McMaster University

Family Medicine

David Braley Primary Care Research Collaborative

Research data preservation and sharing

Watch a recording of the session here

Isaac Pratt, PhD

November 23nd,2021

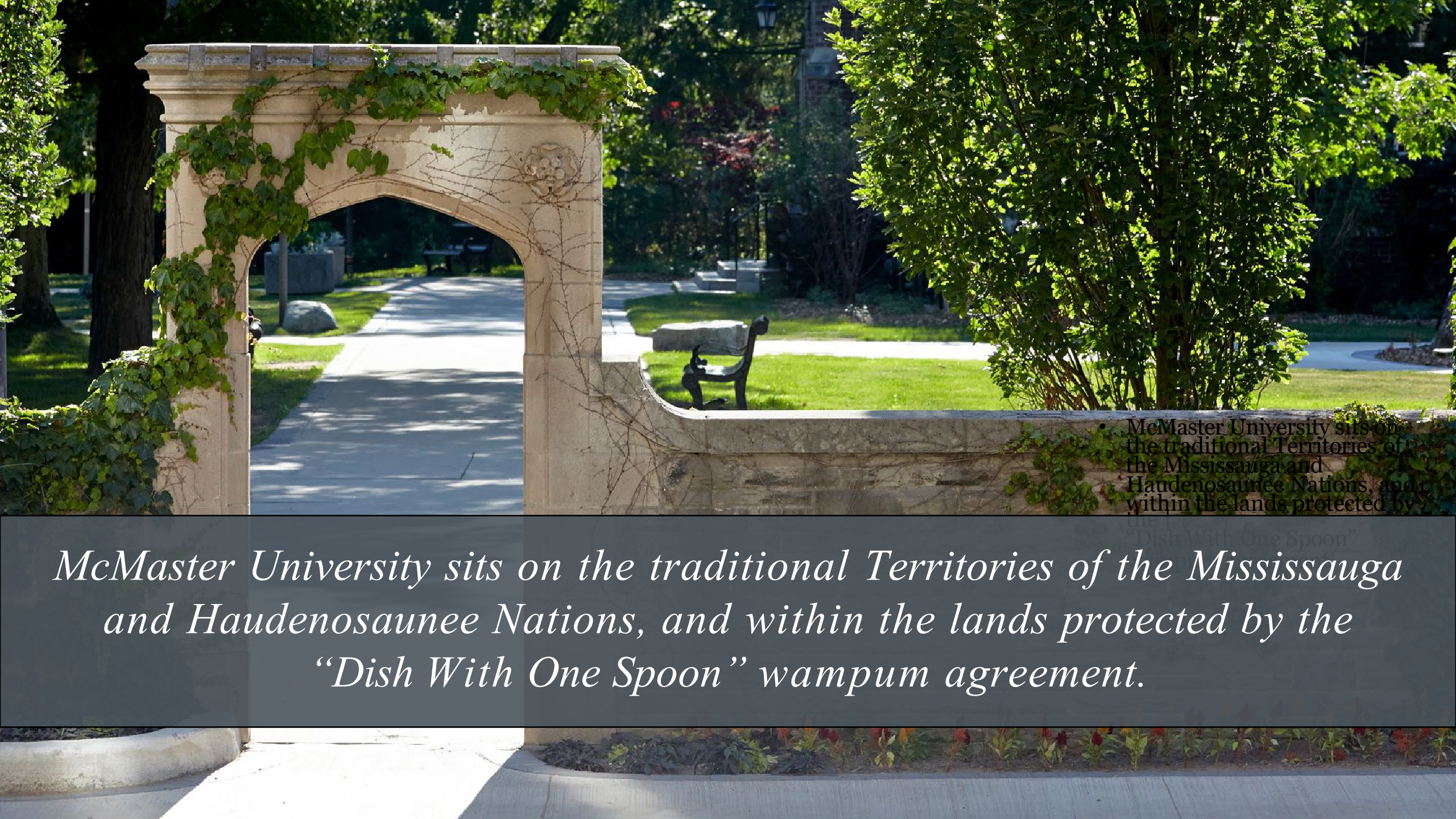
@McMasterFamMed

Michael G. DeGroote SCHOOL OF MEDICINE

Research data preservation and sharing

Isaac Pratt, PhD

November 23rd, 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"

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.

LEARNING OBJECTIVES FOR TODAY

By the end of the session you should be able to:

1. **Understand** the advantages and disadvantages of sharing research data
2. **Determine** whether you are required to deposit or share my data
3. **Prepare** your data for deposit or sharing

A NOTE ON TERMINOLOGY

Created by mpanicon
from Noun Project

Depositing data means uploading research data to a purpose built online research data repository. Depositing data does **not** mean that the data is openly accessible.

WHY DEPOSIT DATA?

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

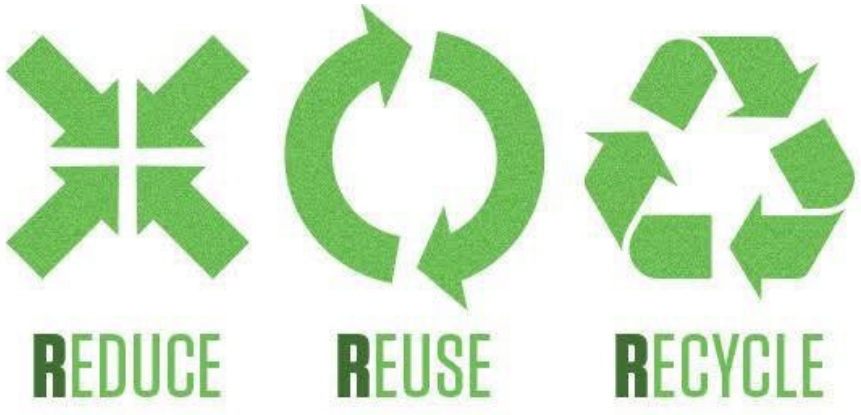
You may want to deposit your data for the following reasons:

- 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?

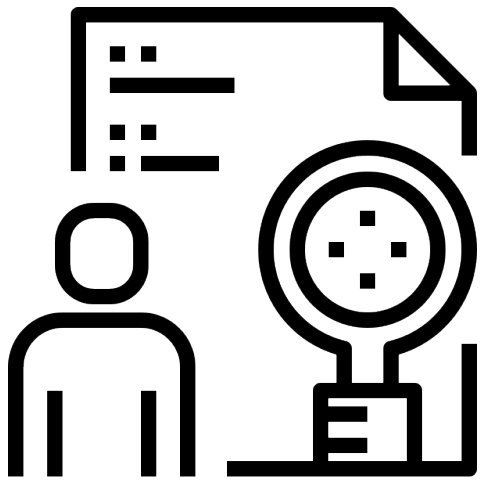
Openly sharing your data is good for:

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



Avoid ‘Single use data’

WHY SHARE DATA?



Created by Unlimiticon
from Noun Project

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

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:

> "**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"

TRI-AGENCY DATA DEPOSIT REQUIREMENTS

The policy also states:

> "Grant recipients are **not required** to share their data."

And:

> "the agencies expect researchers to provide **appropriate access** to the data where ethical, cultural, legal and commercial requirements allow."

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.

JOURNAL/PUBLISHER DATA REQUIREMENTS

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

BMJ https://authors.bmj.com/policies/data-sharing/

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

FAIR PRINCIPLES FOR DATA

Sharing data on its own isn't sufficient. The FAIR principles describe a set of guidelines to help researchers improve datasets for sharing and re-use.

https://www.go-fair.org/fair-principles/

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

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

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.

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

SHARING SENSITIVE DATA

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

If you want to publish or share sensitive data, there are 3 main options:

1. **De-identify or anonymize the dataset:** remove, replace, or redact all sensitive information from datasets prior to upload.
2. Create a restricted access protocol for data access
3. Deposit data on a restricted access platform

1. DE-IDENTIFICATION/ANONYMIZATION

De-identification is the process of removing any identifiable data from a dataset. Typically this is done by removing or replacing personal identifiers.

- Some datasets cannot be fully de-identified without compromising the usefulness of the data.

Datasets that cannot be completely anonymized **should never** be shared openly.

2. DATA DOESN'T HAVE TO BE OPEN

Data can be shared through closed access portals with restricted access mechanisms and **Data Sharing/Transfer Agreements**

- MILO can help you prepare a Data Sharing Agreement
- Examples of this kind of web portal include ICES and CIHI, although these portals can be set up for smaller data sets as well.
- Downside: often custom build which means there is a cost. You are responsible for sustainability of the data.

3. SENSITIVE DATA IN A REPOSITORY

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

One great option to access the discoverability of repository platforms is to publish a **'metadata only' dataset**, with instructions on how to contact you to set up a **Data Transfer** or **Data Sharing Agreement**.

3. SENSITIVE DATA IN FRDR

The Federated Research Data Repository (FRDR) is working on a pilot project to add optional zero-knowledge encryption to the repository.

"**Zero-knowledge encryption**" means that FRDR will never be able to access your data directly. All datasets that are deposited are encrypted and their keys deposited into a separate researcher managed platform.

This allows you to store your data in a trusted repository for archival but still control a restricted access protocol.

interested in piloting this service? email me: rdm@mcmaster.ca

CONSENT FOR DATA SHARING

In general you must have patient/participant consent to share data

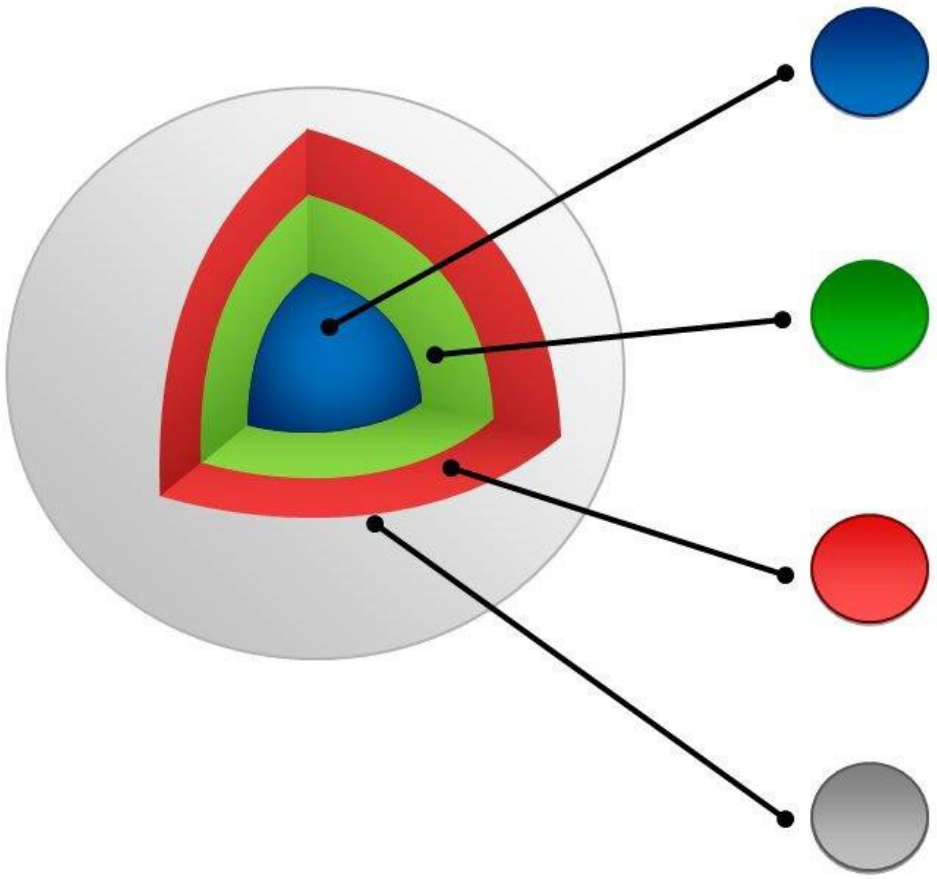
- Portage's "Can I share my data" decision tree
- Portage's Research Data Management Language for Informed Consent

Anonymity by design

- Never collect direct identifiers if you don't need to

PREPARING A DATASET FOR SHARING & PRESERVATION

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

DE-IDENTIFICATION

De-identification requires more than just removing direct identifiers from the dataset.

- **Direct Identifiers** are data points that can immediately identify a person, and include things like name, addresses, phone number, username, email, IP address, etc.
- **Quasi-identifiers** are data that are not direct identifiers but can be combined to uniquely identify an individual. Quasi-identifiers include things like age, gender identity, occupation, ethnicity, geographic information, etc.

DE-IDENTIFICATION

Quasi-identifiers can be combined to identify individuals with a unique combination of variables, especially when have less common characteristics. Ie:

A patient in your database who is Polynesian, female, 80+ may be easier to identify than someone who is white, male, 30-40.

Quasi-identifiers may also be used to identify individuals in the dataset by linking it to other data sources.

DE-IDENTIFICATION

k-anonymity is a mathematical approach to anonymizing a dataset.

A dataset has k-anonymity when a particular individual in the dataset cannot be distinguished from k other individuals in the dataset.

'k' is a number set by the researcher - most commonly set to 5. This means it should not possible to isolate a group of fewer than 5 identical individuals.

For a more comprehensive overview see the Portage Network's Reducing Risk Webinar:
https://www.youtube.com/watch?v=X3MKP_-FrWE
https://portagenetwork.ca/wp-content/uploads/2020/07/ReducingRisk-PortageWebinar.pdf

DE-IDENTIFICATION

Two main methods for achieving k-anonymity:

Suppression: deleting individual cases or responses.

- For example, if there is only one individual in a particular age range of a specific ethnicity, the ethnicity response for that individual could be deleted to preserve the ethnicity category as a whole.

Generalization: grouping specific values into categorized ranges.

- For example, grouping specific ages into age ranges or merging categories into larger groups.

DE-IDENTIFICATION

We recommend two software packages for anonymization:

Amnesia https://amnesia.openaire.eu/

- Amnesia is a user friendly software package that can be run online from a browser.

sdcMicro https://cran.r-project.org/web/packages/sdcMicro/index.html

- sdcMicro is an open source R package with a graphical interface.

REDCap also has some de-identification features

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 LICENSES FOR OPEN SHARING

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

MACSPHERE

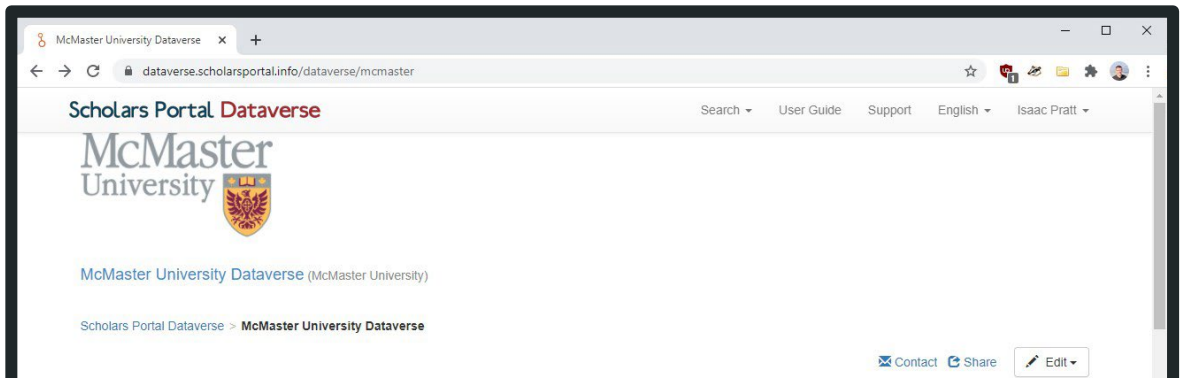
https://macsphere.mcmaster.ca/

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

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

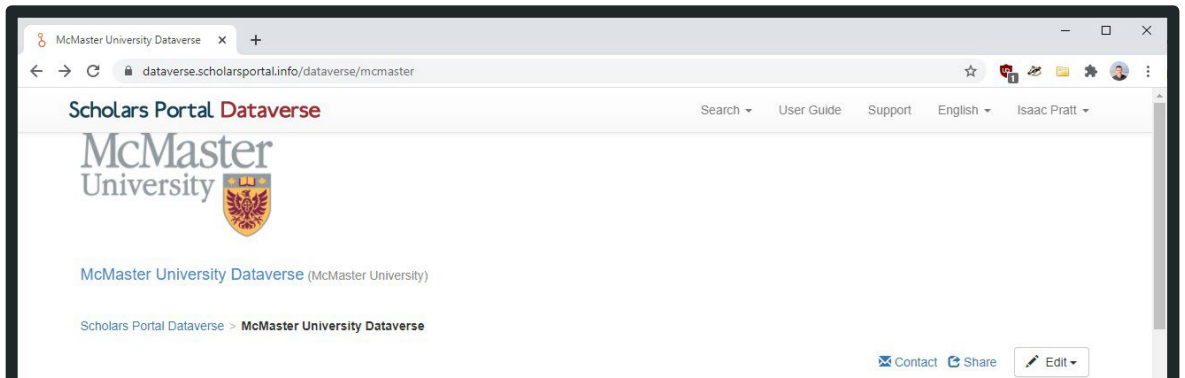


McMaster University | Research & High Performance Computing

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

THANK YOU!

For more information:

Visit: rdm.mcmaster.ca

Check out our upcoming webinars at: https://scds.ca/events/dmds/2021-2022-2/

Contact us: rdm@mcmaster.ca

RDM @McMaster

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