Depositing & Sharing Data Online with McMaster Dataverse

Isaac Pratt, PhD
February 15th, 2022

McMaster University | Research & High Performance Computing | McMaster University | Library

- McMaster University sits on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the "Dish With One Spoon"

McMaster University sits on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the "Dish With One Spoon" wampum agreement.

SESSION RECORDING

- This session is being recorded with the intention of being shared publicly via the web for future audiences.
- In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.
- Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.

CODE OF CONDUCT

- The Sherman Centre and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.
- As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

- Please refer to our code of conduct webpage for more information:

  scds.ca/events/code-of-conduct/

OUTLINE FOR TODAY

1. Why should I share data?
2. Where should I share data?
3. How do I share data?

A NOTE ON TERMINOLOGY

Created by mpanicon
from Noun Project

I use **depositing data** to mean uploading research data to a purpose built online research data repository.

Depositing data does **not** mean sharing data.

WHY DEPOSIT DATA?

Depositing data helps to ensure that data are **securely preserved** and **accessible** (to you) in the long term.

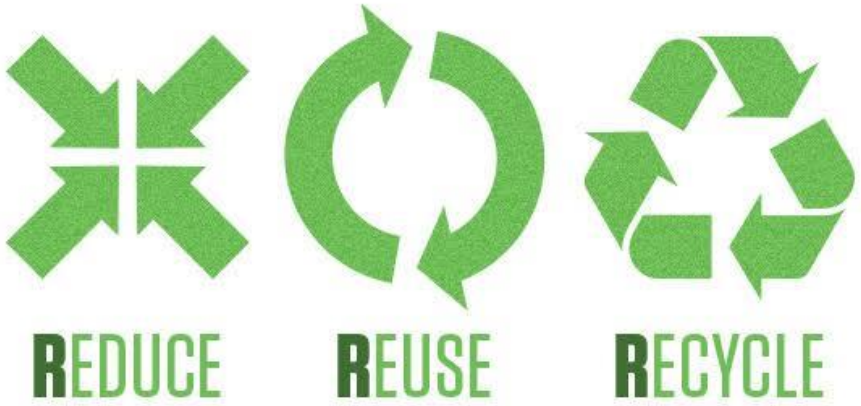
You may want to deposit your data for the following reasons:

- To comply with potential audits
- A journal may request the data to verify or reproduce your results
- Your funder may require it
- To prevent data loss and keep data organized

WHY SHARE DATA?

Openly sharing your data is good for:

- **Society**
- The **academic research community**
- **Your research profile** and reputation



Avoid ‘Single use data’

WHY SHARE DATA?

Improve the **quality** of research

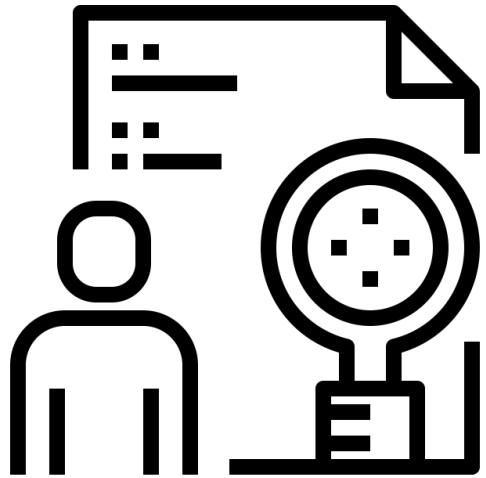
- Allow for verification/reproduction of results by peers
- Potential of combining datasets

Improve the **value** of your research

- Avoid duplication of data collection

Improve the **impact** of your work

- Increases the visibility of research
- Can lead to new collaborations and partnerships



Created by Unlimiticon
from Noun Project

WHY SHARE DATA?

Studies show that **publications with open data are cited more.**

- Publications in PLOS and BMC journals with open data have up to 25% higher citation impact compared to those that don't share data.
  - Collavazi et al, 2020 PLOSOne The citation advantage of linking publications to research data https://doi.org/10.1371/journal.pone.0230416
- Publications of gene expression microarray data have higher citation impact when the data is shared.
  - Piwowar & Vision, 2013 PeerJ Data reuse and the open data citation advantage https://doi.org/10.7717/peerj.175

TRI-AGENCY DATA DEPOSIT REQUIREMENTS

The new Tri-Agency Research Data Management Policy states that:

- "**Grant recipients are required to deposit into a digital repository all digital research data**, metadata and code… in journal publications and pre-prints"
- "The deposit must be made by time of publication"

These new requirements have **not yet been phased in** and there is no current date for when they will take effect.

EXISTING TRI-AGENCY DATA REQUIREMENTS

CIHR:

- **Deposit bioinformatics, atomic, and molecular coordinate data** into the appropriate **public database**
- **Retain original data sets** for a minimum of **five years** after the end of the grant.

SSHRC:

- **Preserve and make available for use by others** all research data collected with the use of SSHRC funds. This must occur within **two years** of the completion of the research project

JOURNAL/PUBLISHER DATA REQUIREMENTS

Many journals are starting to require data sharing or at least **data availability statements**, including:

PLOS https://journals.plos.org/plosone/s/data-availability

Nature https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards

NEJM https://www.nejm.org/about-nejm/editorial-policies

Journals with higher Impact Factors are more likely to have data sharing policies

David B. Resnik et al (2019) Effect of impact factor and discipline on journal data sharing policies, Accountability in Research, 26:3, 139-156, DOI: 10.1080/08989621.2019.1591277

SHARING SENSITIVE DATA

Sensitive data is any data that would cause harm if released openly. This includes personally identifiable information and personal health information.

You cannot share sensitive data openly. If you want to publish or share sensitive data, you have two main options:

1. **Anonymize the dataset:** remove, replace, or redact all sensitive information from datasets prior to upload.
2. Deposit data on a restricted access platform with data sharing agreements.

RESEARCH DATA REPOSITORIES

A **data repository** is a web platform and storage space for researchers to deposit data sets associated with their research. Repositories provide:

- long-term storage and access to research data beyond the life of a grant, research project, or individual careers.
- Discoverability and findability for datasets through features like indexing and DOIs.
- Easy to use shared platforms designed for researchers

RESEARCH DATA REPOSITORIES

Institutional Repositories: **McMaster Dataverse**

External Data Repositories:

- Domain specific https://www.nature.com/sdata/policies/repositories
- General: **FRDR**, Zenodo, Figshare, Mendeley Data, etc

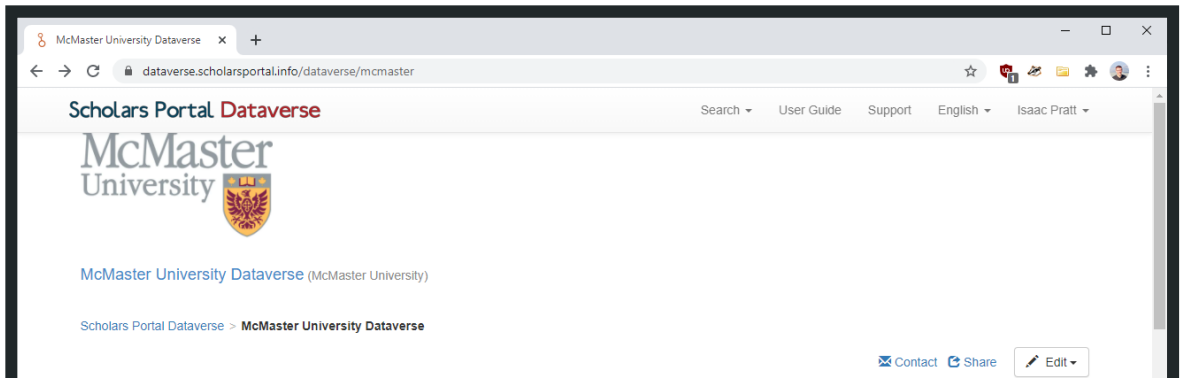
Code repositories: Github, Gitlab, BitBucket, SourceForge

Search for repositories on re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES

MCMASTER DATAVERSE

dataverse.scholarsportal.info/dataverse/mcmaster

- McMaster's Institutional Data Repository is a home for all research data originating from McMaster researchers.
- Provides basic data curation services
- Data is stewarded by professionals at McMaster
- Contains tools for tabular data exploration and analysis

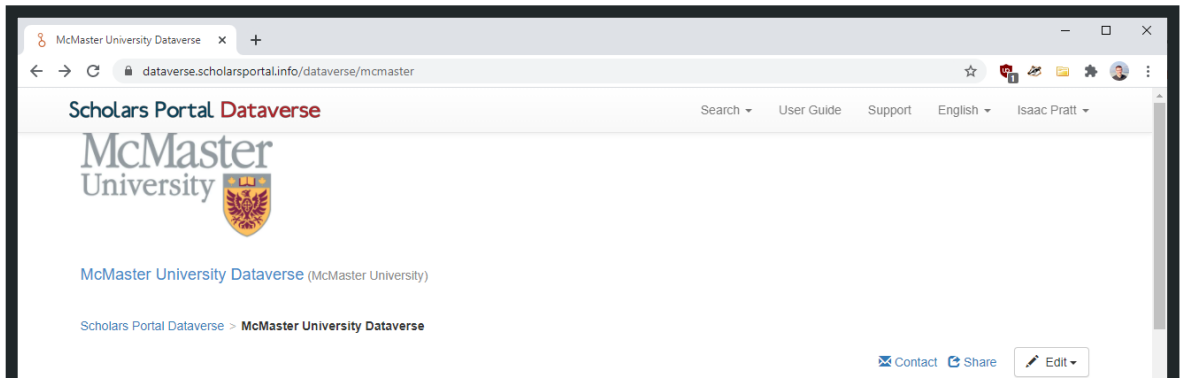


McMaster University | Research & High Performance Computing

McMaster University | Library

MCMASTER DATAVERSE

- Researchers can control what license they use for data sharing
- Researchers can choose whether to share their datasets openly or through limited access.
- Researchers can monitor statistics about the use of their data.
- Deposits can be set up anonymously for double blind reviews.



McMaster University | Research & High Performance Computing

McMaster University | Library

FEDERATED RESEARCH DATA REPOSITORY (FRDR)

https://www.frdr-dfdr.ca/repo/

- Available to any researcher affiliated with a Canadian institution
- Built for large (1 TB+) datasets
- Datasets are actively curated by professional staff at FRDR
- Datasets must be open access but can be embargoed for a one year period

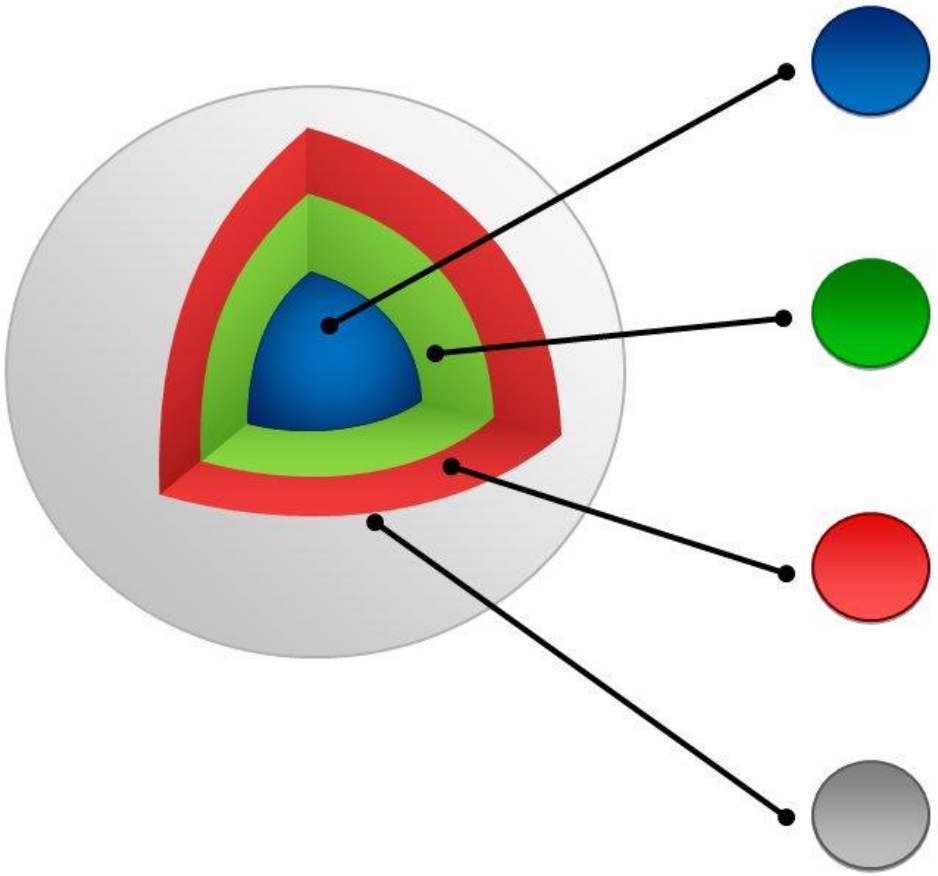
McMaster University | Research & High Performance Computing | McMaster University | Library

PREPARING A DATASET FOR DEPOSIT

Raw data isn't easy to understand. When sharing data, it's important to build your data into a "digital package" using:

- Documentation and Metadata
- Sustainable file types
- Good file organization and naming
- An open copyright license

DATASET AS A DIGITAL 'PACKAGE'



Research output (data/code)
The data is surrounded by layers of information to make it FAIR

Identifiers
Persistent Unique Identifiers such as DOIs and ORCiDs help find, track, and cite data

Standards
Open standard file formats help others access and reuse data

Metadata
Rich metadata and data documentation helps others find and understand datasets

README FIRST

A **readme** file is a document that describes the contents and organization of your dataset. They have the following characteristics:

- Simple text document (.txt or .pdf or .md)
- Includes basic project description, contact information and links to associated publications and data sources
- Explains file organization and naming schemes
- Describes folders and files in the data set

METADATA

Include **metadata** with your data deposit:

- Your contact information and affiliation
- Link to the associated publication (if there is one) and its DOI
- A clear description of the data and **keywords**

Other metadata that might be relevant:

- Geospatial coverage of the data
- Time period covered by the data

SUSTAINABLE FILE FORMATS

Other researchers may not have access to any proprietary software you use, so data and metadata should ideally be stored in **sustainable formats**. Look for formats that are:

- Standardized
- Well documented
- In common usage
- Uncompressed

Research Instrument files may be manufacturer specific and should be converted to a sustainable format when possible.

See https://site.uit.no/dataverseno/deposit/prepare/#what-are-preferred-file-formats

FILE NAMING SCHEMES

A good file name makes it easy to find data and keep track of versions. File names should:

- Describe the file contents
- Include the Date created as YYYYMMDD or YYYY_MM_DD
- Avoid special characters such as & , * % # * ( ) ! @$ ^ ~ ‘ { } [ ] ? < > –
- Be short

testdata.csv vs **2020_12_01_MercuryTestData.csv**

DATA LICENSES

If you don't have a license for your data or code, it falls under the default copyright laws. This means nobody can legally copy, distribute, or modify your work without permission.

Open licenses come in a few flavors:

- **Public Domain** means that you are releasing your data with no restrictions.
- **Attribution** ("BY") licenses only require that anyone using the data gives you credit and links to the original dataset.
- Other restrictions can include Non-Commercial ("NC") and Share alike ("SA")

WHAT LICENSE SHOULD I USE?

The most common open licenses are **Creative Commons** (creativecommons.org)

- Public domain - **CC0**
- Build your own license by combining the following clauses:
  - Require attribution - **BY**
  - Non-commercial – **NC**
  - Share alike - **SA**

COMMUNITY NORMS

For data there are also **community norms.** Dataverse and Open Data Commons community norms include:

- Share your work too
- Credit and Cite datasets you use
- Maintain anonymity of human research participants
- Encourage others to reuse data
- Use open formats
- Don’t use DRM

- https://dataverse.org/best-practices/dataverse-community-norms
- https://opendatacommons.org/norms/

DEMO

For more information:

Contact Isaac at: rdm@mcmaster.ca

RDM resources online: https://rdm.mcmaster.ca/
Sign up for upcoming webinars at: https://scds.ca/events/rdm/2021-2022/

McMaster University | Research & High Performance Computing | McMaster University | Library