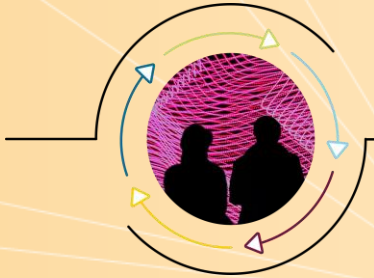
Depositing & Sharing Data Online with McMaster Dataverse

Isaac Pratt, PhD

January 25th, 2023



Research Data Management Services

McMaster University

McMaster University | Library

Lewis & Ruth Sherman Centre for Digital Scholarship

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.

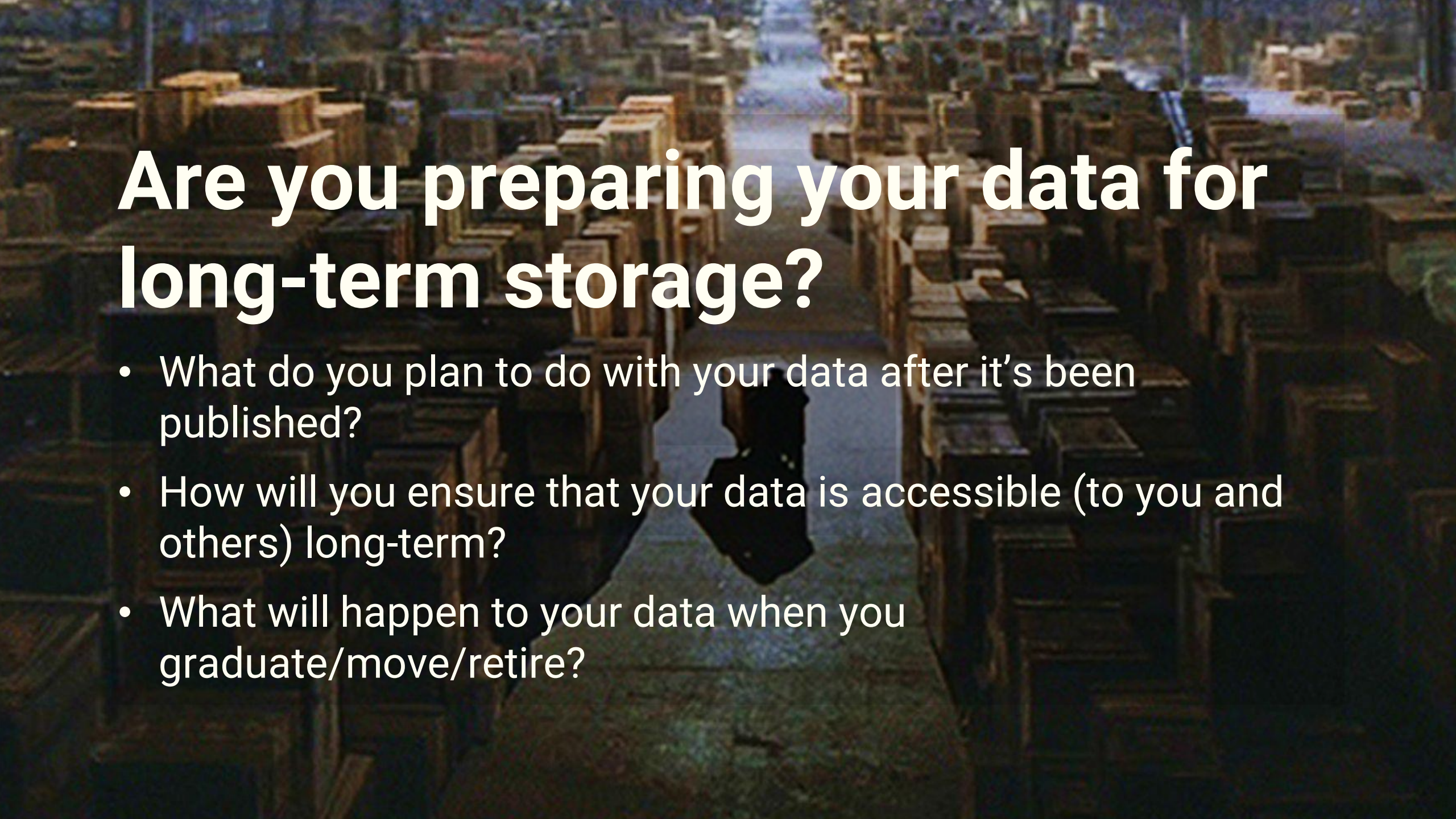
Certificate Program

The Sherman Centre offers a Certificate of Completion that rewards synchronous participation at 7 workshops. We also offer concentrations in Data Analysis and Visualization, Digital Scholarship, and Research Data Management.

Learn more about the Certificate Program: https://scds.ca/certificate-program

If you would like to be considered for the certificate, verify your participation in this form: https://u.mcmaster.ca/verification

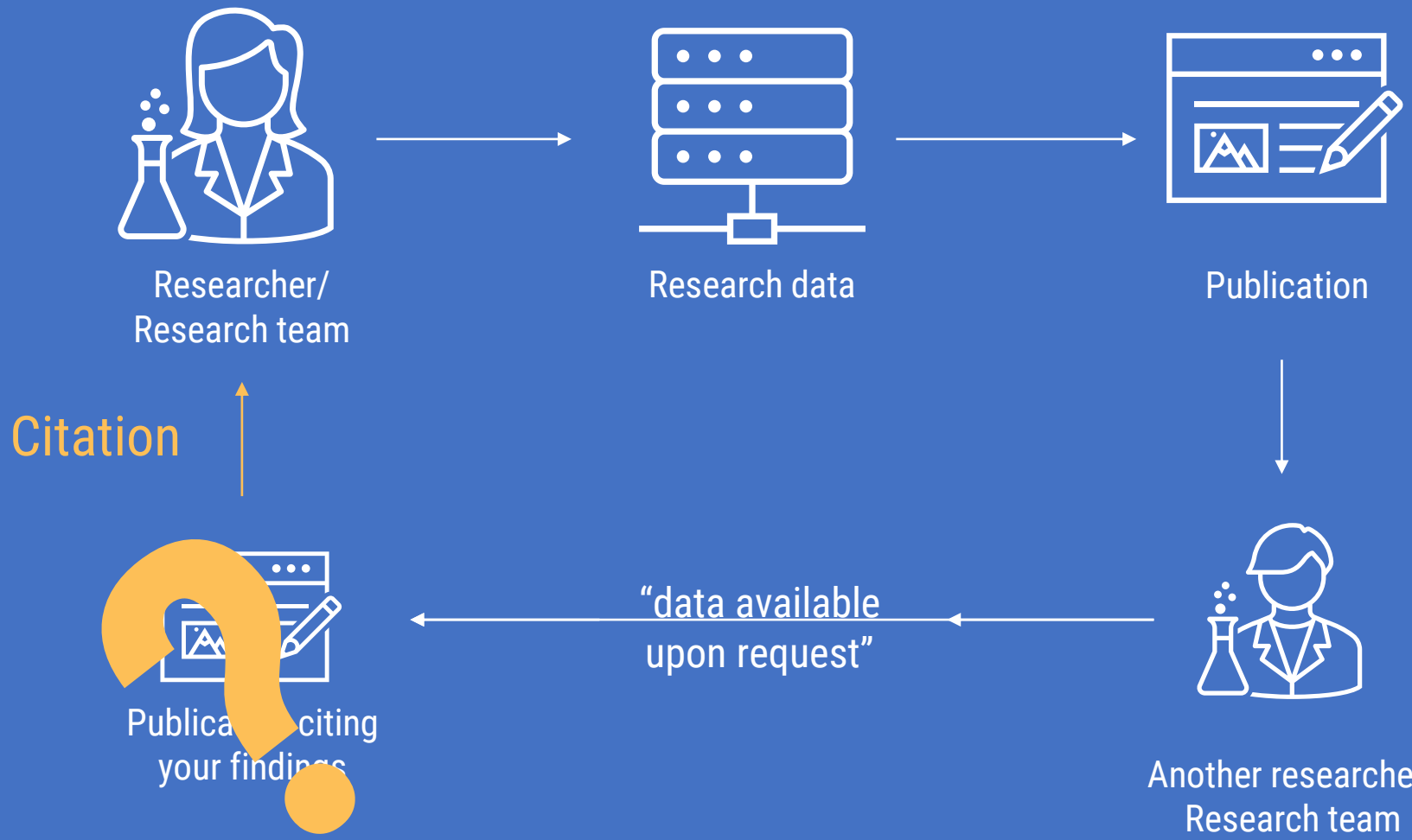
At an unspecified point during the workshop, a code will be read aloud. This is the answer to the third question of the form.



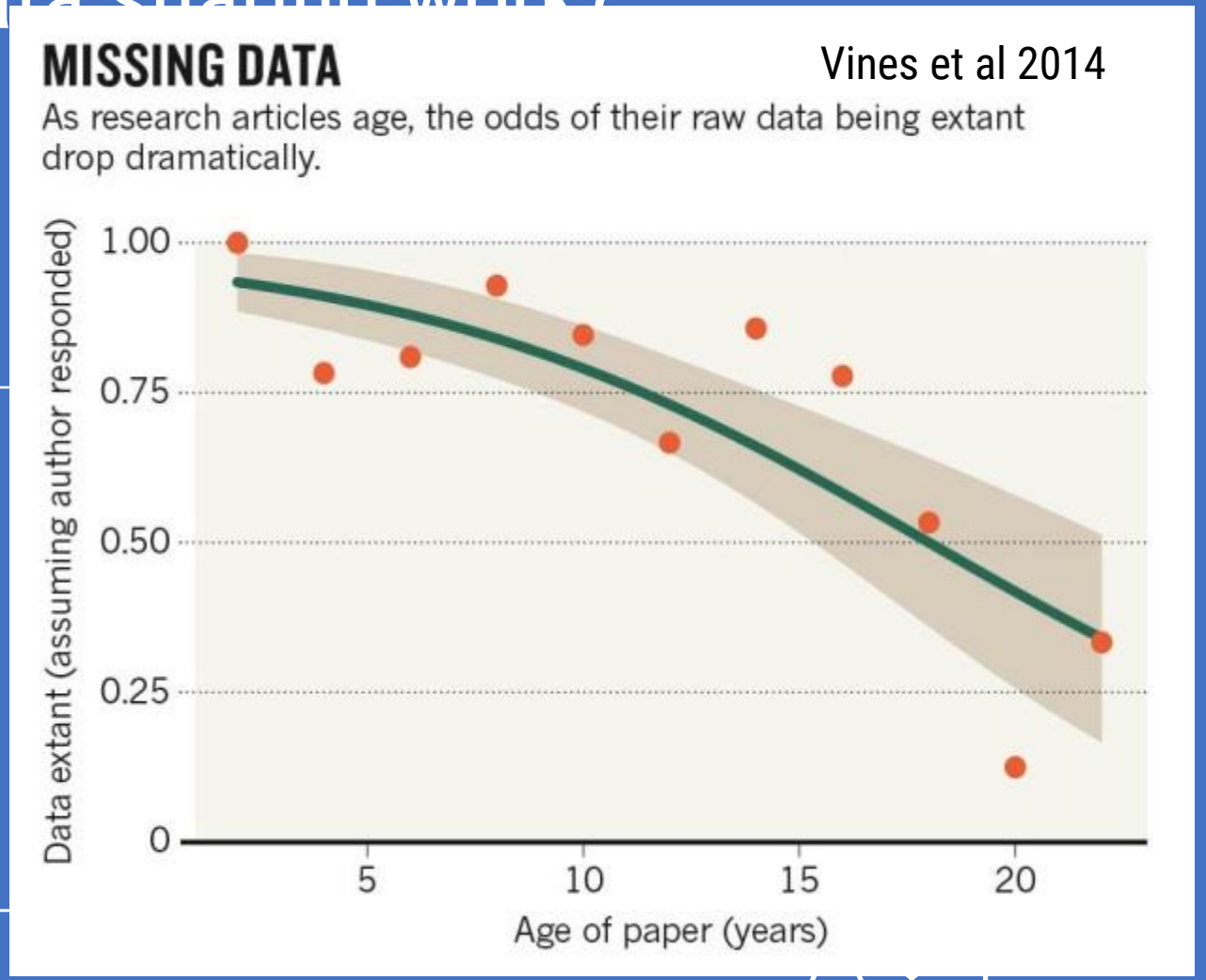
Are you preparing your data for long-term storage?

- What do you plan to do with your data after it’s been published?
- How will you ensure that your data is accessible (to you and others) long-term?
- What will happen to your data when you graduate/move/retire?

How does data sharing work?



How does data sharing work?



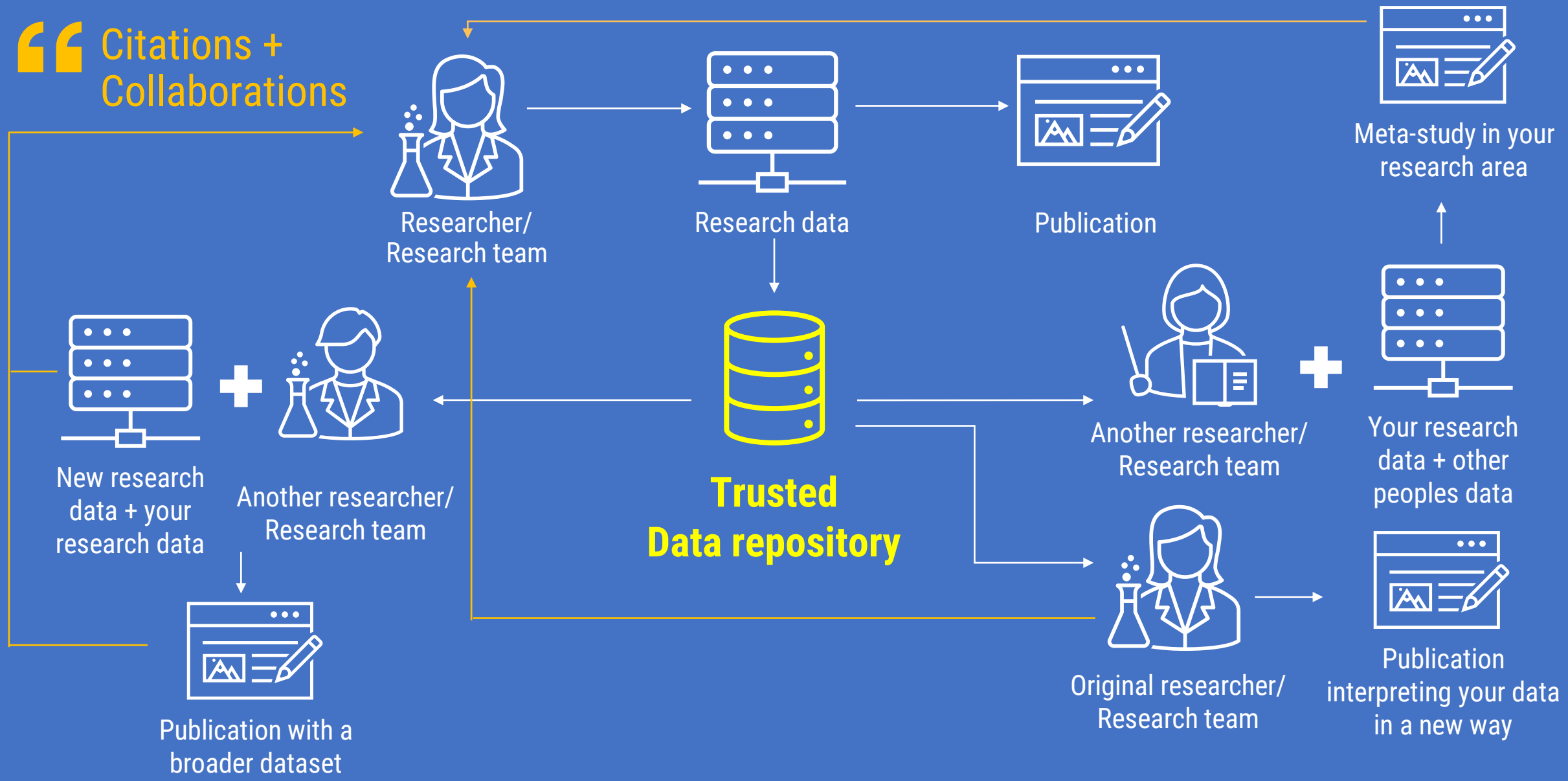
Researcher/
Research team

Citation

Publica citing
your findings

Another researcher/
Research team

What is data sharing?



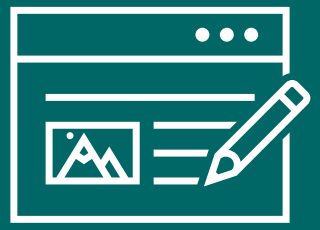
What can expanded open data sharing enable?

Researchers

- Increased citations + research impact
- Meet increased data requirements
- Increased collaborations and partnerships
- Avoid retractions
- Preservation of data

Research community

- More confidence in research results
- Increased ability to build on previous results
- Culture of reproducible research



Publishers

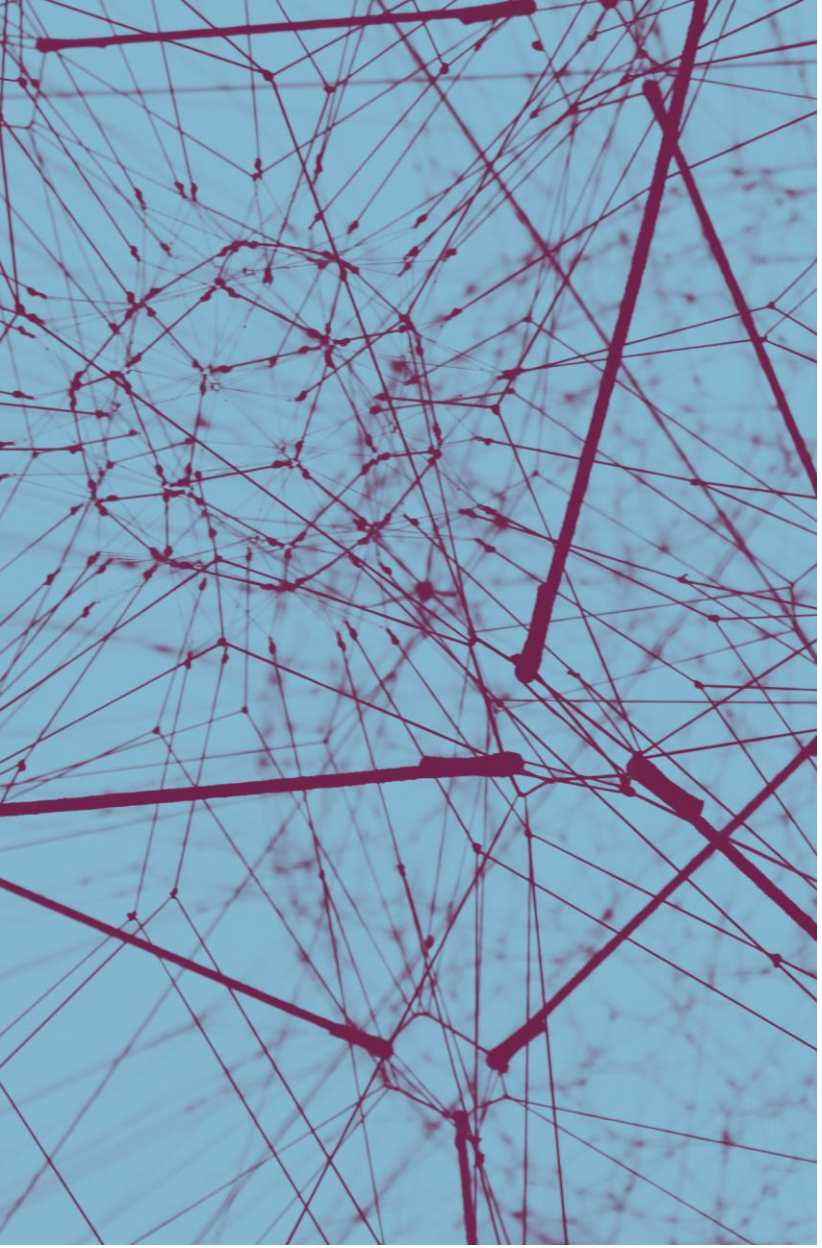
- Reproducible results
- More confidence in published results
- Fewer retractions

Funders

- Maximizes value of funding dollars
- Research excellence
- Alignment with international research community

Society

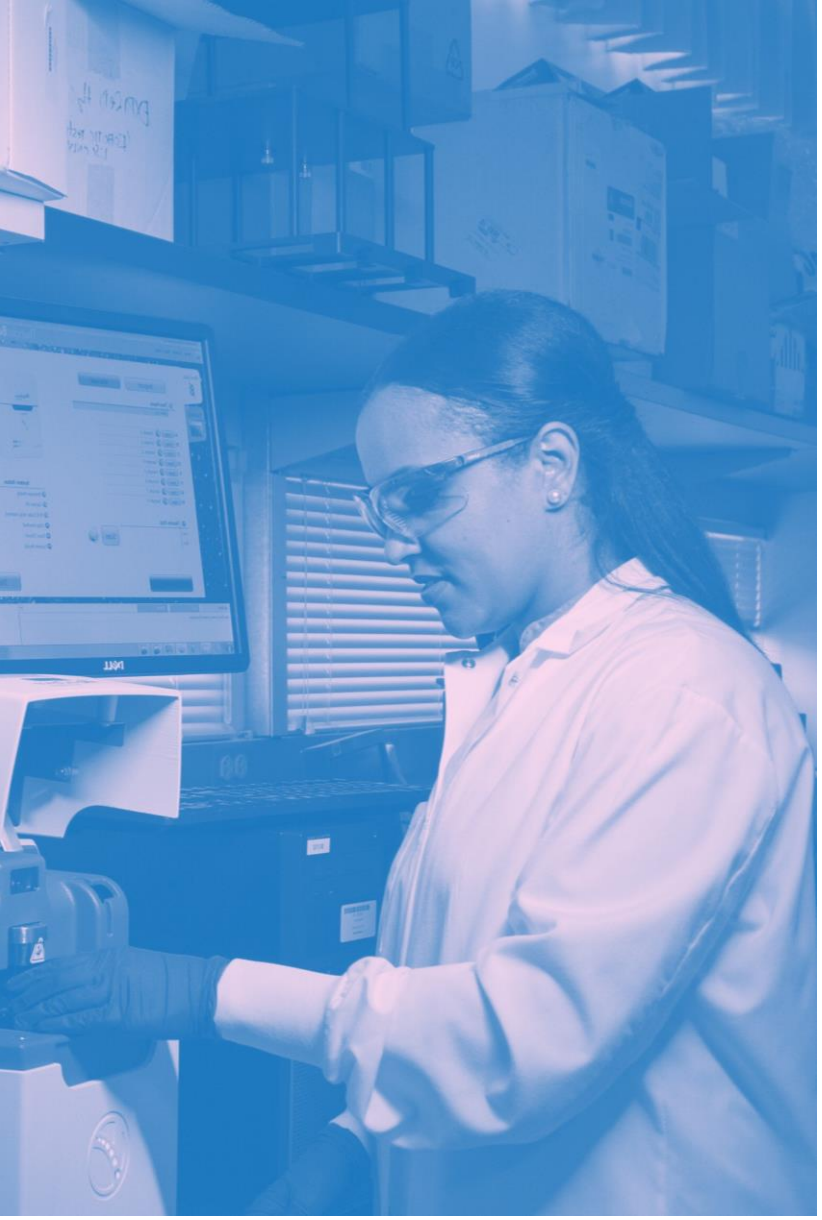
- Faster + greater benefits from research
- Increased public confidence
- Access outside of academia (journalists, NGOs, citizens)



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? Funder requirements

Tri-Agency Data Management Policy: "**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"

NIH Policy for Data Management & Sharing "researchers will maximize the appropriate sharing of scientific data"

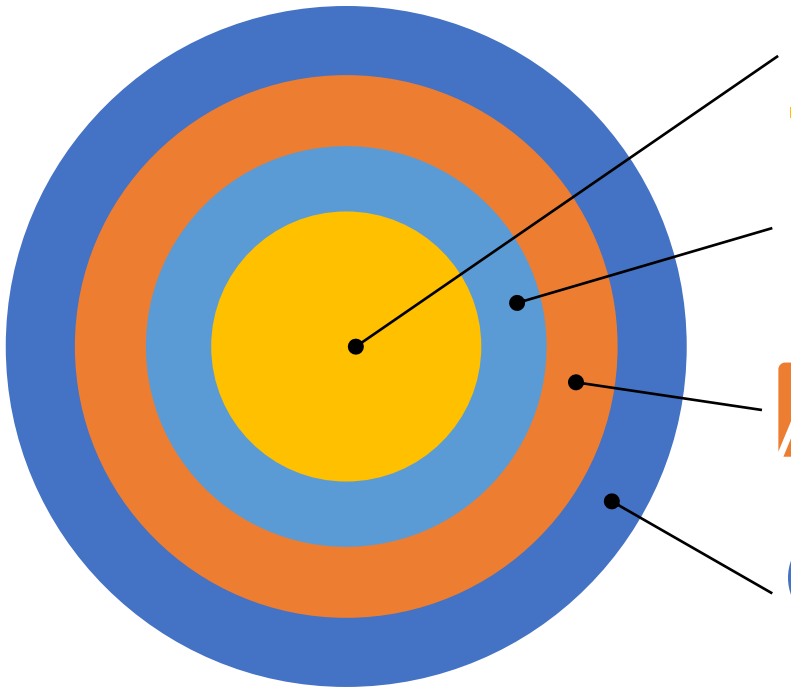
⚠ Not all data can be open.

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

You cannot share sensitive data openly. If you want to publish or share sensitive data, you have two main options:

1. **Anonymize the dataset:** remove, replace, or redact all sensitive information from datasets prior to upload.
2. Deposit data on a restricted access platform with data sharing agreements.

Preparing a dataset for deposit



Research output (the data and code): data is the core and we surround it with layers of information.

Identifiers: Persistent Unique Identifiers such as DOIs and ORCiDs help find, track, and cite data.

Standards: sustainable file formats help others access and reuse data.

Metadata: Rich metadata and data documentation help others find and understand datasets.

Project: Kristin's important chemistry project
Date: June 2013-April 2014
Description: Description of my 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

Kristin Briney, "README.TXT," Data Ab Initio, February 25, 2014 *http://dataabinitio.com/?p=378*

Documentation

- **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.

Metadata

Include metadata with your data deposit:

- Your contact information and affiliation
- Link to the associated publication (if there is one) and its DOI
- A clear description of the data and keywords

Other metadata that might be relevant:

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

Persistent identifiers

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

- Digital Object Identifiers (DOIs): Platforms create DOIs for **publications**, **datasets**, and other digital objects.
- Open Researcher & Contributor ID (ORCiD): Unique identifier for **researchers**. Distinguish yourself from scholars with the same name; connect your datasets, code, and publications.

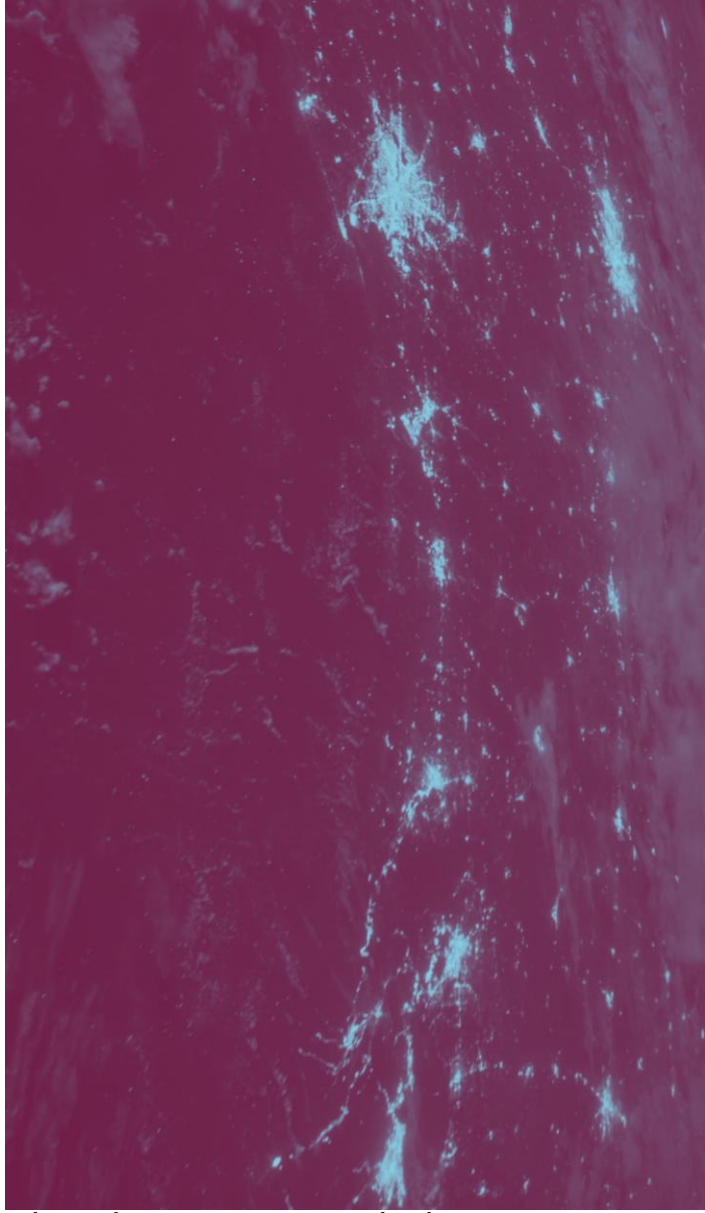


Photo by Nasa on Unsplash.

ORCID
Connecting research and researchers

McMaster University | Library | Lewis & Ruth Sherman Centre for Digital Scholarship

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

? 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 of violating your copyright.

Open sharing needs an open license, which come in a few flavors:

Most open licenses are from **Creative Commons (CC) -** (creativecommons.org)

- **Public Domain** means that you are releasing your data with no restrictions.
- **Attribution** ("CC-BY") licenses add a requirement that anyone using the data gives you credit and link to the original dataset.
- Other clauses include **Non-Commercial** ("NC") and **Share-alike** ("SA") restrictions

Community Norms

Data repositories also have **community norms.**

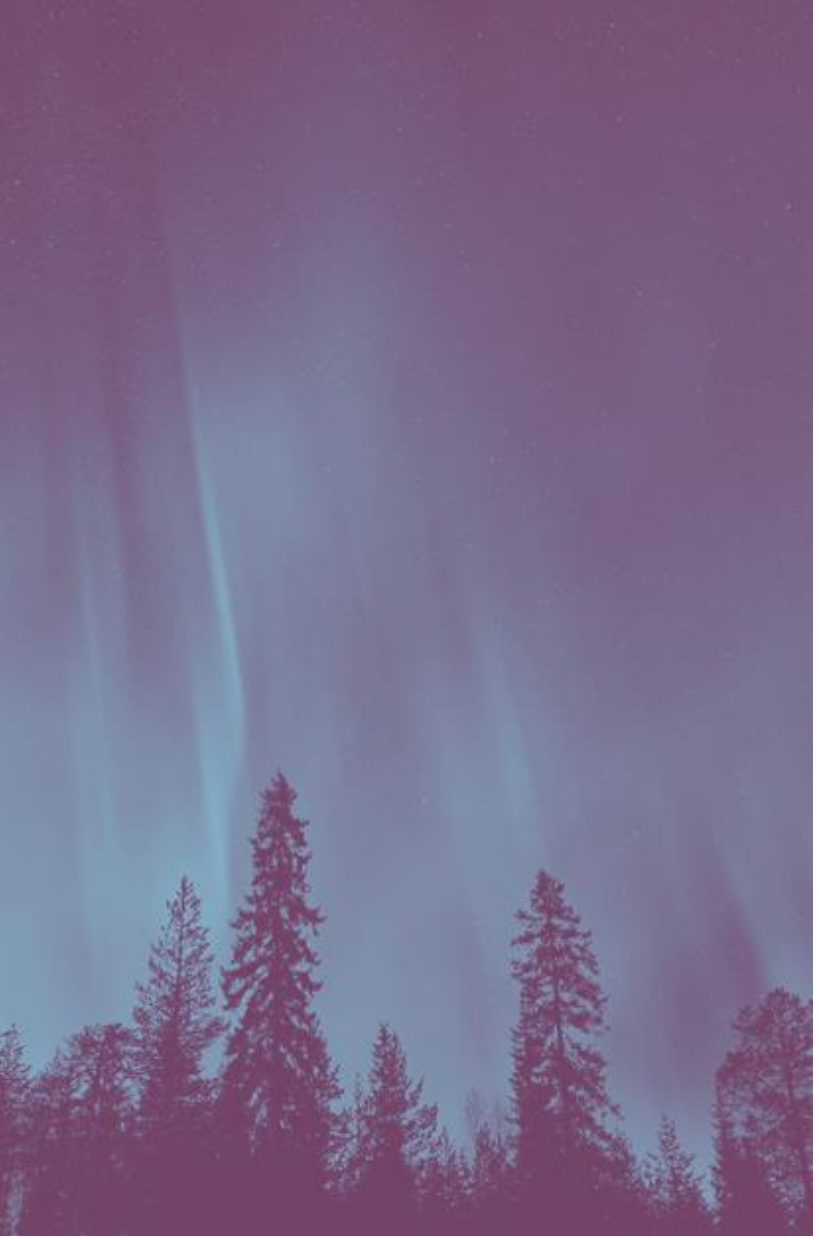
Dataverse and Open Data Commons community norms include:

- 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/

McMaster University | Library | Lewis & Ruth Sherman Centre for Digital Scholarship



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.

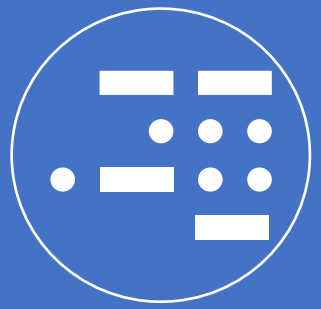
McMaster University | Library | Lewis & Ruth Sherman Centre for Digital Scholarship

Data Repositories

Publishing data in a recognized data repository is the best way to share data. There are thousands of data repositories.

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 code-specific 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://repositoryfinder.datacite.org

McMaster University | Library | Lewis & Ruth Sherman Centre for Digital Scholarship

Recommended Research Data Repositories

Institutional Repositories: **McMaster Dataverse**

External Data Repositories:

- Domain specific: look through:
  - https://www.nature.com/sdata/policies/repositories and
  - https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/repositories-for-sharing-scientific-data
- General: FRDR, Zenodo, Figshare, Mendeley Data, etc

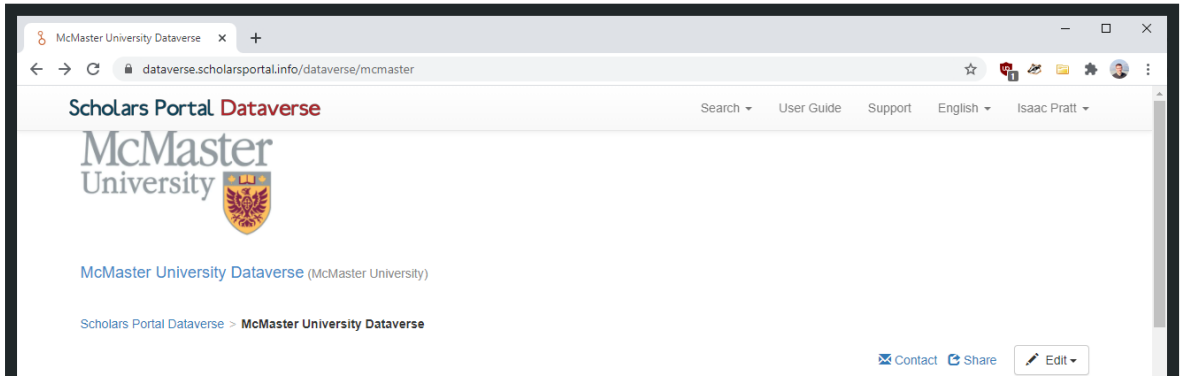
Code repositories: Github, Gitlab, BitBucket

Search for repositories on re3data.org

McMaster Dataverse

https://borealisdata.ca/dataverse/mcmaster

- McMaster's Institutional Data Repository is a home for research data created by McMaster researchers. *(Not recommended for sensitive data)*
- Provides basic data curation services
- Data is stewarded by professionals at McMaster
- Contains tools for tabular data exploration and analysis

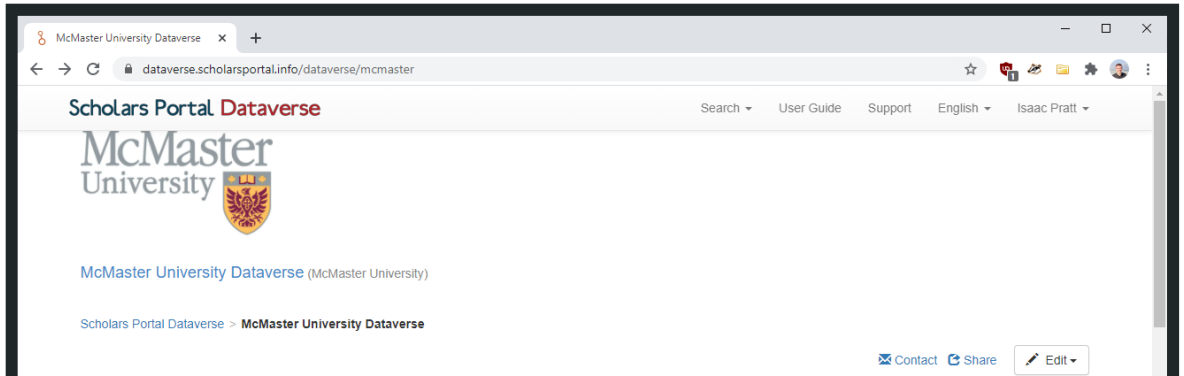


McMaster University | Library

Lewis & Ruth Sherman Centre for Digital Scholarship

McMaster Dataverse

- Demo instance for testing and learning the platform
- 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 | Library | Lewis & Ruth Sherman Centre for Digital Scholarship

Dataverse Demo

For more information:

Contact us at: rdm@mcmaster.ca

RDM Services: https://rdm.mcmaster.ca/

McMaster University | Library

Lewis & Ruth Sherman Centre for Digital Scholarship