request: 40
- Available upon request: 36
- Available upon request via application: 21
- Available via a website: 13
- Available at a future date: 12
- Data will be made available: 4

n=354 [not mutually exclusive]
Data sharing method: Available in the article...

Only 3% actually had data in the supplement

n=354 [not mutually exclusive]
Data sharing method: Available data

Data is “available” but not readily accessible

<table>
<thead>
<tr>
<th>Available method</th>
<th>Frequency (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Available via a repository</td>
<td>89</td>
</tr>
<tr>
<td>Available within the contents of the article</td>
<td>84</td>
</tr>
<tr>
<td>Data sharing is not applicable/possible</td>
<td>72</td>
</tr>
<tr>
<td>Available upon reasonable request</td>
<td>40</td>
</tr>
<tr>
<td>Available upon request</td>
<td>36</td>
</tr>
<tr>
<td>Available upon request via application</td>
<td>21</td>
</tr>
<tr>
<td>Available via a website</td>
<td>13</td>
</tr>
<tr>
<td>Data will be made available at a future date</td>
<td>12</td>
</tr>
</tbody>
</table>

n=354 [not mutually exclusive]
Data sharing method: Available upon request

Data is “available” but not readily accessible

- Available via a repository: 89
- Available within the contents of the article: 94
- Available in the supplementary files: 72
- Data sharing is not applicable/possible: 40
- Available upon reasonable request: 36
- Available upon request: 21
- Available upon request via application: 13
- Available via a website: 12
- Data will be made available at a future date: 4

n=354 (not mutually exclusive)
Lack of clarity and detail

Data Availability Statement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.
Lack of clarity and detail

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.
Lack of clarity and detail

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Data availability
The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request. All data generated or analysed during this study are included in this published article (and its supplementary information files).
Lack of clarity and detail

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

To maintain participant privacy, the minimal anonymized dataset is available by request from the study whose authors may be contacted at csdewa@ucdavis.edu. The results of this manuscript may be replicated in a similar context and participants. There is no data access committee from which the data can be accessed upon request.
Data sharing method: Websites

Most websites did not work or did not take us directly to the data

n=354 [not mutually exclusive]
The data availability decline...

A

B

C

D

The data availability decline...

Restricted or licensed data is still valuable data
The policy says...

“all digital research data, metadata and code that directly support research conclusions...”
The value of making harder to access (and find) research data more discoverable
What are the benefits?

Locate data that may be undiscoverable otherwise

Researchers can remain platform agnostic

Can develop better standards to index:
- Research data available in various locations
- Research data with different access restrictions

Gain a more complete picture of Canadian research data sharing
Challenges

Really hard to do!

How to track data that is everywhere?

Requires:

- Constant monitoring of data
- Close connection to research communities who place restrictions on data access
- Stronger relationships with repositories
- Improved APIs
POLL

Do you see value in making Canadian research data that is not easily accessible (e.g., licensed data, restricted data, data available in external repositories/websites) more discoverable?
Who is responsible for making this research data discoverable?
The case for Portage and FRDR: Reporting Tool
The case for Portage and FRDR: More Context
The case for Portage and FRDR: Discoverability
The case for libraries: Front line opportunity
The case for libraries: Front line opportunity

Shares or restricts access to research data

Researcher

Institutional Workflow
The case for libraries: Front line opportunity

- Shares or restricts access to research data
- Library indexes data regardless of where it is stored
- Institutional Workflow
The case for libraries: Front line opportunity

Shares or restricts access to research data

Library indexes data *regardless of where it is stored*
Libraries and Portage working together

- Shares or restricts access to research data
- Library indexes data *regardless of where it is stored*
In its current state, would your library have the ability to track all of the research data shared by your research community (regardless of where it is stored)?
Real world example: Data Discovery Collaboration

[Image: Data Discovery Collaboration website]

**How the Data Catalog Works**

The Data Catalog provides a simple search interface to discover biomedical research data at the local institution where that data is located. Because we do not require researchers to share their data in a repository, much of the research data you will find in each respective Data Catalog is not available anywhere else.

**Locate Research Data Not Discoverable by Other Means**

The Data Catalog allows researchers to describe their data in order to make it discoverable, but it does not require them to share the data. The Data Catalog allows users to make requests for data directly through the author, an administrator, or a repository.

[Image: National Survey of U.S. Smartphone Owner Health App Usage Dataset]

**Description**

This dataset contains results of a cross-sectional survey of 1,604 smartphone users from the United States. The survey consists of 36 items assessing sociodemographic characteristics, history of health app use, reasons for use, perceived effectiveness of health apps, reasons for stopping use and general health status. Individuals included in the survey are over 18, speak English and own a smartphone. Respondents in the survey are 50% female, 50% having completed high school or fewer years of education, 60% earning less than $50,000 per year, 30% Hispanic/Latino, 30% Black, 30% White and 10% Asian or other.

**Timeframe**

2015 - 2015

**Geographic Coverage**

National

**Subject Domain**

Delivery of Health Care, Health Status, Population Characteristics

[Image: Access Instructions]

Access via Administrator

**Access Requirements**

Application Required

Author approval required

**Access Instructions**

Application is required, using the Study Proposal Form. For access to the form, please email the administrator, Nicole Contaxis. Once the form is completed and returned, the authors will respond with any additional questions to the data request.


[Image: Data Type]

Surveys

[Image: Partner logos]

https://www.datacatalogcollaborationproject.org/
Data Management Plans

Data Deposit

“all digital research data, metadata and code that directly support research conclusions...”
Who should be responsible for tracking the whereabouts of Canadian research data?
How do we improve the reporting of data sharing in the published literature?
Reporting practices are poor

<table>
<thead>
<tr>
<th>BMJ</th>
<th>Nature</th>
<th>Wiley</th>
<th>Taylor and Francis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please state what the data are (e.g. deidentified participant data), who the data are available from, their publishable contact details (e.g. a generic lab email address or an individual’s ORCID identifier – please ensure you have permission) and under what conditions reuse is permitted. Is there additional information available (e.g. protocols, statistical analysis plans)?</td>
<td>The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.</td>
<td>The data that support the findings of this study are available from the corresponding author on reasonable request.</td>
<td>The data that support the findings of this study are available from the corresponding author, [author initials], upon reasonable request.</td>
</tr>
</tbody>
</table>
# Reporting practices are poor

**Data Availability Statement**

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

<table>
<thead>
<tr>
<th>BMJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please state what the data are (e.g. deidentified participant data), who the data are available from, their publishable contact details (e.g. a generic lab email address or an individual's ORCID identifier – please ensure you have permission) and under what conditions reuse is permitted. Is there additional information available (e.g. protocols, statistical analysis plans)?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nature</th>
</tr>
</thead>
<tbody>
<tr>
<td>The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wiley</th>
</tr>
</thead>
<tbody>
<tr>
<td>The data that support the findings of this study are available from the corresponding author upon reasonable request.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Taylor and Francis</th>
</tr>
</thead>
<tbody>
<tr>
<td>The data that support the findings of this study are available from the corresponding author, [author initials], upon reasonable request.</td>
</tr>
</tbody>
</table>
## Data Availability Statement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

<table>
<thead>
<tr>
<th>BMJ</th>
<th>Nature</th>
<th>Wiley</th>
<th>Taylor and Francis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please state what the data are (e.g. deidentified participant data), who the data are available from, their publishable contact details (e.g. a generic lab email address or an individual’s ORCID identifier – please ensure you have permission) and under what conditions reuse is permitted. Is there additional information available (e.g. protocols, statistical analysis plans)?</td>
<td>The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request</td>
<td>The data that support the findings of this study are available from the corresponding author upon reasonable request.</td>
<td>The data that support the findings of this study are available from the corresponding author, [author initials], upon reasonable request.</td>
</tr>
</tbody>
</table>
Polanin JR, Terzian M. A data-sharing agreement helps to increase researchers’ willingness to share primary data: results from a randomized controlled trial. Journal of Clinical Epidemiology. 2019 Feb 1;106:60-9.

Participants who received a data-sharing agreement were more willing to share their dataset.

A member of the control group is 24% more likely to share her dataset should she receive the data-sharing agreement.

Template data sharing agreements included within publication.
Develop better accessibility metadata

Data Availability

Data have been provided under agreement with the Canadian Multicentre Osteoporosis Study (CaMos). The CaMos has developed an Ancillary Project and Data Release Policy that governs ancillary project approval and access to the data. The Design Analysis and Publications (DAP) Committee acts to oversee the selection and implementation of ancillary projects. They review each project application considering feasibility, priority and its impact on CaMos and will, after due consultation with the applicants, make a decision regarding the acceptability of the project. Ancillary projects may be undertaken in any one of CaMos’ nine regional centres, or as a collaboration among investigators, at least one of which must be a CaMos Centre Director. A formal proposal must be submitted to the DAP Committee for review. Following approval, the authors of the proposal will be notified of the release of data, and will sign an agreement, stating that they will only use the data for the purpose described, will follow the timeline specified for the analysis, and will destroy the data files by a given date. The CaMos principal investigators are Dr. David Goltzman and Dr. Nancy Krieger. To obtain a copy of the Ancillary Project and Data Release Policy and the requirements for proposal submission, or to obtain further information from the CaMos principal investigators about data access, please send an e-mail to info@camos.org.
Better reporting: Work with authors & publishers

Make data sharing reporting the norm for authors (pre-publication)

Work with journals to improve reporting requirements (e.g., ICMJE)

Develop a list of required documentation that are necessary to accompany shared research data in publications
Better reporting: Work with authors & publishers

Make data sharing reporting the norm for authors (pre-publication)

Work with journals to improve reporting requirements (e.g., ICMJE)

Develop a list of required documentation that are necessary to accompany shared research data in publications
Documentation (or lack thereof)

- Figures and/or tables: 201
- Data files: 44
- Transparent reporting form: 42
- Study Protocol: 20
- Data collection instruments: 14
- Data analysis plan/documentation: 12
- Software Code: 11
- Videos: 8
- Data dictionary/codebook: 7
- Readme files: 6
- Image files: 4
- Preservation formats for structured data: 3

n=354 [not mutually exclusive]
The potential health benefits from sharing participant-level clinical research data for the purpose of secondary analysis or meta-analysis have been widely touted. Although some researchers remain wary about sharing data, recent policies and proposals by funders, scientific journals, research institutions, and international health organizations mean that data sharing, in one form or another, is inevitable. Now is therefore the time to focus on developing practices for data sharing that are effective, efficient, equitable, and ethical. In the process, we may need to question the assumption that more is better. Simply making more data openly available may not lead to analyses that are relevant and that are actually applied to improve health.

Reality check on reproducibility

A survey of Nature readers revealed a high level of concern about the problem of irreproducible results. Researchers, funders and journals need to work together to make research more reliable.

25 May 2016

Is there a reproducibility crisis in science? Yes, according to the readers of Nature. Two-thirds of researchers who responded to a survey by this journal said that current levels of reproducibility are a major problem. The ability to reproduce experiments is at the heart of science, yet failure to do so is a routine part of research. Some amount of irreproducibility is inevitable: profound insights can start as fragile signals, and sources of variability are infinite. But, the survey suggests, there is a bigger issue—and something that needs to be fixed. One-third of the survey respondents said that they think about the reproducibility of their own research daily, and more than two-thirds discuss it with colleagues at least monthly.

More related stories
• The pressure to publish pushes down quality
• Research data: Silver lining to irreproducibility
• Statistical issue weighing over misuse of P-values

https://datascience.codata.org/articles/10.5334/dsj-2020-003/

Data Without Software Are Just Numbers

Abstract

Great strides have been made to encourage researchers to archive data created by research and provide the necessary systems to support their storage. Additionally, it is recognised that data are meaningless unless their provenance is preserved, through appropriate meta-data. Alongside this is a pressing need to ensure the quality and archiving of the software that generates data, through simulation, control of experiment or data-collection and that which analyses, modifies and draws value from raw data. In order to meet the aims of reproducibility, we argue that data management alone is insufficient: it must be accompanied by good software practices, the training to facilitate it and the support of stakeholders, including appropriate recognition for software as a research output.

https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

https://www.nature.com/news/reality-check-on-reproducibility-1.19961
How to improve reporting practices?

Incorporate data sharing reporting guidelines into DMP Templates
Summary

- It is essential that we continue to learn about our community’s data sharing practices
  - Would similar studies of NSERC and SSHRC data sharing practices be beneficial to inform future discovery efforts?
Summary

- It is essential that we continue to learn about our community’s data sharing practices
  - Would similar studies of NSERC and SSHRC data sharing practices be beneficial to inform future discovery efforts?
- Our work suggests that research data is shared via multiple methods and platforms, with varying levels of access permitted
Summary

● It is essential that we continue to learn about our community’s data sharing practices
  ○ Would similar studies of NSERC and SSHRC data sharing practices be beneficial to inform future discovery efforts?
● Our work suggests that research data is shared via multiple methods and platforms, with varying levels of access permitted
● Indexing research data regardless of where it is stored will improve the discovery of Canadian research data
Summary

- It is essential that we continue to learn about our community’s data sharing practices
  - Would similar studies of NSERC and SSHRC data sharing practices be beneficial to inform future discovery efforts?
- Our work suggests that research data is shared via multiple methods and platforms, with varying levels of access permitted
- Indexing research data regardless of where it is stored will improve the discovery of Canadian research data
- Important to establish protocols for reporting data sharing methods and including necessary documentation in the sharing workflow
The research project

Study documentation and raw data available:

- Read KB, Ganshorn H, Rutley S, Scott DR. Surveying the landscape of CIHR-funded research data sharing practices: An analysis of the published literature. 2020. [https://osf.io/n9jv5](https://osf.io/n9jv5)

Research team:

- Kevin Read, MLIS, MAS, University of Saskatchewan
- Heather Ganshorn, MLIS, University of Calgary
- David Scott, MLIS, MA, University of Lethbridge
- Sarah Rutley, MLIS, MA, University of Saskatchewan
How do we improve the discovery of data that is available but not immediately accessible?

What methods do you think could improve data sharing reporting practices?
Questions?

kevin.read@usask.ca
References

Baker M. 1,500 scientists lift the lid on reproducibility. 2016;533(7604):452-454.


Noun Project Images:

- https://thenounproject.com/search/?q=researcher&i=3011359
- https://thenounproject.com/search/?q=metadata&i=2563280
- https://thenounproject.com/search/?q=library&i=1244000
- https://thenounproject.com/search/?q=metadata&i=1677708
- https://thenounproject.com/search/?q=data%20repository&i=1630978
- https://thenounproject.com/search/?q=lock&i=3362860
- https://thenounproject.com/search/?q=ecosystem&i=1637681

Polanin JR, Terzian M. A data-sharing agreement helps to increase researchers’ willingness to share primary data: results from a randomized controlled trial. Journal of clinical epidemiology. 2019 Feb 1;106:60-9.