# Best Practices for Managing Data in your Research

Isaac Pratt, PhD

Do More with Digital Scholarship Workshop Series

October 22nd, 2021

Lewis & Ruth Sherman Centre for Digital Scholarship

scds.ca

McMaster University | Library

# 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/](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 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.

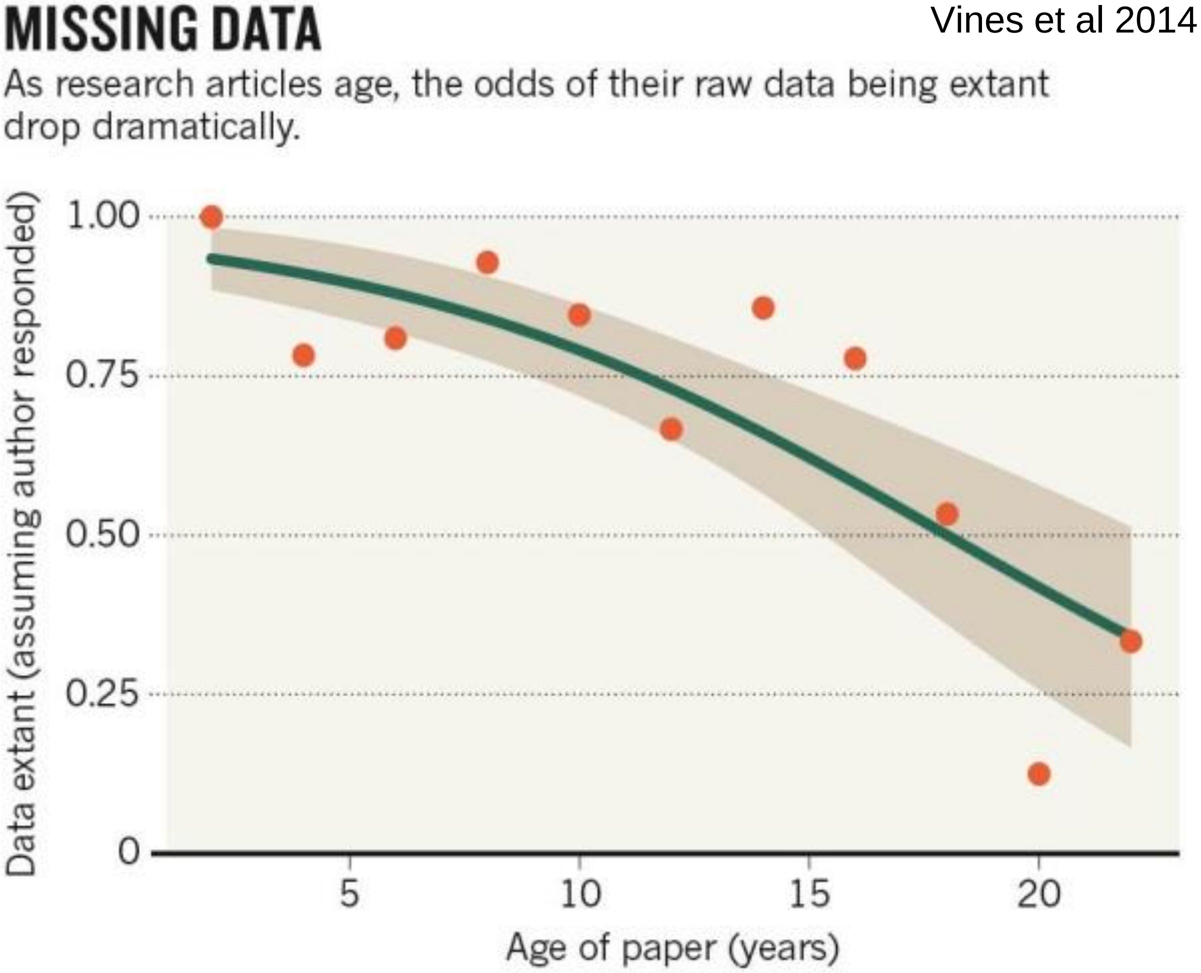
# *Learning objectives*

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

# Is your da

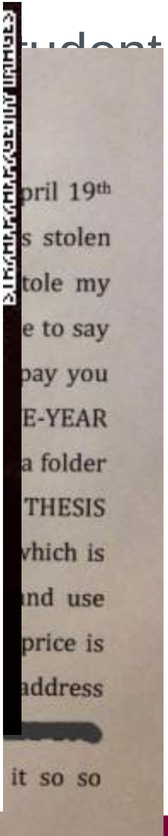
- If your
  tudent or



MISSING DATA

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

Vines et al 2014



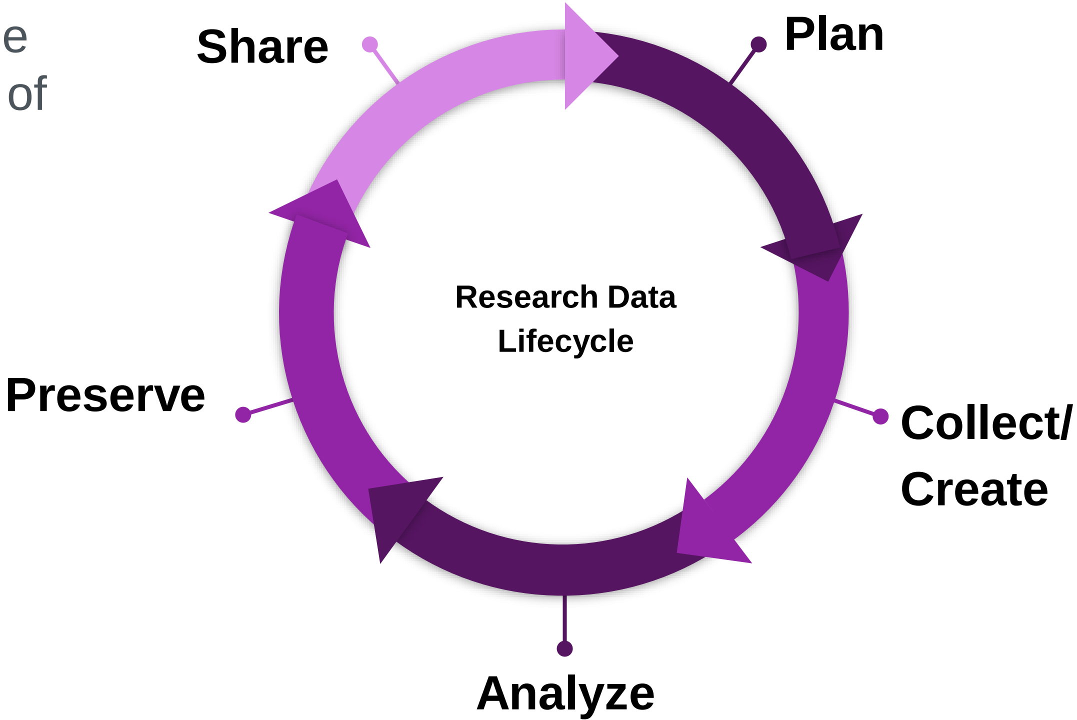
STRICKER/AFP/GETTY IMAGES

pril 19th
s stolen
tole my
e to say
pay you
E-YEAR
a folder
THESIS
which is
nd use
price is
address

it so so

# *What is Research Data Management anyways?*

**Research Data Management** is the active organization & maintenance of data throughout the research data lifecycle. It is a series of research practices we follow to ensure the **security**, **accessibility**, **usability**, and **integrity** of data



# Lets look at an example:

- “Dave” is a graduate student working in Biomedical Science, focused on x-ray imaging of bone tissue samples. Dave’s data is made up of 3 major components:
  - Image files – microCT scan images, microscope images
  - Software/hardware configuration files – instrument specific files, scripts, text files
  - Measurement data files – spreadsheet files

# Lets look at an example:

- Dave’s data is stored separately in a few places:
  - Image files are large (2+ TB) and stored on lab computers and a collection of miscellaneous external hard drives accumulated over the years.
  - The other files are smaller (10 GB) and stored on a personal laptop and a cloud storage platform (OwnCloud).

- Data is not consistently documented

- Data is not published or shared outside the research group except by direct request. No time or energy is put into archiving the research data.

# What went wrong for Dave

- One of the external drives fails, leading to the loss of some of Dave's data. This data loss is not discovered for several weeks. **There is no back up of this data**.

- This leaves Dave with two choices:
  - Extend his degree while he recollects that data, or
  - Publish what he can, even though the explanatory and statistical power of the study has been reduced

# What could Dave do better in the future?

- Dave should:

1. Make a plan for his data
2. Store and back-up his data securely
3. Organize and document his data consistently
4. Make sure his data is ready for archival and sharing

# *Data Management Planning*

A **Data Management Plan (DMP)** is a living document describing your plan for how you will manage your research data.

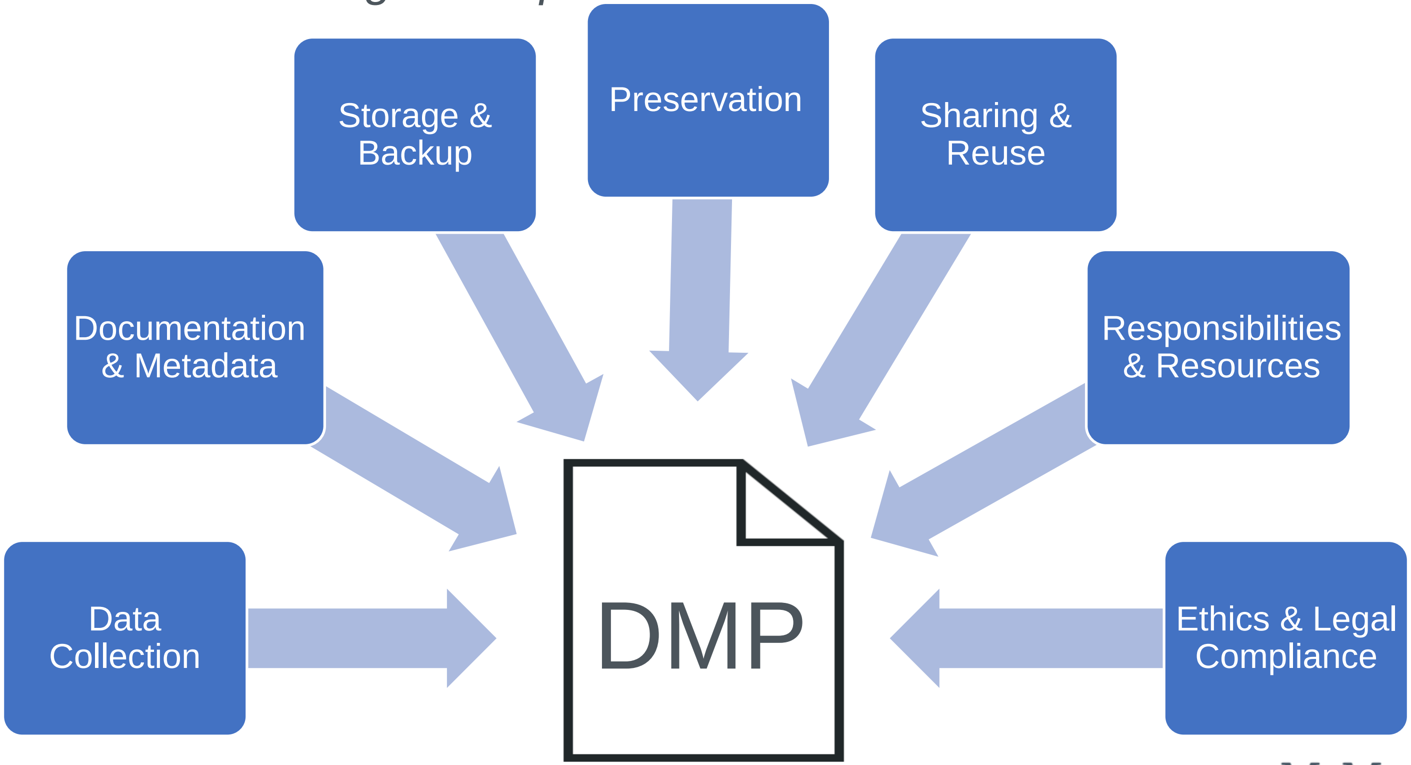
Building a DMP is a structured process that helps you plan and organize your research data.

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?*



Data Collection

Documentation & Metadata

Storage & Backup

Preservation

Sharing & Reuse

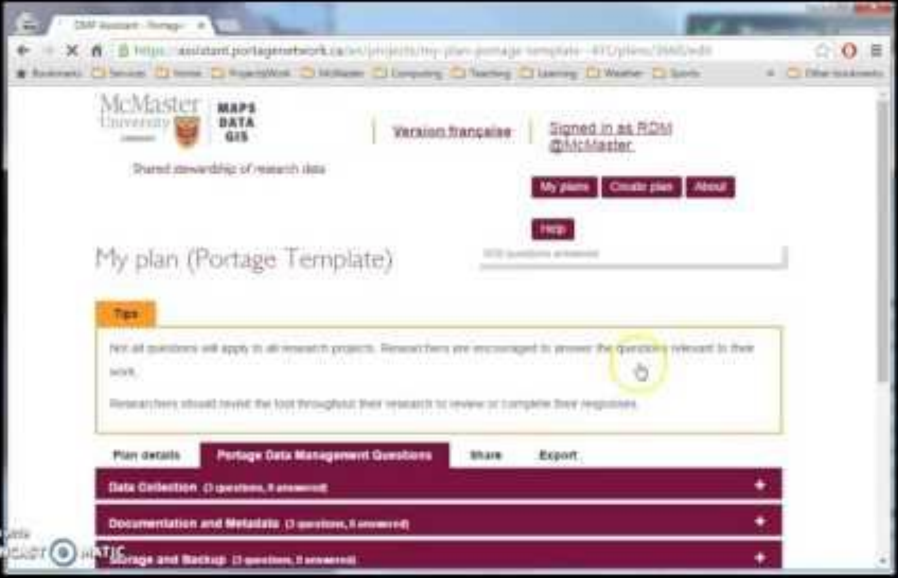
Responsibilities & Resources

Ethics & Legal Compliance

DMP

- 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



McMaster University

MAPS DATA GIS

Version française | Signed in as RDM @McMaster

Shared stewardship of research data

My plans | Create plan | About

Help

My plan (Portage Template)

Tips

Not all questions will apply to all research projects. Researchers are encouraged to answer the questions relevant to their work.

Researchers should revisit the tool throughout their research to review or complete their responses.

Plan details | Portage Data Management Questions | Share | Export

Data Collection (3 questions, 0 answered) +

Documentation and Metadata (3 questions, 0 answered) +

Storage and Backup (3 questions, 0 answered) +

https://assistant.portagenetwork.ca/

# *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.

# *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

# *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 should I protect my data?*

Enable **Multi-Factor Authentication (MFA)** when you can

- Also known as 2 Factor Authentication (2FA)
- MFA is when you need more than one code or ‘Factor’ to login – typically 2 factors: password and a security code sent to your phone number or generated by a linked authenticator app
- MFA can be enabled for your McMaster Microsoft account here https://office365.mcmaster.ca/mfa/
- Enter the getMFA challenge here: https://cto.mcmaster.ca/boost-your-cyber-security-awareness-and-win-prizes-this-october/

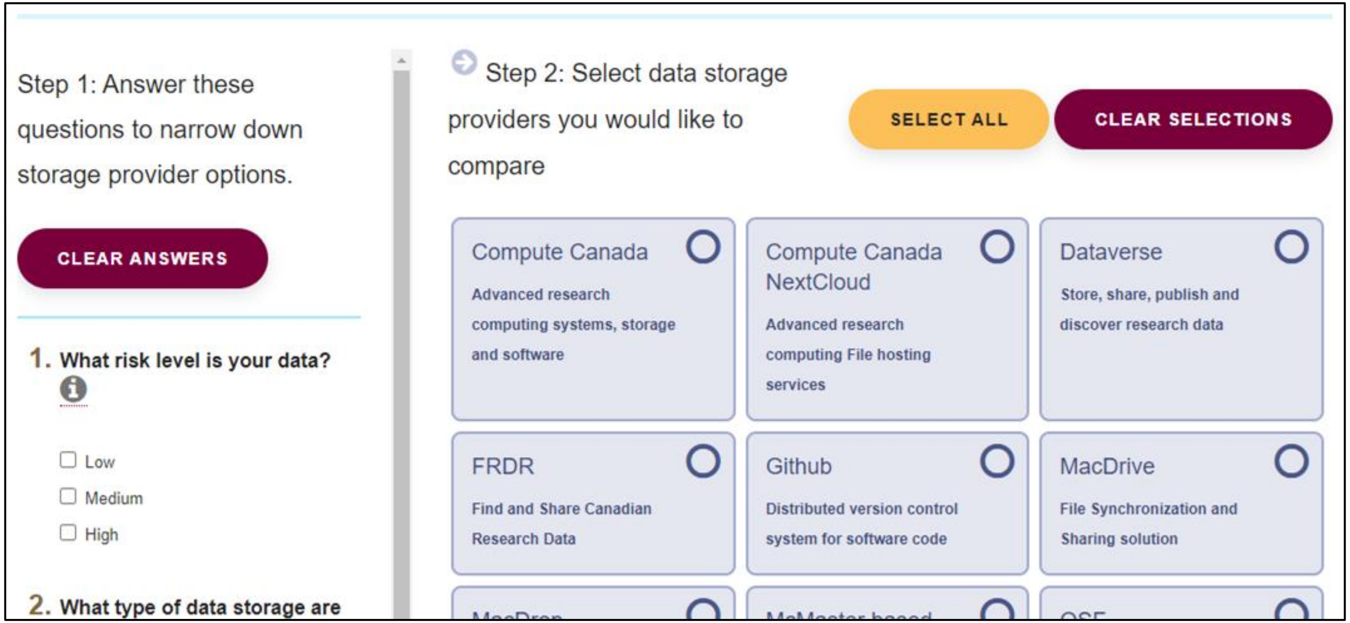
# *How should I protect my data?*

Follow good password practices everywhere:

- Choose a new **unique** password for each important website/service
- Make a **strong** password by combining a series of numbers, letters, and symbols
  - The longer the better
  - Try to combine them into something memorable – like L1br@ryt1pS
- If you're worried about forgetting passwords, consider using a password manager
- **Never share your password** with anyone or send it in an email
- Use a strong password on your computer and phone

# The Research Data Storage Finder Tool

http://u.mcmaster.ca/storagefinder



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

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

○ Compute Canada
Advanced research computing systems, storage and software

○ Compute Canada NextCloud
Advanced research computing File hosting services

○ Dataverse
Store, share, publish and discover research data

○ FRDR
Find and Share Canadian Research Data

○ Github
Distributed version control system for software code

○ MacDrive
File Synchronization and Sharing solution

○ MacDrop

○ McMaster based

○ OSF

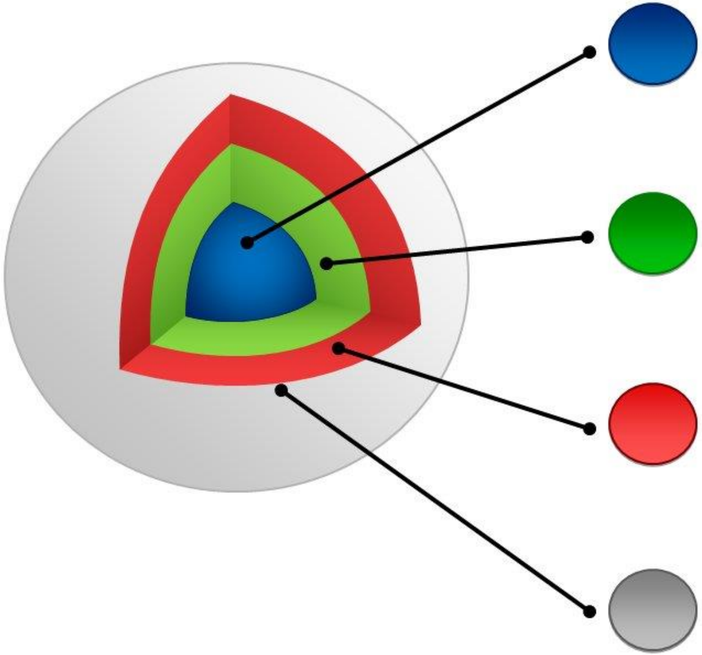
# *Documenting data*

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

Data should be:

- Findable
- Accessible
- Interoperable
- Reusable

# Datasets as digital objects



**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

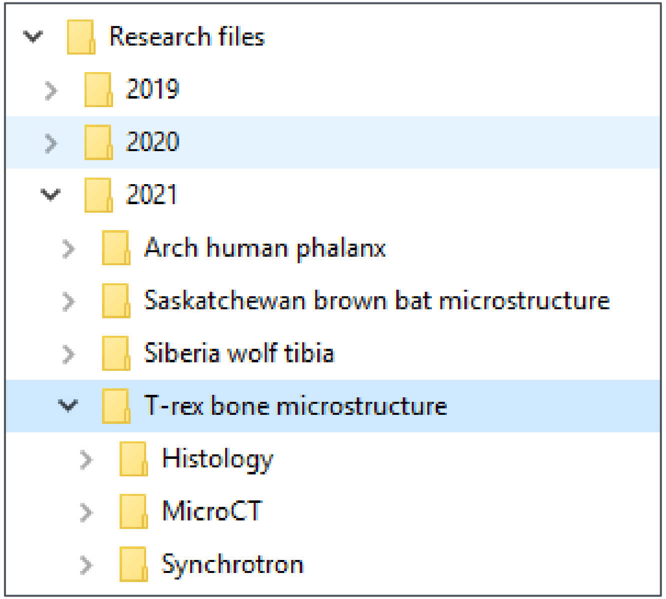
# *Keeping files organized makes it easier to find things*

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)



- Research files
  - 2019
  - 2020
  - 2021
    - Arch human phalanx
    - Saskatchewan brown bat microstructure
    - Siberia wolf tibia
    - T-rex bone microstructure
      - Histology
      - MicroCT
      - Synchrotron

# *Give your files good names*

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

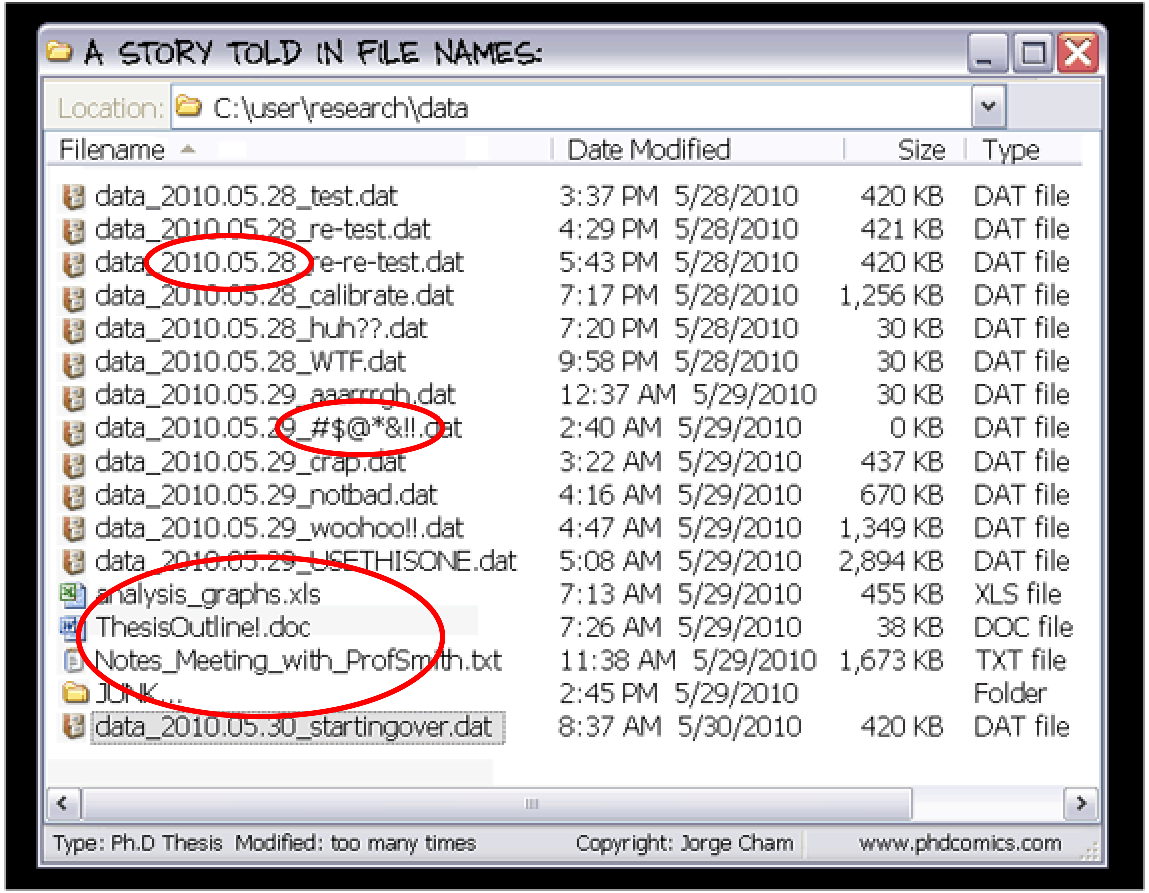
testdata.csv vs 2020_12_01_MercuryTestData.csv

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?



A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

| Filename | Date Modified | Size | Type |
|---|---|---|---|
| data_2010.05.28_test.dat | 3:37 PM 5/28/2010 | 420 KB | DAT file |
| data_2010.05.28_re-test.dat | 4:29 PM 5/28/2010 | 421 KB | DAT file |
| data_2010.05.28_re-re-test.dat | 5:43 PM 5/28/2010 | 420 KB | DAT file |
| data_2010.05.28_calibrate.dat | 7:17 PM 5/28/2010 | 1,256 KB | DAT file |
| data_2010.05.28_huh??.dat | 7:20 PM 5/28/2010 | 30 KB | DAT file |
| data_2010.05.28_WTF.dat | 9:58 PM 5/28/2010 | 30 KB | DAT file |
| data_2010.05.29_aaarrgh.dat | 12:37 AM 5/29/2010 | 30 KB | DAT file |
| data_2010.05.29_#$@*&!!.dat | 2:40 AM 5/29/2010 | 0 KB | DAT file |
| data_2010.05.29_crap.dat | 3:22 AM 5/29/2010 | 437 KB | DAT file |
| data_2010.05.29_notbad.dat | 4:16 AM 5/29/2010 | 670 KB | DAT file |
| data_2010.05.29_woohoo!!.dat | 4:47 AM 5/29/2010 | 1,349 KB | DAT file |
| data_2010.05.29_USETHISONE.dat | 5:08 AM 5/29/2010 | 2,894 KB | DAT file |
| analysis_graphs.xls | 7:13 AM 5/29/2010 | 455 KB | XLS file |
| ThesisOutline!.doc | 7:26 AM 5/29/2010 | 38 KB | DOC file |
| Notes_Meeting_with_ProfSmith.txt | 11:38 AM 5/29/2010 | 1,673 KB | TXT file |
| JUNK... | 2:45 PM 5/29/2010 |  | Folder |
| data_2010.05.30_startingover.dat | 8:37 AM 5/30/2010 | 420 KB | DAT file |

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

# *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 most important aspect of documentation is doing it.

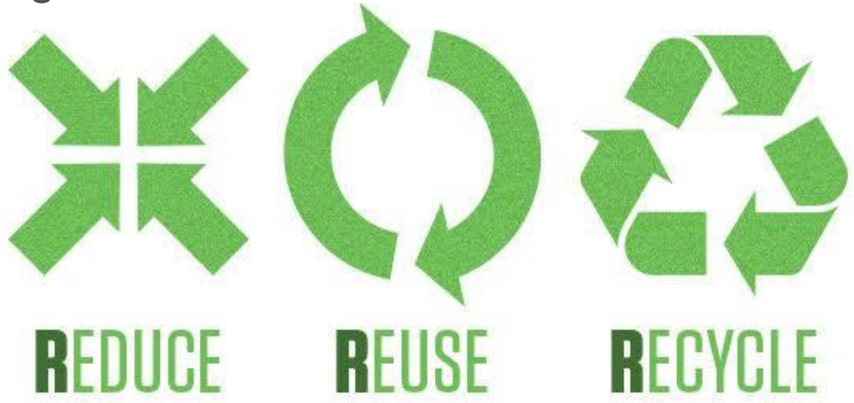
Whatever file naming and organization scheme you choose, make sure it’s **descriptive**, use it **consistently** and **document** it (in a readme.txt file).

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.



# *Movement towards openness*

Sharing data openly is a critical element in pushing academia towards openness and transparency. Open and free data sharing supports research ideals of **verification**, **reproducibility, collaboration,** and maximizes the impact and visibility of research.

# *Why should I share my data?*

Improve the **quality** of your research

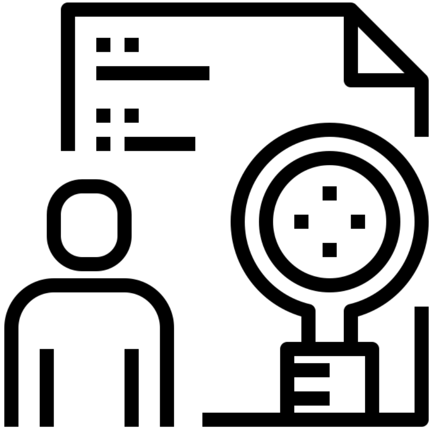
- Allow verification of results/code by peers
- Potential of ‘mega’ datasets

Improve the **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 should I share my 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

# *Why should I share my data?*

Your journal or funder may require data sharing. The Tri-Agencies will in the future require researchers to: “deposit into a digital repository **all digital research data**.” There are currently some agency specific requirements:

- CIHR currently requires researchers to “deposit **bioinformatics, atomic, and molecular coordinate data** into the appropriate public database.”
- SSHRC requires researchers to “**make available for use by others** all research data collected with the use of SSHRC funds”
- See the [Tri-Agency Data Management Policy ]for details

# *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)
- 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?*

## MacSphere

https://macsphere.mcmaster.ca/

- Institutional repository for **scholarly works**:
- A home for all research documents, including publications, presentations, conference proceedings, theses, reports, etc
- When you graduate you will upload your thesis here

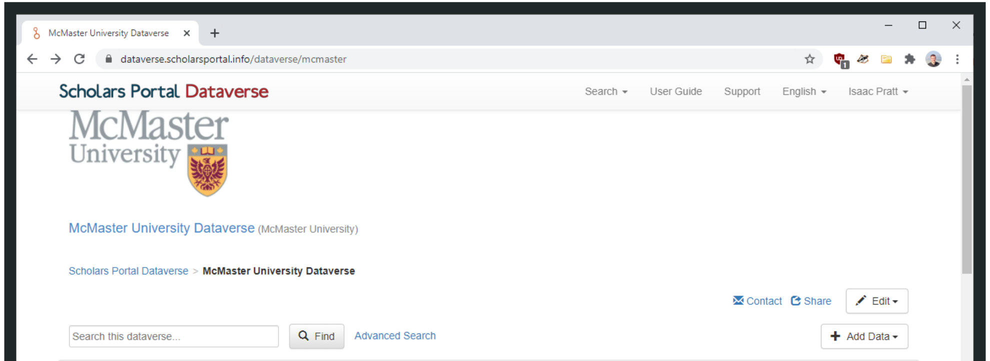
MacSphere
McMaster University Libraries Institutional Repository

# *Ok so where do I put everything?*

## McMaster Dataverse

[dataverse.scholarsportal.info/dataverse/mcmaster](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 (McMaster University)

Scholars Portal Dataverse > **McMaster University Dataverse**

# *Ok so where do I put everything?*

**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

# *Ok so where do I put everything?*

## Other 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

# *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

ORCiD

Connecting research and researchers

# *Publishing data*

Data should be stored in sustainable file formats and media

- Have you ever saved data on a CD, DVD, or BluRay? How about a zip disk or HD 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

# *Publishing data*

Data should be stored in sustainable file formats and media. Use formats that are:

- Standardized
- Open
- Well documented
- In common usage
- Unencrypted
- Uncompressed

# *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: rdm@mcmaster.ca

RDM
@McMaster

Lewis & Ruth Sherman Centre for Digital Scholarship
scds.ca

McMaster University | Library