

Best Practices for Managing Data in your Research

Isaac Pratt

Do More with Digital Scholarship Workshop Series

March 16, 2021

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/

Session Recording and Privacy

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.



McMaster University sits on the traditional Territories of the Mississauga and Haudenosaunee Nations.

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

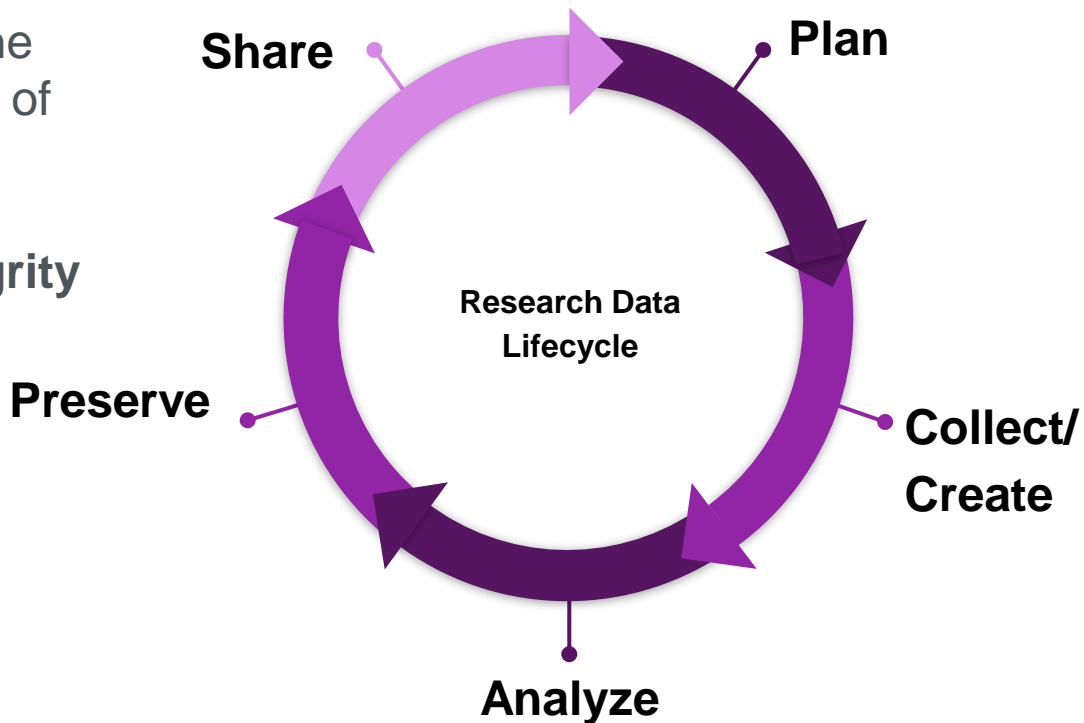
Learning objectives

At the end of this presentation, you should:

- Understand what Research Data Management is and why it is important
- Be ready to integrate a few RDM practices into your own research
- Be prepared to ensure the long term viability and availability of your data

What is Research Data Management anyways?

Research Data Management is the active organization & maintenance of data throughout the research data lifecycle to ensure its **security**, **accessibility**, **usability**, and **integrity**



Consider:

- If your supervisor asked you to share your data with another student, would they be able to make sense of your work?
- If you needed to locate your data files from 5 years ago, how easy would they be to find and use?
- What will happen to your data when you graduate/move/retire?

“One who does not plan long ahead will find trouble at his door”
- Confucius

Consider the ‘standard’ approach to data management:

- Data is stored on laptop or desktop hard drives and backed up to a collection of miscellaneous external hard drives accumulated over the years.
- Data is not consistently documented
- Data is not published or shared outside the research group except by direct request.

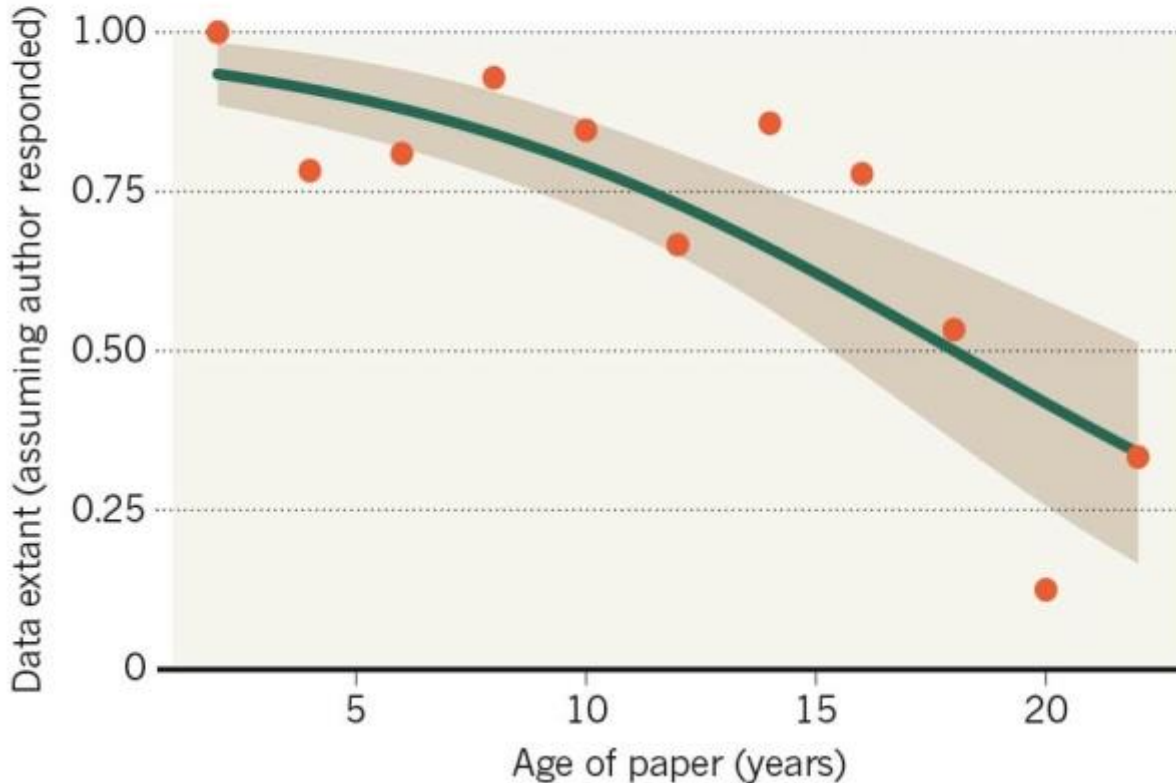
This approach is **vulnerable** to data loss and makes working with the data frustrating

Is your data

MISSING DATA

Vines et al 2014

As research articles age, the odds of their raw data being extant drop dramatically.



April 19th
s stolen
tole my
e to say
pay you
E-YEAR
a folder
THESIS
which is
and use
price is
address
it so so

Data Management Planning

A **Data Management Plan (DMP)** is a living document describing your plan for how you will create, store, organize, document, secure, preserve, and share your research data.

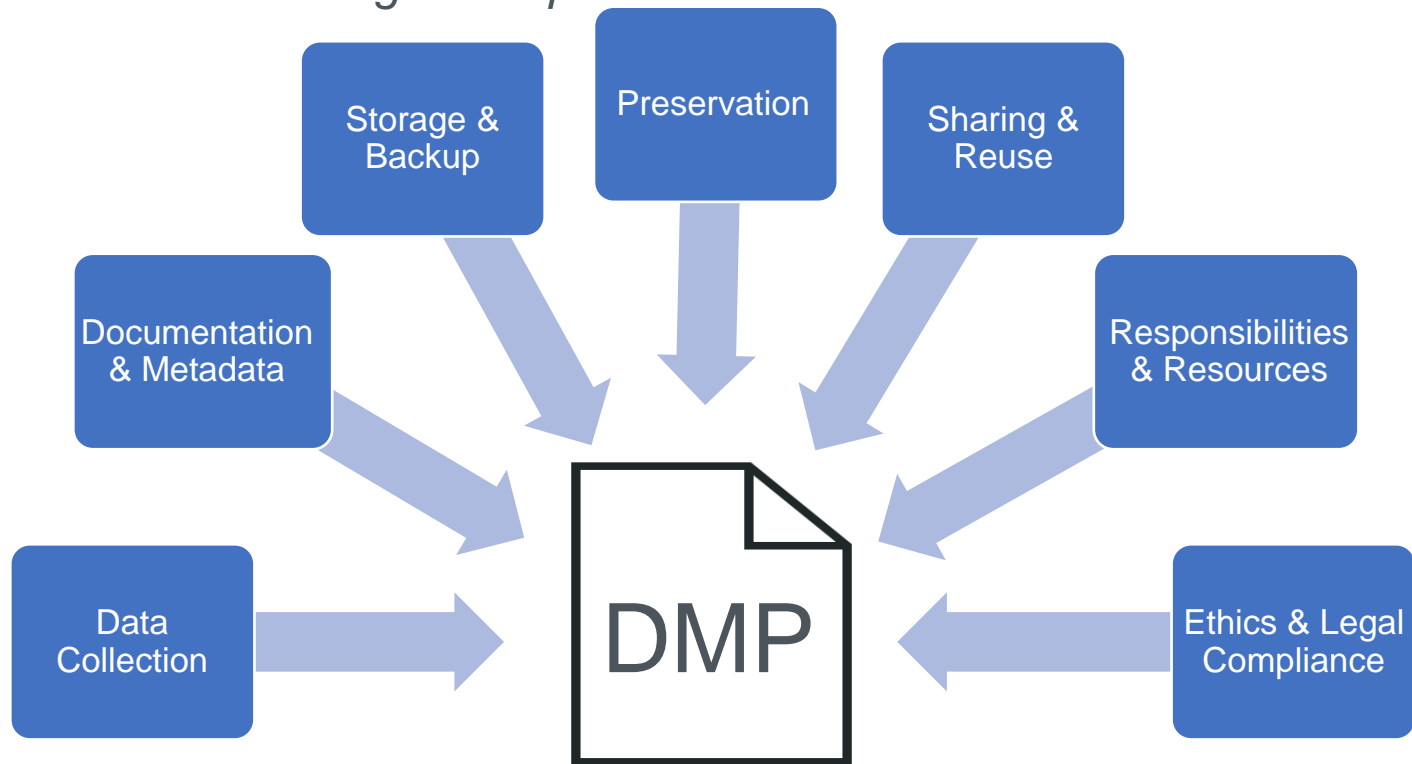
Building a DMP is a structured process that helps you plan and organize your research data.

Creating your own DMP is straightforward using web tools such as the [Portage DMP Assistant](#)

Some research funders require grant applicants to submit a DMP – NSF, NIH, Wellcome Trust, Tri-Agency (starting 2022)

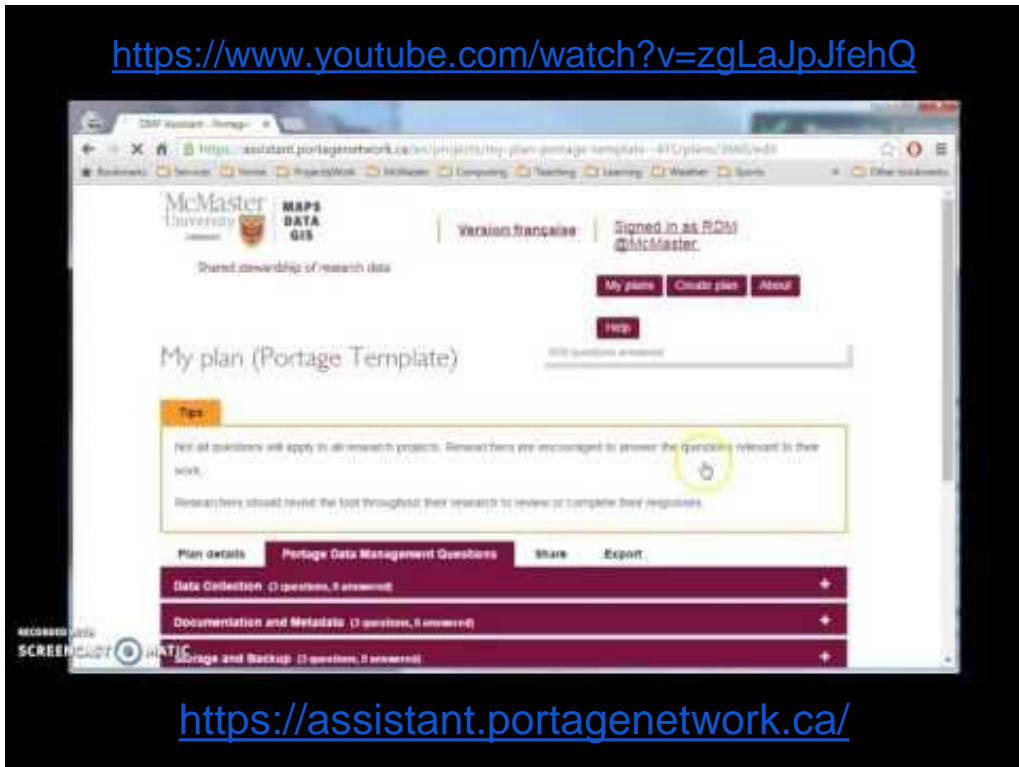
Planning

What goes in a data management plan?



- a web-based, bilingual data management planning tool
- available to all researchers in Canada
- a guide for best practices in data stewardship
- exportable data management plans

<https://www.youtube.com/watch?v=zgLaJpJfehQ>



<https://assistant.portagenetwork.ca/>

Documenting data at collection/creation

Have you ever gone to analyse data or publish a paper only to find that some critical piece of information was not recorded?

Documentation is for your benefit but also for others, including co-workers, collaborators, reviewers, and supervisors.

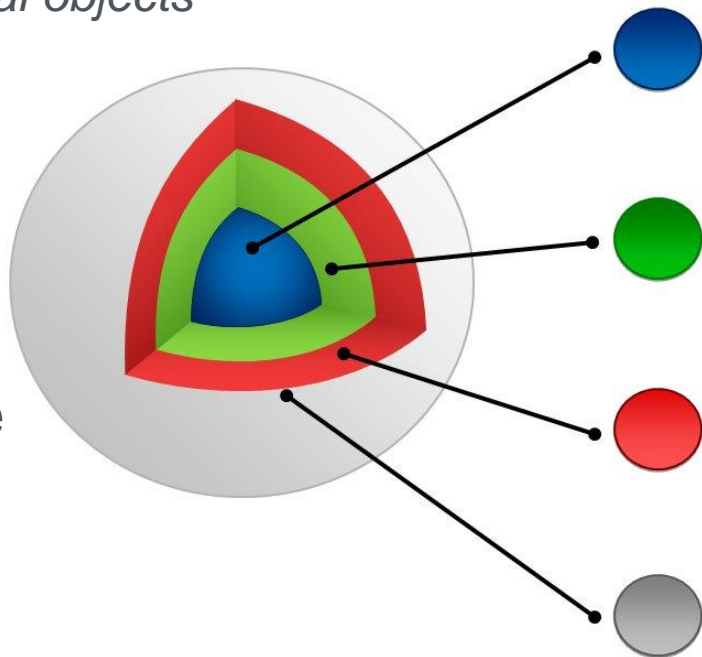
An **Electronic Laboratory Notebook (ELN)** can make documentation easier and more reliable:

- Easy to search, copy, and archive
- Information can be shared with other lab members and collaborators
- Files and data can be linked

FAIR Data

Datasets as digital objects

Findable
Accessible
Interoperable
Reusable



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

Find existing data

Integrate existing datasets into your research:

[FRDR – the Federated Research Data Repository](#) provides a dataset search function which indexes Canadian research datasets.

[Google Dataset Search](#) indexes research datasets hosted across the web.

The [McMaster Library Data Service](#) provides access to restricted government data including Statistics Canada microdata.

How should I store my data?

A good data storage plan needs to balance **accessibility** and **convenience** against **security** and **reliability**.

3-2-1 Backup Strategy:

- **3** copies of your data where
- **2** copies each in a different storage system
- **1** copy is in a trusted off site location

- Example: 1 copy stored locally on hard drive for analysis, 1 copy stored on cloud storage platform, 1 copy stored in a secure campus drive

Don't forget to back up everything else as well!

How do I decide where to store data?

Features to look for when deciding on a storage platform:

- Version control
- File recovery
- Security features (2FA, encryption)
- Collaboration features
- Storage provided
- Cost
- Storage location


Special considerations: Sensitive data, indigenous data, computational needs, code

The Research Data Storage Finder Tool

<http://u.mcmaster.ca/storagefinder>

Step 1: Answer these questions to narrow down storage provider options.

CLEAR ANSWERS

1. What risk level is your data?


Low
 Medium
 High

2. What type of data storage are

Step 2: Select data storage providers you would like to compare

SELECT ALL **CLEAR SELECTIONS**

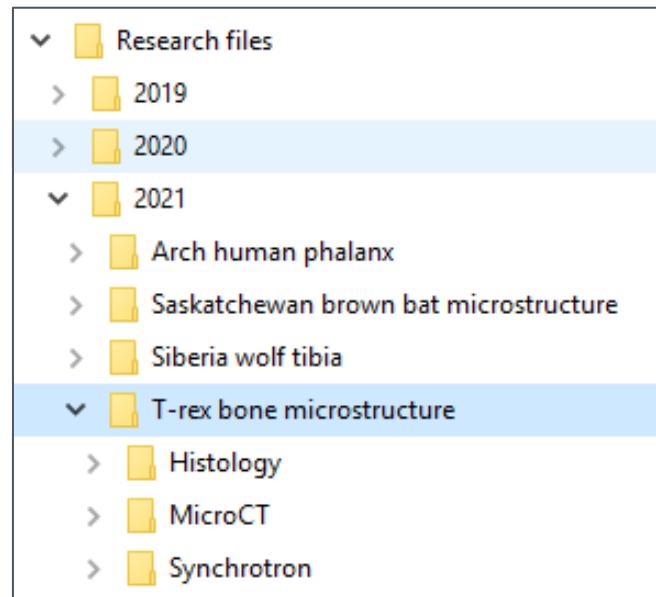
<p>Compute Canada <input type="radio"/></p> <p>Advanced research computing systems, storage and software</p>	<p>Compute Canada NextCloud <input type="radio"/></p> <p>Advanced research computing File hosting services</p>	<p>Dataverse <input type="radio"/></p> <p>Store, share, publish and discover research data</p>
<p>FRDR <input type="radio"/></p> <p>Find and Share Canadian Research Data</p>	<p>Github <input type="radio"/></p> <p>Distributed version control system for software code</p>	<p>MacDrive <input type="radio"/></p> <p>File Synchronization and Sharing solution</p>
<p>MacDrop <input type="radio"/></p>	<p>McMaster-based <input type="radio"/></p>	<p>OSF <input type="radio"/></p>

Keeping files organized makes it easier to find things

The key to organizing files is to make it a habit. Make it easy to know files go.

File organization schemes can include:

- By project
- By researcher
- By experiment type
- By date (often year)
- By some combination of the above
(ie a two level structure of year -> project)



Give your files good names

A good file name makes it easy to find data and keep track of versions

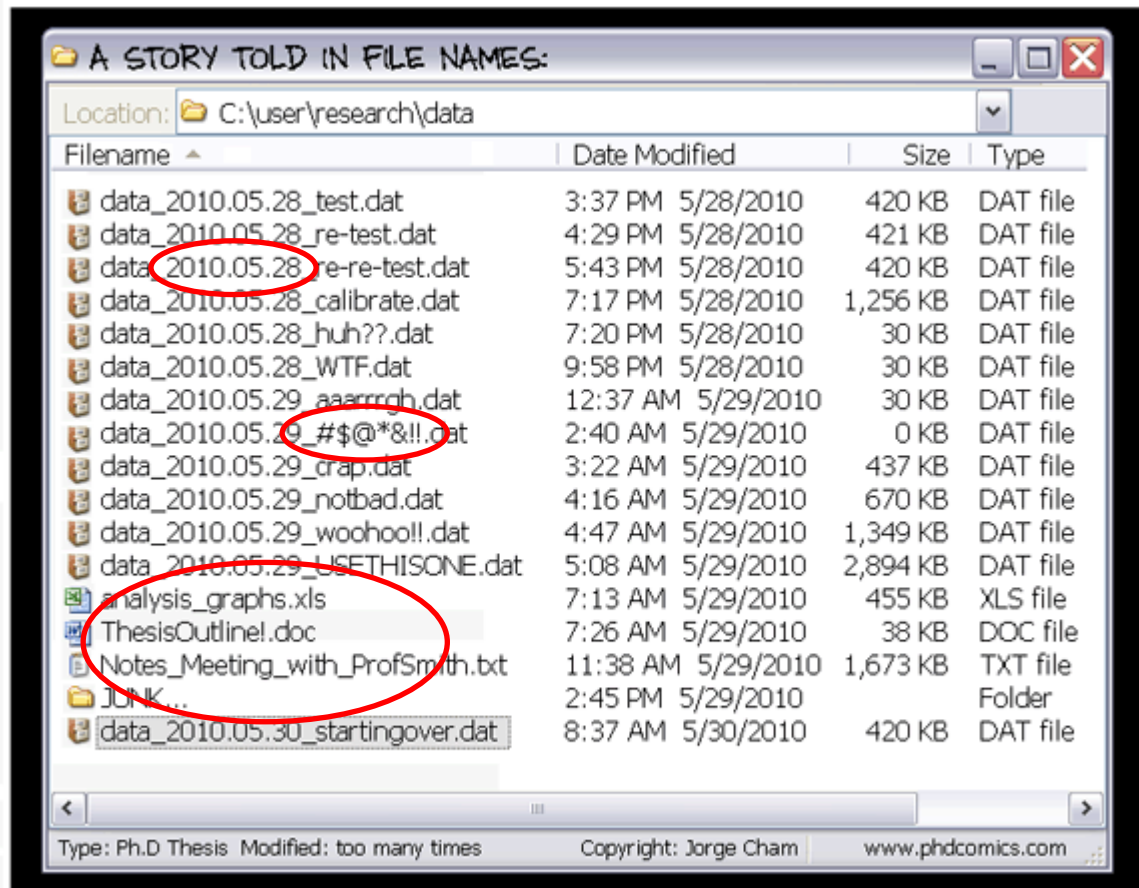
testdata.csv vs 2020_12_01_MercuryTestData.csv

File naming schemes should include:

- Short description of file contents
- Date created as YYYYMMDD or YYYY_MM_DD
- Project name or acronym
- Initials of researcher (if working on a collaborative file)
- Avoid special characters such as & , * % # * () ! @\$ ^ ~ ' { } [] ? < > -
- Try to keep names short

Do you have files named like this?

Is this a good file name system?



Keep documentation (metadata) with your data

If you needed to use data you collected 5 years ago, how easy would they be to find and use?

- Would you know what each variable is?
- Would you have information about when/where/how the data was collected?

Document your data using **readme** files, **codebooks**, and **data dictionaries**

readme.txt first

A **readme** file is a simple text document that describes the contents and organization of your data files.

- .txt or .md open format
- Starts with a basic project description including contact information and location of associated publications and data sets
- Explains file organization and naming schemes
- Describes data 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.

A data dictionary typically includes:

- 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

Build a documentation scheme you will use

The more important aspect of documentation is doing it.

Whatever file naming and organization scheme you choose, make sure it's **descriptive**, use it **consistently** and **document** it (in a readme.txt file).

Collaboration software like Electronic Lab Notebooks, Reference Management software, or the Open Science Framework platform can help.

Publishing data

What do you plan to do with your data once it's been published? How will you ensure that your data remains accessible (to you and others) long-term?

Consider the advantages of publishing your datasets in an online repository for preservation and sharing.



Why should I share my data?

Improve the **quality** of your research

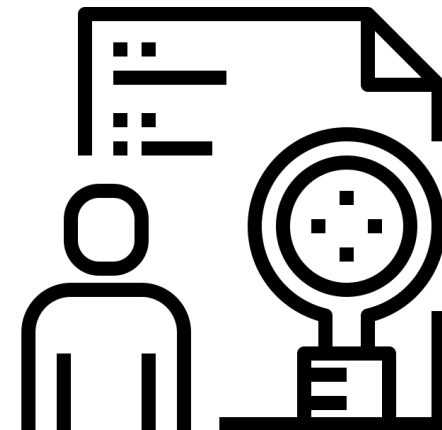
- Allow verification of results/code by peers
- Potential of 'mega' datasets

Improve the **impact** of your work

- Increases the potential visibility of research
- Can lead to new collaborations and partnerships
- Creates a lasting record of your work

Improve the **value** of your research

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



Created by Unlimiticon
from Noun Project



Why should I share my data?

Your journal or funder may require data sharing:

The Tri-Agencies currently have some recommendations and encourage all data to be shared (where possible).

- CIHR and SSHRC currently require some research data to be shared
- See the Tri-Agency Data Management Policy for details
http://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html

Open access publishing

Tri-agency funded research *must* be published open access.

We encourage all research to be published open access when possible!

Online Repositories

- Final manuscripts can be deposited in an institutional or disciplinary repository (such as [arXiv.org](https://arxiv.org))
- Researcher is responsible to navigate copyright requirements of the journal

Journals

- Journal provides open access to the article (within 12 months)
- Most journals will charge open access fees

Ok so where do I put everything?

Institutional Repository: **MacSphere** <https://macsphere.mcmaster.ca/>

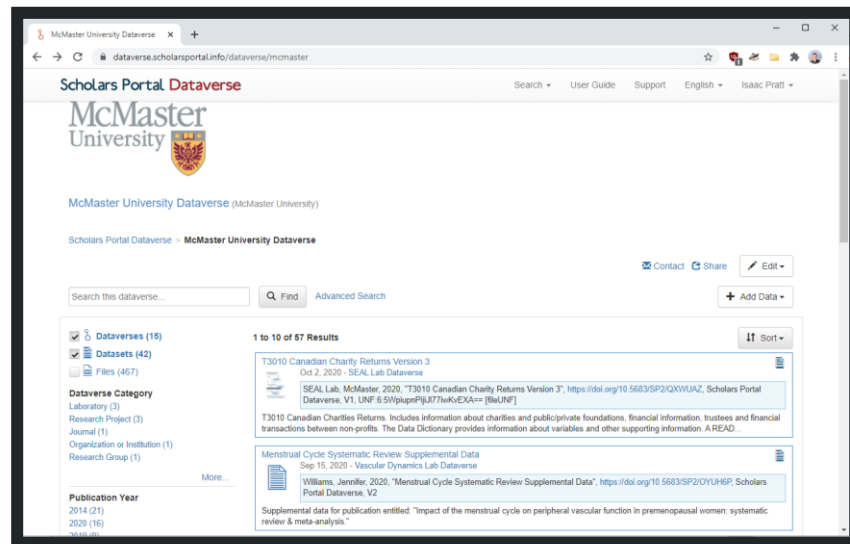
- Built for scholarly works: all kinds of research documents, including publications, conference proceedings, theses, reports, etc
- When you graduate, you will have to upload your thesis here



Ok so where do I put everything?

Institutional Data Repository: **McMaster Dataverse**
<https://dataverse.scholarsportal.info/dataverse/mcmaster>

- Built for datasets
- Contains tools for tabular data exploration and analysis
- Allows researchers to control how they license and share their datasets



Ok so where do I put everything?

Federated Research Data Repository (FRDR)

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

- Built for large (1 TB+) datasets
- Datasets are actively reviewed by FRDR staff
- Datasets must be fully open but can be embargoed for a one year period



Ok so where do I put everything?

External Data Repositories:

- Domain specific vs General
- Zenodo, Figshare, Mendeley Data, FRDR, etc
- Search for repositories on re3data



Persistent Unique Identifiers help keep track of everything

Citing datasets and code is made easier by using **Digital Object Identifiers (DOIs)**

- A DOI is a persistent link to a digital object.

Datasets and code can be linked to ORCiDs, your unique personal researcher identifier



Publishing data

Data should be stored in non-proprietary formats



- Corel WordPerfect was a word processing application. From 1989 to 1992 WordPerfect had almost 50% market share, above Microsoft Word
- Have you ever saved data on a DVD?
- Do you use an online document processing software like Google Docs or Prezi where all your documents are stored online on a proprietary platform in a proprietary format? What would you do if that platform closed down?
- Adobe Flash was shut down December 31st 2020



Do I need a license for my data?

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

Not having an explicit license restricts others from using your code or data, and causes confusion.

What license should I use?

Creative Commons (creativecommons.org)

- CC0 – public domain dedication
- CC-BY – require attribution
- There are further restrictions that can be added such as NC



Open Data Commons (opendatacommons.org)

- Similar licenses to CC but built for data
- PDDL - Public Domain Dedication and License
- ODC-By – require attribution
- ODbL – attribution and share alike

What license should I use?

Dataverse and Open Data Commons also expect researchers to adhere to community norms including:

- 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/>

Top 4 ideas for improving your research data management

1. Make a **plan** for data management
2. Create a **file organization scheme** (and use it)
3. Ensure your data is safely **stored** and backed up
4. **Share** your data openly



Created by Maxim Kulikov
from Noun Project

Thank You.

For more information:

Visit: library.mcmaster.ca/services/rdm

Contact me at: rdmgmt@mcmaster.ca

RDM
@McMaster