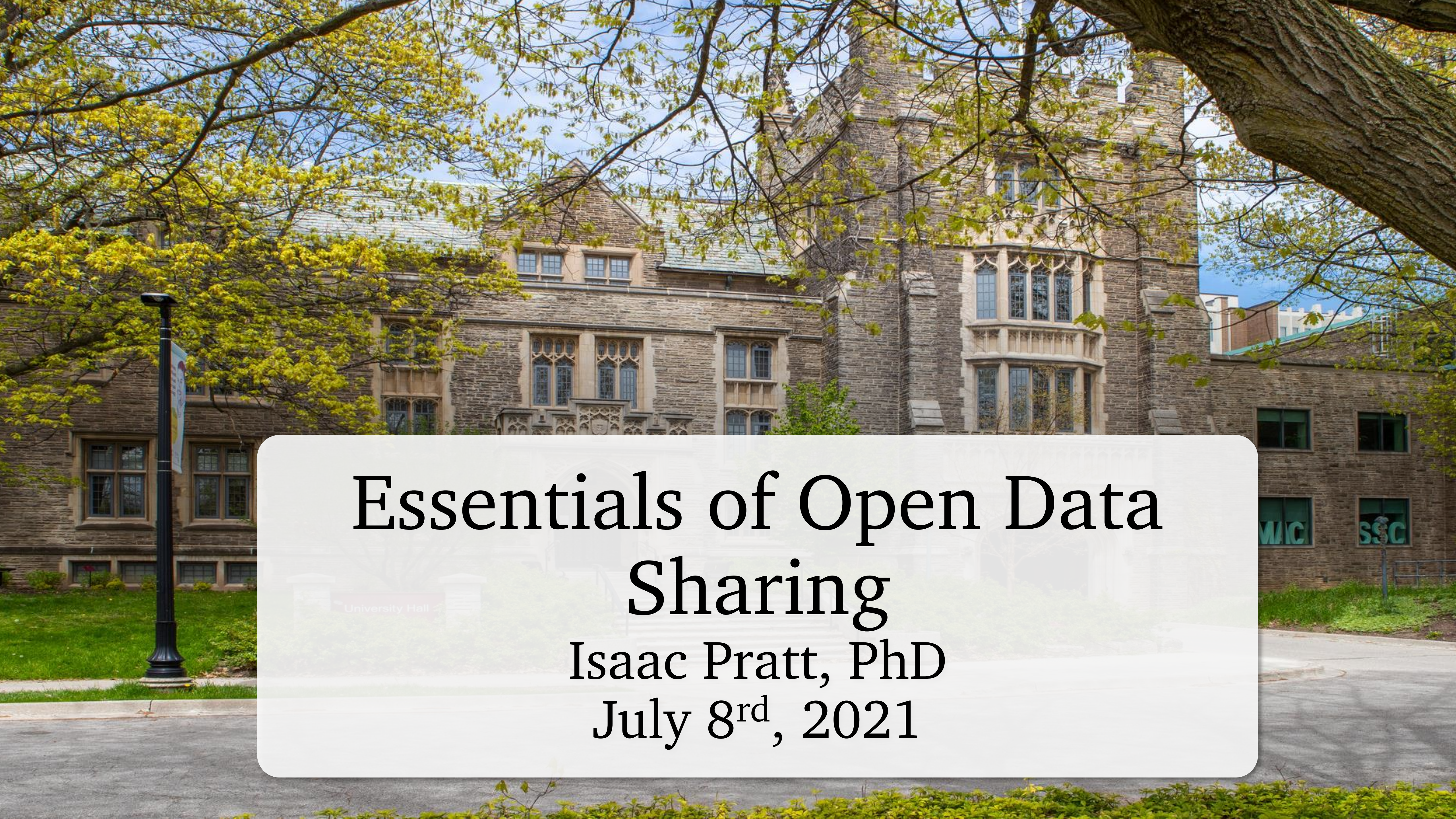


Essentials of Open Data Sharing

Isaac Pratt, PhD

July 8rd, 2021

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.

- 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/

McMaster University | Research & High Performance Computing

McMaster University | Library

HELLO! A BIT ABOUT ME:

- I am a Research Data Management Specialist working in the University Library and Research and High Performance Computing (RHPCS).
- My background is in Biological Anthropology, Medical Imaging, and Human Anatomy.
- I have a PhD in Anatomy & Cell Biology from the University of Saskatchewan.
- Email me with any RDM question or to set up a consultation: pratti@mcmaster.ca

OUTLINE FOR TODAY

1. **Why** should I deposit or share my research data?
   - Am I required to deposit or share my data?
2. **How** do I prepare my data for deposit or sharing?
3. **Where** should I deposit my data?

A NOTE ON TERMINOLOGY

I use **depositing data** to mean uploading research data to a purpose built online research data repository. Depositing data does **not** require you to share the data.

Created by mpanicon
from Noun Project

WHY DEPOSIT DATA?

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

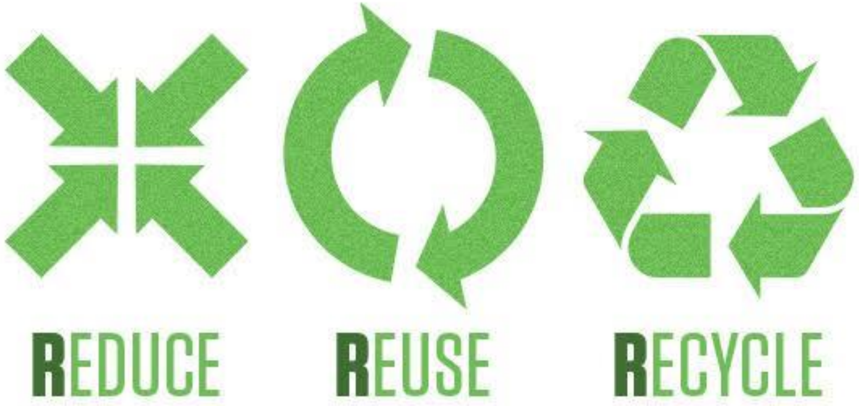
You may want to deposit your data for the following reasons:

- 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 your research

- Allow verification of results/code by peers
- Potential of 'mega' datasets

Improve the **value** of your research

- Avoid duplication of data collection or programming
- Maximizes use of your data/code

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 that arise from agency-supported research"

And:

> "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.

CIHR DATA REQUIREMENTS

CIHR has some specific data related requirements which are currently in force. Researchers are required to:

- **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 DATA REQUIREMENTS

SSHRC also has some specific data related requirements which are currently in force. Researchers are required to:

- **Preserve and make available for use by others** all research data collected with the use of SSHRC funds. This must occur within “a reasonable period of time”

- SSHRC considers "a reasonable period" to be within **two years** of the completion of the research project for which the data was collected.

SHARING SENSITIVE DATA

Most data repositories **will not** accept sensitive data – especially data that contains personally identifiable information.

If you are looking to publish or share sensitive data, you will need to remove, replace, or redact such information from datasets prior to upload.

One option is to publish a 'metadata only' dataset, with instructions on how to contact you to set up a **Data Transfer** or **Data Sharing Agreement**.

For more information on anonymizing or de-identifying data, see the Portage De-Identification Guidance document https://zenodo.org/record/4270551

PREPARING A DATASET FOR DEPOSIT

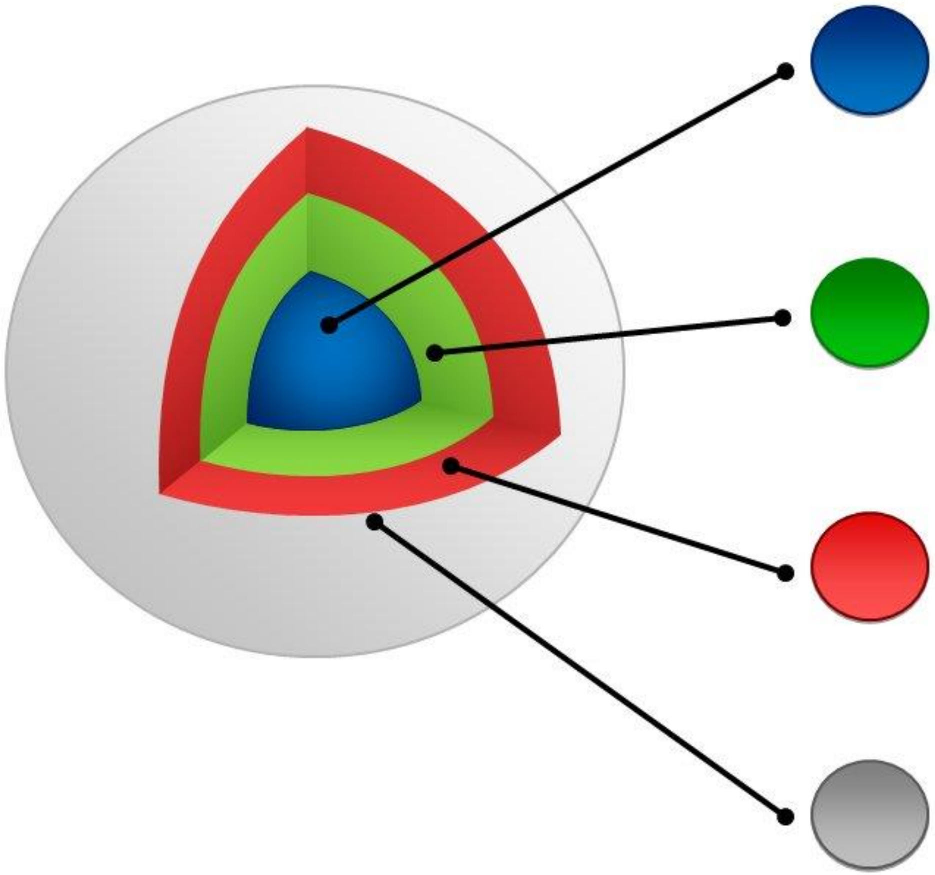
PREPARING A DATASET FOR DEPOSIT

Raw data isn't easy to understand. To make it easier to understand, include descriptive metadata and follow the four FAIR principles:

- **Findable**
- **Accessible**
- **Interoperable**
- **Reusable**

Check out our Dataverse deposit guide at: https://library.mcmaster.ca/sites/default/files/2021_05_mcmaster_dataverse_data_deposit_guidelines.pdf

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

PERSISTENT IDENTIFIERS (PIDs)

Datasets and code can be given **Digital Object Identifiers** (DOIs), unique links that persist over time.

Datasets and code can be linked to **ORCiDs**, unique personal researcher identifiers.

Github code repositories can be given a DOI by publishing releases on **Zenodo**

- https://guides.github.com/activities/citable-code/

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**

METADATA

Include metadata in your data deposit. Best practices call for including:

- Your contact information and affiliation
- Link to the associated publication (if there is one) and it's DOI
- Description of the data and keywords
- Publication date

Other metadata that can be useful includes:

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

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

DATA DICTIONARIES DEFINE YOUR DATA

A **Data dictionary** or **codebook** is a document describing the data and its variables. They typically include:

- Variable names and definitions
- Variable units and format
- Category and coded value definitions and meanings
- Known issues with the data including missing values
- Meaning of null values
- Minimum and maximum values

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 licenses are:

Creative Commons (creativecommons.org)

- **CC0** – public domain dedication
- **CC-BY** – require attribution

Open Data Commons (opendatacommons.org)

- **PDDL** - public domain dedication and license
- **ODC-By** – require attribution

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/

RESEARCH DATA REPOSITORIES

RESEARCH DATA REPOSITORIES

- Institutional Repositories: **MacSphere** & **McMaster Dataverse**
- External Data Repositories:
  - Domain specific https://www.nature.com/sdata/policies/repositories
  - **FRDR**, Zenodo, Figshare, Mendeley Data, etc
- Code repositories: Github, Gitlab, BitBucket, SourceForge
- Search for repositories on re3data.org

REGISTRY OF RESEARCH DATA REPOSITORIES

MACSPHERE

https://macsphere.mcmaster.ca/

- Institutional repository for **scholarly works**:
- A home for all research documents, including publications, presentations, conference proceedings, theses, reports, etc

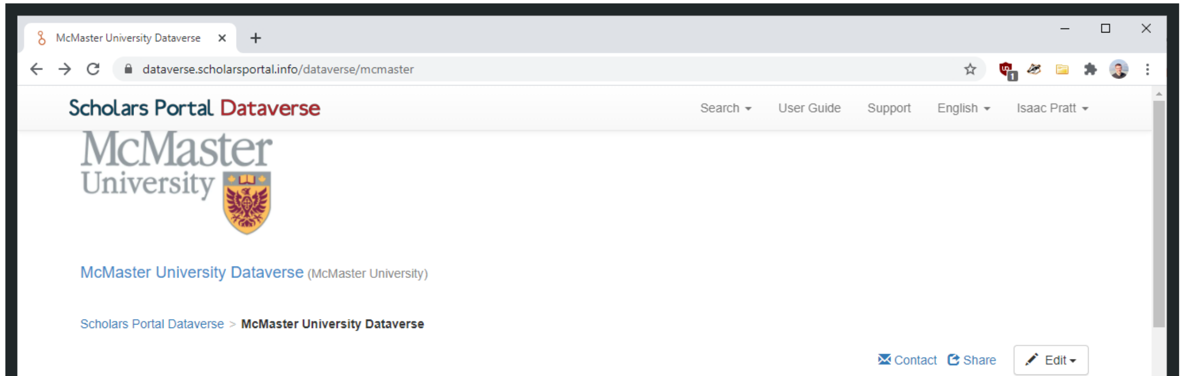
MacSphere

McMaster University Libraries Institutional Repository

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 Dataverse
dataverse.scholarsportal.info/dataverse/mcmaster
Scholars Portal Dataverse
Search
User Guide
Support
English
Isaac Pratt
McMaster University
Contact
Share
Edit

McMaster University Dataverse (McMaster University)

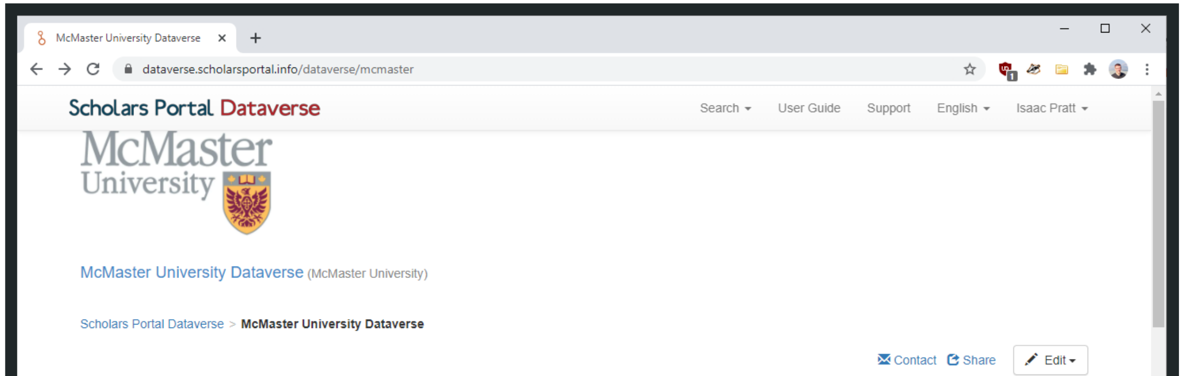
Scholars Portal Dataverse > McMaster University Dataverse

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 Dataverse
dataverse.scholarsportal.info/dataverse/mcmaster
Scholars Portal Dataverse
Search
User Guide
Support
English
Isaac Pratt
McMaster University
Contact
Share
Edit

McMaster University Dataverse (McMaster University)

Scholars Portal Dataverse > McMaster University Dataverse

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

FRDR
FEDERATED RESEARCH DATA REPOSITORY

THANK YOU!

For more information:

Visit: library.mcmaster.ca/services/rdm

Check out the whole webinar series at: https://scds.github.io/intro-rdm/

Contact us at: rdm@mcmaster.ca

RDM @McMaster

McMaster University | Research & High Performance Computing

McMaster University | Library