

Tampa Bay Open Science Workshop

Dr. Chris Anastasiou and Dr. Marcus Beck

2022-09-19T10:44:37-04:00

Table of contents

Course synopsis	4
Agenda	5
Instructors	5
1 Basics of open science	7
1.1 Lesson Outline	7
1.2 Goals and motivation	7
1.3 Why open science?	7
1.4 Learning and speaking the language of open science	11
1.5 The FAIR principles	13
1.6 Schools of thought	14
2 Open science for collaboration	16
2.1 Lesson Outline	16
2.2 Goals and motivation	16
2.3 Essential elements of collaboration	16
2.3.1 Workflow management	17
2.3.2 Version control	19
2.3.3 Data repositories	21
2.3.4 Code of Conduct	22
2.4 Tools for collaboration	22
2.4.1 Slack	22
2.4.2 Trello	23
2.4.3 Google Drive	24
2.4.4 Office 365	24
2.4.5 GitHub	25
2.5 Balancing and managing expectations	25
3 Open science for impactful products	27
3.1 Lesson Outline	27
3.2 Goals and motivation	27
3.3 Data as the foundation for open science	27
3.4 Principles of tidy data	29
3.5 Data dictionaries	32
3.6 Importance of metadata	34

4 Lowering barriers to inclusion and addressing key critiques	37
4.1 Lesson Outline	37
4.2 Goals and motivation	37
4.3 Learning curves	37
4.4 Fear of exposure	40
4.5 What does it mean to be open?	41
4.6 Something is better than nothing	42
Resources for continued learning	43
Open Science Websites	43
Data Management Tools	43
TBEP R Trainings	43
R Lessons & Tutorials	44
R eBooks/Courses	44
Git/Github	44
References	45

Course synopsis

OPEN SCIENCE

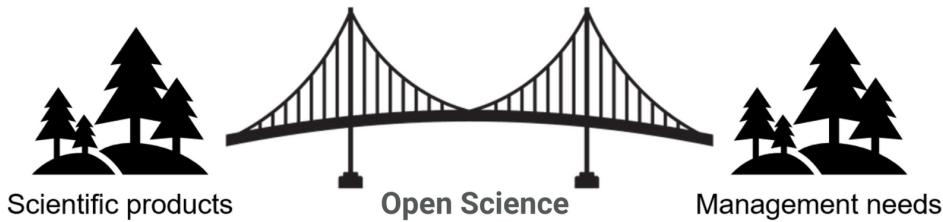
CORE CONCEPTS FOR IMPACTFUL RESEARCH AND RESOURCE MANAGEMENT



Dr. Chris Anastasiou



Dr. Marcus Beck



Half-day Workshop
February 2023, Tampa, Florida



Welcome to the Tampa Bay open science workshop: core concepts for impactful research and resource management. Open science (OS) has been advocated as an effective approach to create reproducible, transparent, and actionable research products. However, widespread adoption among the coastal and estuarine research 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 a half-day of 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 solid 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.

Agenda

1. [The basics of open science \(30 min\)](#)
2. [Open science for collaboration \(90 min\)](#)
3. [Open science for impactful products \(90 min\)](#)
4. [Lowering barriers to inclusion and addressing key critiques \(30 min\)](#)

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

Instructors

[Dr. Chris Anastasiou](#) is a Chief Water Quality Scientist and the Seagrass Mapping Program Lead for the Southwest Florida Water Management District. He holds a PhD in Marine Science from the University of South Florida and has been working in and around the springs of Florida for more than 25 years.

[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.

This book is licensed under a Creative Commons Attribution 4.0 International License.

1 Basics of open science

1.1 Lesson Outline

- Goals and motivation
- Why open science?
- Learning and speaking the language of open science
- The FAIR principles
- Schools of thought

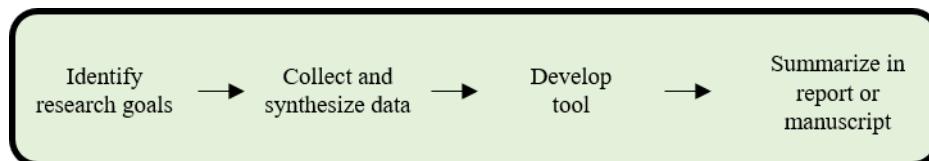
1.2 Goals and motivation

This is the first module in our four hour 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!

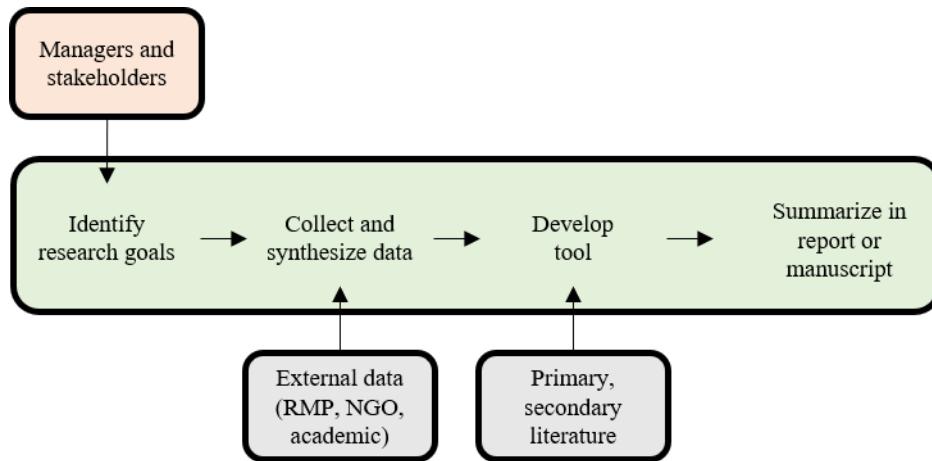
1.3 Why open science?

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?

As I’m sure we’re aware, 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.



Scientific products



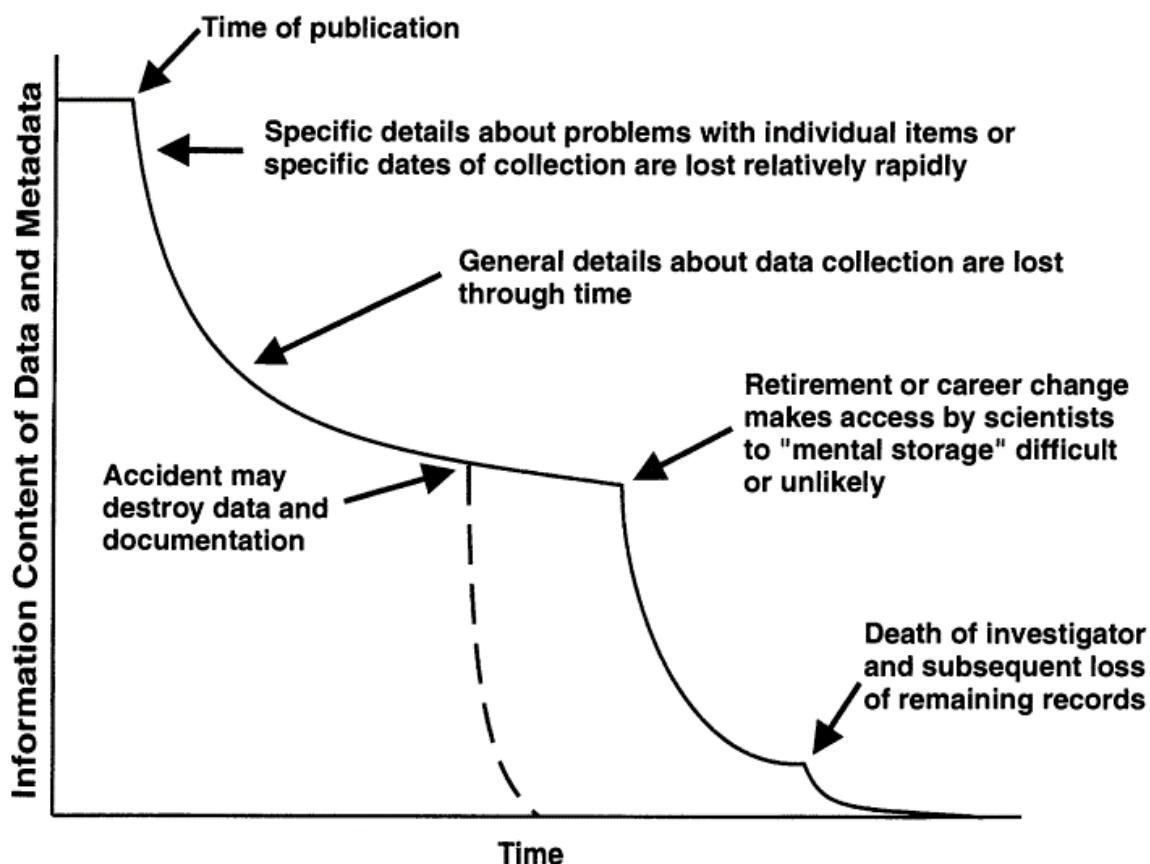
Communication barriers
Irreproducible results
Information loss
Inaccessible data
Opaque workflows



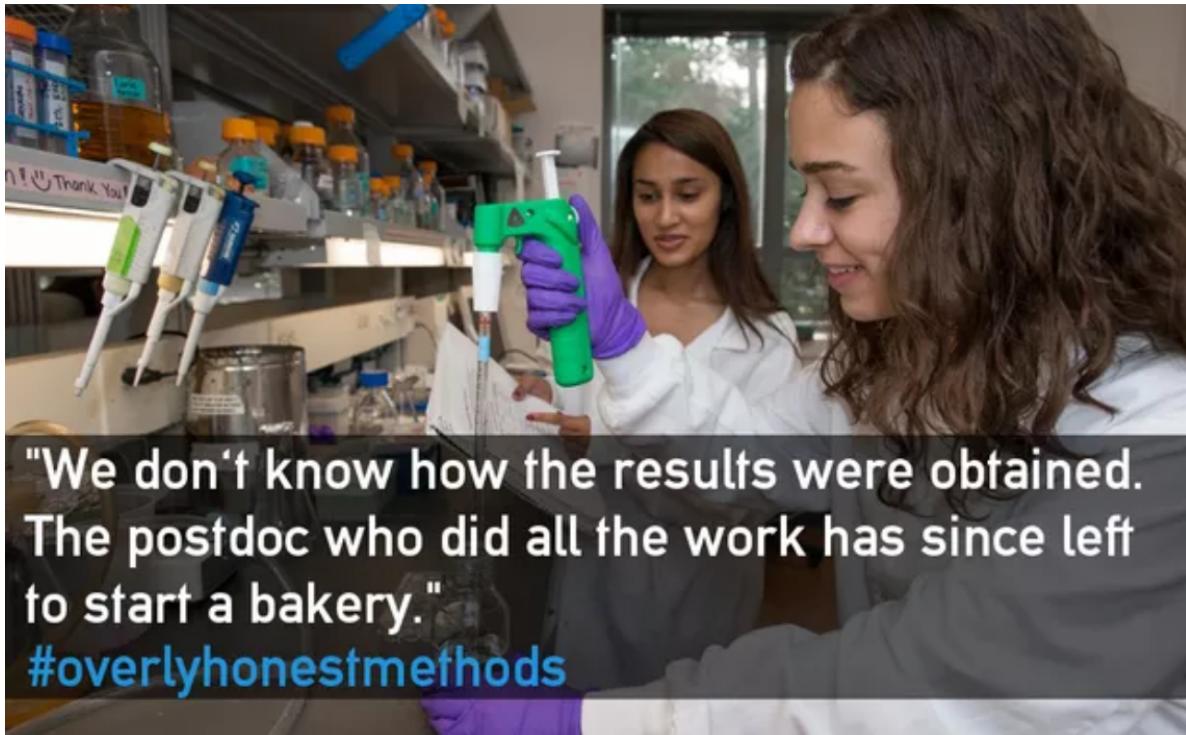
Management needs

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.

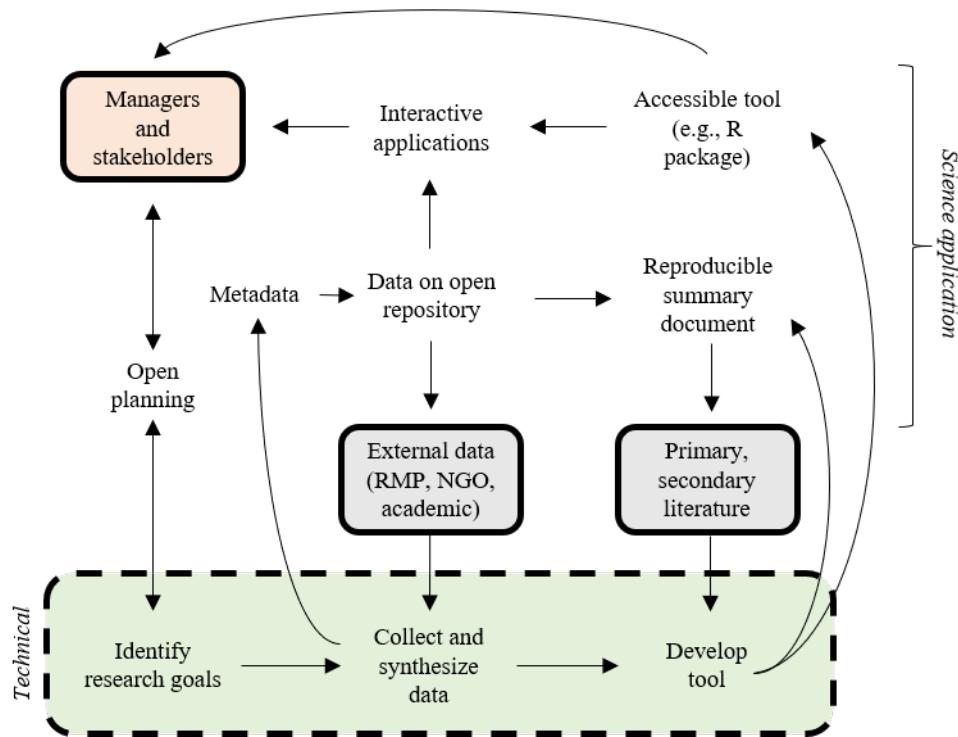


So how do we make changes to our workflows to ensure we can achieve truly applied science using open tools and philosophies?

1.4 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.

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. More importantly, the process is *iterative* and not static.

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 highlighted in bold. There are very specific tools in the open science toolbox that speak to each of these key words. We'll cover some of these later.

From these key words, we can breakdown this definition 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 and open science often focuses on how the two can leverage and support one another despite the latter being a relatively new addition to society. People often describe web-based tools as synonymous with open science.

1.5 The FAIR principles

Advocates of open science also describe the use of 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.

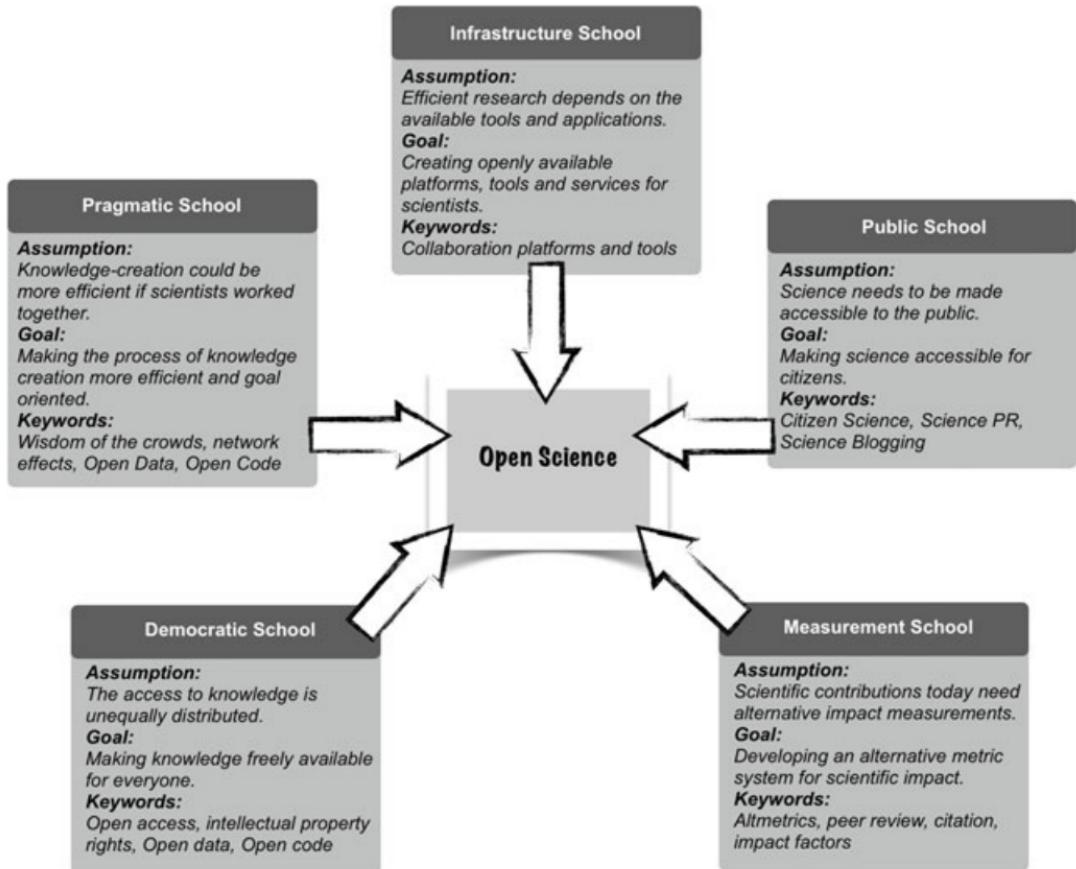
Simply, what this means is: 1) each dataset has a name that doesn't change and can be found with minimal effort using that name, 2) once it's found, you can actually get your hands on it (e.g., not behind a paywall), 3) once you have it, you can use readily available tools to work with the data (e.g., 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 does not achieve FAIR status, 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.

1.6 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 be talking about ideas and tools from the pragmatic, infrastructure, and democratic schools of the 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.

A Exercise and discussion

How do you currently define open science? 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 Lesson Outline

- Goals and motivation
- Essential elements of collaboration
- Tools for collaboration
- Balancing and managing expectations

2.2 Goals and motivation

This is the second module in our four hour 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. We'll close with a discussion on balancing and managing expectations for collaboration in the rapidly evolving online arena.

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

2.3 Essential elements of collaboration

We start our deep dive into open science by focusing on collaboration as the single most important 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.

2.3.1 Workflow management

How do you organize your work each day? How do you make sure pressing deadlines are met on schedule? 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 that was 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?

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.

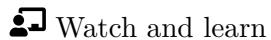
Collaborative work can be facilitated through workflow management that helps you break out of old habits. We'll introduce some specific internet-based tools below to facilitate workflows either for yourself or, better yet, working with others. These can help propel you towards open science.

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:

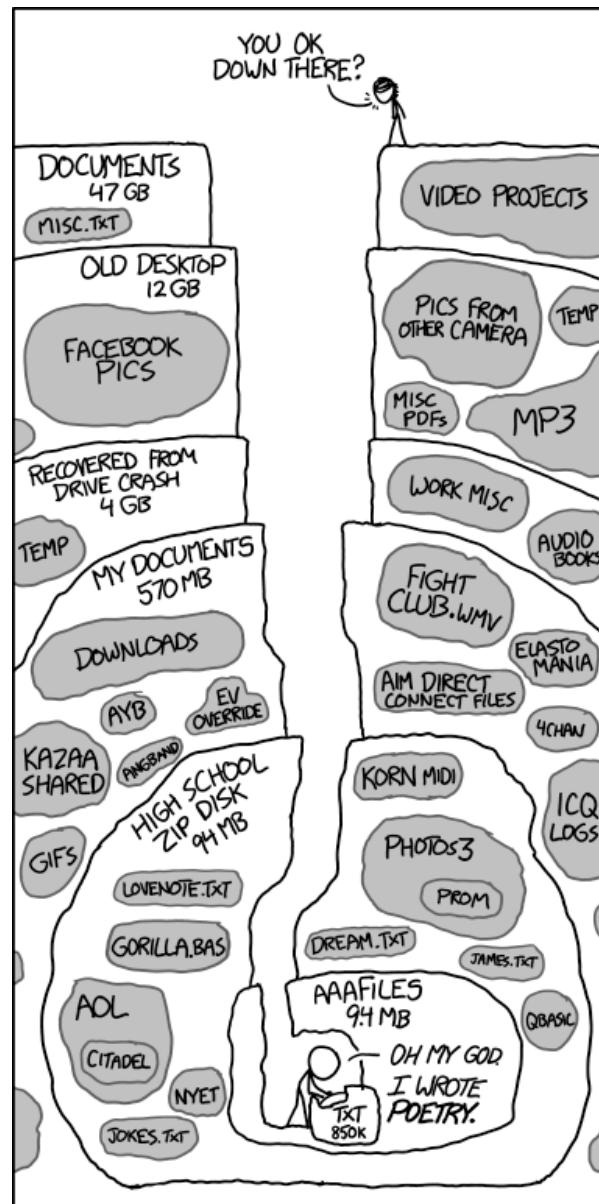
As shown, this approach can work as a literal, physical board or as one used digitally through a web browser or other software. Every Kanban board has the following elements that allow you to work in a more informed manner:

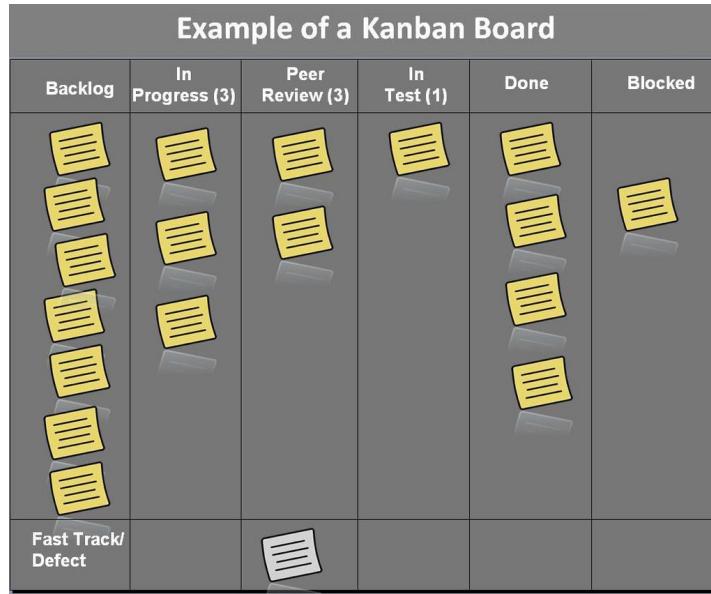
- Provides a “big picture” of progress
- Organizes progress by discrete steps
- Establishes cards as specific tasks

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.



Workflow management in the real world - using GitHub to collaborate.





2.3.2 Version control

A specific problem for workflow management that can be solved by open science tools is file management. Workflows can greatly be 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.

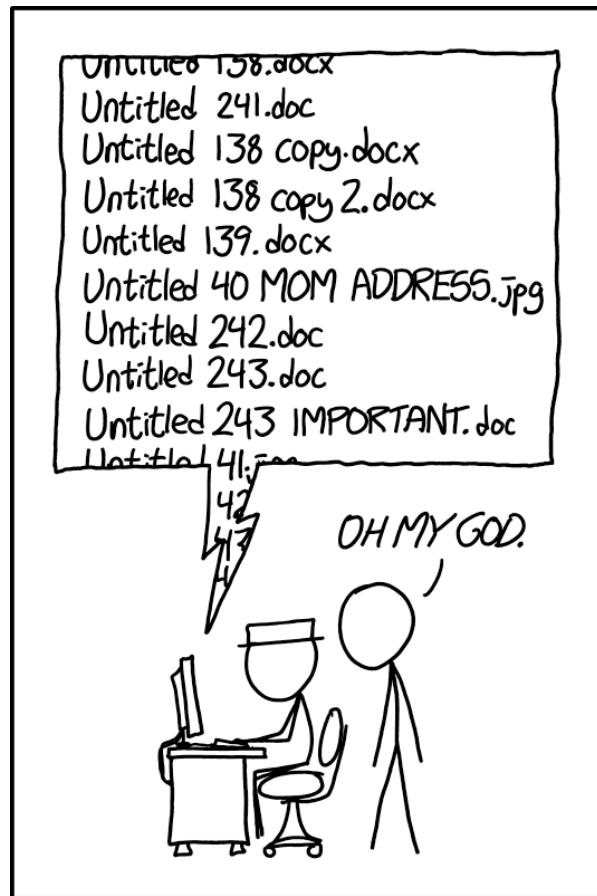
I'm sure many of you have fallen into this trap:

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!)

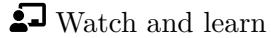
It's more than saving files. Documenting changes with a set of commands that follow strict rules provides a transparent record for yourself and others, and establishing permanency ensures that any of the changes that are made can be vetted and accessed as needed. Think of it as an insurance plan for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

By far, 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](#).



PROTIP: NEVER LOOK IN SOMEONE
ELSE'S DOCUMENTS FOLDER.

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 through better collaboration.



Watch and learn

Now we'll demonstrate how to setup a version control project with RStudio, Git, and GitHub.

2.3.3 Data repositories

How data are treated as living, dynamic pieces of information is critical to the whole 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 all 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 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 live forever so long as the internet exists. 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).

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

2.3.4 Code of Conduct

Every great collaborative team does not begin work before a Code of Conduct is created. 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.

Exercise and discussion

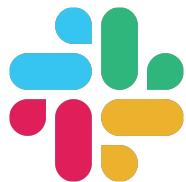
Develop a code of conduct for your group in a shared workspace. Items to consider:

1. What are reasonable expectations for the group members?
2. How is inclusion defined and encouraged?
3. How are similarities and differences recognized?
4. How will conflicts be managed?

2.4 Tools for collaboration

Now we introduce some specific web-based tools that you can use to improve collaboration and openness. 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 to leverage open science.

2.4.1 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 only those individuals in a workspace.

2.4.2 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”.

2.4.3 Google Drive



<https://google.com/drive>

?

What

Cloud-based platform for sharing documents, worksheets, slides, etc. Follows a familiar file-based GUI 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.

2.4.4 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 (should be most folks).

👎 Cons

Requires a Microsoft account and access can be tricky depending on institution. Maintains dependency on expensive Microsoft products that don't facilitate reproducibility. Very often used in closed workflows.

2.4.5 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.

💻 Watch and learn

Setting up a project management system with GitHub.

✍ Exercise and discussion

In small groups, setup a shared workspace using GitHub and create a project management board.

2.5 Balancing and managing expectations

Adopting new workflows and tools to facilitate collaboration are steps in the right direction for open science. Clearly, change from established norms and practices is not easy. It's important to develop a realistic expectation of how this change may play out in the wild.

First, realize that these changes are not going to happen overnight. Breaking free of entrenched workflows is like quitting bad habits. It's not easy because they're comfortable, familiar, and

often habitual. Start small and work gradually towards adopting new ideas. Incremental progress is the name of the game.

Getting others on board is another serious challenge. The sources of frustration you might have at the personal level apply to anyone else working in your team. Even more so, institutional roadblocks may exist. New software can raise red flags for IT support staff. It's important to work with them to develop trust for the software and an accepted process for installation.

Finally, the tools above are temporary. The single constant in open science is change. Use the tools with full realization that they may be appropriate for now, but something (hopefully) better will replace it in the future.

3 Open science for impactful products

3.1 Lesson Outline

- Goals and motivation
- Data as the foundation for open science
- Principles of tidy data
- Data dictionaries
- Importance of metadata

3.2 Goals and motivation

This is the third module in our four hour 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 concepts and tools to make sure your data have a life 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

3.3 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 several 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.

**Data sets
in tutorials**



**Data sets in
the wild**



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.

Anybody who has ever worked with data knows that it comes in many shapes and sizes, most more like the right side of the above picture. Poor data management occurs for many reasons, but here's a few of the common reasons:

- 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 the actual analysis and interpretation part 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 and future collaborators. 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 help you work towards putting FAIR data into practice. In this module, we'll go a step further to discuss how data structure, including metadata, can produce a FAIR dataset.

3.4 Principles of tidy data

At their core, tabular data allow you to store information as observations in rows and variables in columns, yet it's very common to try to make a data table 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 the abuse of tables, often to the detriment of science (Ziemann, Eren, and El-Osta 2016).



slate

@PleaseBeGneiss

excel: is that a date?

me: 57.39 is very much not a date

excel: strong date vibes to me

me: h-how

excel: fixed it

me: 57/39/2020?

excel: you're welcome

Excel allows you to abuse your data in many ways, such as adding color to cells, embedding formulas, and automatically formatting cell types. The problem occurs when this organization becomes 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. What do we mean by this?

A **rectangular** file:

	A	B	C	D	E	F
1	OTBT gmtpm& mnpm					
2		OTBT Rzpsmr	OTBT N/B	Tsct OTBT NB		
3		11434	64037	200000		
4	Osrkzt vglmOzs mnpm					
5		Lsst 12 Ognths sszs snd rzpsmr zxpqsmrz pzmrgd cslcmstmgn				
6		Tgtst bssz svzrsqz rmntmOz	Bzigrz 12 0	R/Orth -1	R/Orth -2	
7	yslzs bmmlt:	1412 11.97	1400	0	0	
8	Tgtst rzpsmr bmmlt:	306 4.36	0	2	23	
9	ATT bmmlt mnpm	272 4.01	0	1	15	
10	Bzndgr 2 bmmlt mnpm	0 #DIV/0!	0	0	0	
11	Bzndgr 3 bmmlt mnpm	34 7.18	0	1	8	
12	RzAsmR mNAUT					
13		Rzpsmr mssgz Rzpsmrzd DOs	Rzpsmrzd Wsrr	Rzpsmr pgpmstmgn		
14		Tgtst Rzpsmr	NgrOsl Rzpsmr Lgcs1 Rzpsmr			
15	306	2	40	264	0	
16	Argdmct mnTgrOstmgn					% cmrrznt mnstslzd bssz sTtzr Tmtmrz ssz
17		Argdmct NsOz ytylz NsOz	Rzvzw Dstz	Ognths scmvz svzr Tmtmrz rmntmOz		
18		DzhamOmdm1sB6065903CT	25/06/1998	76.8	27	
19	Asrt NmObzr mnTgrOstmgn					Ognth/dsylyzsr
20		sDy Asrt NmObz yzt-mp Dstz	ytsndsr dgst	Ognthly mssgz 12 Ognth's mssgz		
21		43588136-00.1/02/1992	675.92	9.7	116.4	
22			Ognth/dsylyzsr			
23						

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 **flat** file:

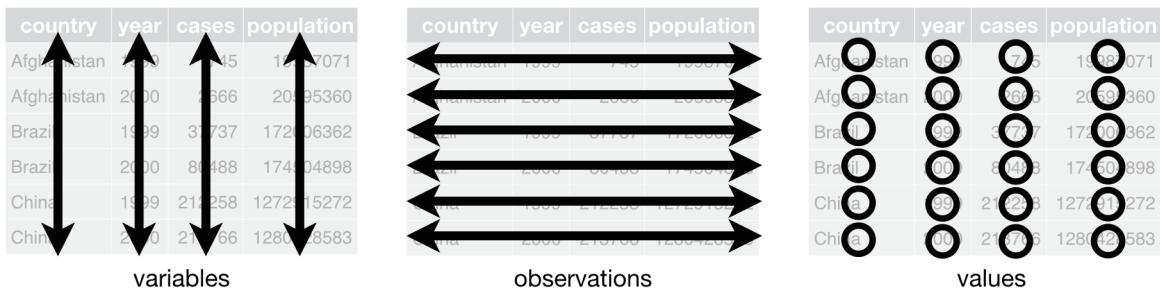
No cell formatting, no embedded formulas, no multiple spreadsheets in the same file, and data entered only as alphanumeric characters.

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 for data storage 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 each column.

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.

Tidy table 1

Station	Latitude	Longitude
A	27.89	-82.48
B	27.86	-82.47

Tidy table 2

Station	Day	Temperature
A	Monday	78.2
A	Tuesday	81.3
B	Monday	68.5
B	Tuesday	68.2

Joined tidy tables by station key

Station	Latitude	Longitude	Day	Temperature
A	27.89	-82.48	Monday	78.2
A	27.89	-82.48	Tuesday	81.3
B	27.86	-82.47	Monday	68.5
B	27.86	-82.47	Tuesday	68.2

Watch and learn

Making an [untidy](#) dataset [tidy](#) using Excel.

Watch and learn

Making an [untidy](#) dataset [tidy](#) using R.

```

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') %>%
  separate(col = `acres/category`, into = c('acres', 'category'), sep = '/')
dat

```

A Exercise and discussion

Now try it on your own. Download the [untidy](#) dataset and make it tidy using your preferred software.

3.5 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 more 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.

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

- Column names

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

- 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.
- It may also be useful to identify a “plot name” for each column name that uses proper spelling, punctuation, and units (if applicable). Using a column name directly in a graphic is generally a bad idea since their meaning outside of the dataset may not be obvious.
- Data types
 - Describe the type of data in each column, e.g., numerical measurements, categorical descriptors, or counts of observations. 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? Also make note of the units that were used.
 - For categorical descriptors, identify all possible categories that are acceptable values for the column, e.g., small, medium, or large for a qualitative descriptor of size.
 - For dates, make note of the format, e.g., YYYY-MM-DD. For time, identify the timezone.

A Exercise and discussion

Create a data dictionary for the [tidy](#) dataset from the previous example.

3.6 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 the intent of preparing your data for archival in a long-term repository. 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 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 (°C), 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

A 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.

4 Lowering barriers to inclusion and addressing key critiques

4.1 Lesson Outline

- Goals and motivation
- Learning curves
- Fear of exposure
- What does it mean to be open?
- Something is better than nothing

4.2 Goals and motivation

We finish our workshop with a discussion of what it means 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.

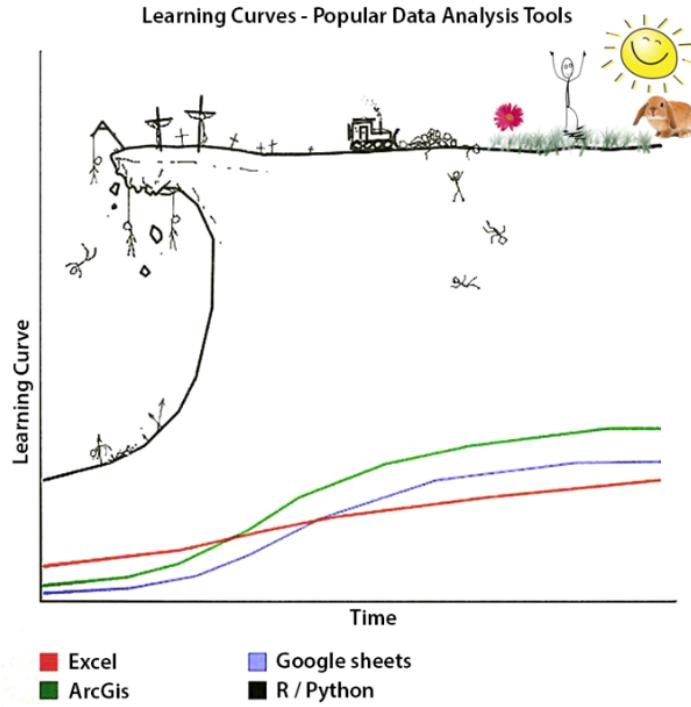
- **Goal:** Understand common hurdles in adopting open science and how to overcome them
- **Motivation:** Become the “open science” expert at your institution!

4.3 Learning curves



It's hard to learn new tools!





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.

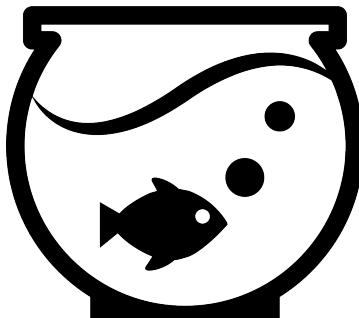
It's also helpful to think about the broader community that can support you along this journey. Learning alone can be discouraging and we strongly recommend that you tap into the diverse community of educators, mentors, bloggers, and friends that can help. Even you can create a community of practice!

Exercise and discussion

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



4.4 Fear of exposure



💡 Challenge

Being open makes me nervous!

👎 Solution

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?

4.5 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 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?

4.6 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.

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 new task, the more likely you'll be able to attempt additional tasks in the future.

Exercise and discussion

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

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.

Open Science Websites

- NCEAS Open Science for Synthesis 2017 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

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

TBEP R Trainings

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

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

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

Git/Github

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

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. "Non-geospatial Metadata for the Ecological Sciences." *Ecological Applications* 7 (1): 330–42. [https://doi.org/10.1890/1051-0761\(1997\)007%5B0330:NMFTE%5D2.0.CO;2](https://doi.org/10.1890/1051-0761(1997)007%5B0330:NMFTE%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>.