

Best practices for managing data in your research.

Isaac Pratt, PhD + Danica Evering, MA January 18, 2023









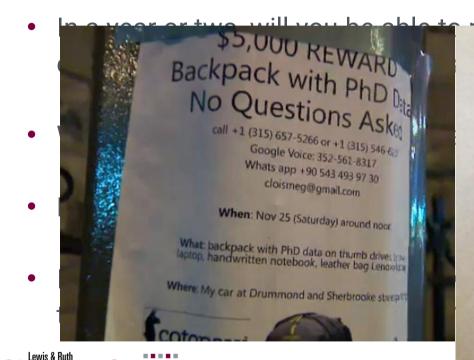
McMaster University is located on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the "Dish With One Spoon" wampum agreement.

Laslovarga, "Webster Falls in Winter, Waterdown, Hamilton, Ontario, Canada - Spencer Gorge / Webster's Falls Conservation Area," 23 January 2011, Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Waterdawn_Webster_Falls_in_Winter8.ipg

Is your data vulnerable?

for Digital Scholarship

scds.ca





I am a Rutgers Chemistry 5th year PhD student. On April 19th afternoon, my LENOVO THINKPAD T420S laptop was stolen from room 203 of Wright-Rieman building. If you stole my laptop and now you are reading this letter, I would like to say that you can keep the computer and I would like to pay you money for my data under D drive. The data is my FIVE-YEAR work. I really need the data under the D drive, there is a folder named RESEARCH, under RESEARCH folder, there is a THESIS folder. I only need that folder for my thesis defense, which is coming very soon. I would like to pay you \$1000 and use whatever way you offer to send you the money. The price is negotiable. My laptop password is 850713zd, my email address. and phone number is PLEASE contact me and I would appreciate it so so much!!!

ble

Library

Is your data vulnerable?



Is your data vulnera

 In a year or two, wi experiments? Have

What will happen t

How much work w

If you needed to she to understand it will

University of Manitoba Psychology

the details of your ails?

/move/retire?

your data?

Winnipeg Free Press

tor, would they be able







Vines et al 2014

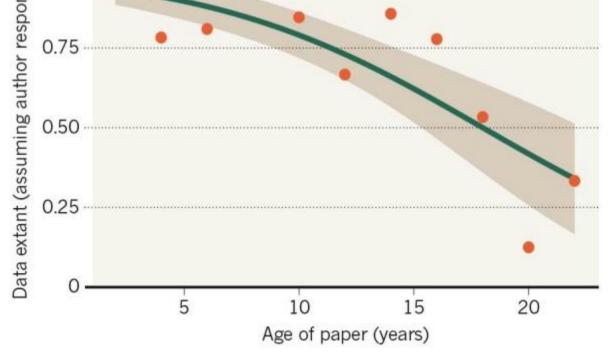
As research articles age, the odds of their raw data being extant drop dramatically.







• If you not to under



f your

€?

ey be able



Library

Sherman Central School

Winnipeg Free Press

Let's 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 x-ray scan images, microscope images
 - Software/hardware configuration files instrument specific files, scripts, text files
 - Measurement data files spreadsheet files





Example practices

- 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?

- Make a plan for your data
- 2. Organize and document your data consistently
- 3. Store and back-up your data securely
- ready for archival and sharing



Avoid duplication of research and increases use of existing data

Save time and resources





Avoid loss of data

Protect research participants









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 proactive process that helps you anticipate and identify opportunities and challenges in managing your data before those opportunities and challenges emerge.

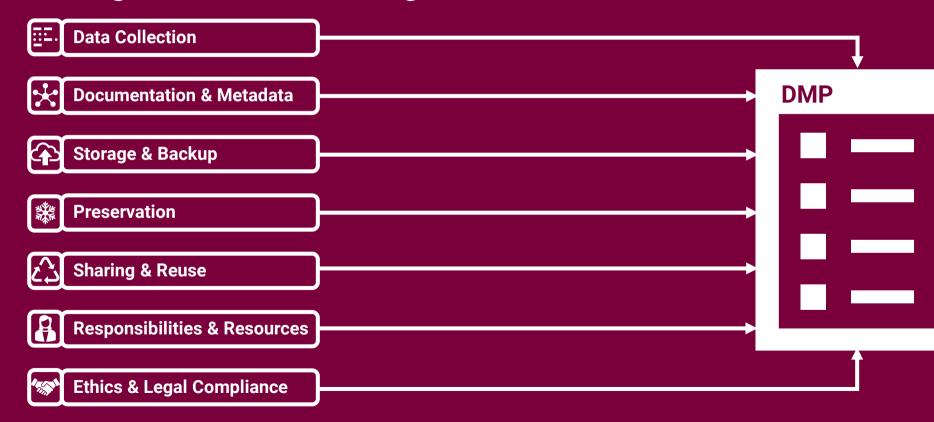
A document which speaks to the management of data both **during** the active phases of your research and **after** the completion of the research project.

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



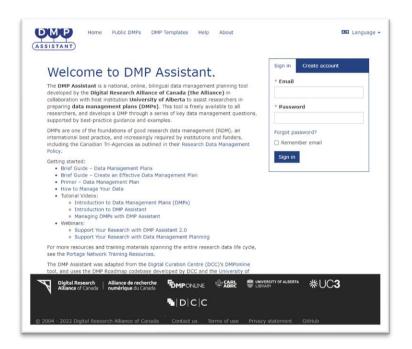


What goes in a Data Management Plan?





Alliance de recherche numérique du Canada



A web-based, bilingual data management planning tool

Available to all researchers in Canada

Walks you through relevant questions for data management

Exportable data management plans

assistant.portagenetwork.ca/





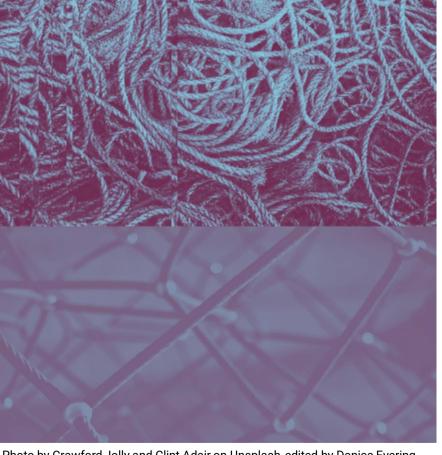


Photo by Crawford Jolly and Clint Adair on Unsplash, edited by Danica Evering.



Organize and document your data

Raw data isn't easy understand and re-use. Data documentation supports **interoperability** and **reusability**.

- If you needed to use data you collected a year ago, how easy would they be to find and use?
- Would you know what every variable is?
- Would you have information about when/where/how the data was collected?
- Have you ever gone to analyze data or publish a paper only to find that some critical piece of information was not recorded, or you don't remember where you wrote it down?





Soil radiometrics: Field and remote data sets for model building and va

242.3KB

Contributors: Cassia Read, David H. Duncan, Chiu Yee Catherine Ho, Matt D. White, Per Date created: 2017-05-02 09:40 PM | Last Updated: 2018-06-14 11:46 PM

Category: Project

Description: Repository for model training and testing data sets for the article: Read CI White M, Vesk PA. Useful surrogates of soil texture for plant ecologists from airborne & Ecol Evol. 2017;00:1–10. https://doi.org/10.1002/ece3.3417

Wiki

This project is home to the soil data for north-west Victoria, Australia used by Read & Useful surrogates of soil texture for plant ecologists from airborne gamma-ray dete-

See the respective data set wiki pages for further information on provenance and

Files

Cassia Read, David H Duncan, Chiu Y C Ho, Modified Matt D White, and Peter A Vesk, "Soil model ... Radiometrics: Field and Remotely-Sensed Data Sets for Model Building and Validation," OSF, June 15, 2018,

Research Project Management

- Collaboration: Google Docs and Microsoft Office let your team work on documents together in real-time, avoiding multiple versions and copies sent by email.
- Reference Management: Zotero, Mendeley, or Endnote support collaboration through shared citation libraries.
- Notetaking software: Obsidian, Evernote, OneNote, Notion, or an Electronic Lab Notebook allow you to create organized, linked notes that you can use to document your research practices
- Open Science Framework (OSF): This free open platform for research collaboration lets you manage files, data, code, and protocols in one central location.

Learn more at rdm.mcmaster.ca/organize



osf.io/uac6x.



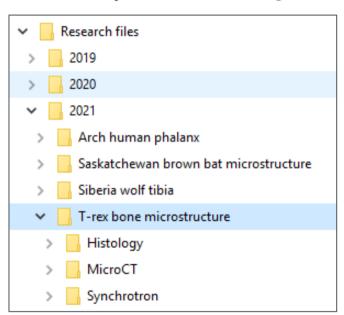
Library

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!

Organizing your research files makes your data easier to understand, share, and archive—both now and in the future. A good file organization system should be **descriptive**, **standardized**, and **implemented consistently**.

File Naming: Include date, project name, short description, initials of researcher, version number, and other metadata (like location). Good file names make your files searchable.
 e.g. 2022_10_01_LakeMercury_TestData1_TM_v3.csv

 Update your team: Once you've established a file organization and naming system, describe it in a README file for your research project and make sure everyone on your research team knows the system. **Date**: 2022_10_01 (collection date)

Project Name: LakeMercury **Short Description**: TestData1 **Name**: TM (Tracy MacDern)

Version Number: v3

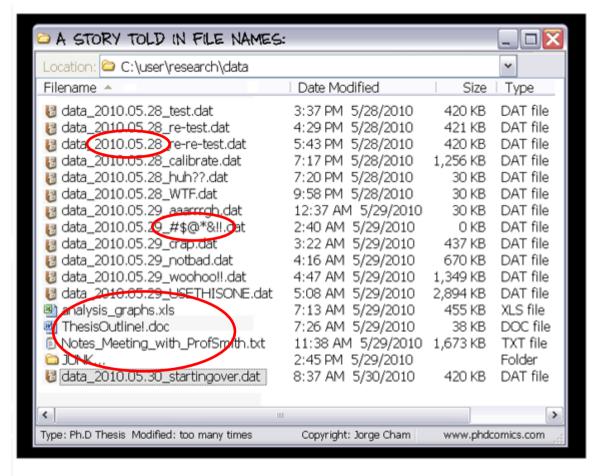




File naming

Do you have files named like this?

Is this a good file name system?







Project: Kristin's important chemistry project

Date: June 2013-April 2014

Description: Description of mu awesome project here Funder: Department of Energy, grant no: XXXXXX Contact: Kristin Briney, kristin@myemail.com

ORGANIZATION

All files live in the 'ImportantProject' folder, with content organized into subfolders as follows:

- 'RawData': All raw data goes into this folder, with subfolders organized by date
- 'AnalyzedData': Data analysis files
- 'PaperDrafts': Draft of paper, including text, figures, outlines, reference library, etc.
- 'Documentation': Scanned copies of my written research notes and other research notes
- 'Miscellaneous': Other information that relates to this project

NAMING

Raw data files will be named as follows:

"YYYYMMDD_experiment_sample_ExpNum" (ex: "20140224 UVVis KMnO4 2.csv")

STORAGE



"README.TXT," Data Ab Initio, nttp://dataabinitio.com/?p=378

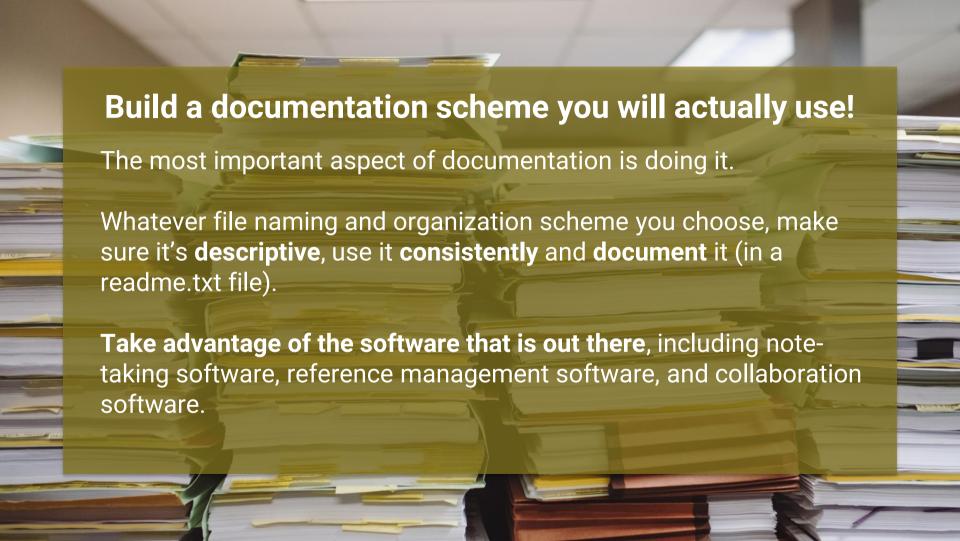
2014

Kristin Briney,

Documentation files

- **README**: A simple text document (.txt) that describes project information, folder hierarchy and file organization, description of important file contents.
- **Data Dictionaries**: A document for tabular data that describing names, labels, units, and constraints.
- **Codebooks:** Like data dictionaries but for survey or statistical data—includes the survey layout and structure, and codes for questions and answers.





Storing your data safely

- 1. Avoiding data loss
 - Theft or loss of devices, accidental damage or destruction
- 2. IT Security
 - Computer viruses, malware, ransomware





Library



Backup Strategies (3-2-1)

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



Copies of your data (at least!)

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

2

Copies are on-hand (easily accessible) on different systems (internal hard drive, cloud storage, etc.)

- a "production" (working) copy
- a "production backup" copy



Copy is in another location ("off-site") from the others with a *trusted* service provider



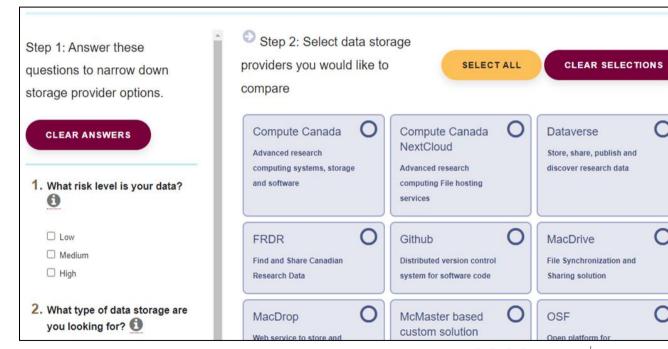


Research Data Storage Finder Tool

http://u.mcmaster.ca/storagefinder

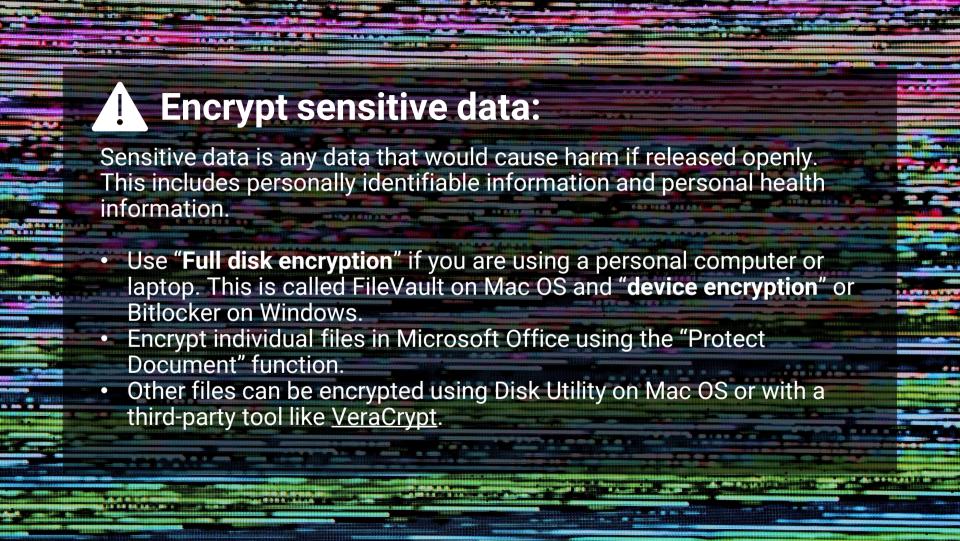
McMaster RDM Services
has a **Data Storage Finder**, an interactive tool
to help you find a vetted
storage provider
depending on risk, volume,
and other needs

This tool also allows you to compare feature sets of selected options.









How should I protect my data?

Enable Multi-Factor Authentication (MFA)

- Also known as 2 Factor Authentication (2FA)
- Requires 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
- Many other web services (Gmail, Dropbox, etc) provide MFA



an't use my Microsoft Author





Password Best Practices

Make sure your online information is secure by ensuring your password is:

- ✓ **Strong**: Make a strong password by combining a series of numbers, letters, and symbols into a long series of words. Try to combine them into something memorable like L1br@ryt1pS.
- ✓ Unique: Use a different password for every website/service
- ✓ **Secret**: Never share your passwords with anybody, even if you trust them. Keep your passwords secret by storing them only in your head. Never send them in an email!
- ✓ Fresh: Change your passwords at least once every three months
- ✓ Devices: Use a strong password on your computer and phone, too



Tip: Remembering multiple passwords can be difficult. Use a trusted password manager to keep track of your passwords for you. Some examples are BitWarden and 1Password.







Common Password Mistakes

Here are some common mistakes people make when creating passwords:

AutoSave Off H Test Example's Passwords.xlsx • Saved >		
File	Home Insert Draw	Page Layout Formulas Data
Service	Email	Password
MacID	example@mcmaster.ca	football
DropBox	example@mcmaster.ca	football
Google	test.example@gmail.com	football
Instagram	test.example@gmail.com	football1
ORCID	example@mcmaster.ca	football
Zotero	example@mcmaster.ca	football

Have you made any of these mistakes before?





Written Down: Passwords written down on a piece of paper or stored in plain text on a computer may be stolen by somebody with malicious intent or easily lost.

Too Simple: Simple or common passwords are easy to guess or brute force. Examples: apple, rowboat, bumblebee, blizzard,

password

Same Password: Using the same password for multiple websites is like having one key for multiple locks; if it's stolen, the thief can open them all.

Publishing Data

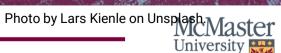
for Digital Scholarship

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.



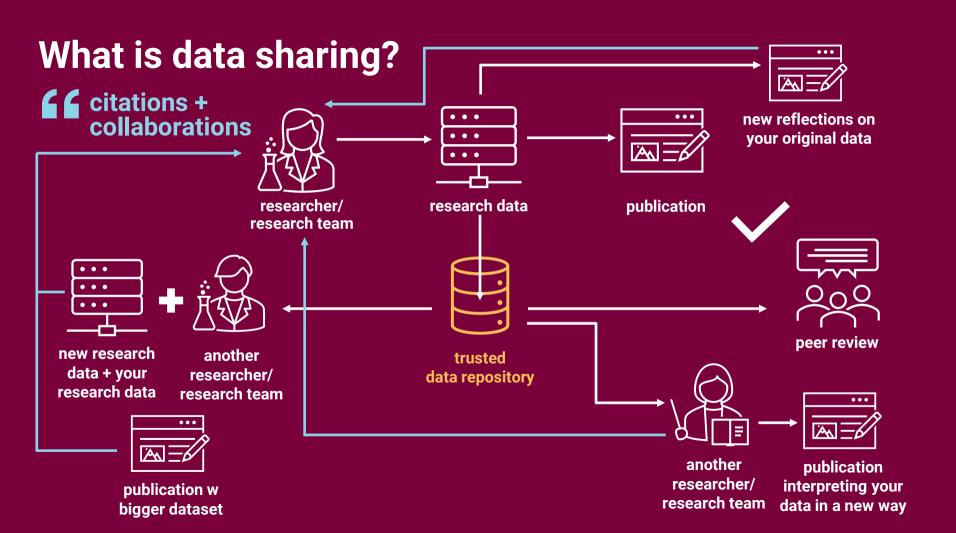


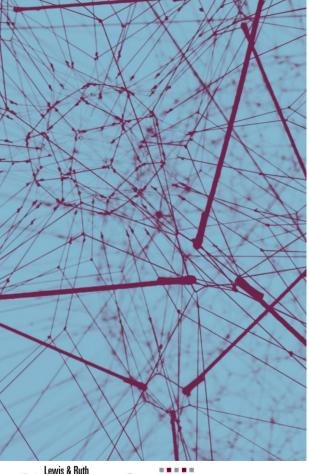




Open Research open data open-source software open access open government

What is data sharing? citation another publication citing researcher/ your findings research team ••• . . . "data available on publication researcher/ research data request" research team peer review





Why share data? Citation Impact

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 share data? Journal and Publisher 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.





Why share data? Tri-Agency Requirements



Government of Canada

Gouvernement du Canada

Search Canada.ca

MENU •

Home > Interagency research funding > Policies and Guidelines > Research Data Management

Research Data Management

Tri-Agency Statement of Principles on Digital Data Management

Open Letter

Tri-Agency Research Data Management Policy

Public Consultation Summary

Frequently Asked Questions

Tri-Agency Research Data Management Policy

1. Preamble

The <u>Canadian Institutes of Health Research (CIHR</u>), the <u>Natural Sciences and Engineer Research Council of Canada (NSERC</u>), and the <u>Social Sciences and Humanities Researc Council of Canada (SSHRC</u>) (the agencies) are federal granting agencies that promote support research, research training, knowledge transfer and innovation within Canada

The agencies expect the research they fund to be conducted to the highest professior and disciplinary standards, domestically and internationally. These standards support research excellence by ensuring that research is performed ethically and makes good of public funds, experiments and studies are replicable, and research results are as accessible as possible. Research data management (RDM) is a necessary part of resea excellence.

The agencies believe that research data collected through the use of public funds sho be responsibly and securely managed and be, where ethical, legal and commercial

"Grant recipients are required to deposit into a digital repository all digital research data, metadata and code... in journal publications and pre-prints."

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 <u>Tri-Agency Data Management Policy</u> for details.





Library

Open Access (OA) 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 <u>arXiv.org</u>)
- 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





Persistent Identifiers help keep track of everything and make research findable.

Persistent Identifiers (PIDs) are unique links that will never expire.

- <u>Digital Object Identifiers (DOIs)</u>: Can be minted for publications, datasets, lab equipment, research software and more.
- Open Researcher & Contributor ID (ORCiD): Unique identifier for researchers. Distinguish yourself from scholars with the same name; connect your datasets, code, and publications.

Connecting research and researchers





Library



Sustainable File Formats

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

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





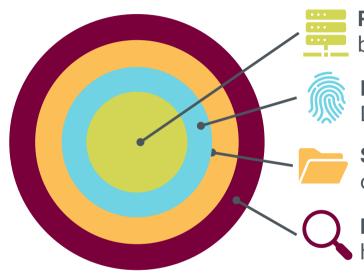








Datasets as a digital objects.



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







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.





Data Licensing

- Creative Commons (CC) (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
- Traditional Knowledge (TK): In addition to the First Nations Information Governance Centre's OCAP® training, communities may also use TK licenses localcontexts.org/licenses

Images from Sg'éwlets, "Traditional Knowledge Labels," http://digitalsgewlets.ca/traditionalknowledge_connaissances_traditionnelles-eng.php and Creative Commons, fixed by Quibik.





In our Stó:lō culture, certain types of knowledge are restricted in some way. This knowledge is considered sacred, secret, potent and/or private, and only certain people or families can and should have access to them. We call this xa:xa in our language. This label indicates that there is additional knowledge about a certain subject that cannot be shared on the website.





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 Digital Rights Management (DRM)

https://dataverse.org/best-practices/dataverse-community-norms https://opendatacommons.org/norms/ nmunity Meeting 2022 1 June 14, 15, and 16



Speakers and Chairs Registration



¿Healy's professional background is multis worked in numerous health capacities at , and international levels. Actively involved ways of knowing, Bonnie's passion is to iities and provide them with tools that they port communities in information data esearch methodologies. Bonnie's irst Nations information systems gives her rong passion for using data as a tool for

#Dataverse2022

The annual Dataverse Community Meeting is an opportunity to build, grow, and enrich the global community. Like the open-source Dataverse product itself, the activities of the Dataverse Community Meetings are community-driven. Over three days of presentations, workshops, and working group meetings we aim to promote and learn about behavioral and technical solutions and standards for curating, sharing, and preserving data that can be discovered and reused across disciplines to reproduce and advance research.

The Dataverse Community Meeting is hosted by Harvard's <u>Institute for Quantitative Social Science</u>. Learn more about The Dataverse Project at our dataverse.org site.









Ok, so where do I put everything?

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 made for research.





Data Repository Decision-Making

Publishing data in a recognized data repository is the best way to share data. There are thousands of data repositories. <u>MacSphere is our institutional repository for publications</u>.



Domain Specific Repositories

Focus on certain types of data such as genomic information or astronomical information.



General Repositories

Accept broader types of research data. ex. McMaster Dataverse (part of Borealis) and Canada's Federated Research Data Repository (FRDR).



Code Repositories

There are also codespecific repositories like Github, Gitlab, BitBucket, SourceForge



Repository Finder

This tool by DataCite can help you find an appropriate repository to deposit your research data: https://repositoryfind er.datacite.org







Ok, so where do I put everything?

MacSphere

https://macsphere.mcmaster.ca

- Institutional repository for scholarly documents (not data).
- A home for all research documents, including publications, presentations, conference proceedings, theses, reports, etc.
- When you graduate you will upload your thesis here.

Link your dataset DOI!



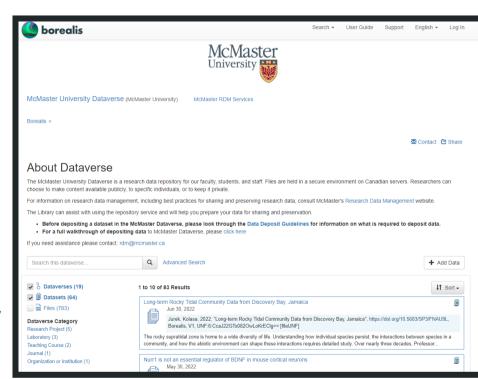




McMaster Dataverse

https://borealisdata.ca/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
- Choose whether to share data openly or through an application process



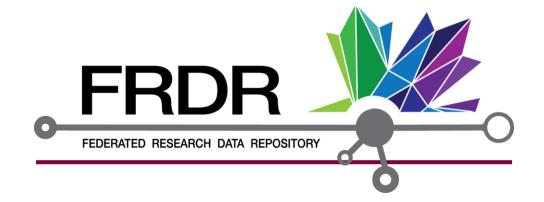




Federated Research Data Repository (FRDR)

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 Zero knowledge encryption for sensitive data

- 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. All datasets are encrypted, and their keys stored in a separate researcher-managed platform. This allows you to deposit your data in a trusted repository for archival but maintain complete control over access to the data.
- Interested in piloting this service? Email us: rdm@mcmaster.ca







Top 4 ideas for improving your research data management:



Make a plan for data management.



Create a file organization scheme (and use it).



Ensure your data is safely **stored** and backed up.



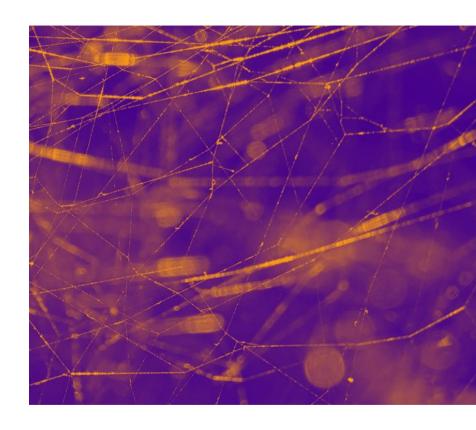
Share your data openly.





RDM Community of Practice

- Monthly meetings of people interested in RDM at McMaster –
 Thurs, Jan. 26 – 11 AM!
- Guest presentation by Dr. Antonio Paez, Earth Environment and Society on Reproducible Research in Graduate Studies.
- Connect with other researchers practicing RDM across the university!
- https://u.mcmaster.ca/rdmcommunity









McMaster RDM webpage: <u>rdm.mcmaster.ca</u>

Contact RDM services at: rdm@mcmaster.ca

Upcoming RDM webinars: rdm.mcmaster.ca/events

Recorded RDM webinars: <u>u.mcmaster.ca/learn-rdm</u>

Make an appointment with a Research Data Management Specialist: u.mcmaster.ca/rdm-appointments



