Workflow for the Week

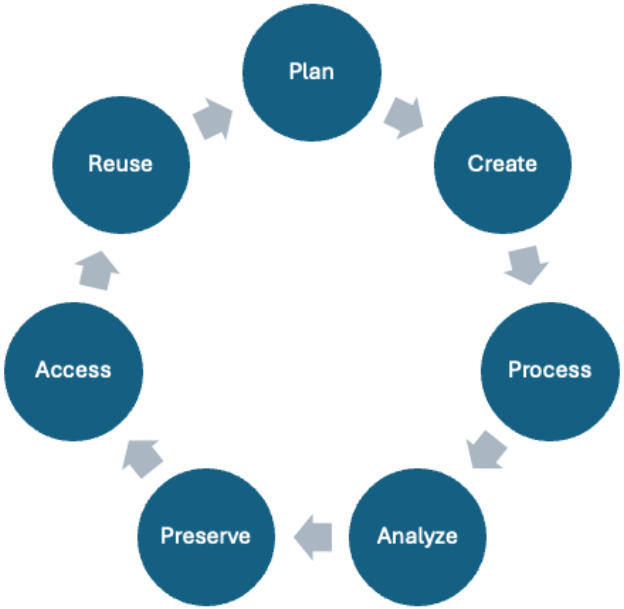
Nick Rochlin, University of Victoria
RDM Jumpstart
May 12, 2025

Key Tools & Concepts

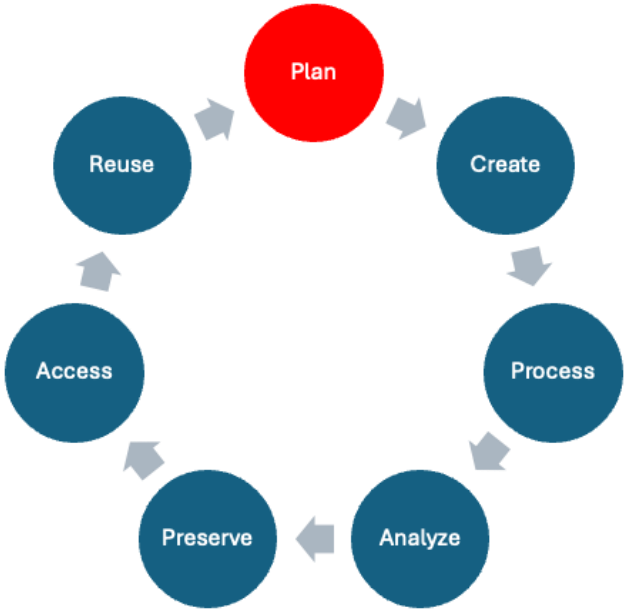
Overview

- Before collecting data
  - Data management planning
  - Open Science Framework (OSF)
- After collecting data
  - OSF
  - R, RStudio, & Rmarkdown
- At project completion
  - Borealis/Dataverse

The Research Data Lifecycle



The Research Data Lifecycle

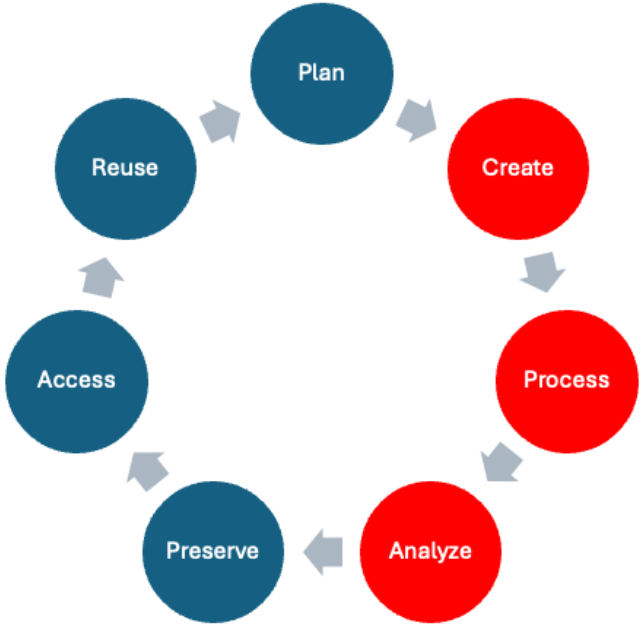


Data Management Plans – Covered in day 1

- Questions and considerations about your research data prior to starting a project.
- Will be covered in more detail this afternoon!

The Data Science Workflow

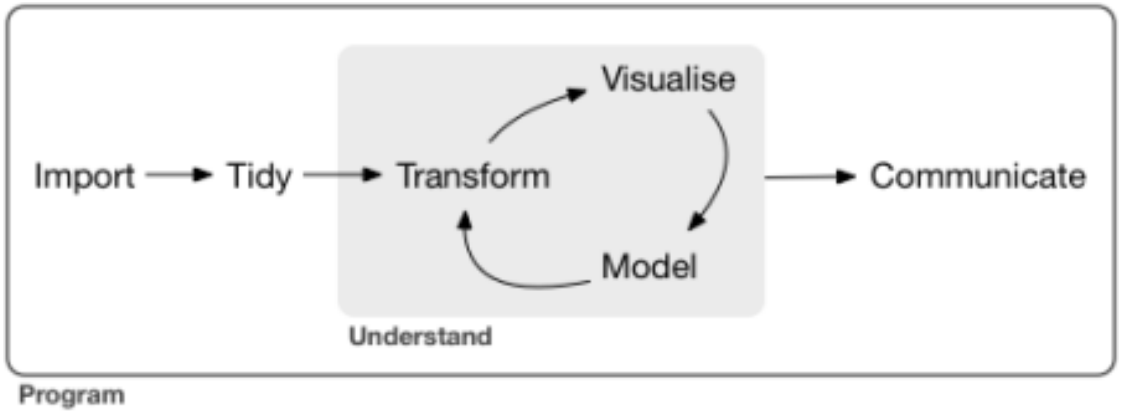
The Data Science Workflow



The Data Science Workflow

- A framework to support transparent and open research processes.
- A granular roadmap to support the “active” stages of the research data lifecycle.
- While the research data lifecycle is focused on making research processes and outputs transparent and reproducible, the data science workflow focuses on manipulating and asking questions from data in an open and reproducible way.

The Data Science Workflow



source

R

- An open-source programming language designed for data manipulation, statistical analysis, and visualization.

RStudio

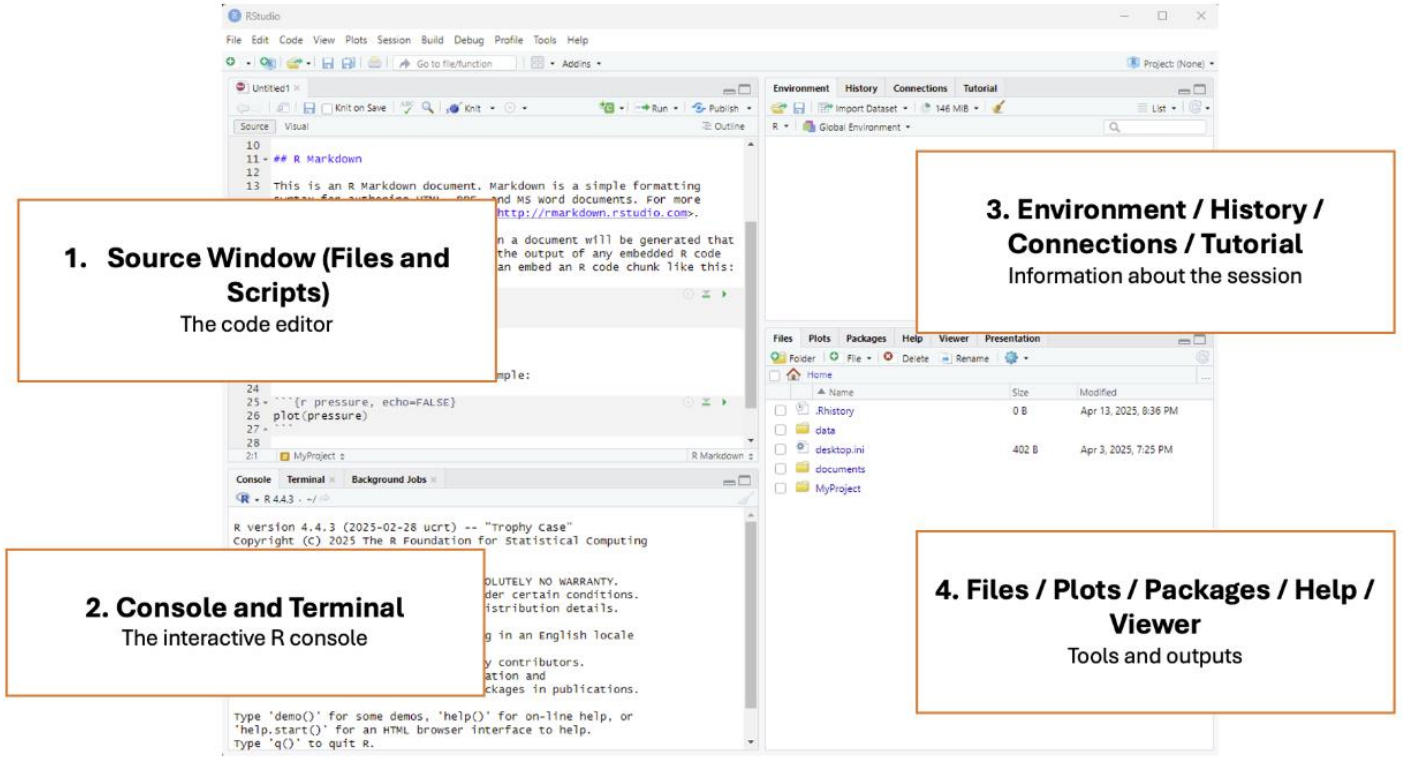
- While R is the programming language, RStudio is an Integrated Development Environment (IDE), which is a fancy term for a software application that makes writing scripts R much more user-friendly.

RStudio

- A analogy:
  - You are writing a document in Microsoft Word. In this case, the plain human language that you are typing would be R, and Microsoft Word, which is the software that allows you to view, format, add visuals, and save your work, is RStudio.

- In the morning of day 2 we'll do a session on using Excel vs. a coding language like R, and then we'll jump into R and RStudio in the afternoon.

RStudio

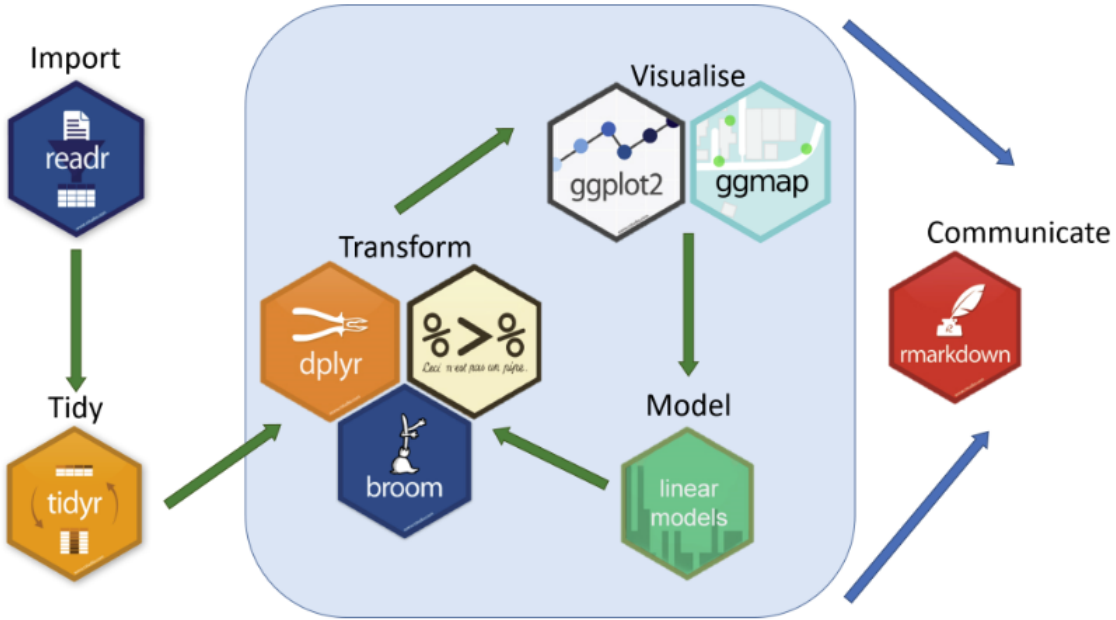

1. Source Window (Files and Scripts)
The code editor
2. Console and Terminal
The interactive R console
3. Environment / History / Connections / Tutorial
Information about the session
4. Files / Plots / Packages / Help / Viewer
Tools and outputs

Across the Tidyverse

The Tidyverse

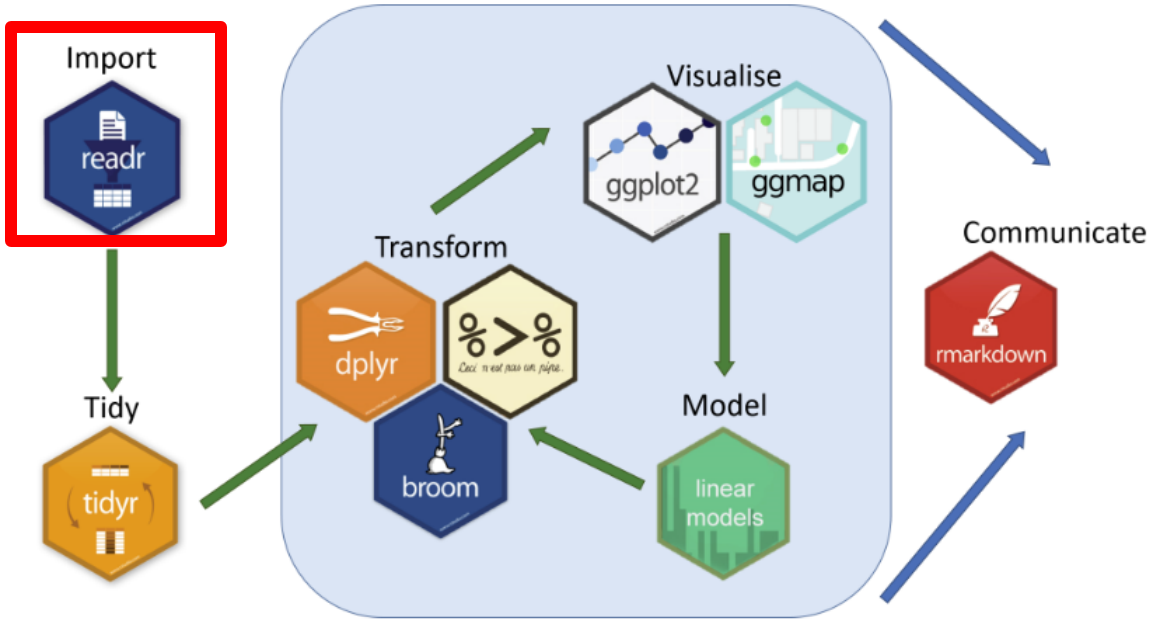
- A group of R “packages” for data manipulation, exploration, and visualization.
- Focused on the connections between the activities of reproducible workflows.
- Provides a good reference point for those new to coding to engage with coding.

The Data Science Workflow with Tidyverse



source

Import – Covered in day 2

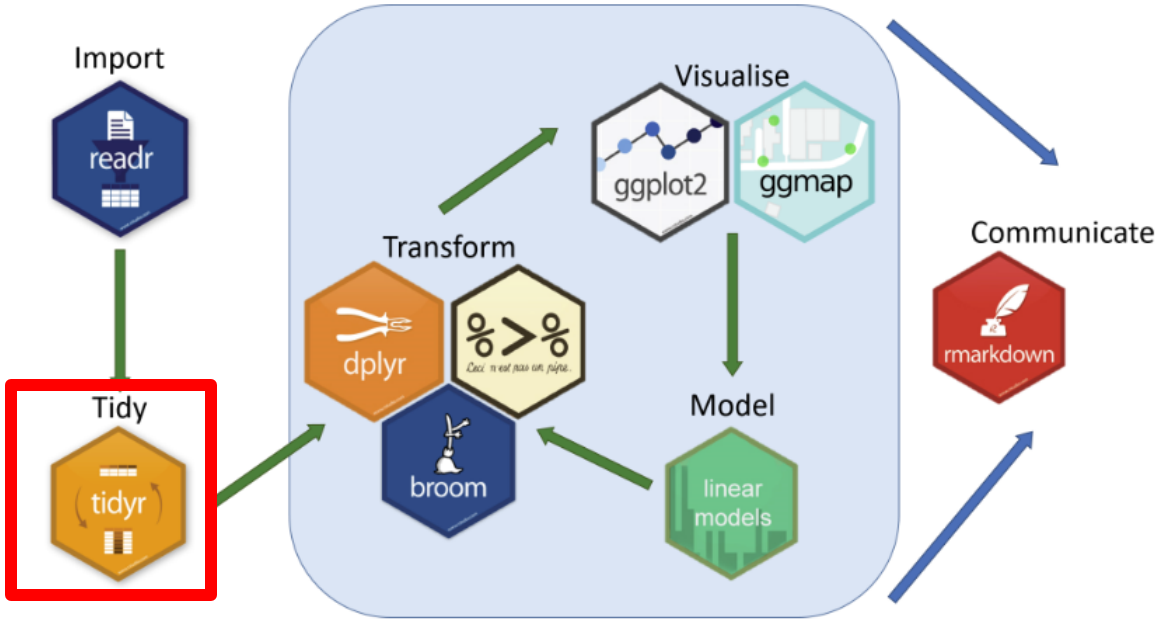


source

Import – Covered in day 2

- Unlike Excel, which can immediately start working with spreadsheet data, for R to be able to work with data, the data first needs to be brought in, or imported, to the coding environment.

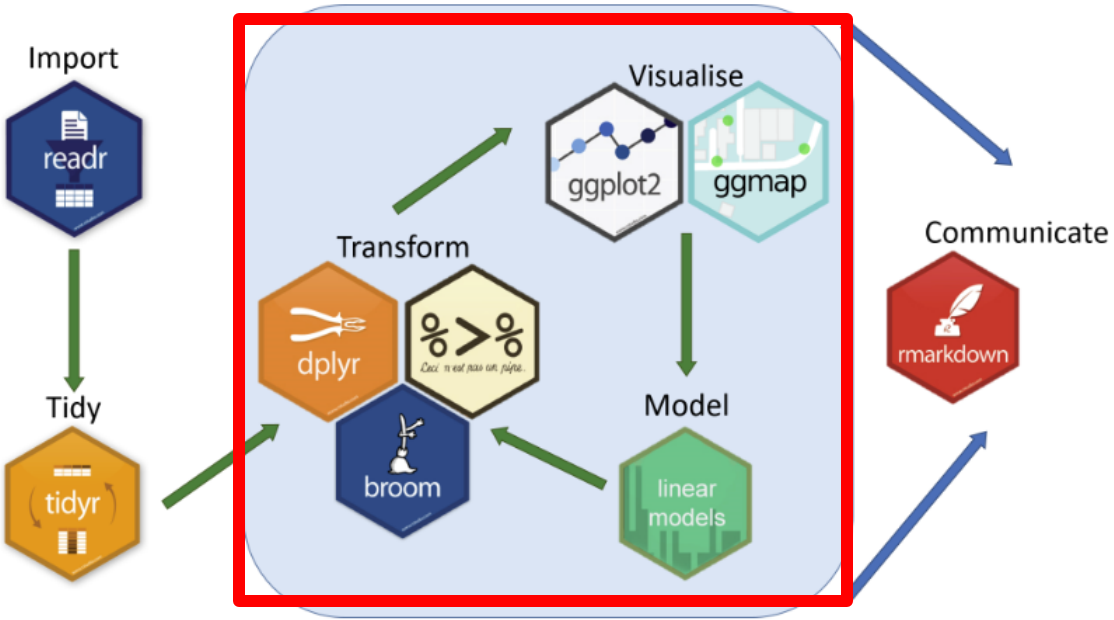
Tidy – Covered in day 3



source

Tidy – Covered in day 3

- Most of the data we collect or retrieve is quite messy, and needs to be cleaned up, or tidied, before we can start asking questions from it.
- This can include things like:
  - Changing the names of columns
  - Accounting for missing values / missing data
  - Re-formatting data
  - ...


Import
readr
Tidy
tidyr
Transform
dplyr
%>%
broom
Visualise
ggplot2
ggmap
Model
linear models
Communicate
rmarkdown

source

Transform – Covered in day 3

- Once data has been tidied, we can start playing around!
- Data transformations, such as:
  - Filtering rows and columns that meet certain criteria
  - Creating new columns based on the values of existing columns
  - Subsetting specific portions of the data
- Transforming data can be a good way to set up data visualizations or modeling, but can also be conducted after visualization and modeling if you notice something interesting in the data.

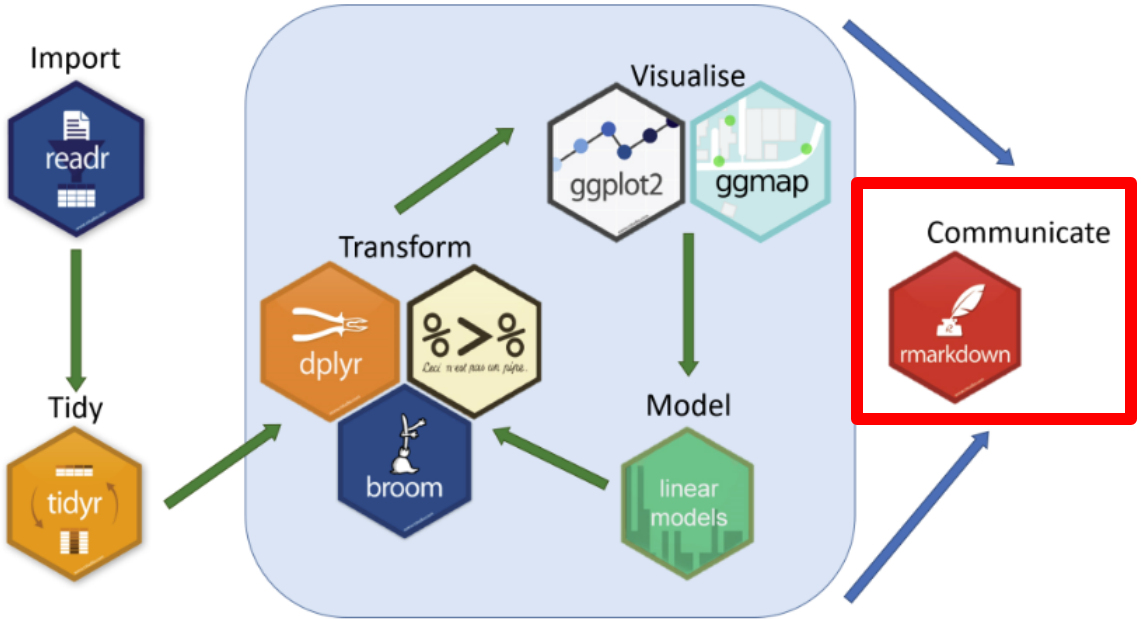
Visualize – Covered in day 4

- Excel can be pretty good at creating visualizations quickly, but it can also be frustrating due to rigid structures and it not doing what you want.
- Visualizations in R require a bit more upfront work, but have much more flexibility and power in what you can create.
- Visualizations can be used to communicate your findings, but can also be a valuable tool in exploring data and discovering interesting trends.

Model – Out of scope

- Applying statistical models to data is the real “science” of data science, but is a bit beyond the scope of this series.
- However, it’s worth mentioning here as a key part of the data science workflow, and we encourage you to use the skills developed throughout the week as a foundation to start entering this area.
- It’s also possible that we may build out a more advanced level of the series, and if that’s of interest, please let us know!

Communicate – Covered in day 2-5



source

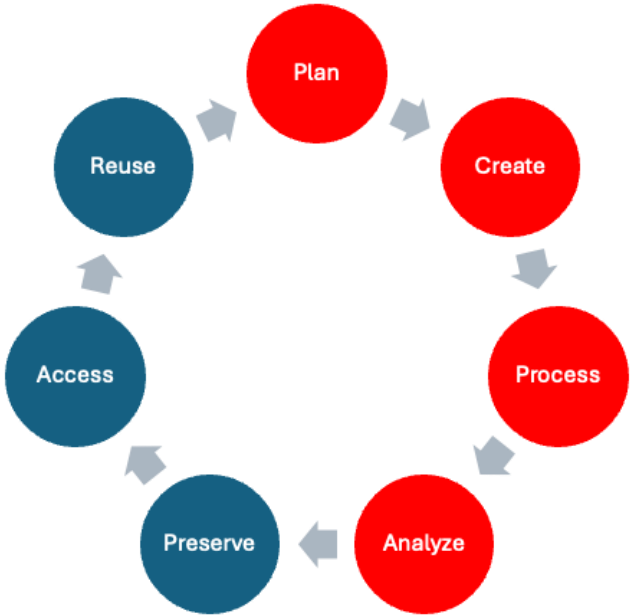
Communicate – Covered in day 2-5

- Communicating your findings is the most fulfilling part of the data science workflow (it's finally over!), but to do this in an open and transparent way, it needs to be considered throughout the entire research life cycle.

Communicate – Covered in day 2-5

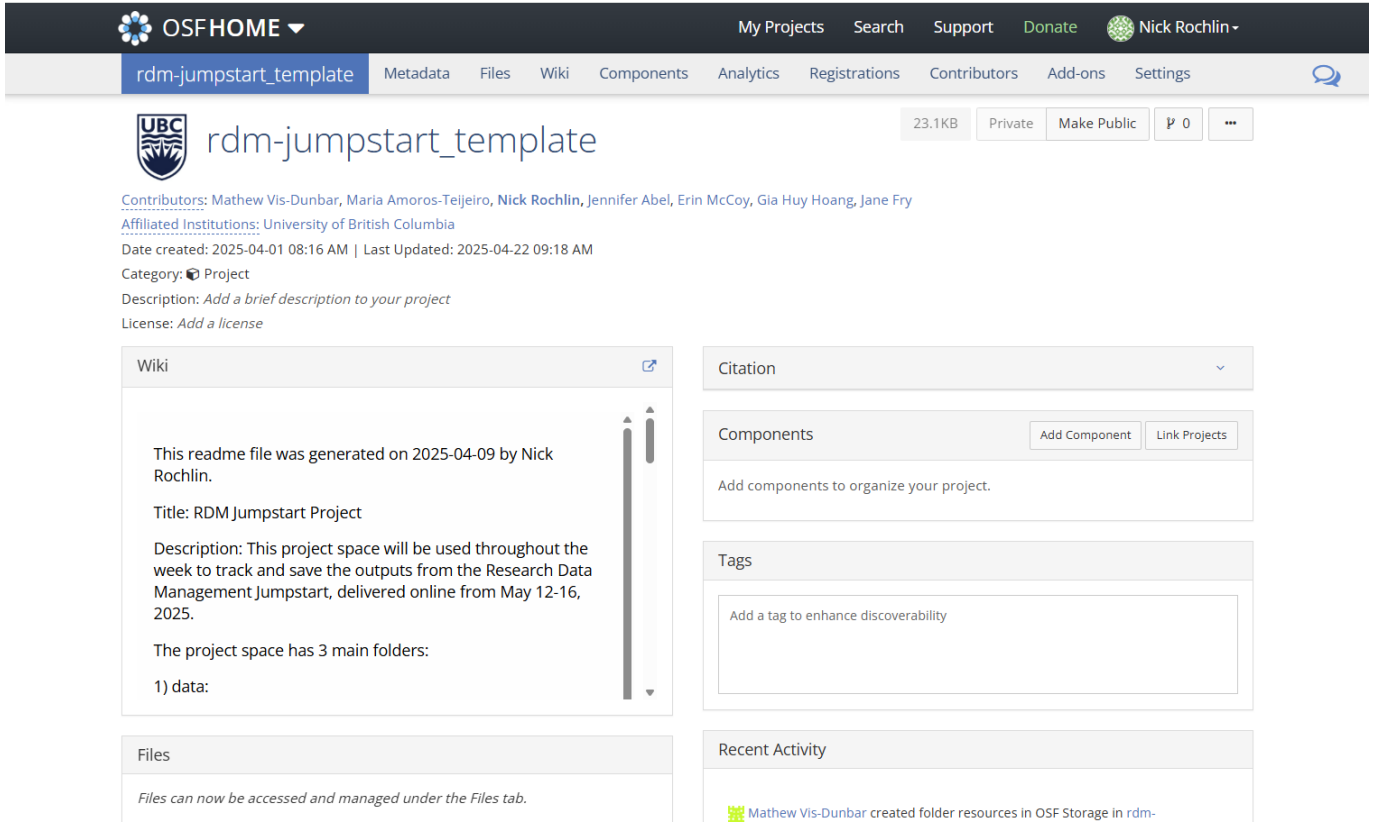
- **Rmarkdown** is a plain text format that combines R code, results, and explanatory text into a single, shareable, reproducible document.
- Allows you to create various file formats like PDF, Word, HTML, and presentations.

Storage & Backup

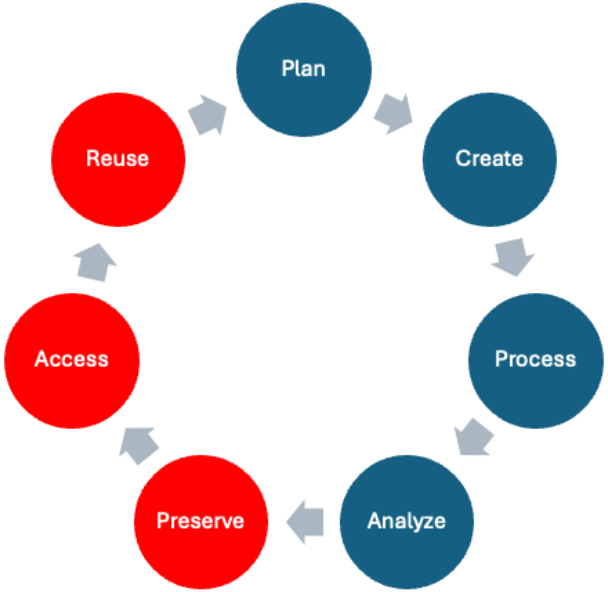


Open Science Framework (OSF) – Whole week

- Cloud-based platform designed to promote transparency, collaboration, and reproducibility in research.
- For the purpose of the Jumpstart, it will be used to:
  - Version control documents
  - Backup and share working documents
- OSF will be introduced in further detail this afternoon and used throughout the week.

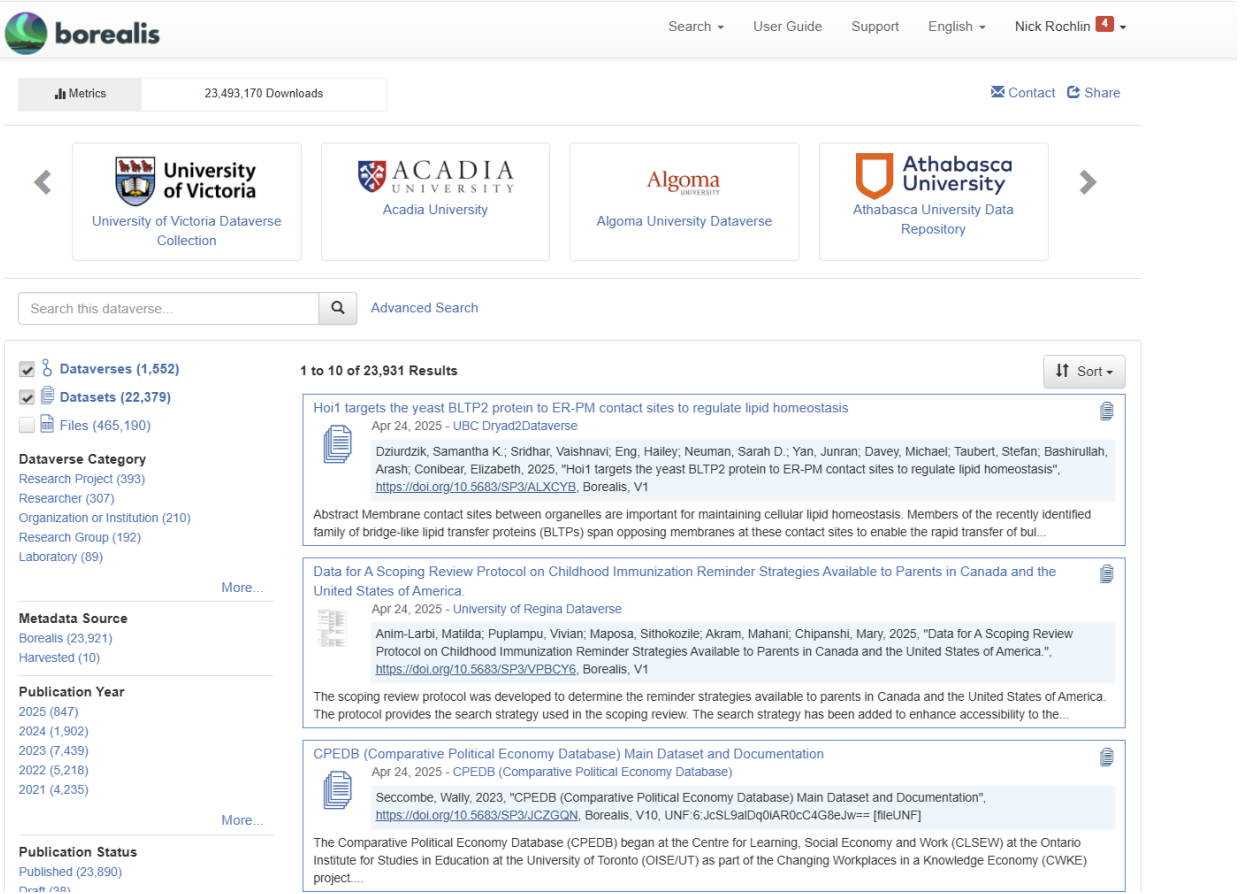

OSF HOME
My Projects
Search
Support
Donate
Nick Rochlin
rdm-jumpstart_template
Metadata
Files
Wiki
Components
Analytics
Registrations
Contributors
Add-ons
Settings
rdm-jumpstart_template
23.1KB
Private
Make Public
0
Contributors: Mathew Vis-Dunbar, Maria Amoros-Teijeiro, Nick Rochlin, Jennifer Abel, Erin McCoy, Gia Huy Hoang, Jane Fry
Affiliated Institutions: University of British Columbia
Date created: 2025-04-01 08:16 AM | Last Updated: 2025-04-22 09:18 AM
Category: Project
Description: Add a brief description to your project
License: Add a license
Wiki
This readme file was generated on 2025-04-09 by Nick Rochlin.
Title: RDM Jumpstart Project
Description: This project space will be used throughout the week to track and save the outputs from the Research Data Management Jumpstart, delivered online from May 12-16, 2025.
The project space has 3 main folders:
1) data:
Citation
Components
Add Component
Link Projects
Add components to organize your project.
Tags
Add a tag to enhance discoverability
Files
Files can now be accessed and managed under the Files tab.
Recent Activity
Mathew Vis-Dunbar created folder resources in OSF Storage in rdm-

Data Preservation & Sharing



Borealis / Dataverse – Covered in day 5

- A Canadian data repository supported by academic libraries and institutions across Canada.
- Facilitates the sharing, preservation, and discovery of Canadian research data.
- Throughout the week we'll be discussing considerations for sharing and preserving data, and on day 5 we'll cover Borealis and will deposit materials from the week into the repository.


borealis
Search
User Guide
Support
English
Nick Rochlin 4
Metrics
23,493,170 Downloads
Contact
Share
University of Victoria
University of Victoria Dataverse Collection
ACADIA UNIVERSITY
Acadia University
Algoma
Algoma University Dataverse
Athabasca University
Athabasca University Data Repository
Search this dataverse...
Advanced Search
Dataverses (1,552)
Datasets (22,379)
Files (465,190)
Dataverse Category
Research Project (393)
Researcher (307)
Organization or Institution (210)
Research Group (192)
Laboratory (89)
More...
Metadata Source
Borealis (23,921)
Harvested (10)
Publication Year
2025 (847)
2024 (1,902)
2023 (7,439)
2022 (5,218)
2021 (4,235)
More...
Publication Status
Published (23,890)
1 to 10 of 23,931 Results
Sort
Hoi1 targets the yeast BLTP2 protein to ER-PM contact sites to regulate lipid homeostasis
Apr 24, 2025 - UBC Dryad2Dataverse
Dziurdzik, Samantha K.; Sridhar, Vaishnavi; Eng, Hailey; Neuman, Sarah D.; Yan, Junran; Davey, Michael; Taubert, Stefan; Bashirullah, Arash; Conibear, Elizabeth, 2025, "Hoi1 targets the yeast BLTP2 protein to ER-PM contact sites to regulate lipid homeostasis", https://doi.org/10.5683/SP3/ALXCYB, Borealis, V1
Abstract Membrane contact sites between organelles are important for maintaining cellular lipid homeostasis. Members of the recently identified family of bridge-like lipid transfer proteins (BLTPs) span opposing membranes at these contact sites to enable the rapid transfer of bul...
Data for A Scoping Review Protocol on Childhood Immunization Reminder Strategies Available to Parents in Canada and the United States of America.
Apr 24, 2025 - University of Regina Dataverse
Anim-Larbi, Matilda; Puplampu, Vivian; Maposa, Sithokozile; Akram, Mahani; Chipanshi, Mary, 2025, "Data for A Scoping Review Protocol on Childhood Immunization Reminder Strategies Available to Parents in Canada and the United States of America.", https://doi.org/10.5683/SP3/VPBCY6, Borealis, V1
The scoping review protocol was developed to determine the reminder strategies available to parents in Canada and the United States of America. The protocol provides the search strategy used in the scoping review. The search strategy has been added to enhance accessibility to the...
CPEDB (Comparative Political Economy Database) Main Dataset and Documentation
Apr 24, 2025 - CPEDB (Comparative Political Economy Database)
Seccombe, Wally, 2023, "CPEDB (Comparative Political Economy Database) Main Dataset and Documentation", https://doi.org/10.5683/SP3/JCZGQN, Borealis, V10, UNF:6:JcSL9alDq0iAR0cC4G8eJw== [fileUNF]
The Comparative Political Economy Database (CPEDB) began at the Centre for Learning, Social Economy and Work (CLSEW) at the Ontario Institute for Studies in Education at the University of Toronto (OISE/UT) as part of the Changing Workplaces in a Knowledge Economy (CWKE) project....

Questions?