Sharing and Deposit

The why, what, where and how

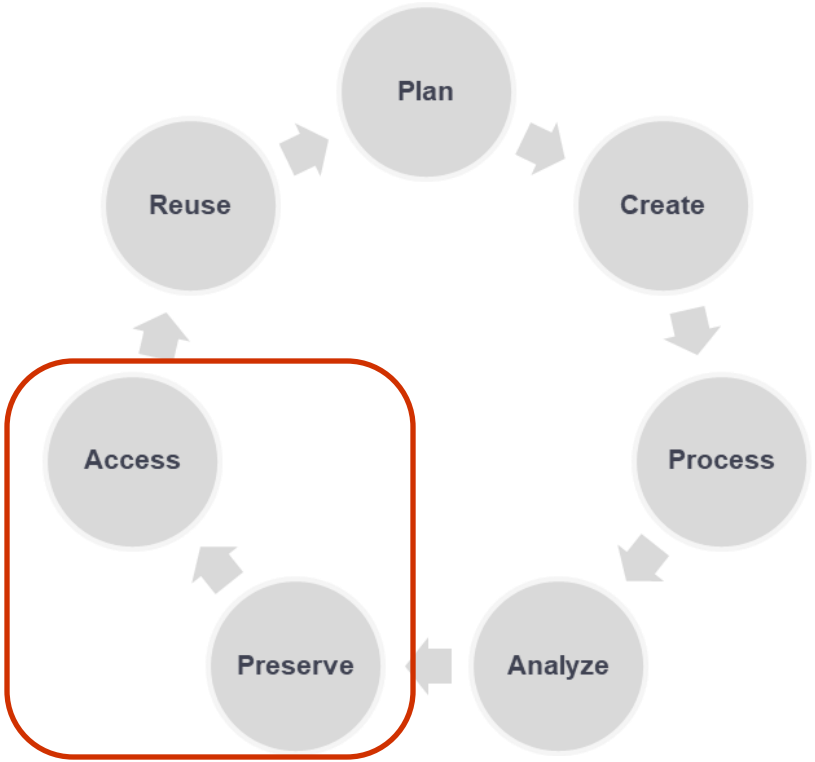
Jennifer Abel, University of Calgary
RDM Jumpstart
May 16, 2025

Sharing Data

Why do we share data?

- Because someone tells us we have to (funders, publishers, supervisors)
- Because it's good research practice
  - Allows others to reproduce our results
  - Allows others to check our work for errors, bias, etc.
  - Allows others to do new work based on ours (with proper attribution)
  - In the spirit of the FAIR principles

When do we share data?



What to share vs. what to keep internally

To share (ideally in non-proprietary formats)

- Master/final data
- Documentation that supports re-use of the data: readmes, data dictionaries, codebooks
- Code used to analyze the data

To keep

- Project management records: internal communications, decision making documents, admin records, ethics information (if appropriate), etc.
- DMPs, standard operating procedures, methodologies
- Analytic data (intermediate steps)
- Analyzed data (graphs, charts, tables, etc.)

- It depends
  - Raw data

IMPORTANT!

Just because you're not sharing something doesn't mean you should get rid of it

- Most institutions will have rules for how long you have to keep research records for after a project is complete
  - Some funders will have these types of rules as well
- Despite what you might have heard, REBs will not always require you to destroy data
  - However, if you tell participants that that's what you're going to do, then you have to do it
- Make these plans/decisions at the beginning of your project!

ALSO IMPORTANT!!

Make sure that everyone involved in the project agrees on what data are being shared (if any), and how and where that will happen.

- This includes your supervisor (if appropriate), any partners you might have at other institutions/in industry/at community organizations, Indigenous communities you may be working with.
- Also includes getting consent from participants in your research.
- If someone says no, you can't share your data.
- Make these plans/decisions at the beginning of your project and document them!

What we're sharing from this project

- Data:
  - The initial data file that you started with on day 2
  - The final data file that you finished with on day 4 (in .tsv and .Rdata)
- Data dictionary
  - The .md and/or PDF for the data we worked with (**not** for the original dataset from Statistics Canada)
- Readme
- Code:
  - Your .Rmd file
  - Your .renv lockfile

Ways to share your data

- On request
- Via a website that you maintain
- Via a data repository
  - This is often called ‘depositing’ your data
  - People sometimes use ‘sharing’ and ‘depositing’ interchangeably, but deposit is just one way of sharing data

Data Repositories

What are data repositories?

- Online database services that provide long-term preservation for data and make them available for discovery and use
- Unlike online storage services (e.g., Dropbox, Box, Google Drive), they're designed to make your data findable by others
- Unlike an institutional repository, they hold data, rather than theses/dissertations, publications, presentations, manuscripts, etc.
- Used **after a project is complete** and data won't be changing on a regular basis

What data repositories aren't

- Active storage for data during the course of a project
  - Either as your main storage or as your backup
- 'Dark storage' for data
  - Most repositories won't take your data if it can't be made findable/accessible in any way
  - Almost always an assumption that you'll make your data findable at minimum
  - In most cases, metadata about the data will be accessible, even if the data themselves aren't

Why use a data repository?

- You don’t have to handle all aspects of data sharing yourself
  - E.g., responding to every request for data, maintaining a website, ensuring that data are FAIR for a long period of time, ensuring that data are preserved for a long period of time…
- You can get support from repository staff if you have particular needs/questions/issues
- You can get metrics about how often your dataset is downloaded/viewed

What to look for in a data repository

- Good data repositories will
  - Have a mandate and plan for data preservation
  - Offer information about the data set that enables people to discover and learn about the data
  - Provide potential data users with either direct access or information on access conditions
  - Ensure each data set has a persistent identifier (PID) that will always take people to information about the dataset and can be cited in publications
- Do a little research before you start to make sure a repository is right for you

Types of data repositories

- Generalist vs. domain-specific
  - Generalist repositories take data from any research discipline
  - Domain-specific repositories are limited to data from particular areas: e.g., genomics, health sciences, social sciences
- Institutional vs. non-institutional
  - Institutional data repositories only accept data from people affiliated with a particular institution
  - Non-institutional data repositories aren't limited to people at a particular institution, but may have other restrictions

Canadian data repository options

Federated Research Data Repository (FRDR): https://www.frdr-dfdr.ca/repo/

- A national generalist repository hosted by the Alliance
- Custom-built for large datasets
- Mainly targeted at faculty researchers (students get access through a faculty member)

Borealis: https://borealisdata.ca/

- A network of 70 Canadian institutional repositories hosted by Scholars Portal/University of Toronto Libraries
- Better for smaller datasets
- Each institution decides who has access, what materials will be accepted, how data curation/review is handled, etc.

Borealis

Benefits of Borealis

- Servers are hosted in Canada
- Has a robust preservation plan (https://borealisdata.ca/preservationplan/)
- Provides a persistent digital object identifier (DOI) for all datasets
- Metadata are harvested by a number of aggregation services, e.g., Lunaris (https://www.lunaris.ca/en), Google Dataset Search (https://datasetsearch.research.google.com/)
- Metrics available for how often datasets/files have been downloaded
- If your institution uses Borealis, there'll be local support available to you

Drawbacks of Borealis

- Not all institutions use it
- Maximum file size of 5GB, so it can’t handle all datasets
- Terminology can be a little strange (e.g., the software/platform is called ‘Dataverse’ but collections are also called ‘Dataverses’)

An important thing to consider: Licensing

- When you share data you create, you should carefully consider your license
  - What do you want other people to be able to do with it, and not do with it?
  - What kind of credit do you want to receive?
- Most data repositories will require you to have some kind of license
  - Often you can select from already existing licenses (e.g., Creative Commons, Open Data Commons)
  - Sometimes you can create your own
- If you're working in a partnership, make sure you discuss this with your partners at the beginning of a project (put it in your DMP)

Creative Commons (CC) Licenses

https://creativecommons.org/

- Tells a re-user what they can do with a copyrighted work
- Notes:
  - You must be the copyright holder to assign one of these licenses
  - Licenses can't be revoked once assigned, so pick wisely! Use the CC License Chooser if you're not sure: https://chooser-beta.creativecommons.org/
- Very commonly used by data repositories (including Borealis)

CC License Types

Abbreviation	Key Feature(s)	What it means
CC BY	By Attribution	Re-users must credit the creator/copyright holder
CC BY-SA	By Attribution, Share-Alike	Re-users must credit the creator/copyright holder; any new material based on this work must be licensed under the same license
CC BY-NC	By Attribution, Non-Commercial	Re-users must credit the creator/copyright holder; the work cannot be used for commercial purposes
CC BY-NC-SA	By Attribution, Non-Commercial, Share-Alike	Re-users must credit the creator/copyright holder; the work cannot be used for commercial purposes; any new material based on this work must be licensed under the same license
CC BY-ND	By Attribution, No Derivatives	Re-users must credit the creator/copyright holder; no derivatives or adaptations of the work are allowed
CC BY–NC-ND	By Attribution, Non-Commercial, No Derivatives	Re-users must credit the creator/copyright holder; the work cannot be used for commercial purposes; no derivatives or adaptations of the work are allowed

CC Public Domain Dedication (CC0)

- Not actually a license: it’s a dedication of your work/data to the public domain
- Anyone can use the work/data and do what they want with it without crediting you in any way
- Unless you’re very altruistic, you probably want a different CC license type

More things to consider when depositing in Borealis

- Whether you can make the deposit yourself, or whether someone else needs to be involved. This will depend on your institution's policies
  - Mediated: you send the dataset to a staff member and they handle the deposit/publishing after review
  - Semi-mediated: you submit the dataset in Borealis and a staff member will review before publishing
  - Unmediated: you can submit and publish yourself
- What your timeline is, in cases of mediated/semi-mediated deposit
  - Important if you need deposit for e.g., peer review, thesis/dissertation submission, grant submission, etc.
  - Leave yourself enough time for review!

More things to consider when depositing in Borealis

- Whether there's a specific collection (aka 'Dataverse') that you're depositing in
  - We'll be using a particular collection today to keep all our deposits together
  - If you think you need your own collection, contact your institutional support
- Large files/datasets will take time to upload/publish
  - You can only upload 1000 files at a time
- .csv and spreadsheet-type files will be converted into a tabular data format
  - The original files will still be available
  - Conversion can take a bit of time - don't worry!

Hands-on: Depositing in the Borealis Demo site

After this session, you can find detailed guidance on how to deposit in the Borealis User Guide (https://learn.scholarsportal.info/all-guides/borealis/)

Creating an account: https://learn.scholarsportal.info/all-guides/borealis/user-accounts/

Depositing data: https://learn.scholarsportal.info/all-guides/borealis/datasets/

Questions?