# Putting your Plan into Practice

## Data management plans, real-life examples, and our workflow for the week

Jennifer Abel, University of Calgary
RDM Jumpstart
May 12, 2025

# Data Management Plans

# What is a Data Management Plan? (DMP)

- A formal (i.e., written) document that details the strategies and tools you'll use to effectively manage data during the active phase of your research.
- It also documents the mechanisms you'll use for preserving and appropriately sharing data at the end of the project.
- Your DMP should be a 'living' document that's modified as changes occur during the course of your project.

# Why use a DMP?

- To organize all those things we talked about in the previous session!
- To demonstrate your RDM practices to various interested parties (e.g., funders, ethics boards, partners and collaborators, supervisors)
- To be an element of documentation and a tool for project memory
- To help support transparency, reproducibility, openness, and FAIRness in your research

# Things to consider when you're creating a DMP

- Who's your audience?
  - E.g., funders will want to see high-level information, while your supervisor/lab colleagues might want to have more detailed information about specific workflows/procedures
- Do you have page limits?
  - E.g., most funders don't want something longer than 3 pages (sometimes 2)
- Do you need to use a specific template?
  - This might depend on your institution/funding opportunity

# How to create a DMP

There are online tools that you can use

- DMP Assistant: https://dmp-pgd.ca/
  - Free, Canadian, bilingual, hosted on Canadian servers, supported by Canadian RDM professionals
  - Has a range of templates you can use; to see examples, visit https://dmp-pgd.ca/public_plans
- DMP Tool (US-focused): https://dmptool.org/
- DMPonline (UK-focused): https://dmponline.dcc.ac.uk/

Or you can create your own document, if there isn't a particular template or tool you need to use

# Our RDM Jumpstart DMP

- The basics of what you'd want in a DMP
- Some elements that are more for external audiences (e.g., funders) than internal ones (e.g., other research project members)
    - Particularly the benefits of using Borealis for depositing data
- If this was an internal-facing DMP, we would probably have more information on specifics of procedures/workflows
    - E.g., the file naming convention, who's responsible for doing what

# Exploring the DMP and Our Project

# Exploring the DMP and our project

What elements that we talked about in the last session are included in each of the four paragraphs?

Reminder: those elements are

- Ethical, legal and commercial issues
- Data collection
- Data documentation
- Storing, accessing and working with data
- Specifics of procedures/workflows
- Long-term data management, discoverability and access

# Exploring the DMP and our project

- Paragraph 1:
  - Data collection; Ethical, legal and commercial issues
- Paragraph 2:
  - Data documentation; Storing, accessing and working with data; Specifics of procedures/workflows (very briefly)
- Paragraph 3:
  - Storing, accessing and working with data
- Paragraph 4:
  - Long-term data management, discoverability and access; Ethical, legal and commercial issues

# Data collection: The dataset we're using

General Social Survey, Cycle 29, 2015 [Canada]: Time Use, Main File: https://odesi.ca/en/details?id=/odesi/doi__10-5683_SP3_RDS0CK.xml

"The General Social Survey (GSS) gathers data on social trends in order to monitor changes in the living conditions and well-being of Canadians over time, and to provide immediate information on specific social policy issues of current or emerging interest. This survey monitors changes in time use to better understand how Canadians spend and manage their time and what contributes to their well-being and stress."

# Ethical, legal and commercial issues: Licensing, risk

Licensing:
The dataset that we're using is licensed under the Statistics Canada Open License (https://www.statcan.gc.ca/en/reference/licence)

- Certain things we can do, like "use, reproduce, publish, freely distribute, or sell" the data or products we generate from it
- Certain things we can't do, like use the data for illegal purposes

Risk:
The risk level is low: this is publicly available data that Statistics Canada has deemed safe to make openly available.

# Sidebar: Some notes on the dataset

1. The dataset is very large: 17000+ observations and 848 variables
2. The documentation around it is somewhat complex and not as FAIR as it could be
   - The main data dictionary file is a 186-page-long PDF
   - It's sometimes hard to find information about basic things like units of measurement; e.g. whether duration is measured in minutes or hours
3. Column names are not transparent: e.g., can you guess what PHSDFLG means? Or DURS200?
4. Missing values are coded in multiple different ways depending on the variable

# We'll work with something a little simpler

1. We'll be looking at a subset of 29 variables
2. We've created a new data dictionary in a non-proprietary format that makes essential information easier to find
3. We'll have an exercise on re-coding the column names to make them more transparent
4. We've re-coded all the missing values consistently

# Back to the DMP: Data documentation

We have two pieces of documentation:

- Data dictionary (as discussed just now)
- Readme

These are in non-proprietary formats, to maximize FAIRness

Let’s have a look at them!

# Storing, accessing and working with data

Discussion of

- The software/platforms we're using (OSF, R)
- Where files are being stored (in OSF on servers in Canada; on personal/institutional devices)
- How often files are being backed up (at the end of the day)
- How much storage will be needed (this is an overestimate, but better than an underestimate)
- Who has access to the files

# Long-term data management, discoverability & access

Discussion of

- What we'll be depositing (data, documentation, code)
- Where we're depositing our data (Borealis demo site)
- Why we've chosen that repository
- What formats the files will be in (non-proprietary)
- What license we'll be applying to our data (Creative Commons 4.0 By Attribution; to be discussed more on Day 5, but see https://creativecommons.org/ for a preview)

# Questions?

# Reminders!

Please install R **and** RStudio before tomorrow (if you haven't already)