Florida AFS 2023 Open Science Workshop

4/11/23

Table of contents

Course synopsis

Welcome to the 2023 Florida AFS open science workshop. Open science (OS) has been advocated as an effective approach to create reproducible, transparent, and actionable research products. However, widespread adoption among the research and management community has not occurred despite its perceived benefits. In the face of major challenges like global warming and sea level rise, the collaborative framework provided by OS is needed now more than ever. This workshop will cover material introducing participants to core concepts of OS. The target audience includes anyone interested in applying OS in their own workflows as part of a larger research and resource management team.

By the end of this workshop, you should have a good understanding of fundamental concepts in open science and how they can be applied to help bridge the research-management divide.

You will also have the skills to understand how collaborative open science tools can be used to increase efficiency and transparency, understand fundamental best practices for working with data to facilitate openness, and be able to apply these lessons within your own teams by effectively addressing barriers to adoption.

Much of the content on this web page was adopted from the TBEP Data Management SOP.

Prepare

Please attend the workshop with a personal laptop and power supply. Make sure your laptop can access publicly available WiFi. *You will also need to install software prior to the workshop, visit the setup page for full instructions.* We will have limited capacity to help with installation issues the day of the workshop, so please come prepared. The setup instructions will guide you through the following.

1. Install R: link
2. Install RStudio: link
3. Install Quarto: link
4. GitHub create account: link
5. Install Git: link

We also assume some knowledge about R. Please visit this page for a crash course if you need to brush up on your R skills.

Agenda

1. The basics of open science: 12:30 - 1:00
2. Open science for collaboration: 1:00 - 2:00
3. Open science for impactful products: 2:15pm - 3:30pm

Each module uses a set of common icons to orient you to specific tasks or experiences during this workshop. These include the following:

Exercise and discussion

Watch and learn

Description of a collaborative tool

Pros of a collaborative tool or solution to an open science challenge

Cons of a collaborative tool

Challenge to overcome for open science

Instructor

Dr. Marcus Beck is the Program Scientist for the Tampa Bay Estuary Program and is developing data analysis and visualization methods for Bay health indicators. He received his PhD in Conservation Biology from the University of Minnesota in 2013. Marcus has experience researching environmental indicators and developing open science products to support environmental decision-making. Marcus is also an open source software and dashboard developer to facilitate science application. CV, Google Scholar, GitHub

---

This version of the website was built automatically with GitHub Actions on 2023-04-11.

Part I

Modules

1 Basics of open science

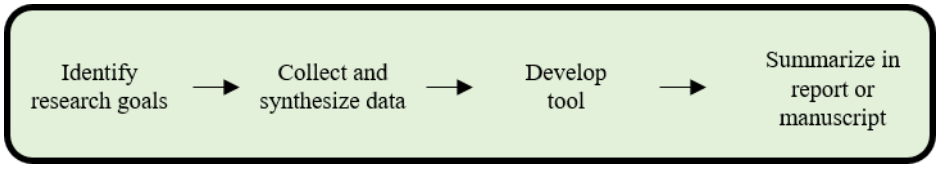
1.1 Goals and motivation

This is the first module in our workshop on open science. This module describes the need for open science, how it can improve research applications, and exposes you to common ideas and terminology that we'll be using throughout the day. Consider this your 30,000 foot view of open science. Our later modules will provide more detail on specific topics in open science that you can use for continued learning.

- **Goal**: get comfortable with key ideas and concepts for understanding open science
- **Motivation**: This is the first step in your open science journey!

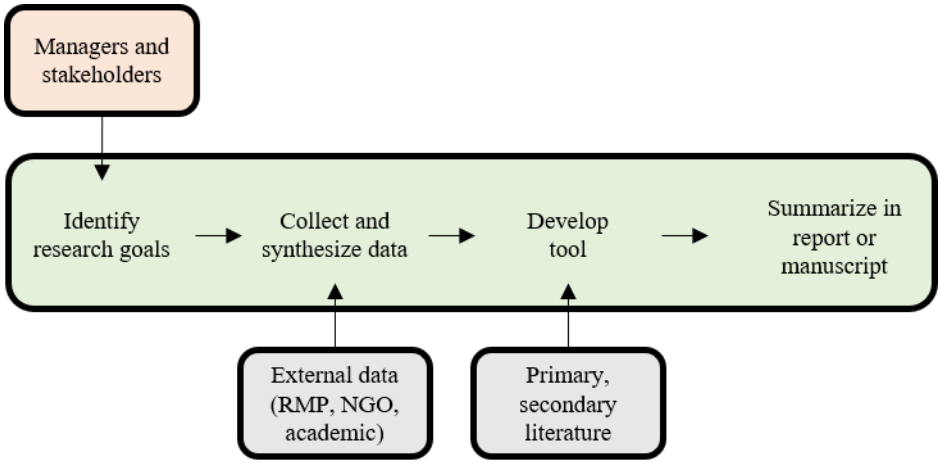
1.2 Why open science?

Let's start with revisiting the scientific process. I'm sure this looks familiar to all of you. This is geared towards an applied research question.



Our basic scientific approach to discovery is motivated by a question or research goal, developing a hypothesis for the question, collecting data based on the hypothesis, developing a tool that can be used for decision-making, and summarizing the results in a conventional format.

In a little more detail, your workflow may look something like this.



Many scientists, especially early career researchers (my past self included), may assume that this is sufficient to affect change. We write the report, send it out into the world, and move on to the next project. This is a common mentality:

> "This 500-page report will answer all of their questions!"

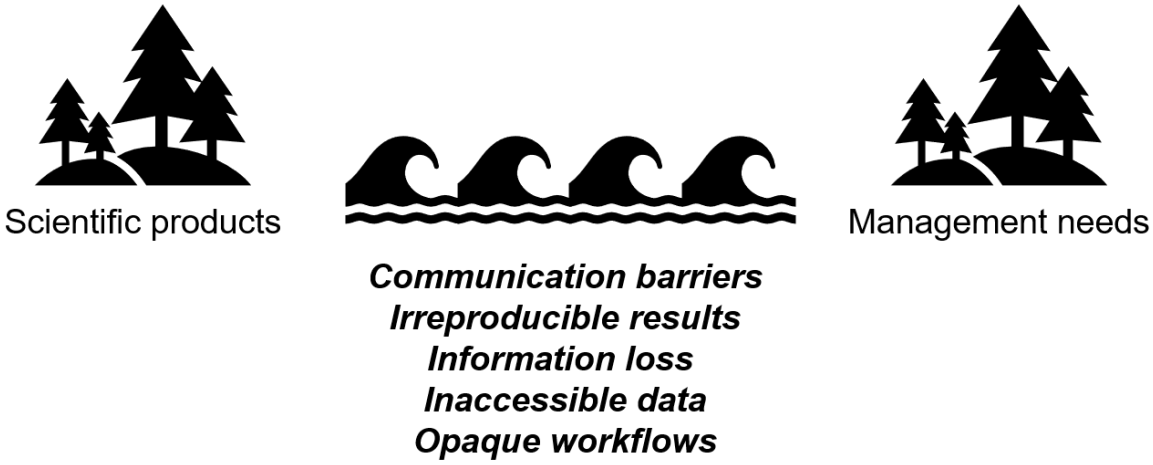
From the other side, such as the manager or policy-maker, the report may be received like this:

> "This 500-page report answers none of my questions!"

It's dense, inaccessible, and there are probably questions about the underlying data and methods used to achieve the results. More importantly, it doesn't present the information in an easily digestible format to quickly make the right decision. Sometimes, if you think you're doing applied science, it may just be *implied* science that falls short of application.

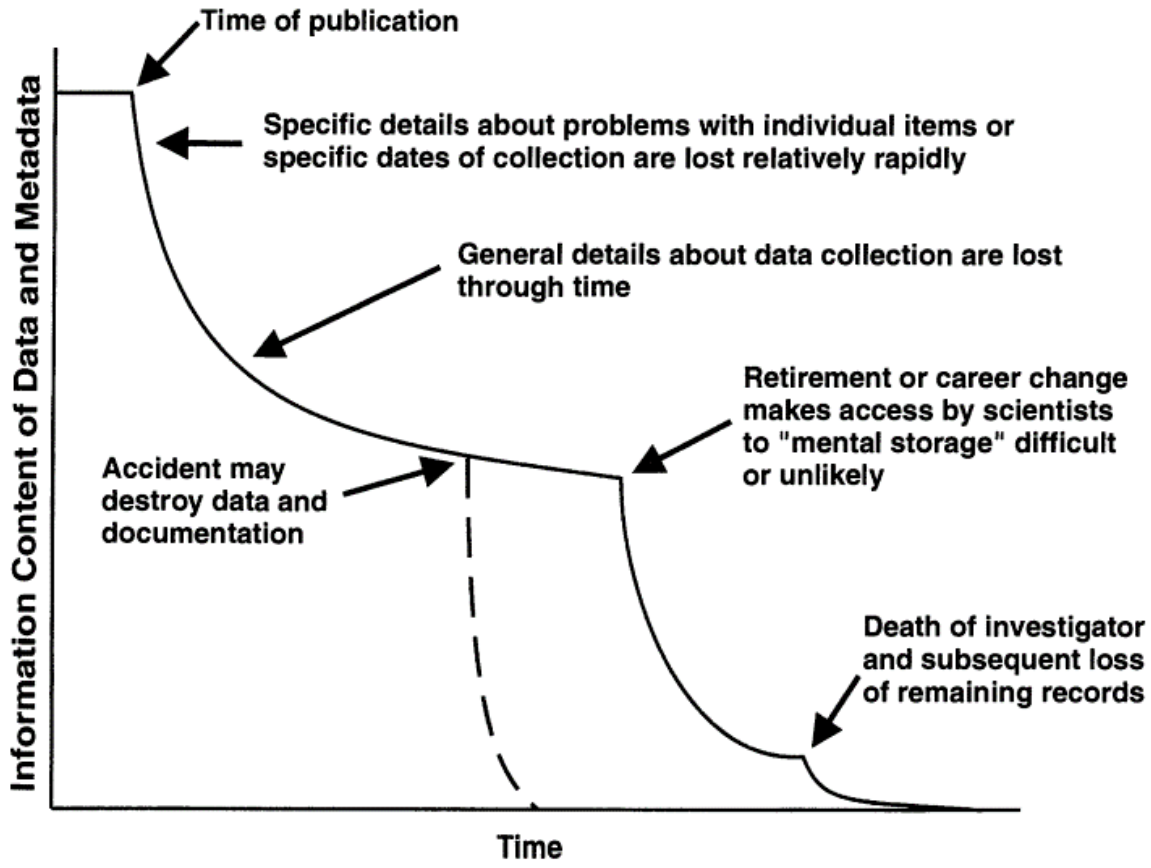
Why is this conventional approach to science ineffective at seeding change?

The environmental management community is often siloed with each branch doing their own thing and speaking their own language. Between the research (typically academic) and management community, we call this the research-management divide.

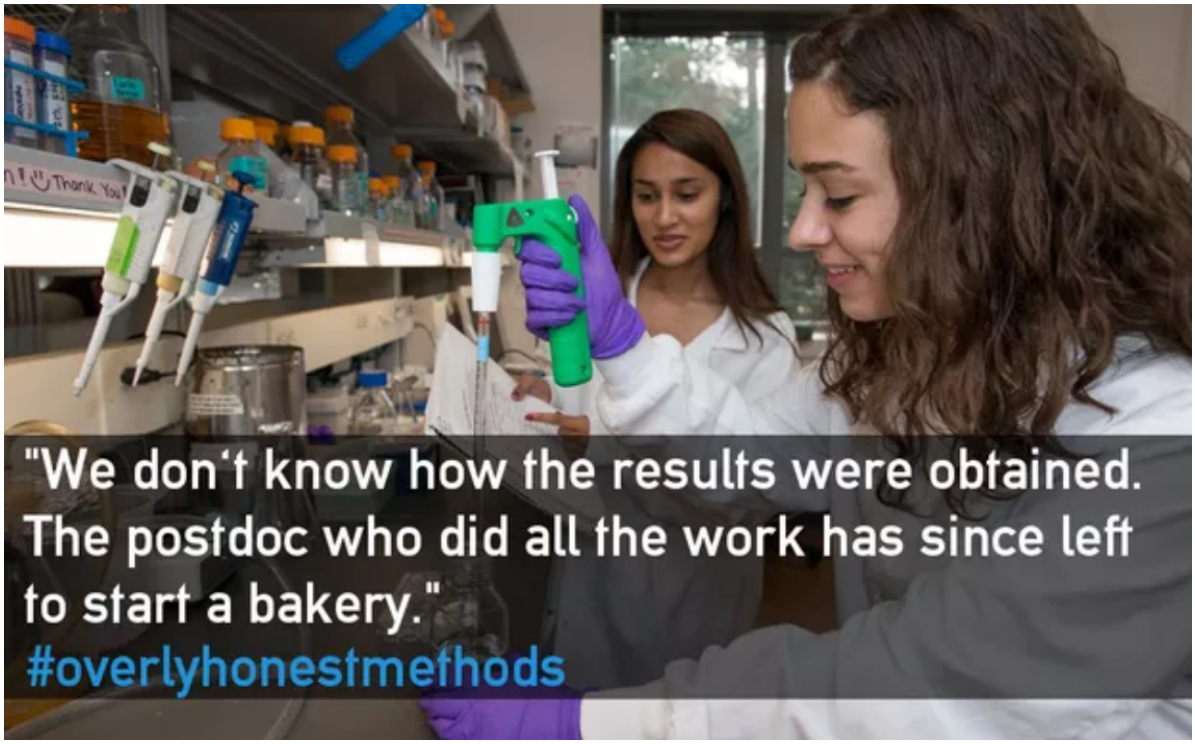


A distinct gap exists between how scientific products are developed and how they can be used to meet management needs. This is often the result of communication barriers, irreproducible results, information loss with poor documentation, inaccessible data, and opaque workflows known only to the analyst.

These barriers can occur at any stage of the research process. This compelling graphic from Michener et al. (1997) describes the atrophy of information in a closed approach to creating science.



The last part is especially morbid. Sometimes, this is called the "bus factor". What would happen to your important work and life achievements if you were hit by a bus? Would others be able to pick it up? Research products with a high bus factor are at risk of being lost if critical team members are no longer available. This is a very real problem for continuity of science.

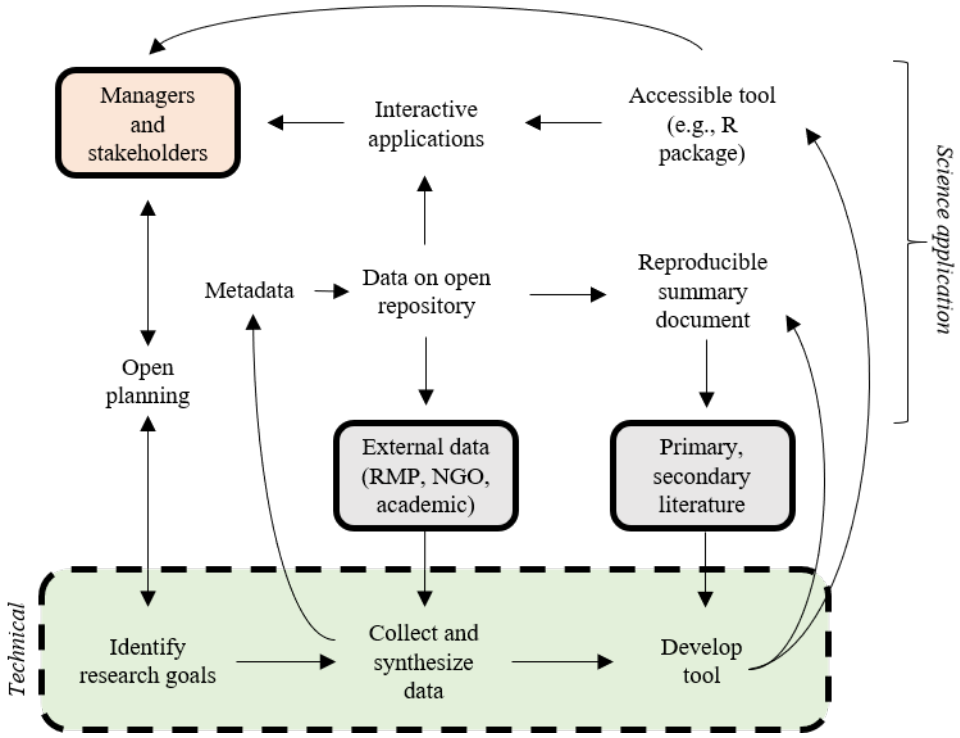


So how do we make changes to our workflows to ensure we can achieve truly applied science using open tools and philosophies?

1.3 Learning and speaking the language of open science

The tools and broader philosophy behind open science can help us bridge the research-management divide. It involves a fundamental shift in how we approach the scientific process, both for your own internal workflows and how you can engage others in the process. By others, we mean not just researchers, but specifically those that need the information to make informed decisions. This also includes your *future self*.

Before we present a formal definition, let's describe a modification of the conventional workflow that includes an open process to discovery and implementation (Beck et al. 2020; modified from Hampton et al. 2015).



This workflow is similar to the original scientific method, but the technical components are open to managers and stakeholders, we're treating data differently by using metadata and archiving, we're creating summary documents that include source code with text, and we're producing decision-support tools to meet the needs outside of the research community. Importantly, the process is also *iterative* and dynamic.

Throughout this workshop, we'll learn about some open science tools that can be used in this generalized workflow to achieve better science in less time (Lowndes et al. 2017).

Now let's settle on a definition for open science (from Open Knowledge International, http://opendefinition.org/, https://creativecommons.org/):

> "The practice of science in such a way that others can *collaborate* and *contribute*, where research data, lab notes and other research processes are *freely available*, under terms that enable *reuse*, *redistribution* and *reproduction* of the research and its underlying data and methods."

Key words from this definition are italicized. There are very specific tools in the open science toolbox that enable each of these key words. We'll cover some of these later.

Similarly, the current administration has declared 2023 the Year of Open Science. Their definition is:

> "The principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility, and equity."

We can can breakdown these definitions into key principles.

1. **Open data**
    - Public availability of data
    - Reusability and transparent workflows
    - Data provenance and metadata
2. **Open process**
    - Iterative methods using reproducible workflows
    - Collaboration with colleagues using web-based tools
    - Leveraging external, open-source applications
3. **Open products**
    - Interactive web products for communication
    - Dynamic documents with source code
    - Integration with external networks for discoverability

You'll notice that web-based tools and open science are often discussed at the same time. Science existed before the internet. Open science often focuses on how the two can leverage and support one another despite the latter being a relatively new addition to society. We often describe web-based tools as synonymous with open science.

1.4 The FAIR principles

Advocates of open science also use the FAIR principles (Wilkinson et al. 2016) as a vehicle for achieving the former. It's important to understand what they mean so that you can be fluent in both. The FAIR acronym is described as follows:

- **F**indable: The data have a globally unique and persistent identifier, including use of "rich" metadata.
- **A**ccessible: Once found, the data can be retrieved using standardized communications protocols that are open, free, and universally implementable.
- **I**nteroperable: The ability of data or tools from non-cooperating resources to integrate or work together with minimal effort.
- **R**eusable: If the above are achieved, the data and metadata are described in a way that they can be replicated and/or combined in different settings.

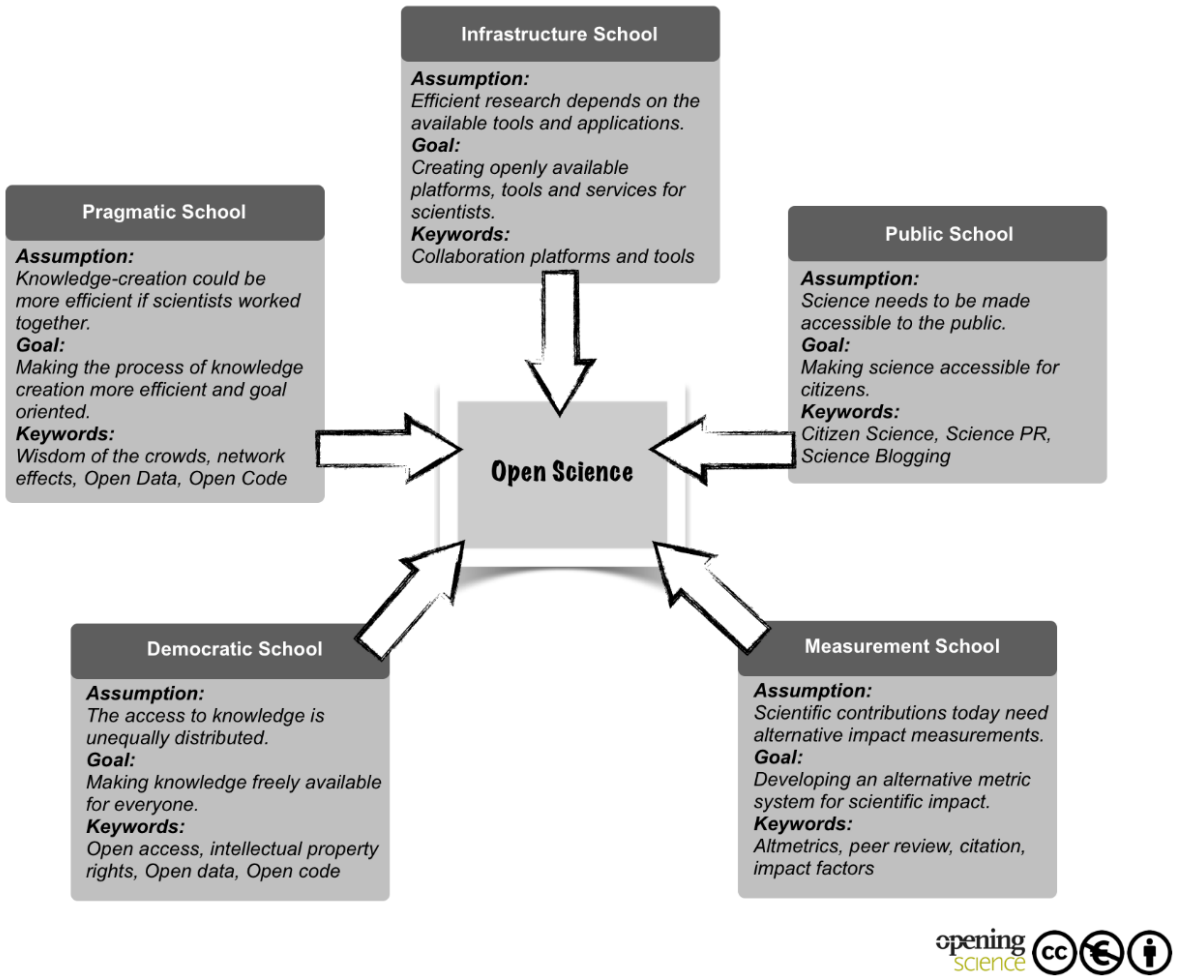
Simply, what this means is: 1) each dataset has a name that doesn't change and can be found with minimal effort using that name, 2) once it's found, you can actually get your hands on it (e.g., not behind a paywall), 3) once you have it, you can use readily available tools to work with the data (e.g., not using proprietary software), and 4) you can actually apply the data for your own needs because it has sufficient context, including its reproduction, given that the first three principles are met.

For our purposes, think of these ideas as general guidelines you can ask yourself when doing science. If you find that your work is not FAIR, then you're probably not being as open as you could be. We'll of course provide some tools to help you be FAIR and open.

1.5 Schools of thought

Finally, it's useful to make a distinction of how different people may talk about open science. This can help you better navigate conversations and become an advocate for open science in your own right.

A useful paradigm is provided by Fecher and Friesike (2014) to describe open science as five distinct schools of thought:



These are of course only conceptual boxes and there's considerable overlap across all schools when open science is used in practice. For our purposes, we'll mostly be talking about ideas and tools from the pragmatic, infrastructure, and democratic schools of thought. The end goal is to provide you with the means to create more efficient and impactful science that can more readily be used by others in a collaborative setting.

Exercise and discussion

Take a few minutes to jot down your individual answers to the following questions. When you're done, share amongst your peers at your table.

1. How do you currently define open science, if at all?
2. What tools do you use in your job that facilitate collaboration or openness?
3. Has any of the above changed your understanding of what open science means for research and/or resource management?

2 Open science for collaboration

2.1 Goals and motivation

This is the second module in our workshop on open science. This module will explore some open science tools to help you and your team become better collaborators and to better engage your science with external partners. We'll introduce some essential elements of collaboration and discuss some readily available tools for doing so.

- **Goal**: understand methods of collaboration and the pros/cons of various tools
- **Motivation**: start building the tools for your open science toolbox

2.2 Essential elements of collaboration

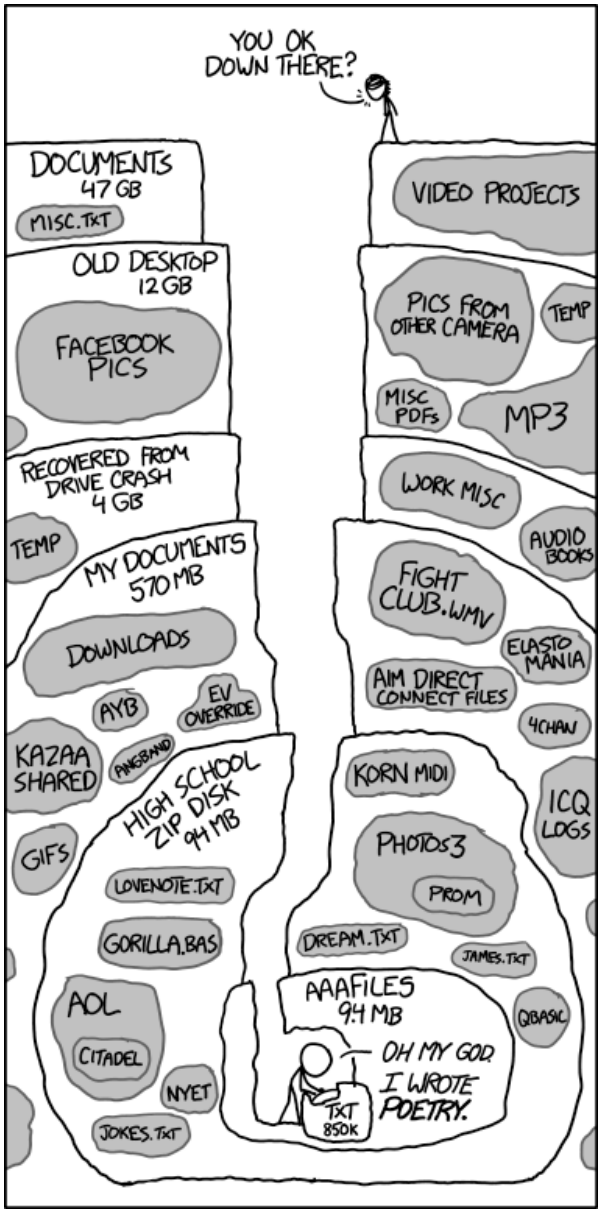
We start our deep dive into open science by focusing on collaboration as a fundamental activity that can be enhanced through transparent, efficient, and reproducible tools. Having effective tools to work together is a critical theme of many open science practices. There are many tools in the toolbox and we need to introduce some core concepts before we demonstrate how to implement them in practice.

2.2.1 Workflow management

How do you organize your work each day? How do you make sure projects are on schedule and pressing deadlines are met? How do you plan for short-term and long-term goals? Do you have a five-year, ten-year, or longer career plan?

Work to achieve goals cannot be accomplished without a systematic approach to organizing tasks. Chances are, we each have our own system that works for us and was probably developed through trial and error. Although everyone has familiar workflows, they are often idiosyncratic and deeply entrenched by habit. That can be in direct conflict with collaboration when we try to mesh internal workflows with those of others.

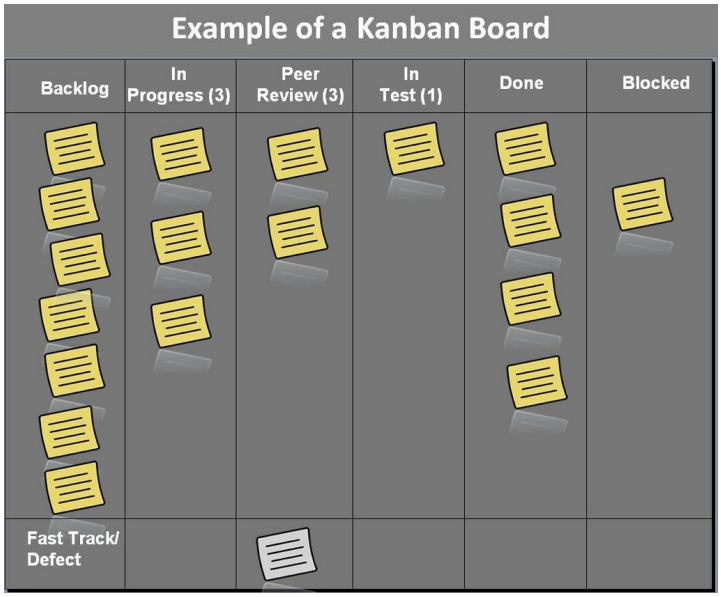
Does this look familiar?


YOU OK DOWN THERE?
DOCUMENTS 47 GB
MISC.TXT
OLD DESKTOP 12 GB
FACEBOOK PICS
RECOVERED FROM DRIVE CRASH 4 GB
TEMP
MY DOCUMENTS 570 MB
DOWNLOADS
AYB
EV OVERRIDE
KAZAA SHARED
PAINT.NET
HIGH SCHOOL ZIP DISK 94 MB
GIFS
LOVENOTE.TXT
GORILLA.BAS
AOL
CITADEL
NYET
JOKES.TXT
VIDEO PROJECTS
PICS FROM OTHER CAMERA
TEMP
MISC PDFS
MP3
WORK MISC
AUDIO BOOKS
FIGHT CLUB.WMV
ELASTOMANIA
AIM DIRECT CONNECT FILES
4CHAN
KORN MIDI
ICQ LOGS
PHOTOS3
PROM
DREAM.TXT
JAMES.TXT
AAAFILES 9.1 MB
QBASIC
OH MY GOD I WROTE POETRY.
TXT 850K

Although the above comic from xkcd speaks directly to file management, it hints at a broader problem of personal information management that can seriously complicate working with others. I'm sure we've all struggled to find that one file for that one project from a vague recollection of seeing it a few months ago.

Collaborative work can be facilitated through workflow management that helps you break out of old habits. We'll introduce some specific internet-based tools below to facilitate workflows either for yourself or, better yet, working with others. These can help propel you towards open science.

Here, we introduce the Kanban approach to workflow management. The idea is simple. Create a task-oriented workflow using a card management system organized by progress. It looks something like this:



As shown, this approach can work as a literal, physical board or as one used digitally through a web browser or other software. Every Kanban board has the following elements that allow you to work in a more informed manner:

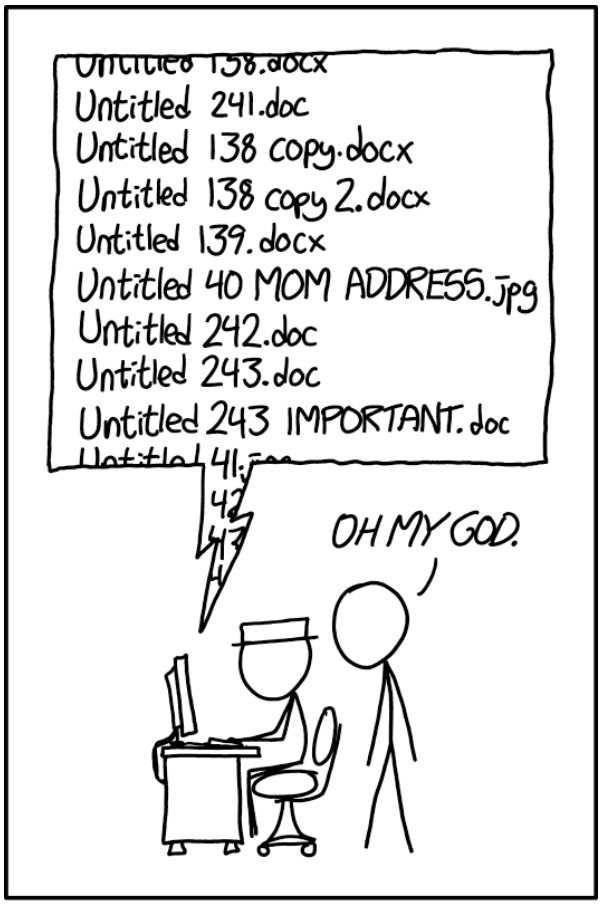
- Provides a "big picture" of progress
- Organizes progress by discrete steps
- Establishes cards as specific tasks

Many of the open science tools we describe below use this system. It is a generalizable format that works in different settings, whether it be general project management or something more formal like software development.

2.2.2 Version control

A specific problem for workflow management that can be solved by open science tools is file management. Workflows can be immensely enhanced by tools that use strict guidelines for tracking changes and allowing a complete view of the evolution of a project. This is where version control comes in.

I'm sure many of you have fallen into this trap:

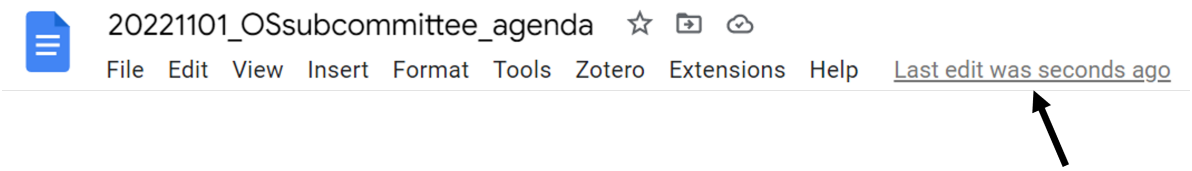


Version control is a way to track the development history of a project. It serves the joint purposes of:

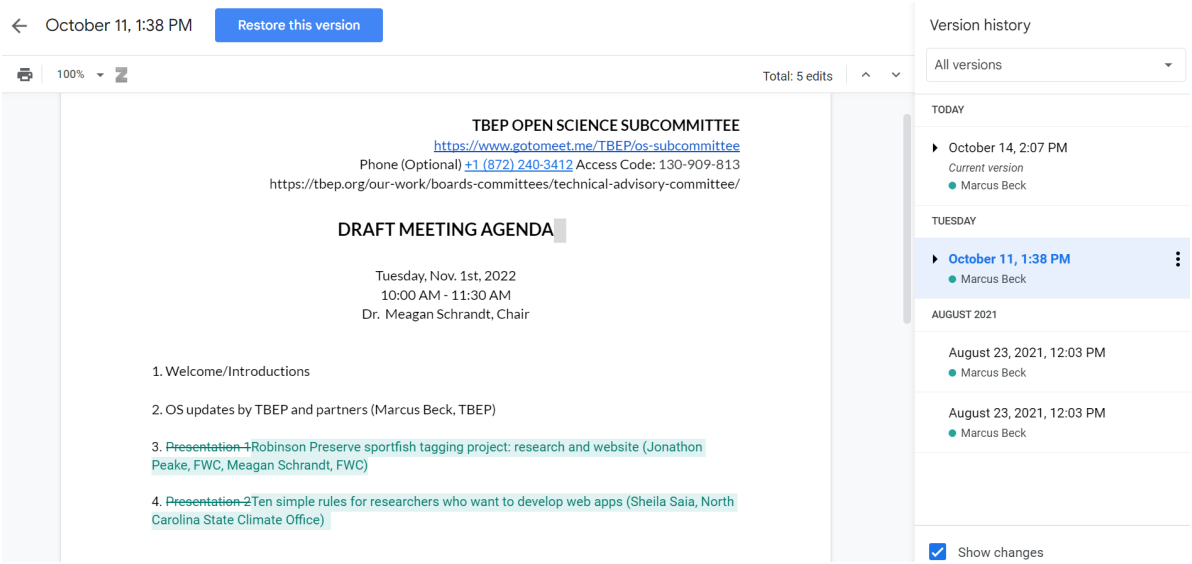
1. Formally documenting the changes that have been made to code or software
2. Making sure that the development history is permanent
3. Providing a system for collaborating across platforms (with friends!)

It's more than saving files. Documenting changes with a set of commands that follow strict rules provides a transparent record for yourself and others, and establishing permanency ensures that any of the changes that are made can be vetted and accessed as needed. Think of it as an insurance plan for your project.

If you've ever used Google Docs, you might have noticed a feature that looks a lot like version control. The Google Drive platform is a great way to start working together and a great way to familiarize yourself with the basics of version control.



For any Google Doc, clicking on the link shown by the arrow will open the Version history pane which shows all of the edits that were made to the document. You can view any of the edits, who made the edits, view the changes (before/after) in the document, or even restore the document to a previous version.



These are the building blocks of version control as demonstrated with Google Docs:

1. No iterative and ambiguous file naming
2. History of changes assigned to each editor
3. Ability to restore a previous version

Perhaps more importantly, these tools are in the cloud and openly accessible (unlike other cloud-based services). File links (via a URL) also do not change if a file is moved to a

different location in the drive. Overall, the Google platform is an accessible means of improving collaboration (but not without it's cons).

2.2.3 Git and GitHub

Although Google products can get you a long way towards better collaboration, they do not use dedicated version control software. These tools become more important as your projects become more complex - those beyond simple documents or spreadsheets.

The most widely used software for version control is Git. Although we do not cover the specifics of this software, it's useful to understand the purpose and what it can do in making your work more open and impactful. Git is integrated with many popular open source development platforms, such as RStudio.

Many people often confuse Git with GitHub. GitHub is an online platform for working collaboratively through Git AND it allows you to be open with your work. We'll provide some examples below of how this can be done. Importantly, you do not need to be an expert in Git to be able to use GitHub. This speaks volumes for how team efficiency can be improved with GitHub through better collaboration.

This recent blog provides a helpful introduction to Git/GitHub for the casual user.

Challenge

Many institutions block access to Google products or GitHub. See some ideas in our last module to overcome this issue.

Watch and learn

Workflow management in the real world - using GitHub to collaborate. Here we present some examples from the Tampa Bay Estuary Program State of the Bay report and water quality report card.

Watch and learn

Now we'll demonstrate how to setup a version control project with RStudio, Git, and GitHub. This example will cover:

1. Creating the project in GitHub

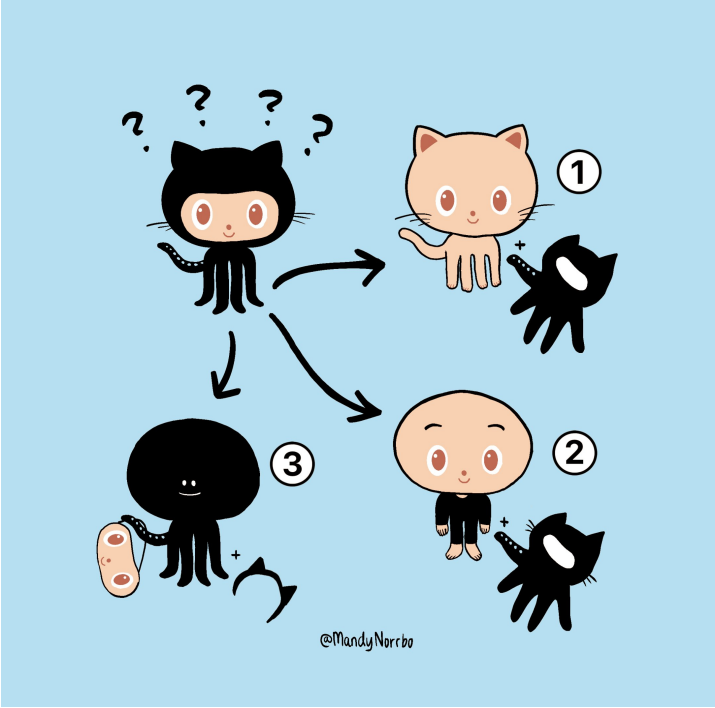


Figure 2.1: Octocat, the strange and loveable mascot of GitHub.

2. Creating a file, adding content, and committing it to the project
3. Setting up issues in GitHub
4. Adding members to the project
5. Creating a Kanban project board to assign tasks

2.2.4 Code of Conduct

Every responsible collaborative team begins work by creating a Code of Conduct. This documents a set of community and social standards within which the work can be completed. It ensures all viewpoints are heard and respected and establishes a means by which conflicts can be resolved.

Here's a great example from our friends at openscapes and one from the ROpenSci community. Some guidance for creating a code of conduct is also provided in this blog post from ROpenSci.

The goal of every code of conduct is to ensure an agreed upon set of norms are used by all team members to help create a safe and positive experience.

2.3 Additional tools for collaboration

Below we introduce additional web-based tools that you can use to improve collaboration and openness, including those described above. We present them as a suite of options to consider based on the pros and cons associated with each tool. This is by no means a comprehensive list, but it should get you started towards better collaboration in an open environment.

2.3.1 Slack

https://slack.com/

What

An online messaging platform for internal communication. Conversations can be organized by topic (via channels) or you can send direct messages to one or more team members. You can have multiple workspaces for different groups.

Pros

Alleviate email overload through quick, informal messaging. Offers a fresh approach to online communication.

Cons

Yet another thing to monitor. Free subscription limits archive of messages. Communication is limited to those in the same workspace.

2.3.2 Trello

https://trello.com/

What

A Kanban style workflow organization platform. Can be used for personal organization or in teams. Card management allows you to assign due dates, add attachments, make checklists, assign tasks to yourself or team members, and label by themes.

Pros

Easy to use and can upgrade with "power-ups" for integration with other services (e.g., Google). Use across locations (e.g., from home or in the office) is easy because it's based in a web browser.

Cons

Not entirely open because it's only visible to yourself or those you explicitly invite. Free version is limited to only a handful of "power-ups".

2.3.3 Google Drive

https://google.com/drive

What

Cloud-based platform for sharing documents, worksheets, slides, etc. Follows a familiar file-based structure that is common to most operating systems.

Pros

Easy to use and can be a very open space for collaboration. Fairly interoperable with different file formats. Some functionality with version control (i.e., ability to "revert" to previous versions and to view changes).

Cons

Requires a Google account and access can be tricky depending on institution. Even though some versioning is provided, the format can encourage poor file management. Who knows what Google is doing with your data.

2.3.4 Office 365

https://www.microsoft.com/en-us/microsoft-365

What

Cloud-based platform for secure sharing of Microsoft documents, worksheets, slides, etc.

Pros

Easy to use and fully supports Microsoft products. Low barrier of inclusion to others that are already using Microsoft products.

Cons

Requires a Microsoft account and access can be tricky depending on institution. Maintains dependency on expensive Microsoft products that aren't reproducible or interoperable. Very often used in closed workflows.

2.3.5 GitHub

https://github.com

What

Cloud-based platform for sharing code with Git version control. Supports sharing of most file types, although code and text-based files are the primary use.

Pros

Collaborative and fully transparent work environment for files under version control. Supports workflow management through issue tracking and Kanban style project boards. Links to third-party platforms for archiving and DOI generation (e.g., Zenodo). Octocat mascot is super cute.

Cons

Learning curve is steep if you want to fully leverage version control. Not a formal data archival service by itself and file sizes are limited.

Exercise and discussion

In small groups, setup a shared workspace using GitHub and create a project management board. Some real world examples of why you might do this were presented in the earlier watch and learn.

1. Open GitHub in a web browser and have one person create a new repository (the big, green "New" button in Repositories). Add each member to the repository after it's created (hint: Settings -> Collaborators)
2. Have that same person create a project board for the repository (Hint: Projects -> New project -> board format)
3. After each person accepts the invitation to the repository (check your email!), each new member create a new file in the repository (Hint: Click "Add file" near the top). Name it something unique, save and commit the changes
4. Assign issues to different members of the repository to do something to the new files (Hint: on the right menu, select "Assignees"). Add the issue to the project board (Hint: on the right menu, select "Projects" and click the new project).
5. Work on the issues until the time is up. Close each issue as they're completed.

3 Open science for impactful products

3.1 Goals and motivation

This is the third module in our workshop on open science. Now we focus on core principles for data management as the foundation for open science. We discuss the role of data management to support decisions using open science. Then, we introduce the concepts of tidy data as a unified format for storing information. We close with a discussion of metadata tools and data repositories that allow your data to live beyond the project.

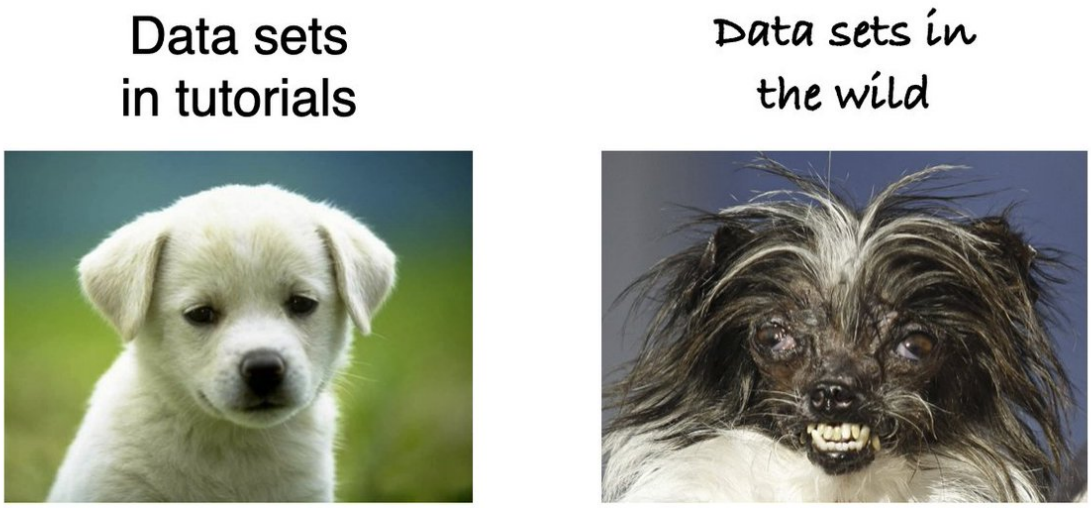
- **Goal**: understand best practices for data management as a key concept for open science
- **Motivation**: cultivate data as a living, shared resource

3.2 Data as the foundation for open science

In the last module, we talked about collaboration as the single most important activity of open science. So, why are we now talking about data management? Understanding the tools of collaboration allows you to better engage with your colleagues and partners, but open engagement will mean nothing if your data look like garbage.

You can probably recall past instances when poor data management has been a challenge for open collaboration. Here are a few real-world examples:

1. A collaborator calls you on the phone asking about a historical dataset from an old report. You spend several hours tracking down this information because you don't know where it is. The data you eventually find and provide to your collaborator has no documentation and they don't know how to use it or use it inappropriately.
2. You receive a deliverable from a project partner that was stipulated in a scope of work. This deliverable comes in multiple formats with no reproducible workflow to recreate the datasets. You are unable to verify the information, eroding your faith in the final product and making it impossible to update the results in the future.
3. An annual reporting product requires using new data each year. The staff member in charge of this report spends several days gathering the new data and combining it with the historical data. Other projects are on hold until this report is updated. Stakeholders that use this report to make decisions do not trust or misunderstand the product because the steps for its creation are opaque.



Data come in many shapes and sizes, most more like the right side of the above picture. Poor data management occurs for many reasons, but these are a few of the common causes:

- What's easy for recording data doesn't usually translate to easy analysis
- Egregious use of Excel as data management software
- Metadata is a chore that is often an afterthought

It's often said that 90% of working with data is cleaning (or "wrangling"), whereas actual analysis and interpretation is a small fraction of your total effort. Using better data management practices will not only help you save time, it's also a service for your colleagues, potential collaborators, and your future self. Therefore, better data management leads to more open science.

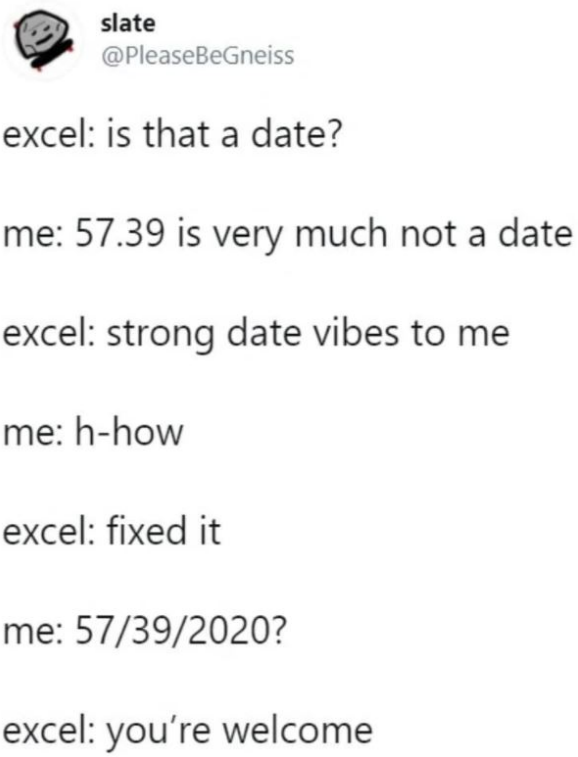
The FAIR principles outlined in the first module are especially useful when working with data for open science applications. Many of the collaborative tools in the second module can facilitate application of FAIR data. In this module, we'll go a step further to discuss how data structure, including metadata, can produce a FAIR dataset.

3.3 Principles of tidy data

Tabular data allow you to store information, where observations are in rows and variables are in columns. It's very common to try to make tabular data more than it should be. Unless you

spend a lot of time working with data, it can be difficult to recognize common mistakes that lead to *table abuse*.

Before we get into tidy data, we need to discuss some of the downfalls of Excel as a data management system. There are many examples that demonstrate how Excel has contributed to costly mistakes through table abuse or outright negligence, often to the detriment of science (Ziemann, Eren, and El-Osta 2016).



Excel allows you to abuse your data in many ways, such as adding color to cells, embedding formulas, and automatically formatting cell types. This creates problems when the organization is ambiguous and only has meaning inside the head of the person who created the spreadsheet. Embedding formulas that reference specific locations in or across spreadsheets is also a nightmare scenario for reproducibility.

If you absolutely must use Excel to store data, the only acceptable format is a rectangular, flat file. This is typically saved as a *.csv* file. What do we mean by this?

A **rectangular** file:

> Store data only in rows and columns in matrix format (e.g., 10 rows x 5 columns), with no "dangling" cells that have values outside of the grid or more than one table in a spreadsheet.

	A	B	C	D	E	F
1	OTBT gmtpmt& mnpmt					
2		OTBT Rzpsmr	OTBT N/B	Tsct OTBT NB		
3		11434	64037	200000		
4	Osrkzt vglmOzs mnpmt					
5		Lsst 12 Ognths sslzs snd rzpsmr zxpgsmrz pzrmgd cslcmlstmgn				
6		Tgtsl bssz	svzrsgz rmntmOz	BzTgrz 12 O	R/Onth -1	R/Onth -2
7	yslzs bmmlt:	1412	11.97	1400	0	0
8	Tgtsl rzpsmr bmmlt:	306	4.36	0	2	23
9	ATT bmmlt mnpmt	272	4.01	0	1	15
10	Bzndgr 2 bmmlt mnp	0	#DIV/0!	0	0	0
11	Bzndgr 3 bmmlt mnp	34	7.18	0	1	8
12	RzAsmR mNAUT					
13			Rzpsmr mssgz		Rzpsmr pgpmlstmgn	
14		Tgtsl Rzpsmr	Rzpsmrzd DOs	Rzpsmrzd Wsrr	NgrOsl Rzpsmr	Lgcsl Rzpsmr
15		306	2	40	264	0
16	Argdmct mnTgrOstmgn				% cmrrznt mnstsllzd bssz sTtzr Tmtmrz ssl	
17		Argdmct NsOz	ytylz NsOz	Rzvmzw Dstz	Ognths sctmvz	svzr Tmtmrz rmntmOz
18		DzhmOmdmi	sB6065903CT	25/06/1998	76.8	27
19	Asrt NmObzr mnTgrOstmgn			Ognth/dsy/yzsr		
20		sDy Asrt NmObz	yzt-mp Dstz	ytsndsrd cgst	Ognthly mssgz	12 Ognth's mssgz
21		43588136-00.	1/02/1992	675.92	9.7	116.4
22			Ognth/dsy/yzsr			
23						

A **flat** file:

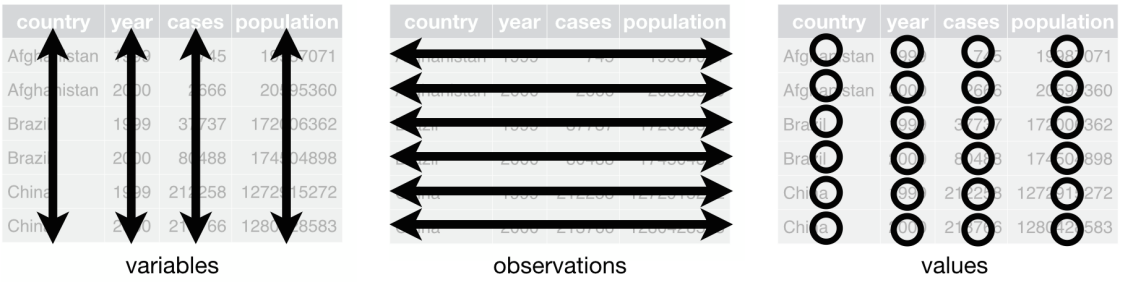
> No cell formatting, no embedded formulas, no multiple spreadsheets in the same file, and data entered only as alphanumeric characters.

Broman and Woo (2018) provide an excellent guide that expands on these ideas. Essentially, these best practices force you to isolate the analysis from the data - many people use Excel to mix the two, leading to problems.

Now we can talk about tidy data. The tidy data principles developed by Hadley Wickham (Wickham 2014) are a set of simple rules for storing tabular data that have motivated the development of the wildly popular tidyverse suite of R packages (Wickham et al. 2019). The rules are simple:

1. Each variable must have its own column;
2. Each observation must have its own row; and,
3. Each value must have its own cell.

Graphically, these rules are shown below (from Wickham and Grolemund 2017):



Using these principles may seem unnatural at first because of a difference between what's easy for entering data versus what makes sense for downstream analyses. For example, dates are often spread across multiple columns, such as having one column for each year of data where the header indicates the year that applies to data in the column.

The following examples show five tables represented in different arrangements. Only one of the tables is tidy - which one?

Habitat	Year	Acres	Category
Seagrass	2019	519	B
Oysters	2019	390	B
Sand	2019	742	C
Seagrass	2020	438	C
Oysters	2020	875	B
Sand	2020	702	A
Seagrass	2021	375	A
Oysters	2021	724	A
Sand	2021	505	C

Figure 3.1: Table 1

Only the first table is tidy - each variable has its own column, each observation has its own row, and each value has its own cell. Table 2 violates the first rule, Table 3 violates the third rule, and tables 4a and 4b violate the first and second rules.

Table 2 is a special case where the data are mostly tidy, but represented as a ***long*** format of the same ***wide*** data in Table 1. Storing data in a long format is not necessarily wrong and, in fact, is a common format for databases. There are pros and cons to both, depending on the intended analysis. Here are two examples of some water quality data you might encounter in the wild, one wide and one long.

Habitat	Year	Attribute	Value
Seagrass	2019	Acres	519
Seagrass	2019	Category	B
Oysters	2019	Acres	390
Oysters	2019	Category	B
Sand	2019	Acres	742
Sand	2019	Category	C
Seagrass	2020	Acres	438
Seagrass	2020	Category	C
Oysters	2020	Acres	875
Oysters	2020	Category	B
Sand	2020	Acres	702
Sand	2020	Category	A
Seagrass	2021	Acres	375
Seagrass	2021	Category	A
Oysters	2021	Acres	724
Oysters	2021	Category	A
Sand	2021	Acres	505
Sand	2021	Category	C

Figure 3.2: Table 2

Habitat	Year	Acres and Category
Seagrass	2019	519/B
Oysters	2019	390/B
Sand	2019	742/C
Seagrass	2020	438/C
Oysters	2020	875/B
Sand	2020	702/A
Seagrass	2021	375/A
Oysters	2021	724/A
Sand	2021	505/C

Figure 3.3: Table 3

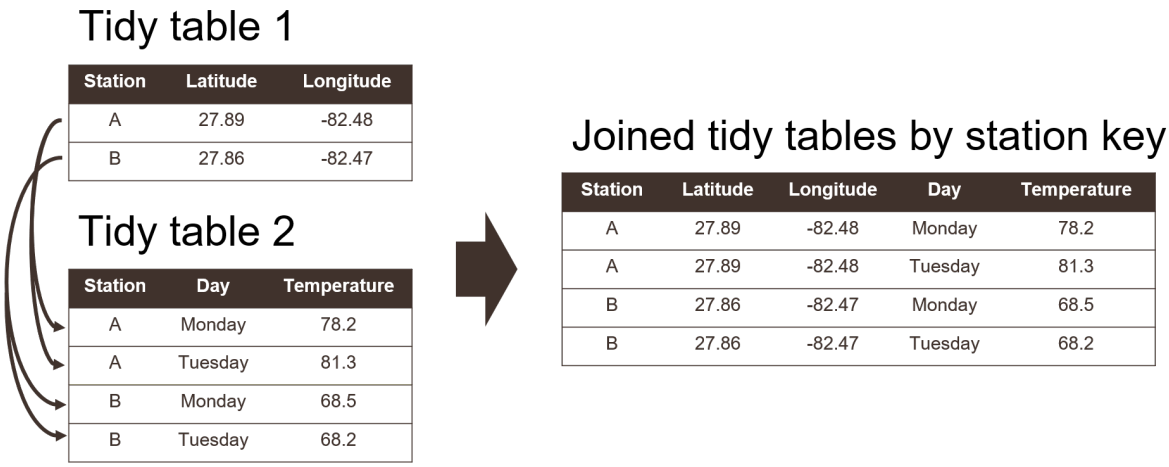
Habitat	2019	2020	2021
Oysters	B	B	A
Sand	C	A	C
Seagrass	B	C	A

Figure 3.4: Table 4a

Habitat	2019	2020	2021
Oysters	390	875	724
Sand	742	702	505
Seagrass	519	438	375

Figure 3.5: Table 4b

Using a tidy format also allows you to more easily join data between tables. This is a common task when you have information spread between different tables because: 1) it might not make sense to keep the data in the same table, and 2) the analysis depends on information from both tables. Tidy data shared between tables can be linked using a "key" as a common identifier.



Watch and learn

Making an untidy dataset tidy using Excel (an irreproducible example).

Watch and learn

Making an untidy dataset tidy using R (a reproducible example).

```
library(readxl)
library(dplyr)
library(tidyr)

# import and wrangle
dat <- read_excel('data/untidy.xlsx', skip = 1) %>%
  fill(Location) %>%
  pivot_longer(cols = `2019`:`2021`, names_to = 'Year', values_to = 'Acres/Category', name
```

```
  separate(col = `Acres/Category`, into = c('Acres', 'Category'), sep = '/', convert = T)
dat
```

Exercise and discussion

Now try it on your own. Download the untidy dataset and make it tidy using your preferred software.

3.4 Data dictionaries

Once you understand the tidy principles, you'll find that analysis is much, much easier. More importantly, this also facilitates documentation for the explicit purpose of open sharing. As responsible data stewards, we need to think of data as a living resource for open collaboration that is not just a means for more conventional scientific products (e.g., a publication). Data are increasingly being documented and cited as unique entities and we need proper documentation to help support the FAIR principles.

A good first step in documentation is to create a data dictionary. This will help us when we start creating metadata. Think of this as the specific description of the contents of a tabular data file. Developing a data dictionary not only helps with metadata, but also helps you think more clearly about your data.

A data dictionary describes column names and the type of data in each column. Simple things like how you name a data column can have larger implications for downstream analysis pipelines or interpretability of a dataset.

Here's an example of a data dictionary for a made up dataset. Without seeing the actual data, you get a good sense of what's included and acceptable values for adding new data.

Column name	Plot name	Data type	Acceptable values	Description
stat	Station	Categorical	A, B	Unique identifier for the sampling station
lat	Latitude	Numeric	-90 – 90	Latitude location of the station in degrees
lon	Longitude	Numeric	-180 – 180	Longitude location of the station in degrees
day	Day of week	Categorical	Sunday through Saturday	Day of the week when sampling was done
temp	Temp. (F)	Numeric	0-100	Measured temperature in Fahrenheit at the station

Here we provide some general guidelines for developing your own data dictionary.

- Column names
  - Be as descriptive as possible while trying to keep the name as short as possible. Really long names with lots of detail can be just as frustrating as very short names with very little detail. The column name should be intuitive to point the analyst in the right direction.
  - Try to avoid spaces or commas in column names since some software may interpret that as the start of a new column. Do not start a column name with a number.
  - It may also be useful to identify a "plot name" column that uses proper spelling, punctuation, and units. Entries in this column can be used for making graphics that have interpretable names (e.g., using "Temp. (F)" instead of "temp" as the name of a plot axis).
- Data types
  - Describe the type of data in each column, e.g., numerical measurements, categorical descriptors, or counts of observations (integer). Never, ever mix data types in the same column.
  - If your data are continuous numeric values, try to identify an acceptable range for the values, e.g., are there minimum or maximum values that would indicate the data are out of range?
  - For categorical descriptors, identify all possible categories (if feasible) that are acceptable values for the column, e.g., small, medium, or large for a qualitative descriptor of size. These may be ordered or unordered.
  - For dates, make note of the format, e.g., YYYY-MM-DD. For time, identify the timezone and if it includes daylight savings or not.

◢ Exercise and discussion

Create a data dictionary for the tidy dataset from the previous example. Use a spreadsheet to create a table to describe the column names, data type, acceptable values, and a description.

3.5 Importance of metadata

How many times have you been sent a dataset without any idea what it contains or why it was created? How are you sure the information is valid and that your analysis takes into account

the limitations of the data? How many times have you willfully sent someone a dataset without fully providing this information?

Without metadata, it's impossible to know critical details about a dataset that can inform its analysis, and more importantly, its use to inform decision-making. Curating data should be synonymous with metadata generation and is an important part of open science. We cannot provide open data in good faith without also providing metadata.

Metadata is literally defined as "data about data". It varies from simple text descriptions of a dataset, such as "who", "what", "when", "where", "why", and "how", to more formalized standards with that prepare your data for archiving in a long-term repository. The data dictionary is only part of a complete metadata description.

A useful definition is provided by Gilliland (2016):

> A suite of industry or disciplinary standards as well as additional internal and external documentation and other data necessary for the identification, representation, interoperability, technical management, performance, and use of data contained in an information system.

Why don't we see more metadata in the wild? Short answer is that it's often an afterthought, if considered at all. Creating metadata is usually tedious and the return on investment is not apparent at onset of a project. However, the collective growth of sciences and its application to real world problems is dependent on metadata.

The US Geological Survey provides a useful document on creating Metadata in "plain language" to distill the basic information contained in a metadata file. It provides a workflow for answering the "who", "what", "when", "where", "why", and "how" questions for metadata. Below is a brief synopsis:

> What does the dataset describe?

Information here would include very basic details about the dataset including a **title**, **geographic extent**, and **period of time** covered by the data. For geographic extent, this may often include explicit coordinates covering the study area. Location is useful for indexing your dataset relative to others, if for example, a researcher wanted to find data for all studies in the geographic extent of Tampa Bay.

> Who produced the dataset?

This would be yourself and anyone else who has made a significant contribution to the development of a dataset. Data are increasingly being used as citable resources and including individuals that were important in its generation ensures proper attribution. If someone has spent hours toiling in the field to collect the data or hours visually scanning a spreadsheet for quality control, include them!

> Why was the dataset created?

Describing why a dataset was created is critically important for understanding context. If others want to use your data, they need to know if it's appropriate for their needs. Here you would describe the goal or objectives of the research for which the data were collected. It should be clear if there are limitations in your data defined by your goals.

How was the dataset created?

Here you would describe the methods used to generate the data, e.g., field sampling techniques, laboratory methods, etc. This information is important so others can know if you've used proper and accepted methods for generating the data. Citing existing SOPs or methods that are recognized standards in your field would be appropriate.

How reliable are the data?

It's also important to explicitly note instances when the data could be questionable or inappropriate to use. Here you could describe any quality assurance or quality control (QAQC) checks that were used on the data. There are often formalized ways to do so, such as codes or descriptors in tabular data defining QAQC values (e.g., data in range, below detection limits, sensor out of service, etc.).

How can someone get a copy of the dataset?

Good metadata has information on who to contact for getting the data. This contact may not be the same as who created the dataset (e.g., IT staff). For archived or publicly available data, this information is more important for who to contact should someone have questions. Information on obtaining a copy of the data should also describe any special software or licensing issues related to accessing the data.

Once you've gathered this information, how do you turn it into literal metadata? It depends on how deep you want to go. At it's simplest, your metadata could be a simple text file with answers to the questions. Or it could be a specific file format used by modern data archive repositories (e.g., EML format).

Here's an example of a bare bones metadata file. One could easily type this up in a text file or spreadsheet.

Ft. DeSoto Continuous Monitoring Buoy Data	
Dataset Type	Non-spatial database
Name of Data Source:	Ft_DeSoto_Buoys_Data
Number of Water Resources Sampled:	2 Sites in Ft Desoto Bay
Datasource Abbreviation (dataset):	Ft_DeSoto_Buoy208, Ft_DeSoto_Buoy209
Description of Datasource:	Continuous water quality data collected by YSI EMM 150 buoys at two locations in Ft DeSoto Bay Pinellas County, Florida. Data is collected every 15 minutes. Depth of collection is fixed at approximately 0.75 meters from water surface.
Data Collection Locations:	Buoy 209 is located at 27.629, -82.710. Buoy 208 located at 27.630 and -82.707.
Data Parameters Collected:	Station ID, Year (yyyy), Time (24 hour, EST), Date (mmddyy), Temperature (ºC), Conductivity (µs/cm), Salinity (ppt), pH, Chlorophyll-a (mg/L), Pheophytin (mg/L), DO % (%), DO (mg/L)
Method of Transferring Data to the Atlas:	Transfer via FTP site
How Often Data is Transferred to the Atlas:	Quarterly
Data Current as of:	2021-01-05 12:35:39 UTC
Disclaimer/Use Constraints:	None
Custodian Information:	Pinellas County Public Works
Contact Name:	Jane Doe
Contact Phone:	(555) 123-4567
Contact E-mail:	jdoe@pinellascounty.org

◢ Exercise and discussion

Create a metadata file for the tidy dataset from the previous example. Just start by using your data dictionary and develop narrative answers for "who", "what", "when", "where", "why", and "how" to describe your dataset. Get creative with your descriptions since this is a made up dataset. Enter this information in a text file or spreadsheet.

3.6 Data repositories

How data are treated as living, dynamic pieces of information is critical to the ethos of open science. This is especially true when the FAIR principles are invoked. Data should not live on your hard drive as something only known to yourself.

Although we will not cover data repositories in depth, it's important to recognize the critical role that data archiving and metadata have in open science. How many times have you thought "wow, it would be great if I could have the data from this paper!" Making data open is a great way to propel science through better collaboration.

The ease of getting a dataset online depends on where you want to put the data. In most cases, your dataset should be tidy and accompanied by metadata. For simple solutions, such as FTP hosting or putting a dataset on Google Drive, all you need to do is upload the data and metadata by hand. However, this doesn't necessarily make it findable and the permanency is uncertain.

The absolute best standard for hosting data online is through a *Federated Data Repository*:

> An online network of connected repositories that use similar standards to collectively store data for discovery and access. Uploading a dataset to one node of a repository will make it available through all other nodes.

Such repositories follow strict but necessary guidelines to ensure your data have permanence and adhere to the FAIR principles. The data are definitely findable (e.g., through a web search), accessible (free to download), and interoperable (accepted standards are ensured). The "reproducible" aspect can be debatable, but that can be solved through other means (e.g., code sharing on GitHub).

Some examples of data repositories, most are domain-specific:

- KNB: Knowledge Network for Biocomplexity, a general purpose repository for ecological data
- HydroShare: Data and models used in hydrology
- OPC: California Ocean Protection Council, marine and coastal datasets
- OSF: The Open Science Foundation, a general location for sharing data and other research tools

◢ Exercise and discussion

Using the tidy dataset, your data dictionary, and your metadata description from the previous exercises, we'll archive the data on the test node of KNB. This workflow is exactly the same for the real KNB repository and is designed to get you comfortable adding data to an archive.

1. Log in to the test archive using your ORCID information: https://dev.nceas.ucsb.edu/
2. Click the submit button at the top to start the process of entering your data.
3. Add the relevant data dictionary and metadata information in the forms (because this is a simple exercise, not all forms need to be filled). The forms will be slightly different from your metadata, but enter the relevant information for the appropriate form.
4. Save the dataset when you're done and marvel at your work.

Part II

Additional content

4 Lowering barriers to inclusion and addressing key critiques

4.1 Goals and motivation

What does it mean to use open science in the real world? It's great to talk about the value of open science and the tools you can use, but it's a completely different ball game when it comes to putting these ideas into practice. Our goal is that you leave this workshop an advocate and early adopter for the ideas we discussed today - spread these ideas to your peers and colleagues! To realistically achieve this goal, we will talk about some of the challenges you will face so you can develop a realistic expectation of what's to come.

- **Goal**: Understand common hurdles in adopting open science and how to overcome them
- **Motivation**: Become the "open science" expert at your institution!

4.2 Learning curves

Challenge

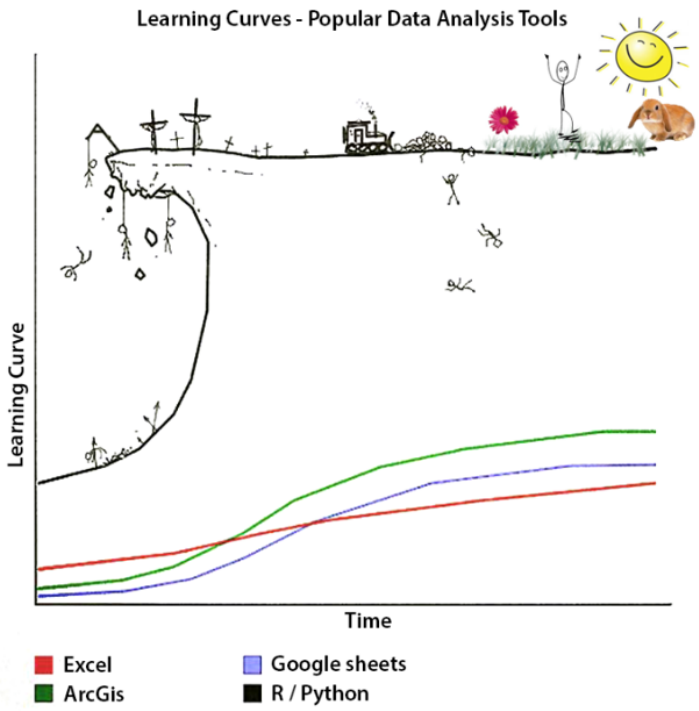
It's hard to learn new tools!

Solution

It's an investment, look to the community!

You've probably seen a graphic like this if you've ever taken a course in R or Python. The hope is that you're able to quickly reach the land of sunshine and bunnies, but the path is treacherous and even insurmountable for some.

A huge obstacle in using open science is that the toolsets can have steep learning curves. More popular platforms, such as Excel, are used by many because they're simple and intuitive. However, as noted earlier, FAIR workflows and tools are sacrificed for ease of use.



Although it's true that adopting new tools will slow forward progress, this is only temporary. Consider your path towards learning new platforms an investment in your future. The immediate benefit may not be apparent, but you'll soon wonder how you ever got by before.

It's also helpful to think about the broader community that can support you along this journey. Learning alone can be discouraging and we strongly recommend that you tap into the diverse community of educators, mentors, bloggers, and friends that can help. Even you can create a community of practice!

Exercise and discussion

How can you engage your peers to develop a shared workspace to learn new tools? What tools will you learn?

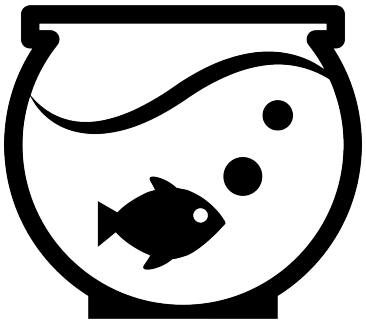
4.3 Fear of exposure

Challenge

Being open makes me nervous!

Solution


maintain
mentors!
friends
developers!
teachers!
bloggers!
contrib
i'm doing a thing all on my own!
code hero
support
community
@allison_horst



Being open helps you collaborate, increases competitiveness, and creates a better scientific product!

Practicing open science can feel like science in a fish bowl. Although this is kind of the goal, many view this transparency as a liability. Many fear having their ideas "scooped" or losing credibility because of greater exposure of mistakes. These are real concerns that require consideration when working towards more open workflows.

In conventional academic settings, competition for resources (e.g., via grant funding) is a real issue and being open can be seen as a risk to the competitive edge. We cannot dismiss this fact, but rather we can think about a lack of openness as a hindrance to forward progress and stifled creativity.

Think about being open as a means to finding your next collaborator. Creating FAIR data opens the door for others to engage with your science. In fact, being open can increase the competitiveness of research proposals by building a stronger team that collaborates and shares data through better workflows.

First time practitioners of open science also worry about the risk of "airing their dirty laundry". By exposing the process and potential mistakes, many worry that their integrity as scientists may be questioned.

These fears are unfounded as the scientific process by definition is iterative. Hypotheses are supported or refuted through trial and error - if you're getting your answer after one pass, you're probably not doing it right. Making the process more transparent can help build trust as your collaborators can better appreciate how decisions and conclusions were made.

Mistakes in research are also very common, much more so than many people realize. By being open, it is true that mistakes are more visible, but this also provides a mechanism for fixing. Being open can lead to a better product by simply having more eyes on the process. It also helps normalize mistakes as part of the process - perfection is an unrealistic expectation.

Exercise and discussion

What are your personal concerns about adopting open science?

4.4 What does it mean to be open?

Challenge

People and institutions define open differently!

Solution

Understand the context and demonstrate the value!

Also realize that open science can mean different things to different people. By extension, this also applies to institutions. We presented the five schools of open science to help conceptualize ideas and tools when we discuss what it means to different groups.

Think about your employer and what they might care about if you advocate for adoption of open science. Do you need to convince them that there is value in being open? What is their value proposition? What are the hurdles to achieving openness at your institution?

For many institutions, being open may come with IT hurdles as you push for alternative software platforms. Working with IT staff to develop trust and comfort for new software may be your burden, but as always, it's an investment in the future.

Maybe there are legal contexts to being open. For example, Florida has the "Sunshine" law that makes all government communications public record. What does this mean for using new workflows in open science? Is this is an improvement or a liability (see previous section)?

If you're an administrator or manager, maybe you're the one that makes the call about being open. It's important for you to create a culture that promotes and supports open science. Allow space and time for your staff to learn new skills. Realize that investing time in open science is an investment in the future.

Exercise and discussion

What does being open mean to you? What do you think being open means to your employer?

4.5 Something is better than nothing

Challenge

Doing all the things is impossible!

Solution

Start small, incremental progress is the name of the game!

First time open science enthusiasts can be overwhelmed by the apparent need to check all the boxes on the open science list. There's often a prevailing sentiment that you're not doing open science unless you do all the things. This is simply not true. Just remember that doing something is a huge improvement over doing nothing.

Openness in science exists on a spectrum. Your goal should be incremental movement away from the completely closed end of the spectrum. Perhaps you set a goal of only accomplishing one open science task for a particular project. Maybe you start by developing a simple metadata text file or developing a data dictionary. Or maybe you make a commitment to try a new communication platform for collaborative engagement.

Channeling this concept, Wilson et al. (2017) discuss "good enough practices" in scientific computing, acknowledging that very few of us are professionally trained in these disciplines and sometimes "good enough" is all we can ask for. Lowenberg et al. (2021) also advocate for simple adoption, rather than perfection, when it comes to data citation practices.

Also, be mindful of complacency (and apathy, at its very worst). Just because you think you've mastered a task doesn't mean you can't continue to learn. Always strive to improve yourself and the tools you use to be open. The fact that the toolbox is constantly evolving makes this a necessity.

So, be kind to yourself when learning new skills and realize that the first step will likely be frustration, but through frustration comes experience. The more comfortable you become with a task, the more likely you'll attempt new tasks in the future. I promise you will see a return on your investment.

◢ Exercise and discussion

What are some simple things you can do to begin adopting open science?

5 Setup for the Workshop

Thanks for your interest in the open science workshop. You will need to do the following, outlined below, before the workshop. The last item is optional, but strongly encouraged.

1. Install R: link
2. Install RStudio: link
3. Install Quarto: link
4. GitHub create account: link
5. Install Git and configure with RStudio (optional): link

Most of these steps will require administrative privileges on a computer. Please work with your IT staff to complete the setup if you do not have these privileges. Please reach out if you have any issues with installation: mbeck@tbep.org

5.1 Install R and RStudio

R and **RStudio** are separate downloads and installations. R is the underlying statistical computing software. RStudio is a graphical integrated development environment (IDE) that makes using R much easier and more interactive. *You need to install R before you install RStudio.*

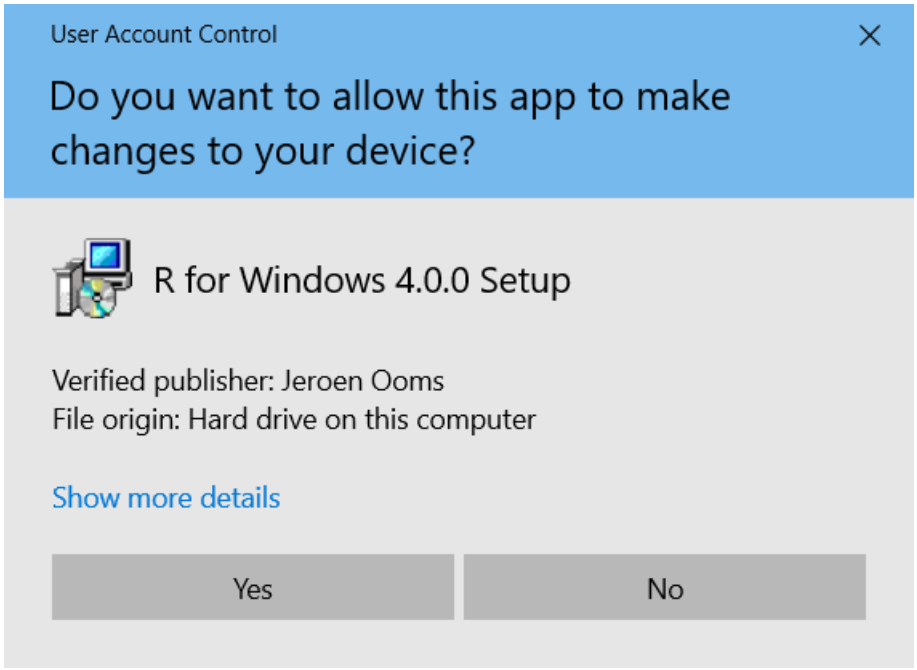
Thanks to the USGS-R Training group and Data Carpentry for making their installation materials available. The following instructions come directly from their materials, with a few minor edits to help you get set up.

5.1.1 *Windows*: Download and install R

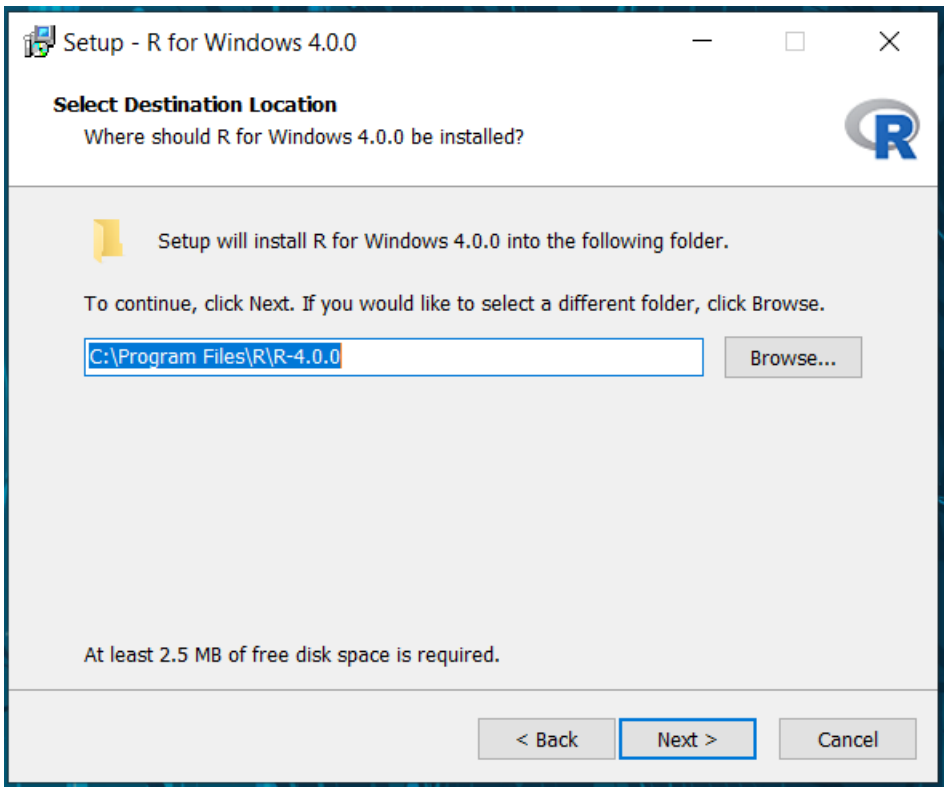
Go to CRAN and download the R installer for Windows. Make sure to choose the latest stable version (v4.2.3 as of April 2023).

Once the installer downloads, Right-click on it and select "Run as administrator".

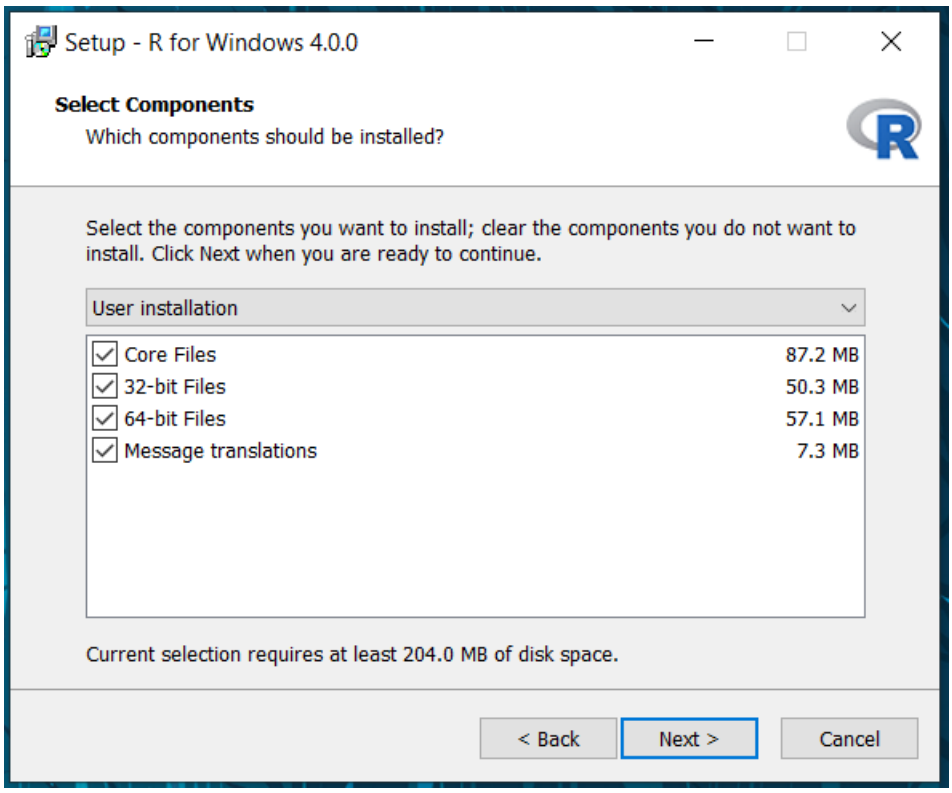
Type in your credentials and click yes (or if you don't have administrator access have your IT rep install with Admin privileges).



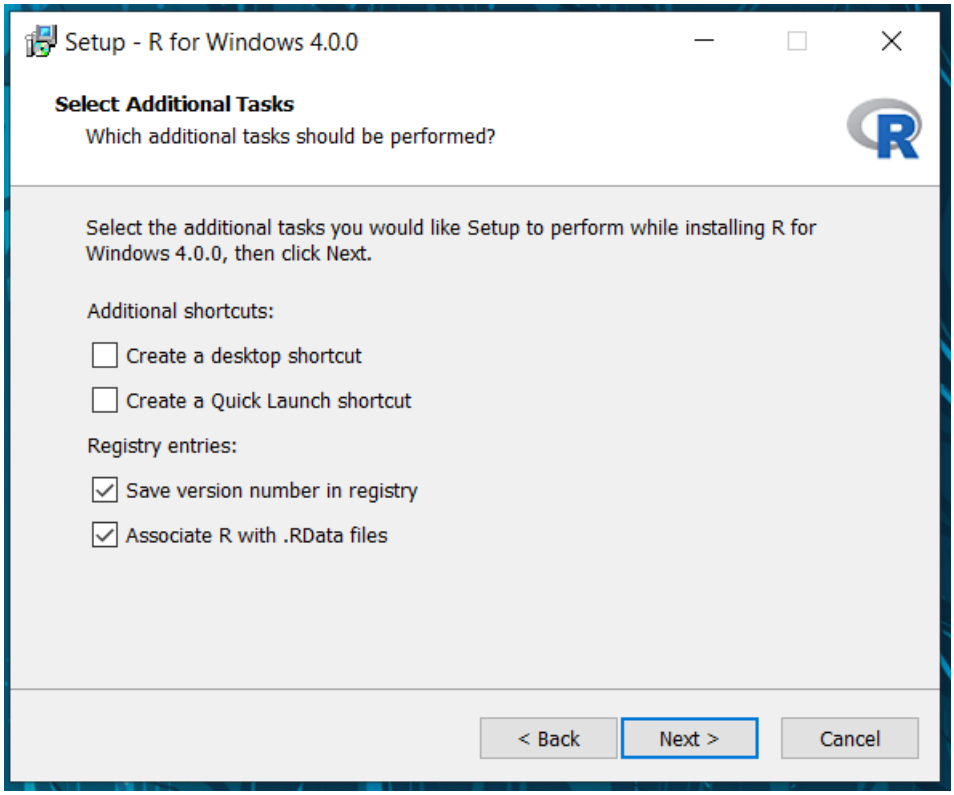
You can click next through the standard dialogs and accept most defaults. But at the destination screen, please verify that it is installing it to C:\Program Files\R



At the "Select Components" screen, you can accept the default and install both 32-bit and 64-bit versions.



At this screen, uncheck 'Create a desktop icon' because non-admin users in Windows will be unable to delete it.

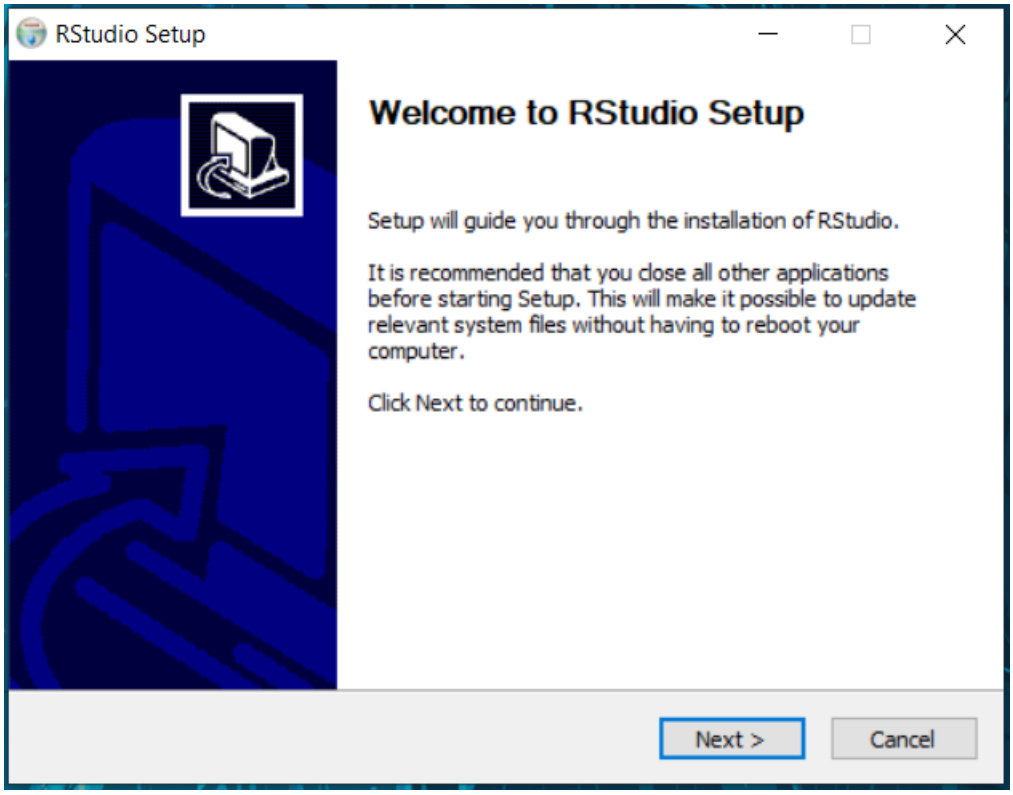


5.1.2 *Windows*: Download and install RStudio

Download RStudio from here.

After download, double-click the installer. It will ask for your administrator credentials to install (you might need to have your IT rep install again).

Accept all the default options for the RStudio install.



5.1.3 *macOS*: Download and install R

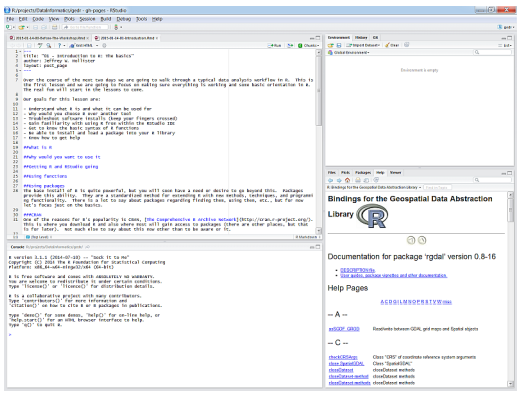
- Download and install R from the CRAN website for Mac here.
- Select the `.pkg` file for the latest R version
- Double click on the downloaded file to install R
- It is also a good idea to install XQuartz (needed by some packages)

5.1.4 *macOS*: Download and install RStudio

- Go to the RStudio download page
- Under Installers select **RStudio x.yy.zzz - Mac OS X 10.6+ (64-bit)** (where x, y, and z represent version numbers)
- Double click the file to install RStudio

5.1.5 Check Install

Once installed, RStudio should be accessible from the start menu. Start up RStudio. Once running it should look something like this:



5.2 Install Quarto

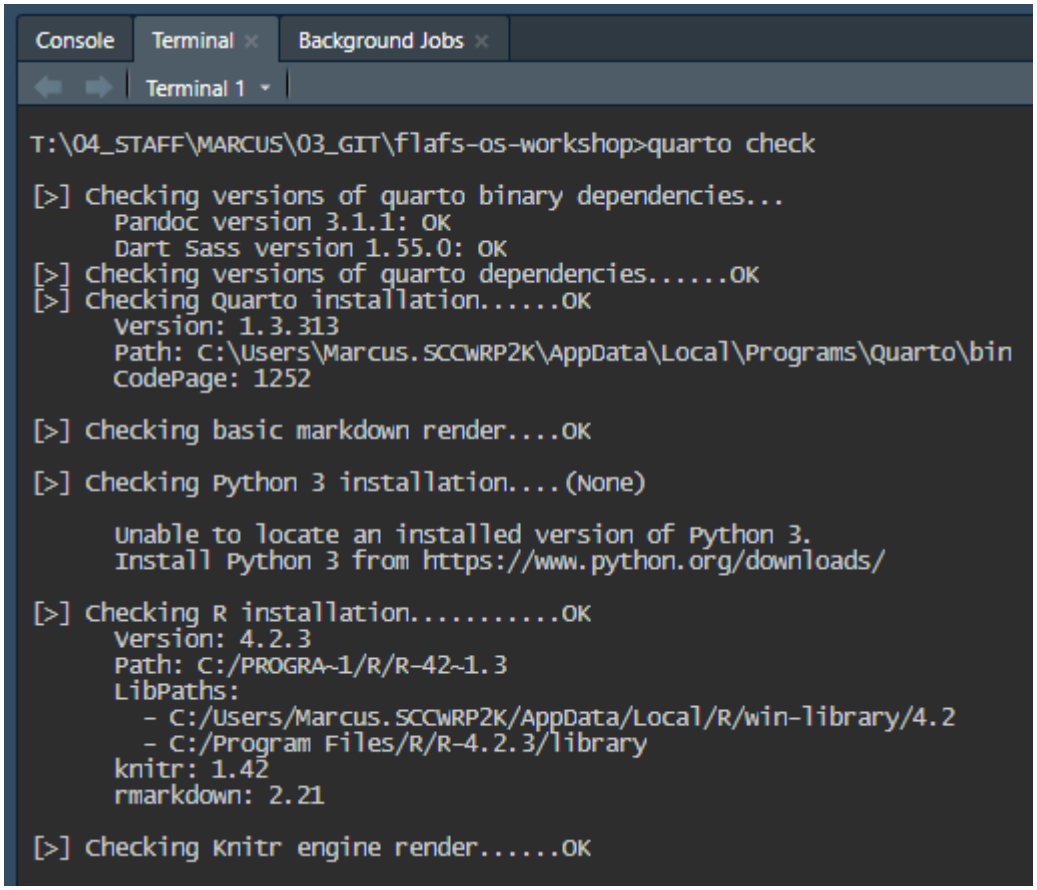
A visual editor for Quarto is installed with RStudio. However, you'll need to install Quarto CLI to make full use of its features.

Navigate to https://quarto.org/docs/get-started/. You'll see a screen that looks like this:

Platform	Download	Size	SHA-256
Ubuntu 18+/Debian 10+	quarto-1.3.318-linux-amd64.deb	84.57 MB	8d48b9c
Linux Arm64	quarto-1.3.318-linux-arm64.deb	84.4 MB	d9ab433
Mac OS	quarto-1.3.318-macos.pkg	105.26 MB	2a2e815
Windows	quarto-1.3.318-win.msi	77.82 MB	374e0bb
	Release notes and more downloads…		

Select the download appropriate for your operating system (Windows is the big blue button). After the executable file is downloaded, navigate on your computer to the folder containing the file, double-click to install, and accept the default settings at the prompts.

After installation is done, open RStudio and select the Terminal tab (usually bottom-left pane, next to the Console tab). Type `quarto check` and press enter at the prompt. You should see something like this if installation was successful.



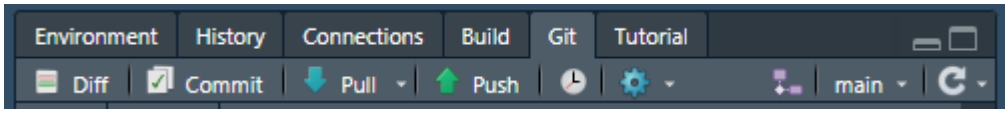
5.3 Create GitHub account

Open a web browser and open the url https://github.com. On the top-right, you should see a button to sign up. Click the button and register an account by choosing an email, username, and password.

5.4 Install Git (optional)

After you've registered a new GitHub account, you also need to install the Git software on your computer. Open the url https://git-scm.com/book/en/v2/Getting-Started-Installing-Git and follow the instructions for your operating system.

After Git is installed, open RStudio to verify the installation. You should see a new "Git" tab usually located in the top-right pane of RStudio.



5.4.1 Make sure RStudio can talk to GitHub via Git (optional)

The next step can be a bit tricky, but is essential if you want to push files from your computer to GitHub using Git. First, install the `usethis` R package.

```
install.packages("usethis")
```

You must let Git know who you are and that you have permission to write to a GitHub repository. First, let Git know who you are, where you enter your user name and email associated with the account from the previous step.

```
usethis::use_git_config(user.name="Jane Doe", user.email="jane@example.org")
```

Next, you need to setup a personal access token (PAT) that defines the permissions to write to a repository. This can be done as follows:

```
usethis::create_github_token()
```

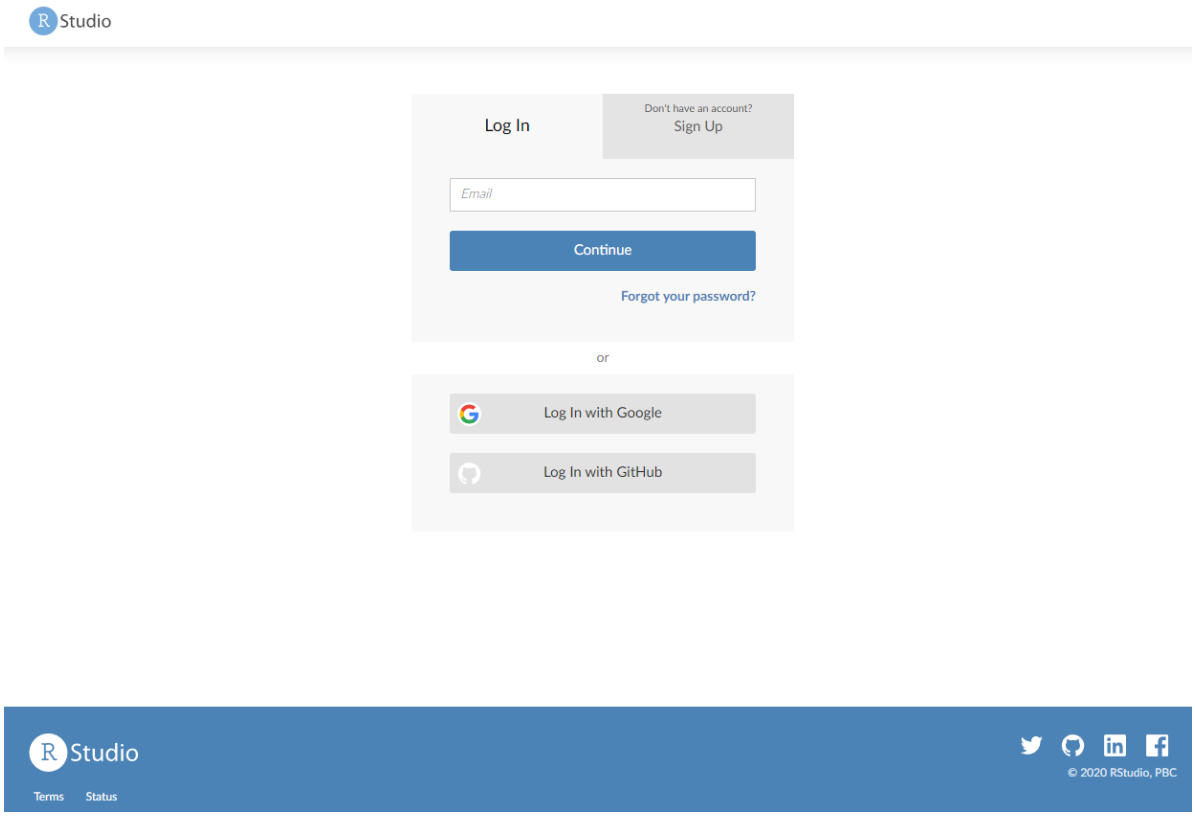
Then follow the remaining prompts to complete the PAT creation. A more thorough explanation can be found here.

5.5 This is hard!

If you have trouble installing any of the software prior to the workshop, you can use use RStudio in the cloud on the Posit website. This is only a backup option and we strongly encourage you to troubleshoot the installation when able.

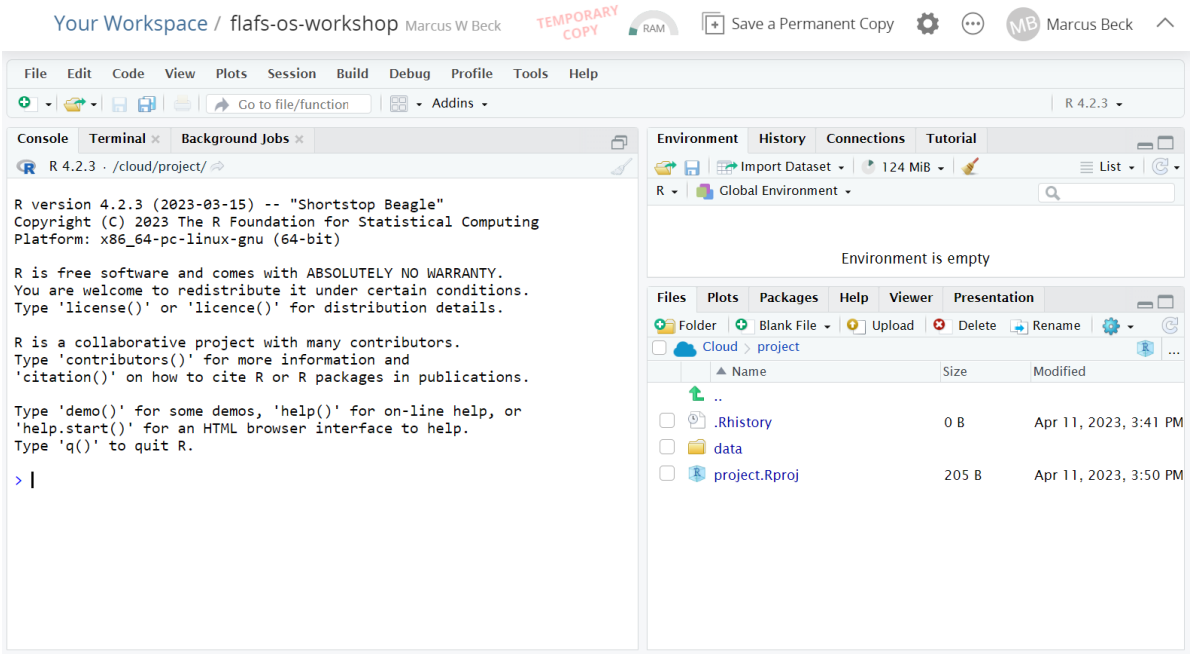
To use RStudio in the cloud, copy this link and paste it in a web browser: https://posit.cloud/content/5775087

If you do not have a Posit Cloud account, you will see this screen when you first visit the URL:



You can setup an account for free using a login you create or through a third-party (Google or GitHub).

After your account is setup, you should a screen that looks something like this:



You'll see that this is a TEMPORARY COPY under your account. Make it permanent by clicking the button on top. This will save any changes you make to this project under your account.

6 Introduction to R

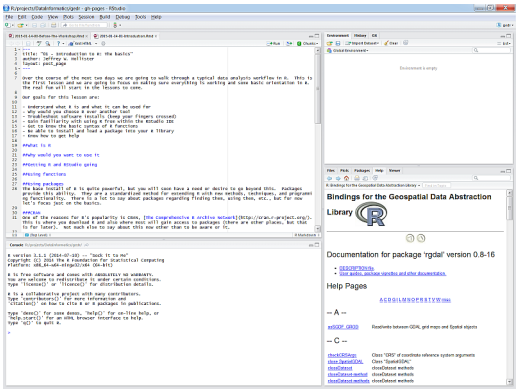
6.1 RStudio

RStudio is the go-to Interactive Development Environment (IDE) for R. Rstudio includes many features to improve the user's experience.

Let's get familiar with RStudio.

6.1.1 Open R and RStudio

Find the RStudio shortcut on your computer and fire it up. You should see something like this:



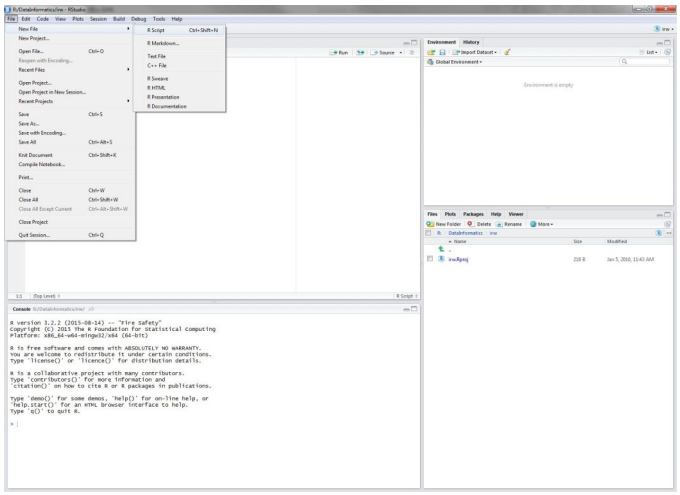
There are four panes in RStudio:

- **Source**: Your primary window for writing code to send to the console, this is where you write and save R "scripts"
- **Console**: This is where code is executed in R
- **Environment, History, etc.**: A tabbed window showing your working environment, code execution history, and other useful things
- **Files, plots, etc.**: A tabbed window showing a file explorer, a plot window, list of installed packages, help files, and viewer

6.1.2 Scripting

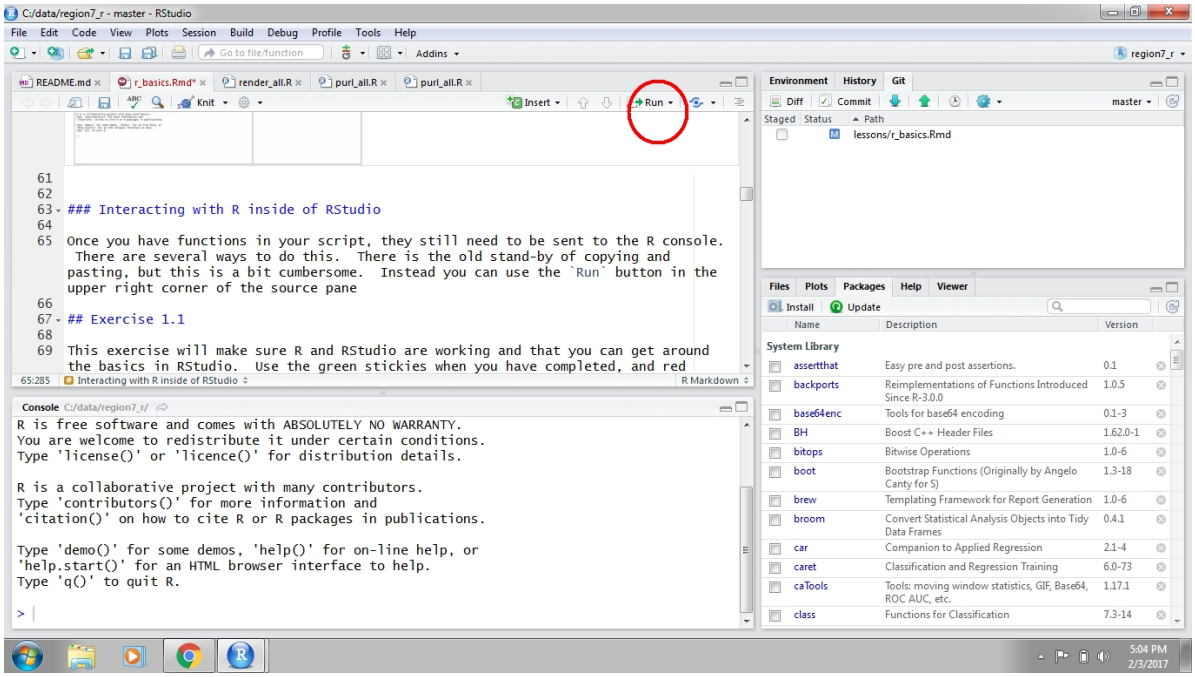
In most cases, you will not enter and execute code directly in the console. Code can be written in a script and then sent directly to the console.

Open a new script from the File menu…



6.1.3 Executing code in RStudio

After you write code in an R script, it can be sent to the Console to run the code. There are two ways to do this. First, you can hit the `Run` button at the top right of the scripting window. Second, you can use `ctrl+enter` (`cmd+enter` on a Mac). Either option will run the line(s) of script that are selected.



6.2 R language fundamentals

R is built around functions. The basic syntax of a function follows the form: `function_name(arg1, arg2, ...)`.

With the base install, you will gain access to many functions (2321, to be exact). Some examples:

```
# print
print("hello world!")
```

```
[1] "hello world!"
```

```
# sequence
seq(1, 10)
```

```
 [1]  1  2  3  4  5  6  7  8  9 10
```

```r
# random numbers
rnorm(100, mean = 10, sd = 2)
```

```
  [1] 10.734358  7.906393  9.589319  6.474589 14.316268 12.542258 10.054952
  [8]  9.608234  7.126527 11.735630  8.268890  7.054800  9.650335  9.766843
 [15]  7.798853 10.776536  6.993406 13.773328 10.982464 11.663124 12.204990
 [22]  9.671264  7.539526 12.783312 11.849491 12.480539  8.652831 10.830023
 [29]  7.478186 10.139869  8.946933 10.813591 10.043339 11.913794  8.955364
 [36] 11.240398 11.579844  6.171031  9.776879 12.519216 10.095212 12.732183
 [43] 10.711508 11.233507  7.339925  5.497580  8.388487  9.305817  9.978814
 [50] 10.978034  8.442679 14.388395  9.314340 11.869395  8.180452 10.593033
 [57]  7.842708  9.561013 10.606765  7.988650 10.159884  8.430801  8.912602
 [64] 12.325345  8.837551 10.491012 11.007816 12.756156 10.031707  9.989608
 [71]  7.486096 10.998964 12.316410  7.633556  8.045620  8.503892 10.626919
 [78] 10.542669 10.796047  6.439851  8.003476  8.381836 11.508068 11.275246
 [85] 10.171023  8.665794  9.767831 10.808862 10.945925 10.769616 13.459629
 [92]  7.519122 11.628666  8.912998 10.063868 10.864173  9.223377  8.157556
 [99]  7.935373 11.281122
```

```r
# average
mean(rnorm(100))
```

```
[1] -0.09615084
```

```r
# sum
sum(rnorm(100))
```

```
[1] 5.10136
```

Very often you will see functions used like this:

```r
my_random_sum <- sum(rnorm(100))
```

The first part of the line is the name of an object that you make up. The second bit, `<-`, is the assignment operator. This tells R to take the result of `sum(rnorm(100))` and store it in an object named, `my_random_sum`. It is stored in the environment and can be used by just executing it's name in the console.

```r
my_random_sum
```

```
[1] -14.55022
```

6.2.1 What is the environment?

There are two outcomes when you run code. First, the code will simply print output directly in the console. Second, there is no output because you have stored it as a variable using `<-`. Output that is stored is saved in the `environment`. The environment is the collection of named objects that are stored in memory for your current R session.

6.3 Packages

The base installation of R is quite powerful. Packages allow you to include new methods for use in R.

6.3.1 CRAN

Many packages are available on CRAN, The Comprehensive R Archive Network. This is where you download R and also where most will gain access to packages. As of 2023-04-11, there are 19322 packages on CRAN!

6.3.2 Installing packages

When a package gets installed, that means the source code is downloaded and put into your library. A default library location is set for you.

We use the `install.packages()` function to download and install a package. Here, we install the readxl package, used below, which is used to upload data from and Excel file.

```
install.packages("readxl")
```

You should see some text in the R console showing progress of the installation and a prompt after installation is done.

After installation, you can load a package using the `library()` function. This makes all functions in a package available for you to use.

```
library(readxl)
```

An important aspect of packages is that you only need to download them once, but every time you start RStudio you need to load them with the `library()` function.

6.4 Data structures in R

Now we can talk about R data structures. Simply put, a data structure is a way for programming languages to handle information storage.

6.4.1 Vectors (one-dimensional data)

The basic data format in R is a vector - a one-dimensional grouping of elements that have the same type. These are all vectors and they are created with the `c` (concatenate) function:

```
dbl_var <- c(1, 2.5, 4.5)
int_var <- c(1L, 6L, 10L)
log_var <- c(TRUE, FALSE, T, F)
chr_var <- c("a", "b", "c")
```

The four types of vectors are `double` (or numeric), `integer`, `logical`, and `character`. The following functions can return useful information about the vectors:

```
class(dbl_var)
```

```
[1] "numeric"
```

```
length(log_var)
```

```
[1] 4
```

6.4.2 Data frames (two-dimensional data)

A collection of vectors represented as one data object are often described as two-dimensional data, like a spreadsheet, or in R speak, a data frame. Here's a simple example:

```
ltrs <- c("a", "b", "c")
nums <- c(1, 2, 3)
logs <- c(T, F, T)
mydf <- data.frame(ltrs, nums, logs)
mydf
```

```
  ltrs nums  logs
1    a    1  TRUE
2    b    2 FALSE
3    c    3  TRUE
```

The only constraints required to make a data frame are:

1. Each column (vector) contains the same type of data
2. The number of observations in each column is equal.

6.5 Getting your data into R

It is the rare case when you manually enter your data in R. Most data analysis workflows typically begin with importing a dataset from an external source. We'll be using `read_excel()` function from the readxl package.

We can import the `ExampleSites.xlsx` dataset as follows. Note the use of a *relative* file path. You can see what R is using as your "working directory" using the `getwd()` function.

```
sitdat <- read_excel("data/ExampleSites.xlsx")
```

Let's explore the dataset a bit.

```
# get the dimensions
dim(sitdat)
```

```
[1] 11  5
```

```
# get the column names
names(sitdat)
```

```
[1] "Monitoring Location ID"        "Monitoring Location Name"
[3] "Monitoring Location Latitude"  "Monitoring Location Longitude"
[5] "Location Group"
```

```
# see the first six rows
head(sitdat)
```

```
# A tibble: 6 x 5
  `Monitoring Location ID` `Monitoring Location Name` Monitoring Location Lati~1
  <chr>                    <chr>                                          <dbl>
1 ABT-026                  Rte 2, Concord                                  42.5
2 ABT-062                  Rte 62, Acton                                   42.4
3 ABT-077                  Rte 27/USGS, Maynard                            42.4
4 ABT-144                  Rte 62, Stow                                    42.4
5 ABT-237                  Robin Hill Rd, Marlboro                         42.3
6 ABT-301                  Rte 9, Westboro                                 42.3
# i abbreviated name: 1: `Monitoring Location Latitude`
# i 2 more variables: `Monitoring Location Longitude` <dbl>,
#   `Location Group` <chr>
```

```
# get the overall structure
str(sitdat)
```

```
tibble [11 x 5] (S3: tbl_df/tbl/data.frame)
 $ Monitoring Location ID       : chr [1:11] "ABT-026" "ABT-062" "ABT-077" "ABT-144" ...
 $ Monitoring Location Name     : chr [1:11] "Rte 2, Concord" "Rte 62, Acton" "Rte 27/USGS, M
 $ Monitoring Location Latitude : num [1:11] 42.5 42.4 42.4 42.4 42.3 ...
 $ Monitoring Location Longitude: num [1:11] -71.4 -71.4 -71.4 -71.5 -71.6 ...
 $ Location Group               : chr [1:11] "Assabet" "Assabet" "Assabet" "Assabet" ...
```

You can also view a dataset in a spreadsheet style using the `View()` function:

```
View(sitdat)
```

6.6 Summary

In this intro we learned about R and Rstudio, some of the basic syntax and data structures in R, and how to import files.

7 Resources for continued learning

The following is a non-exhaustive list of additional resources you can use for continued learning on your open science journey.

7.1 Open Science Websites

- NCEAS Open Science for Synthesis workshop
- NCEAS Reproducible Research Techniques
- Open Science Foundation open science workshop
- Openscapes
- Openscapes Champions Lesson Series
- Supercharge your research: A 10 week plan for open data science
- ROpenSci guidance on creating a Code of Conduct
- NOAA Reproducible Reporting with R
- PeerJ collection on practical data science

7.2 Data Management Tools

- Environmental Data Initiative Data Management Resources
- University of California DMPTool
- US Geological Survey resources for Metadata Creation
- ELIXIR and others Data Stewardship Wizard
- TBEP Data Management SOP

7.3 TBEP R Trainings

- Peconic Estuary Program R training, recording
- TBEP June 2020 R training, recordings
- Writing functions in R
- R package development workflow
- A soft introduction to Shiny

7.4 R Lessons & Tutorials

- Software Carpentry: R for Reproducible Scientific Analysis
- Data Carpentry: Geospatial Workshop
- Data Carpentry: R for Data Analysis and Visualization of Ecological Data
- Data Carpentry: Data Organization in Spreadsheets
- R for Water Resources Data Science
- RStudio Webinars, many topics
- R For Cats: Basic introduction site, with cats!
- Topical cheatsheets from RStudio, also viewed from the help menu
- Cheatsheet from CRAN of base R functions
- Totally awesome R-related artwork by Allison Horst
- Color reference PDF with text names, Color cheatsheet PDF from NCEAS

7.5 R eBooks/Courses

- Jenny Bryan's Stat545.com
- Garrett Grolemund and Hadley Wickham's R For Data Science
- Chester Ismay and Albert Y. Kim's Modern DiveR
- Julia Silge and David Robinson Text Mining with R
- Hadley Wickham's Advanced R
- Hadley Wickham's R for Data Science
- Yihui Xie R Markdown: The Definitive Guide
- Winston Chang R Graphics Cookbook
- Wegman et al. Remote Sensing and GIS for Ecologists: Using Open Source Software
- Lovelace et al. Geocomputation with R
- Edszer Pebesma and Roger Bivand Spatial Data Science

7.6 Git/Github

- Jenny Bryan's Happy Git and Github for the useR
- Git and GitHub for the Casual User
- Coding Club Intro to Github

8 References

Beck, M. W., C. O'Hara nad J. S. S. Lowndes, R. D. Mazor, S. Theroux, D. J. Gillett, B. Lane, and G. Gearheart. 2020. "The Importance of Open Science for Biological Assessment of Aquatic Environments." *PeerJ* 8: e9539. https://doi.org/10.7717/peerj.9539.

Broman, K. W., and K. H. Woo. 2018. "Data Organization in Spreadsheets." *The American Statistician* 72 (1): 2–10. https://doi.org/10.1080/00031305.2017.1375989.

Fecher, B., and S. Friesike. 2014. "Open Science: One Term, Five Schools of Thought." In *Opening Science*, 17–47. Springer, Cham.

Gilliland, A. J. 2016. "Setting the Stage." In *Introduction to Metadata*, 3rd ed. Los Angeles, California: Getty Publications.

Hampton, S. E., S. S. Anderson, S. C. Bagby, C. Gries, X. Han, E. M. Hart, M. B. Jones, et al. 2015. "The Tao of Open Science for Ecology." *Ecosphere* 6 (7): 1–13. https://doi.org/10.1890/ES14-00402.1.

Lowenberg, Daniella, Rachael Lammey, Matthew B Jones, John Chodacki, and Martin Fenner. 2021. "Data Citation: Let's Choose Adoption over Perfection." Zenodo. https://doi.org/10.5281/zenodo.4701079.

Lowndes, J. S. S., B. D. Best, C. Scarborough, J. C. Afflerbach, M. R. Frazier, C. C. O'Hara, N. Jiang, and B. S. Halpern. 2017. "Our Path to Better Science in Less Time Using Open Data Science Tools." *Nature Ecology & Evolution* 1 (0160): 1–7. https://doi.org/10.1038/s41559-017-0160.

Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner, and S. G. Stafford. 1997. "Nongeospatial Metadata for the Ecological Sciences." *Ecological Applications* 7 (1): 330–42. https://doi.org/https://doi.org/10.1890/1051-0761(1997)007%5B0330:NMFTES%5D2.0.CO;2.

Wickham, H. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10): 1–23. https://doi.org/10.18637/jss.v059.i10.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. D'Agostino McGowan, R. François, G. Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, H., and G. Grolemund. 2017. *R for Data Science*. Sebastopol, California: O'Reilly.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (160018). https://doi.org/10.1038/sdata.2016.18.

Wilson, G., J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal. 2017. "Good Enough Practices in Scientific Computing." *PLoS Computational Biology* 13 (6): e1005510.

https://doi.org/10.1371/journal.pcbi.1005510.
Ziemann, M., Y. Eren, and A. El-Osta. 2016. "Gene Name Errors Are Widespread in the Scientific Literature." *Genome Biology* 17 (1): 1–3. https://doi.org/10.1186/s13059-016-1044-7.

9 Contributor Covenant Code of Conduct

9.1 Our Pledge

We as members, contributors, and leaders pledge to make participation in our community a harassment-free experience for everyone, regardless of age, body size, visible or invisible disability, ethnicity, sex characteristics, gender identity and expression, level of experience, education, socio-economic status, nationality, personal appearance, race, religion, or sexual identity and orientation.

We pledge to act and interact in ways that contribute to an open, welcoming, diverse, inclusive, and healthy community.

9.2 Our Standards

Examples of behavior that contributes to a positive environment for our community include:

- Demonstrating empathy and kindness toward other people
- Being respectful of differing opinions, viewpoints, and experiences
- Giving and gracefully accepting constructive feedback
- Accepting responsibility and apologizing to those affected by our mistakes, and learning from the experience
- Focusing on what is best not just for us as individuals, but for the overall community

Examples of unacceptable behavior include:

- The use of sexualized language or imagery, and sexual attention or advances of any kind
- Trolling, insulting or derogatory comments, and personal or political attacks
- Public or private harassment
- Publishing others' private information, such as a physical or email address, without their explicit permission
- Other conduct which could reasonably be considered inappropriate in a professional setting

9.3 Enforcement Responsibilities

Community leaders are responsible for clarifying and enforcing our standards of acceptable behavior and will take appropriate and fair corrective action in response to any behavior that they deem inappropriate, threatening, offensive, or harmful.

Community leaders have the right and responsibility to remove, edit, or reject comments, commits, code, wiki edits, issues, and other contributions that are not aligned to this Code of Conduct, and will communicate reasons for moderation decisions when appropriate.

9.4 Scope

This Code of Conduct applies within all community spaces, and also applies when an individual is officially representing the community in public spaces. Examples of representing our community include using an official e-mail address, posting via an official social media account, or acting as an appointed representative at an online or offline event.

9.5 Enforcement

Instances of abusive, harassing, or otherwise unacceptable behavior may be reported to the community leaders responsible for enforcement at mbeck@tbep.org. All complaints will be reviewed and investigated promptly and fairly.

All community leaders are obligated to respect the privacy and security of the reporter of any incident.

9.6 Enforcement Guidelines

Community leaders will follow these Community Impact Guidelines in determining the consequences for any action they deem in violation of this Code of Conduct:

9.6.1 1. Correction

Community Impact: Use of inappropriate language or other behavior deemed unprofessional or unwelcome in the community.

Consequence: A private, written warning from community leaders, providing clarity around the nature of the violation and an explanation of why the behavior was inappropriate. A public apology may be requested.

9.6.2 2. Warning

Community Impact: A violation through a single incident or series of actions.

Consequence: A warning with consequences for continued behavior. No interaction with the people involved, including unsolicited interaction with those enforcing the Code of Conduct, for a specified period of time. This includes avoiding interactions in community spaces as well as external channels like social media. Violating these terms may lead to a temporary or permanent ban.

9.6.3 3. Temporary Ban

Community Impact: A serious violation of community standards, including sustained inappropriate behavior.

Consequence: A temporary ban from any sort of interaction or public communication with the community for a specified period of time. No public or private interaction with the people involved, including unsolicited interaction with those enforcing the Code of Conduct, is allowed during this period. Violating these terms may lead to a permanent ban.

9.6.4 4. Permanent Ban

Community Impact: Demonstrating a pattern of violation of community standards, including sustained inappropriate behavior, harassment of an individual, or aggression toward or disparagement of classes of individuals.

Consequence: A permanent ban from any sort of public interaction within the community.

9.7 Attribution

This Code of Conduct is adapted from the Contributor Covenant, version 2.0, available at https://www.contributor-covenant.org/version/2/0/code_of_conduct.html.

Community Impact Guidelines were inspired by Mozilla's code of conduct enforcement ladder.

For answers to common questions about this code of conduct, see the FAQ at https://www.contributor-covenant.org/faq. Translations are available at https://www.contributor-covenant.org/translations.