**Why do we care about sharing data?**

**Why we share and resuse data?**

Some reasons to share research data are:

**Avoid unnecessary or costly experiments** by using previous research results.

**Validate research findings:** Independent verification of scientific results and conclusions (by replicating research workflows).

**Repurpose data:** Use the data for new research questions or in combination with other datasets. They are also extremely valuable as educational resources.

**Build upon previous work:** to accelerate scientific discovery and meta-analysis by avoiding duplication of efforts or reliance on irreproducible research.

**Tri-Agency Research Data Management Policy**

The Goverment of Canada promotes RDM in its Tri-Agency Research Data Management Policy.

Through its federal funding agencies, the Government of Canada seeks to implement data

management plans (DMPs) and sharing of research data to maximize the benefits to society.

 **Sharing data is a professional responsibility**

Depositing a dataset in a repository is NOT ONLY an exercise in meeting the requirements of funding agencies and journals. It is an ethical and professional responsibility of researchers to ensure reproducible science, and the access and reuse of scientific data.

**Therefore, research needs to move towards**

Competent researchers in RDM and data analysis.

Standardized approaches to sharing raw data and analysis code to support research findings.

Researchers with a commitment to transparency and best scientific practice practices to ensure research integrity.

**Benefits for different stakeholders**

Efficiency Collaborative work Reproducibility/impact

Rigorous peer review

Validation and

reproducibility

Open science

Transparency Accountability Return on investment

**Current issues with data**

**Data could be in many places**

**Data is not shared**

**Common issues in data repositories**

When shared, more often than not we observed the data:

 Lacks comprehensive metadata and readme file(s) explaining the context, methodology, and structure of the dataset.

 Presents a disorganized structure that makes its reuse impossible.

 Is treated only as a supplement of research articles.

**Principles of sharing data**

**Ensure your data is a valuable, standalone resource**

The following are essential aspects researchers must consider when sharing data:

 Your dataset should be a standalone resource.

 Your dataset should be discoverable and understandable.

 Your dataset must be reusable by the community.

**FAIR principles**

**General guidelines for dataset deposits**

 **General guidelines for data sharing**

1.  Provide a descriptive title, summary and keywords that reflect the content of the dataset.

2.  Define a dataset schema/road.

3.  Write a readme/metadata file.

4.  Organize data folders and scripts/codes folders.

**1. Provide a descritive title, summary and keywords**

The title must reflect the nature and content of the dataset.

Original: PiPaw2.0

Better: Home cage based motor learning platform PiPaw2.0

Original: Foliar Functional Trait Mapping

Better: Foliar Functional Trait Mapping of a mixed temperate forest using imaging spectroscopy

Original: Covariation in Width and Depth in Bedrock Rivers Data Archive

Better: Data archive for width and depth covariation within the bedrock Fraser Canyon, British Columbia, Canada

The title of your dataset IS NOT the same as the title of your research article

**Description (summary)**

The description must reflect the nature, content and methods of the dataset. The use of numerous keywords is recommended to increase its discoverability.

**Keywords**

To find relevant keywords, ask yourself the following question:

What terms can a reuser use in a search field to find my record?

**2. Define a dataset schema/road**

Define an organized scheme for your data at the beginning (best) or during your research (not bad).

 Folders/directory structures

 Think about file types/formats

 Establish logical/descriptive naming conventions

Overall, ensure that the schema is logical and consistent. An external user must be able to understand the directory structure.

**3. The guiding light of a dataset: the README**

The (main)  readme file is a guide to understanding the dataset and enabling its reuse or execution.

**Contents of a readme file**

In general, a dataset readme file shows:

 A dataset identifier showing aspects such as title, authors, date of collection, and geographical information.

 A map of files/folders defining the hierarchy of folders and subfolders and their contents. The user can also define explicit naming conventions.

 The methodological information presents the methods for data collection/generation, analysis, and experimental conditions.

 A set of instructions and software for opening, handling and reproducing research pipelines.

 Sharing and access information detailing permissions and terms of use.

**4. Organize dataset folders**

An organized scheme is the  key to understanding data structure.

 **Diving into the folder tree**

 **Organizing a data folder**

The data  must be organized logically and hierarchically according to the characteristics of each dataset.

 **Input data**

Sharing the input/raw data is a research integrity and data management best practice. The Data\_Input/

 can contain:

This Metadata/  contains information about the listed data files to ensure understanding and usability. It may list:

Guide to data sources: It describes how the data were generated or their provenance. This may include methodological details. and technical metadata.

Codebooks / data dictionaries: Explain the contents of files. (mainly .csv tables). They can be .txt or .csv-xlxs files.

The aim of these resources is to support the reuse of the data by providing a faithful and sufficient description of the variables.

 **Analysis data**

A Data\_Analysis/  contains the processed files, those used to generate the research results.

 **Intermediate data (Optional)**

A Data\_Intermediate/  can contain intermediate processed data, or pre-processed files as part of an analysis pipeline. For example, image ‘masks’ and machine learning classifiers that are used to further process images.

 **Scripting is the way**

Although most scientists may be more comfortable with GUIs, the current research landscape requires the use of scripts and (analysis) code to ensure the reproducibility of research results.

 Coding should be considered an essential skill, as well as other methods such as animal surgery, patch clamp, or flow cytometry.

 **Processing scripts**

A Scripts\_Processing  may contain scripts/code that prepare (or transform) the raw data (images, tables) for analysis Data\_Analysis/ .

Examples of workflows:

Drop variables (subset the dataset)

Generate new variables (Perform computations, calculate averages, etc.)

Combine different sources of information (merge tables or files)

You may want to consider saving the generated intermediate files in the Data\_Intermediate/ folder

 **Keep in mind**

You will create several processing scripts. Logical naming conventions are the key to linking the input/output data to the processing scripts.

 **Analysis scripts**

In general, these scripts import and process the analysis data.

 **A master script?**

The Scripts/  can also contain a master script that executes all other scripts, creating a fully automated pipeline.

 **The output folder**

 **Commitment to reproducibility**

Sharing the output resulting from computations/code is one of the best commitments to open and reproducible science. It is also a way to preserve material for future use in an organized way.

**Data submission checklist**

 **Submitting your data to a repository**

When you submit your data to a repository (FRDR), make sure it meets these characteristics:

1.  Your folders and files are organized in a clear and structured way (understandable to the community): Use standardized file formats (e.g., CSV, TIFF) and check for consistency in naming conventions.

2.  The metadata/readme is as complete as possible and can be understood as a standalone object that provides data collection methods, processing steps, and relevant context.

3.  Verify independent usability: Data must be complete and understandable (including any necessary instructions for data interpretation) without the need for the accompanying research article.

**FAQ**

 **When do I start organizing my data for sharing?**

We recommend implementing RDM practices early and throughout the research process. Organizing data after years of chaotic data management is not a good idea.

 **When do I share my data?**

Your data can be shared at any time during the research process. You do not have to wait until a research article is published to share your data.

 **What if my dataset does not fit into protocols such as TIER 4?**

You do not need to worry about this. The most important thing is that your dataset is well documented, logically organized, and has naming conventions that make it understandable to potential reusers.

**FAQ**

 **Is my data citable?**

Of course it is. Your dataset gets a DOI, which makes it a citable object independent of your research article. In fact, if you publish your dataset before your article, you can even cite your datasets in your research.

 **How can others use my dataset?**

That depends on the license you use. We recommend a CC-BY 4.0 license, which allows broad reuse of the data.

 **Where do I share my data**

You can share your data in specialized or generalist repositores like The Federated Research Data Repository (FRDR) or Borealis.

**In summary**

Be aware that the dataset is a research object that serves the public and the scientific community, and that can be used (and cited) independently of the research article.

Better yet, think of articles as supplements to your dataset!!!

**Canadian generalist repositories**

**The Federated Research Data Repository (FRDR)**

The Federated Research Data Repository (FRDR) is a national platform for Canadian researchers to discover, store, and share research data.

**Our goals**:

 Improve data discoverability (in partnership with Lunaris).  Promote open science practices and the reuse of research data.  Ensure the long-term preservation of valuable research data.

FRDR supports a wide range of disciplines and data types, providing a robust infrastructure for management and dissemination of research data across Canada.

**Benefits of using FRDR**

 FRDR ensures the long-term preservation, accessibility and usability of datasets through its curation and preservation team.

 FRDR supports funding agencies requirements related to open access to data (and research data management plans).

 Promotes dataset visibility and reuse across a wide range of disciplines.

 FRDR supports large datasets, making it an ideal repository for data-intensive research.

 FRDR supports researchers in data management best practices.

FRDR has competent staff to guide researchers and institutions to ensure that datasets are valuable and comply with FAIR principles.

 **Datasets as standalone, reusable objects**

At FRDR, we aim for datasets to be standalone objects (independent of research articles) with potential social, research or educational uses.

**Borealis**

Borealis is a Canadian research data repository supported by academic libraries, research institutions, and the Digital Research Alliance of Canada.

**Features**:

 Built on Dataverse open-source software hosted by Scholars Portal / University of Toronto Libraries.

 Integrated with single sign-on login for Canadian Institutions (Canadian Access Federation).

 Indexed in DataCite search, Google dataset search, Lunaris for discoverability.

**Borealist network in Canada**

**Borealis collections**

**Borealis tools**

**Resources and support**

FRDR documentation

Training resources from the Alliance

Contact us to ensure that your data is well prepared and can be effectively shared with the research

community.

Email: rdm-gdr@alliancecan.ca

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

**Depositing research data**

A primer for scientists

Daniel Manrique-Castano, Ph.D Digital Research Alliance of Canada

Thursday, January 2, 2025

**Data availability statement**

“The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.”

Researchers do not share data

**Datasets as standalone objects**

 Findable

 Persistent identifiers

 Rich metadata

 Indexed in a searchable resource

 Accessible

 Open file formats

 Software requirements

 Interoperable

 Formal, standardized, common language

 Reference to other (meta)data

 Reusable

 Appropriate context and detailed

provenance

 Accurate/descriptive attributes

 Clear license and usage rights

**Think about**

FRDR users can use our [text] or [web] template to generate a readme file for submission to FRDR.

Additional resources are:

- Creating a README file

- Readme.so

- Readme.ai

**To refresh your memory**

 Plan/define directory structures, file formats, and naming conventions.

For example, TIER 4.0 is a systemic template to standardize and increase transparency/reproducibility of research data. The user can download a folder structure and adapt it to specific cases.

**Tip**

Folder tree

Original images (.tiff, .czi)

Measuring device output files (.txt, .csv)

Original registration datasheets (.png, .csv, .xlxs)

**a) Data files (stored in subfolders if necessary)**

Folder tree

**b) A metadata file/folder** 

Like the input data, these files contain a codebook/data dictionary. Also, these files can be

accompanied by a Data\_Appendix files that showcase basic descriptive statistics or show

data distributions.

* Folder tree

The Scripts\_Analisys  folder hosts scripts/code to generate results that may be in the form of:

 Images

 Figures

 Tables

 Statistical models

Folder tree

**Tip**

The Output/  contains subfolders storing the files generated by the analysis scripts in the form

of:

 Images

 Figures

 Tables

 Statistical models

Folder tree

**FRDR is for Canadian researchers**

**FRDR supports researchers and institutions**

Borealis network in Canada

Each institution or group has a top-level collection.

Datasets are deposited into collections or sub-collections.

Some institutions support researchers with own subcollections.

Borealis datasets are organized in collections

 File preview to explore files directly in the browser.

 Data explorer tool to visualize variables in tabular data files

(e.g., SPSS, Excel, CSV). Chart

 Github integration using GitHub actions.

Borealis table viewer

Visit FRDR or Borealis