RESEARCH FILE ORGANIZATION

The Life-Changing Magic of Tidying Up

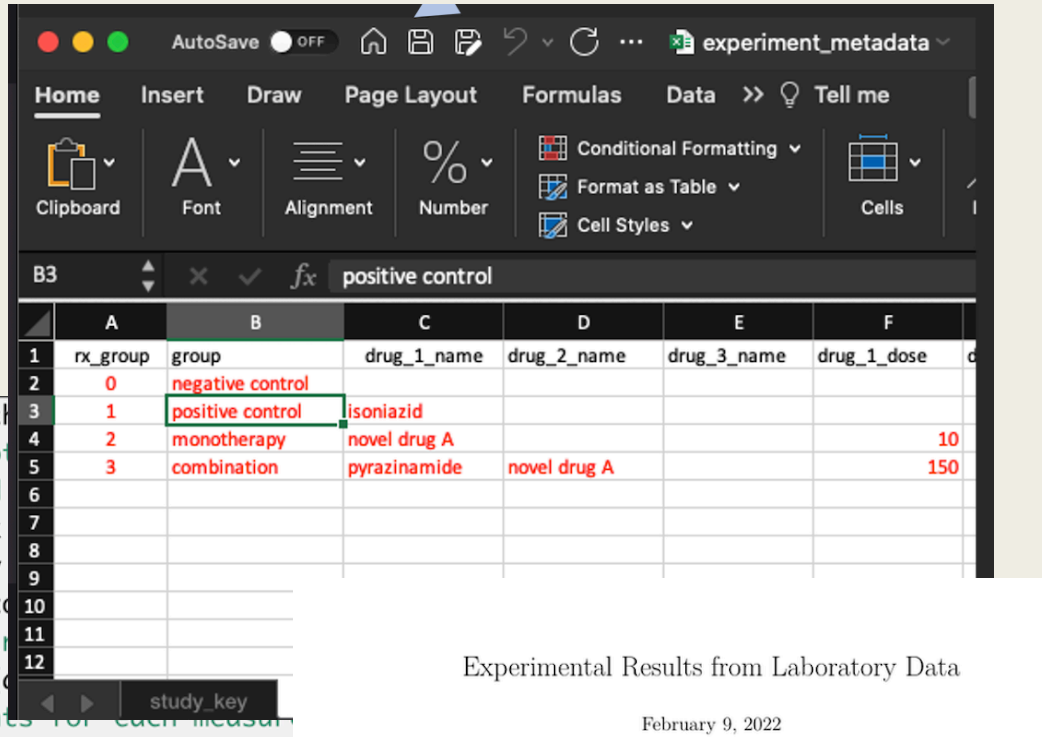
Brooke Anderson

What we'll talk about

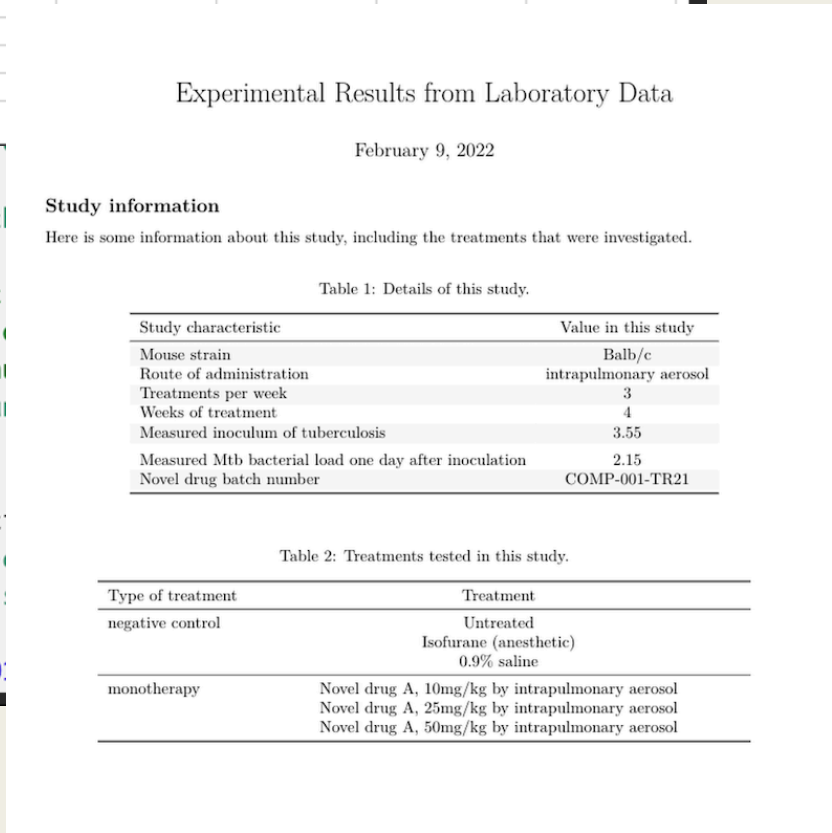
- Why should you organize your research files?
- How should you organize your research files?
- What can you do with well-organized files?

Research files

- Research datasets
- Code scripts for pre-processing and analyzing data
- Reports and presentations




```
ggplot(growth
       # Plot
       # and
       aes(x
           y
           co
  # Add a lin
  geom_line(c
  # Add points for each measur
  geom_point() +
  # Customize the labels for th
  labs(x = "Time (hours)",
       y = "Optical density at
       color = "Growth conditi
       linetype = "Test tube nu
  # Customize the plot appearan
  theme_classic() +
  # Add a title and subtitle
  ggtitle("Growth curve", subt
  # Use a log scale for the y-a
  # from 0.01 to 1 (you can't s
  # as a log transform)
  scale_y_log10(limits = c(0.01
```



Experimental Results from Laboratory Data

February 9, 2022

Study information

Here is some information about this study, including the treatments that were investigated.

Table 1: Details of this study.

Study characteristic	Value in this study
Mouse strain	Balb/c
Route of administration	intrapulmonary aerosol
Treatments per week	3
Weeks of treatment	4
Measured inoculum of tuberculosis	3.55
Measured Mtb bacterial load one day after inoculation	2.15
Novel drug batch number	COMP-001-TR21

Table 2: Treatments tested in this study.

Type of treatment	Treatment
negative control	Untreated Isofurane (anesthetic) 0.9% saline
monotherapy	Novel drug A, 10mg/kg by intrapulmonary aerosol Novel drug A, 25mg/kg by intrapulmonary aerosol Novel drug A, 50mg/kg by intrapulmonary aerosol

Research Paper ■

Issues in Biomedical Research Data Management and Analysis: Needs and Barriers

NICHOLAS R. ANDERSON, MS, E. SALLY LEE, MS, J. SCOTT BROCKENBROUGH, PHD, MARK E. MINIE, PHD, SHERRILYNNE FULLER, PHD, JAMES BRINKLEY, MD, PHD, PETER TARCZY-HORNOCH, MD

Abstract **Objectives:** A. Identify the current state of data management needs of academic biomedical researchers. B. Explore their anticipated data management and analysis needs. C. Identify barriers to addressing those needs.

Design: A multimodal needs analysis was conducted using a combination of an online survey and in-depth one-on-one semi-structured interviews. Subjects were recruited via an e-mail list representing a wide range of academic biomedical researchers in the Pacific Northwest.

Measurements: The results from 286 survey respondents were used to provide triangulation of the qualitative analysis of data gathered from 15 semi-structured in-depth interviews.

Results: Three major themes were identified: 1) there continues to be widespread use of basic general-purpose applications for core data management; 2) there is broad perceived need for additional support in managing and analyzing large datasets; and 3) the barriers to acquiring currently available tools are most commonly related to financial burdens on small labs and unmet expectations of institutional support.

Conclusion: Themes identified in this study suggest that at least some common data management needs will best be served by improving access to basic level tools such that researchers can solve their own problems. Additionally, institutions and informaticians should focus on three components: 1) facilitate and encourage the use of modern data exchange models and standards, enabling researchers to leverage a common layer of interoperability and analysis; 2) improve the ability of researchers to maintain provenance of data and models as they evolve over time though tools and the leveraging of standards; and 3) develop and support information management service cores that could assist in these previous components while providing researchers with unique data analysis and information design support within a spectrum of informatics capabilities.

■ **J Am Med Inform Assoc.** 2007;14:478–488. DOI 10.1197/jamia.M2114.

File organization in practice

One study surveyed over 250 biomedical researchers at the University of Washington.

Research Paper ■

Issues in Biomedical Research Data Management and Analysis: Needs and Barriers

NICHOLAS R. ANDERSON, MS, E. SALLY LEE, MS, J. SCOTT BROCKENBROUGH, PHD, MARK E. MINIE, PHD, SHERRILYNNE FULLER, PHD, JAMES BRINKLEY, MD, PHD, PETER TARCZY-HORNOCH, MD

Abstract **Objectives:** A. Identify the current state of data management needs of academic biomedical researchers. B. Explore their anticipated data management and analysis needs. C. Identify barriers to addressing those needs.

Design: A multimodal needs analysis was conducted using a combination of an online survey and in-depth one-on-one semi-structured interviews. Subjects were recruited via an e-mail list representing a wide range of academic biomedical researchers in the Pacific Northwest.

Measurements: The results from 286 survey respondents were used to provide triangulation of the qualitative analysis of data gathered from 15 semi-structured in-depth interviews.

Results: Three major themes were identified: 1) there continues to be widespread use of basic general-purpose applications for core data management; 2) there is broad perceived need for additional support in managing and analyzing large datasets; and 3) the barriers to acquiring currently available tools are most commonly related to financial burdens on small labs and unmet expectations of institutional support.

Conclusion: Themes identified in this study suggest that at least some common data management needs will best be served by improving access to basic level tools such that researchers can solve their own problems. Additionally, institutions and informaticians should focus on three components: 1) facilitate and encourage the use of modern data exchange models and standards, enabling researchers to leverage a common layer of interoperability and analysis; 2) improve the ability of researchers to maintain provenance of data and models as they evolve over time though tools and the leveraging of standards; and 3) develop and support information management service cores that could assist in these previous components while providing researchers with unique data analysis and information design support within a spectrum of informatics capabilities.

■ **J Am Med Inform Assoc.** 2007;14:478–488. DOI 10.1197/jamia.M2114.

File organization in practice

They noted that, *"a common theme surrounding data management and analysis was that many researchers prefer to utilize their own individual methods to organize data. … Some researchers admitted to having no organizational method at all, while others used whatever method best suited their individual needs."*

Research Paper ■

Issues in Biomedical Research Data Management and Analysis: Needs and Barriers

NICHOLAS R. ANDERSON, MS, E. SALLY LEE, MS, J. SCOTT BROCKENBROUGH, PHD, MARK E. MINIE, PHD, SHERRILYNNE FULLER, PHD, JAMES BRINKLEY, MD, PHD, PETER TARCZY-HORNOCH, MD

Abstract **Objectives:** A. Identify the current state of data management needs of academic biomedical researchers. B. Explore their anticipated data management and analysis needs. C. Identify barriers to addressing those needs.

Design: A multimodal needs analysis was conducted using a combination of an online survey and in-depth one-on-one semi-structured interviews. Subjects were recruited via an e-mail list representing a wide range of academic biomedical researchers in the Pacific Northwest.

Measurements: The results from 286 survey respondents were used to provide triangulation of the qualitative analysis of data gathered from 15 semi-structured in-depth interviews.

Results: Three major themes were identified: 1) there continues to be widespread use of basic general-purpose applications for core data management; 2) there is broad perceived need for additional support in managing and analyzing large datasets; and 3) the barriers to acquiring currently available tools are most commonly related to financial burdens on small labs and unmet expectations of institutional support.

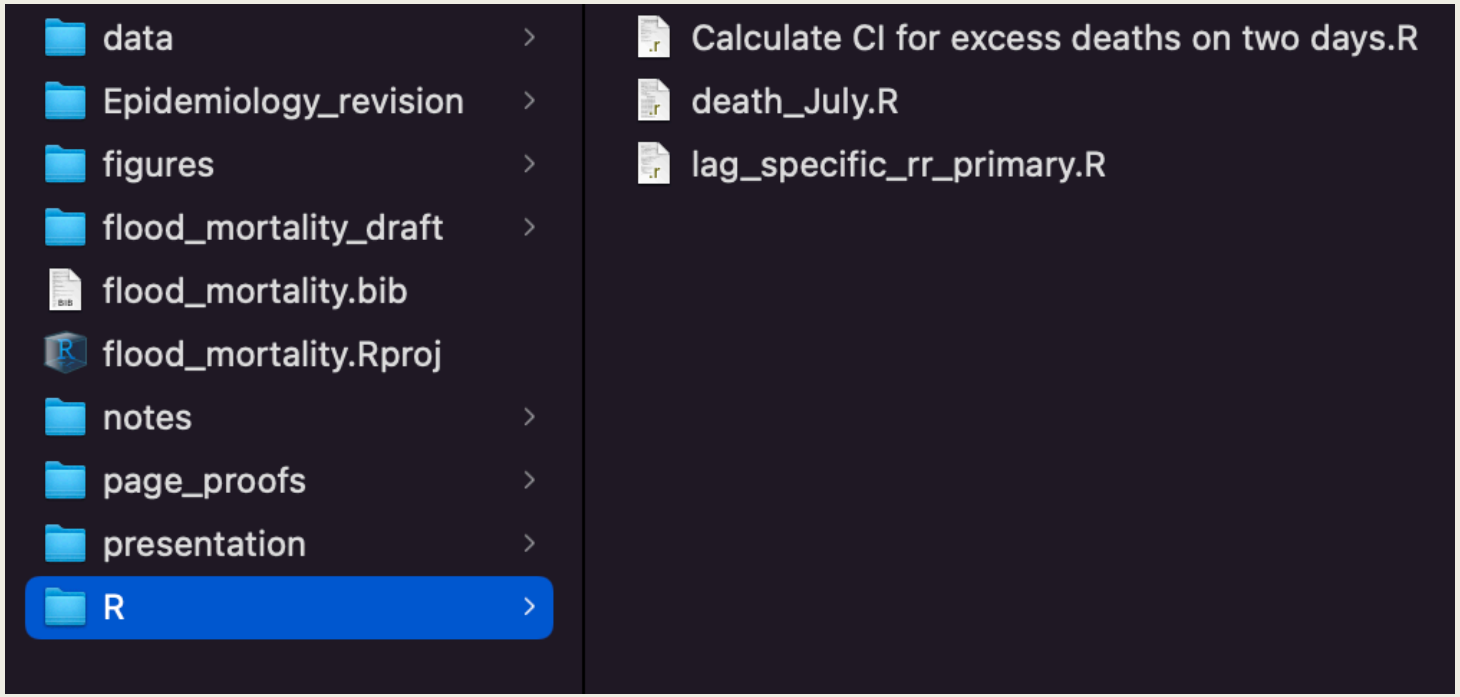
Conclusion: Themes identified in this study suggest that at least some common data management needs will best be served by improving access to basic level tools such that researchers can solve their own problems. Additionally, institutions and informaticians should focus on three components: 1) facilitate and encourage the use of modern data exchange models and standards, enabling researchers to leverage a common layer of interoperability and analysis; 2) improve the ability of researchers to maintain provenance of data and models as they evolve over time though tools and the leveraging of standards; and 3) develop and support information management service cores that could assist in these previous components while providing researchers with unique data analysis and information design support within a spectrum of informatics capabilities.

■ **J Am Med Inform Assoc.** 2007;14:478–488. DOI 10.1197/jamia.M2114.

Respondent answers

- One respondent answers, “They’re not organized in any way---they’re just thrown into files under different projects.”
- Another respondent: “I grab them when I need them, they’re not organized in any decent way.”

WHY ORGANIZE FILES

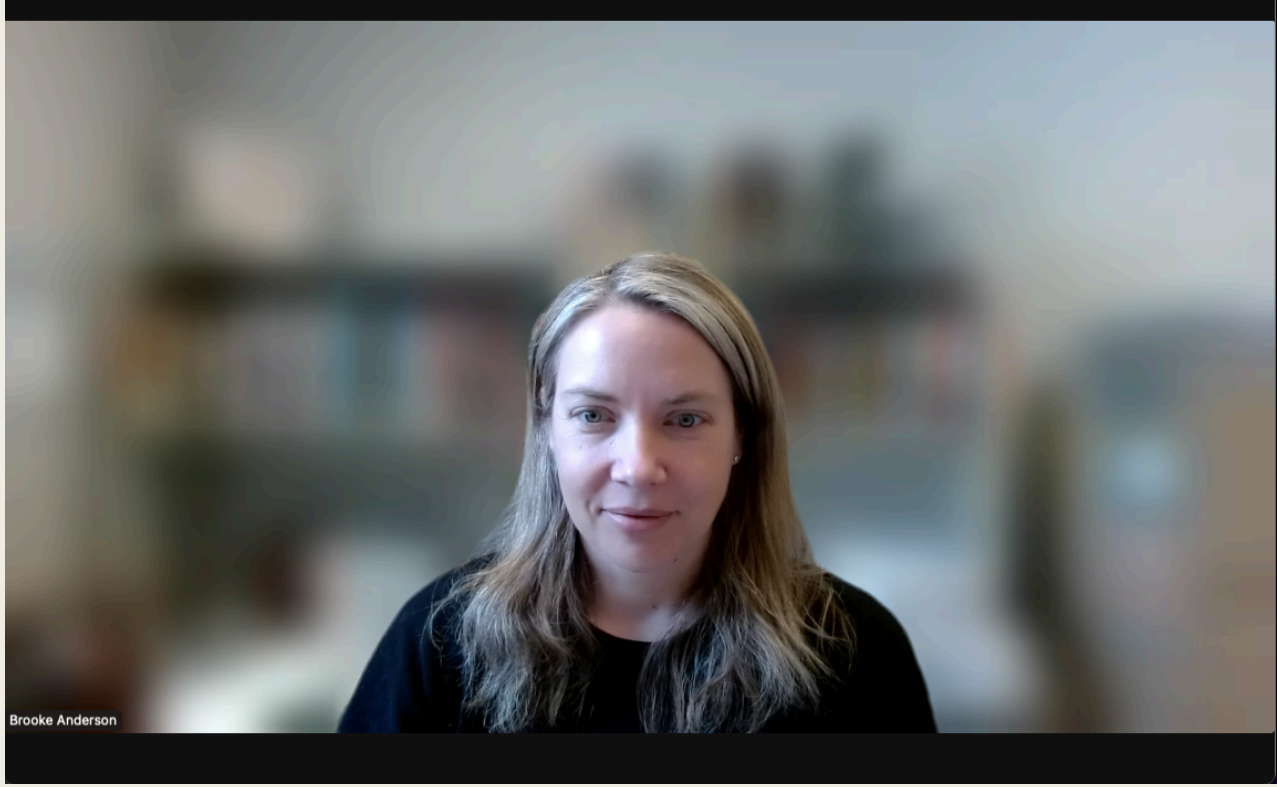


Efficiency

- With good organization, "**Methods and data sections in papers practically write themselves**, with no time wasted in frenzied hunting for missing information." – Baker, 2016
- You can quickly get new people up to speed and efficiently pass along lab research projects as people come and go from the lab group
- You can write tools that work with the structure

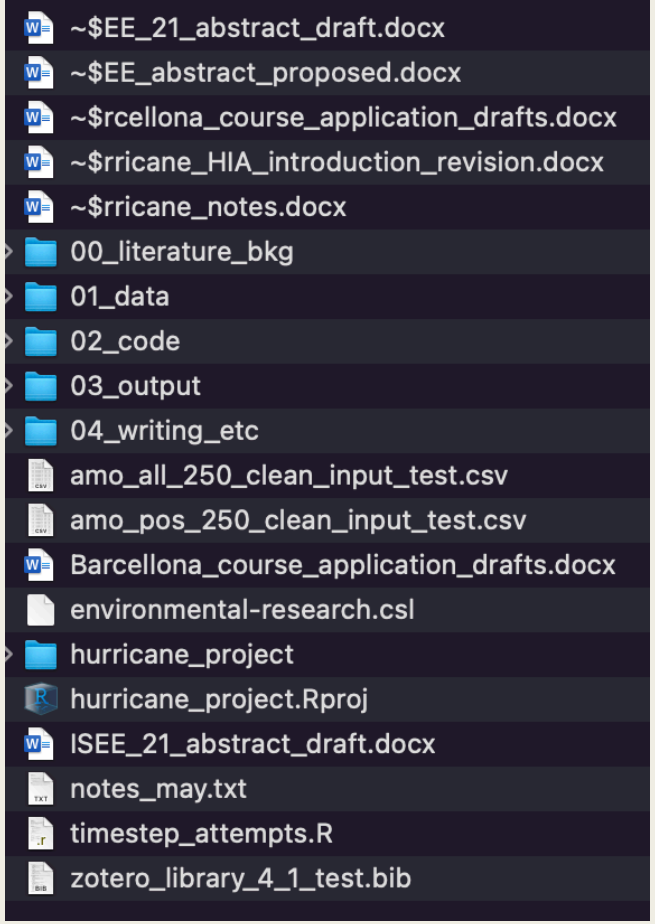
Efficiency

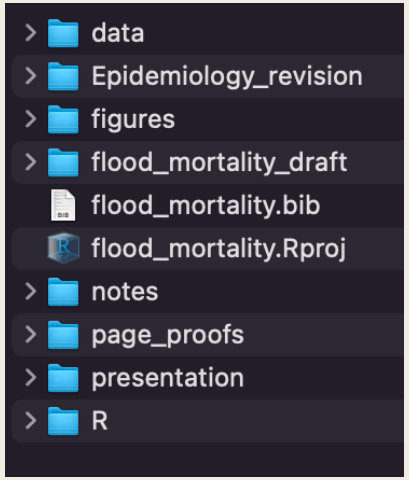
“**Everything you do, you will probably have to do over again.** … If you have organized and documented your work clearly, then repeating the experiment with the new data or the new parameterization will be much, much easier.” -- Noble, 2009



Sharing research files

We’re less likely to share things that aren’t tidy. This is why we often blur our backgrounds for online meetings!



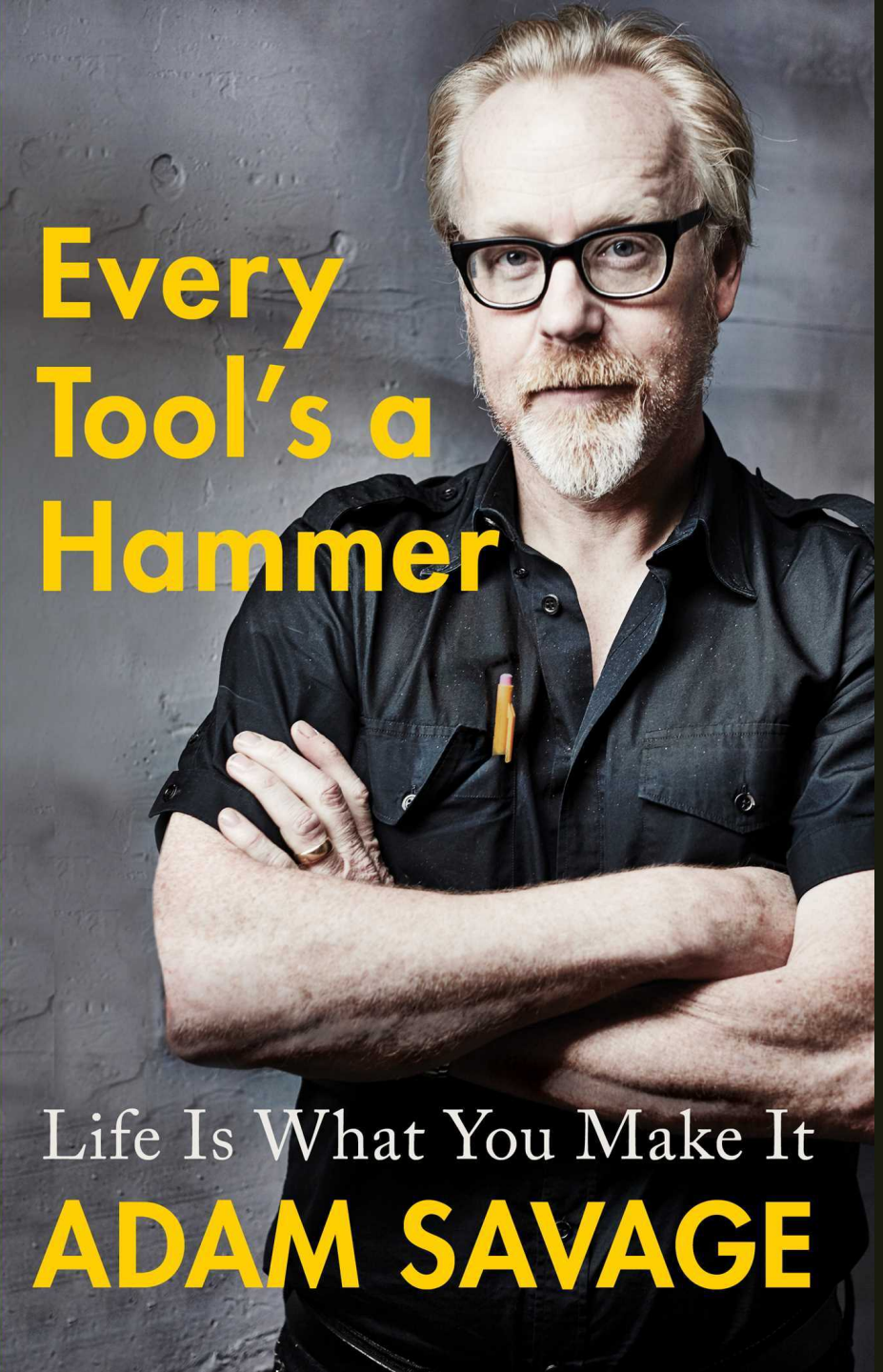


Sharing research files

"Without clear instructions, many researchers struggle to avoid chaos in their file structures, and so are **understandably reluctant to expose their workflow for others to see**. This may be one of the reasons that so many requests for details about methods, including requests for data and code, are turned down or go unanswered."

–Marwick et al., 2018

ADAM SAVAGE

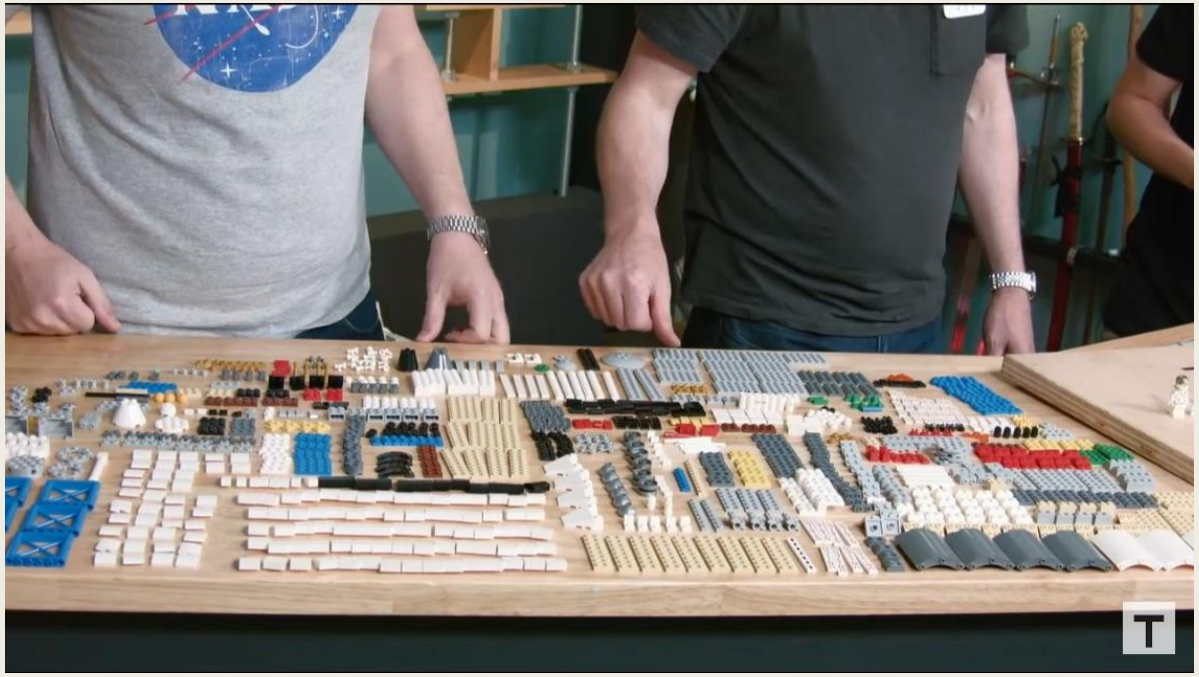


Knolling

Laying things out so that you can see everything at once…

”I started to clean up before heading upstairs at the end of the day, and lo and behold the shop became a **far more efficient** and well-oiled machine to work in.

The freed-up space in my mind and the open work space at my fingertips **allowed me a lot of room, both mental and physical**, to pursue a wide variety of projects, and I finally started to understand how much benefit was to be gained by taking the time to clean.”



KNOLLING LEGOS

Why you should organize research files

- You can do your research more efficiently
- You can introduce new people to the project quickly
- You are more likely to share well-organized files, which is critical for research reproducibility
- You have more mental space to focus on substantive questions as you work

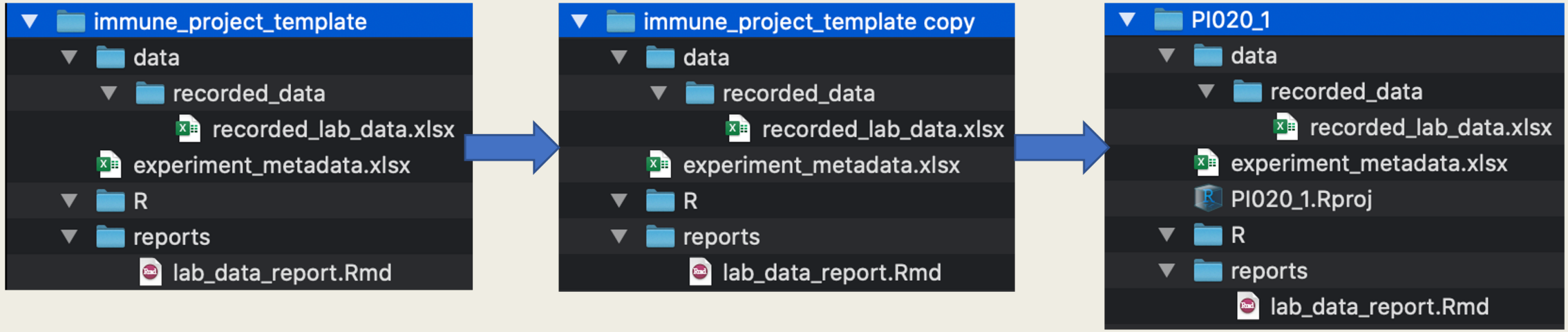
HOW TO ORGANIZE FILES

Using a single directory

- Easy to set up as an R Project
- Easy to set up with version control
- Portable
- Shareable

Blueprint versus physical template

1 Find the project directory template in the file finder program on your computer. Copy the entire directory, paste the copy where you want to store the project directory for your new study, and rename the directory to the name of your new study.



The physical implementation of this is very simple—create an empty example directory and copy and rename for new projects. The harder part is designing a good structure for the directory.

Designing your system

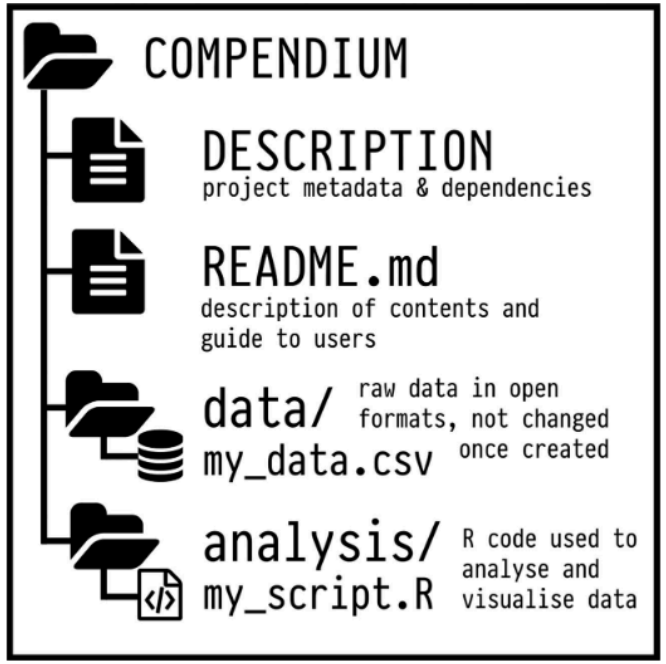
Think about use

Think about reuse

Think about discoverability

Use and Reuse

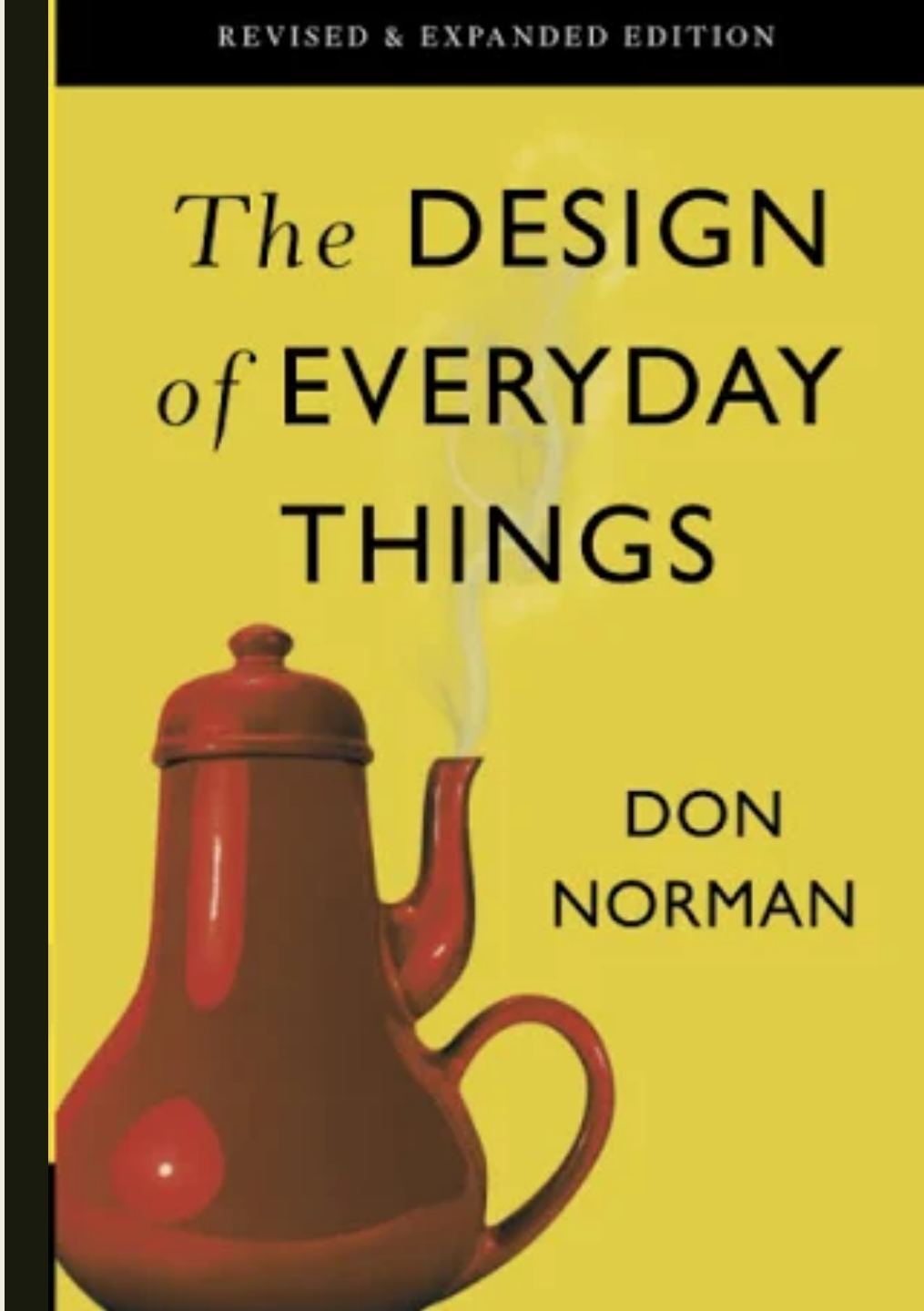
- Group similar types of files together
- Name subdirectories generically



Discoverability

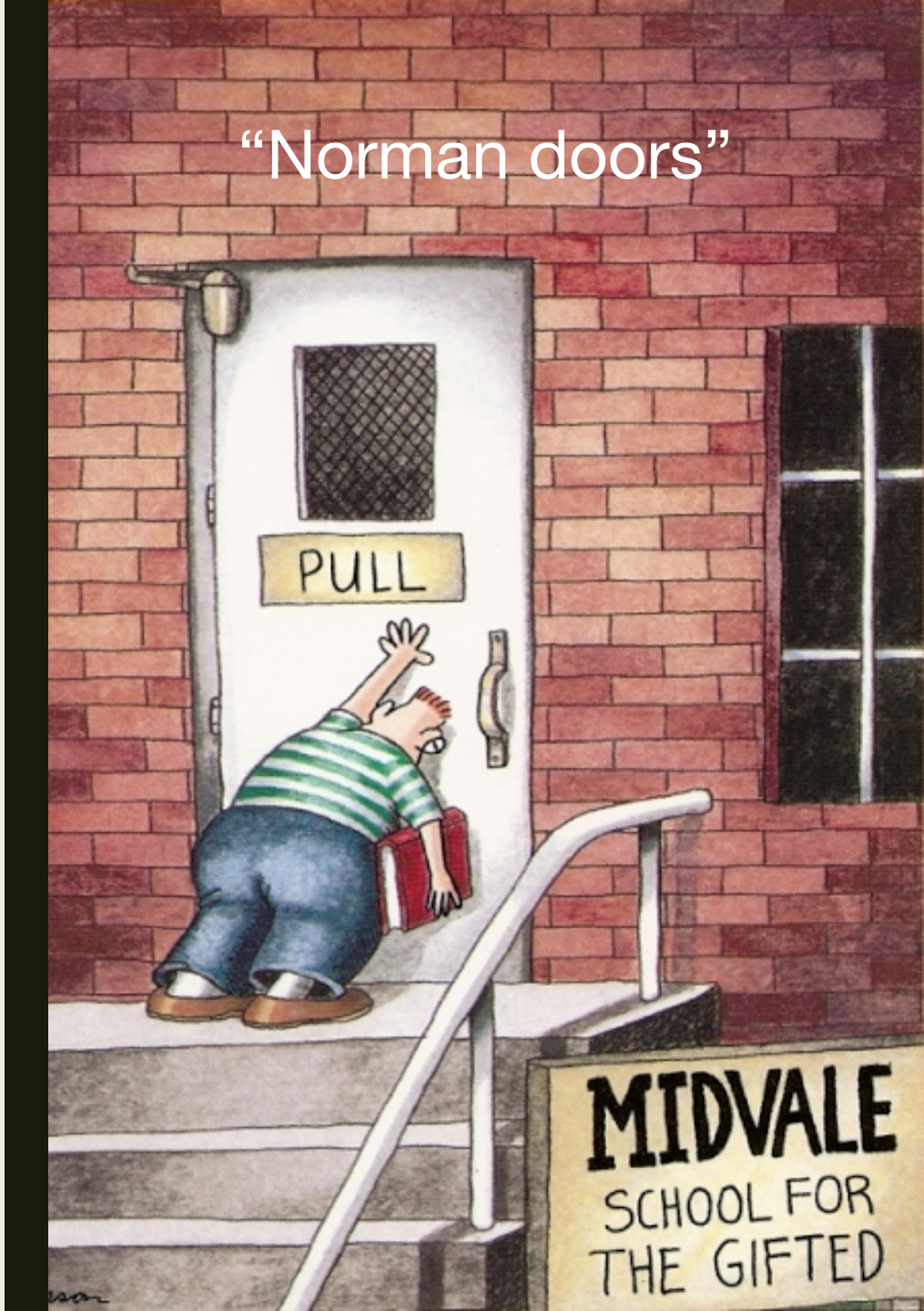
Can a user figure out how to use something quickly, easily, and correctly?

‘Good design is hard to notice… it serves without drawing attention to itself.’



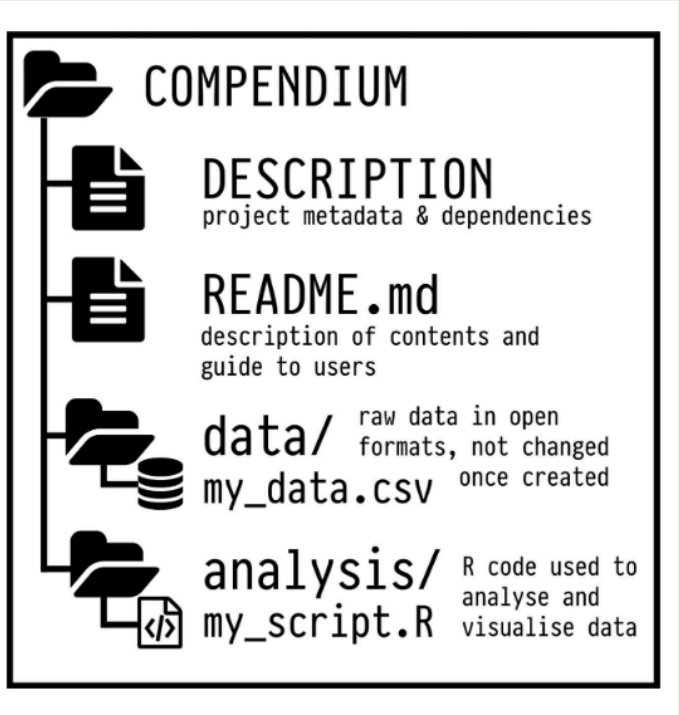
Discoverability

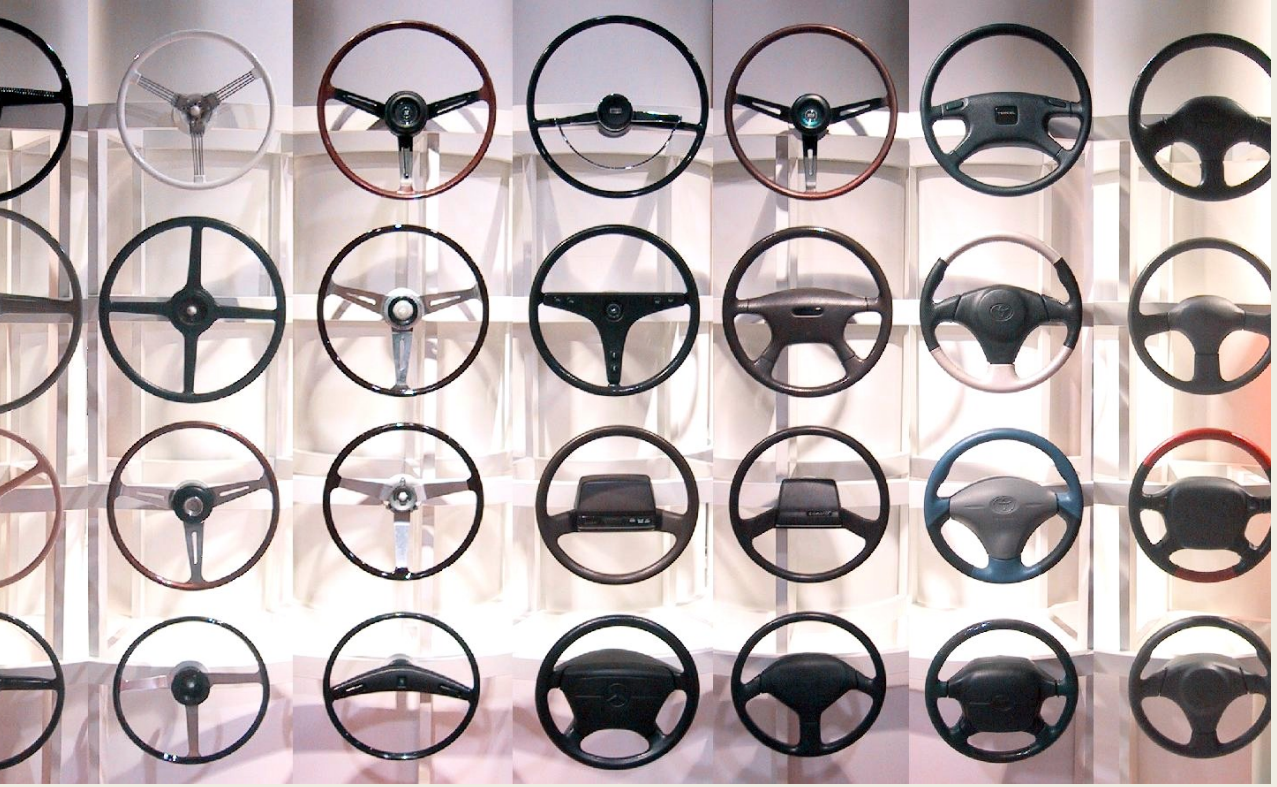
Can a user figure out how to use something quickly, easily, and correctly?



Discoverability

- “The key principle is to organize the compendium so that another person can know what to expect from the plain meaning of the file and directory names.” –Marwick et al., 2018
- “The core guiding principle is simple: Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why.” –Noble, 2009

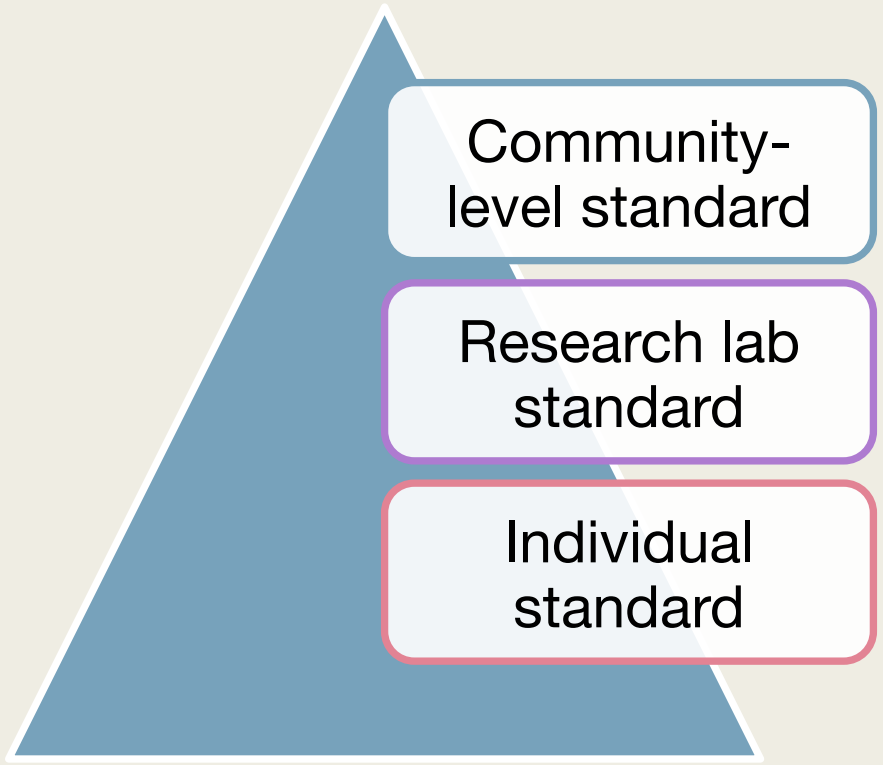




Standards aid discoverability

If you follow a standard of your broader academic community, it makes project files easy for someone else from that community to navigate (Marwick et al., 2018).

Use a standard system to organize files



Marwick—Compendium based on R Package structure

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 80–88
https://doi.org/10.1080/00031305.2017.1375986

Packaging Data Analytical Work Reproducibly Using R (and Friends)

Ben Marwick[a], Carl Boettiger[b], and Lincoln Mullen[c]

[a]University of Washington, Seattle, WA; [b]University of Wollongong, Wollongong, New South Wales; [c]University of California, Berkeley, CA; [d]George Mason University, Fairfax, VA

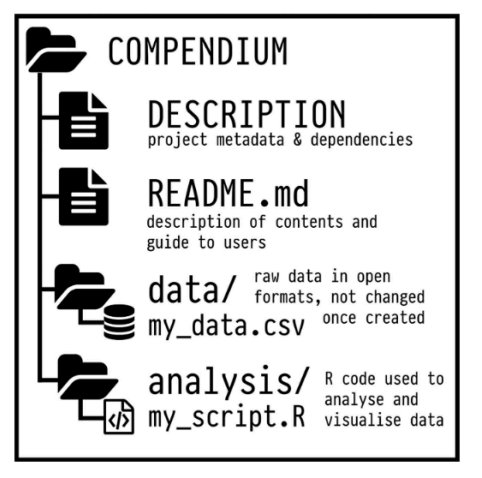
ABSTRACT
Computers are a central tool in the research process, enabling complex and large-scale data analysis. As computer-based research has increased in complexity, so have the challenges of ensuring that this research is reproducible. To address this challenge, we review the concept of the research compendium as a solution for providing a standard and easily recognizable way for organizing the digital materials of a research project to enable other researchers to inspect, reproduce, and extend the research. We investigate how the structure and tooling of software packages of the R programming language are being used to produce research compendia in a variety of disciplines. We also describe how software engineering tools and services are being used by researchers to streamline working with research compendia. Using real-world examples, we show how researchers can improve the reproducibility of their work using research compendia based on R packages and related tools.

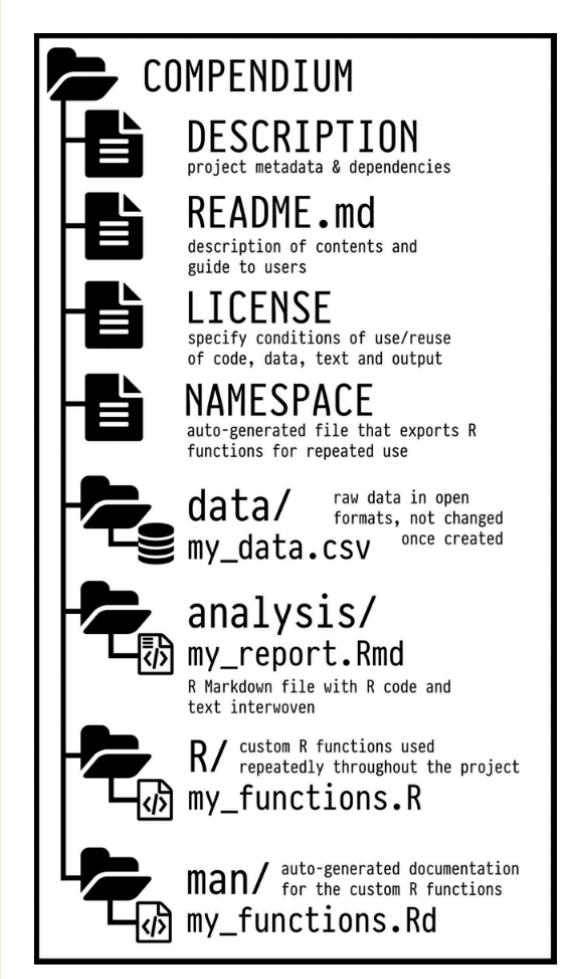
ARTICLE HISTORY
Received May 2017
Revised August 2017

KEYWORDS
Computational science; Data science; Open source software; Reproducible research

Marwick et al. suggest a series of directory structures in their 2018 paper that is based on the structure of R package directories

Marwick Compendium





Their structures range from very simple to more complex to accommodate projects of different complexity

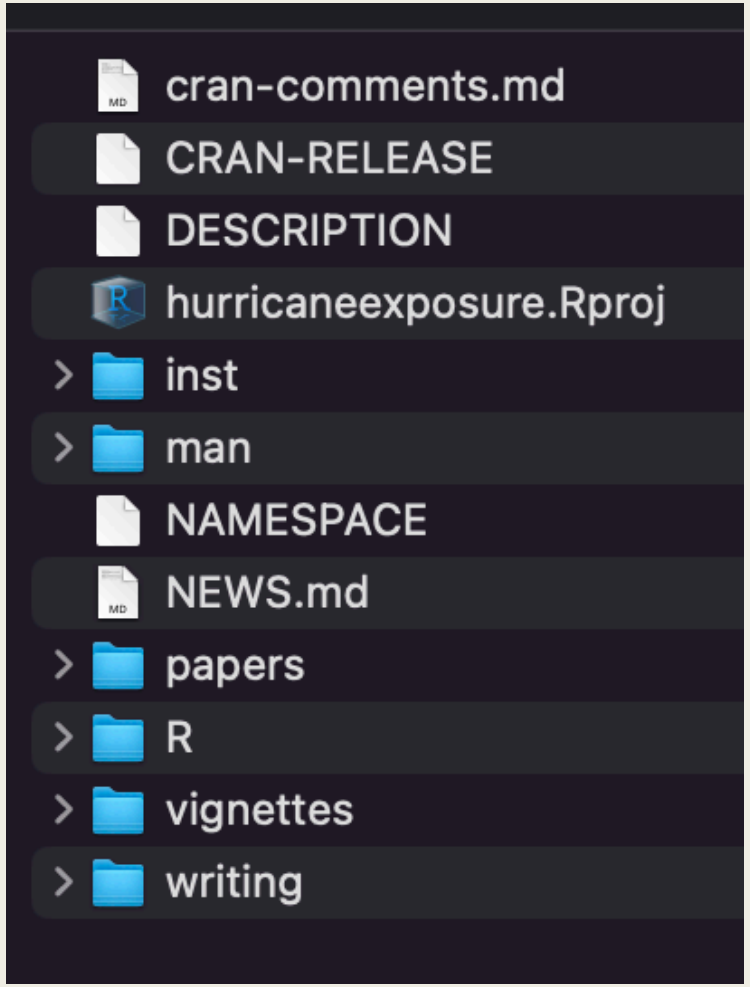
How to organize your research files

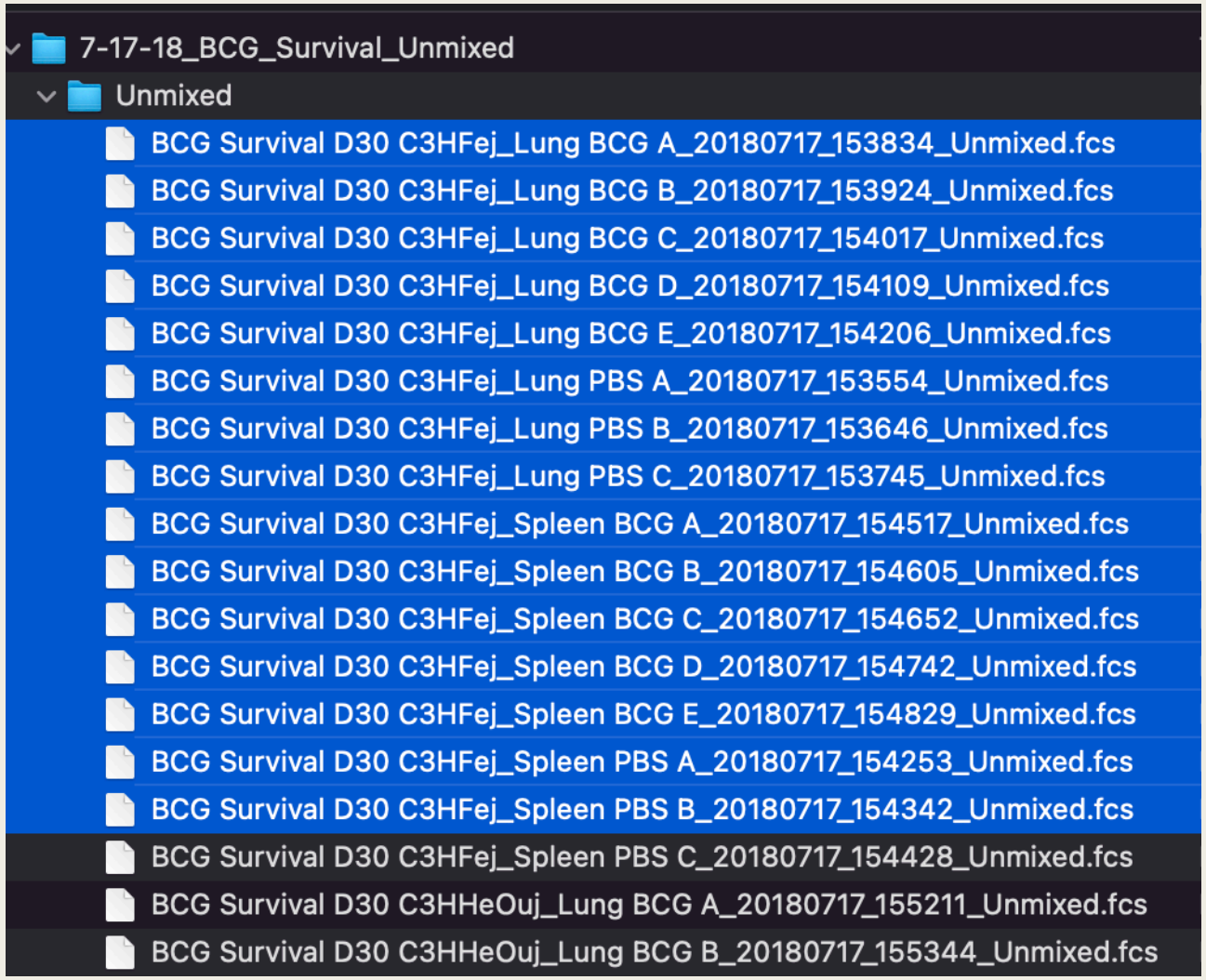
- Organize files for a project within a single computer directory
- Design a system that is easy to use, easy to reuse, and easy to figure out
- Use a standard system to organize your files

ADVANCED TOPICS

Coding to a standard directory structure

“ A research compendium should organize its files according to the prevailing conventions of the scholarly community, whether that be an academic discipline or a lab group. Following these conventions will help other people recognize the structure of the project, and also **support tool building which takes advantage of the shared structure**.” –Marwick et al., 2018

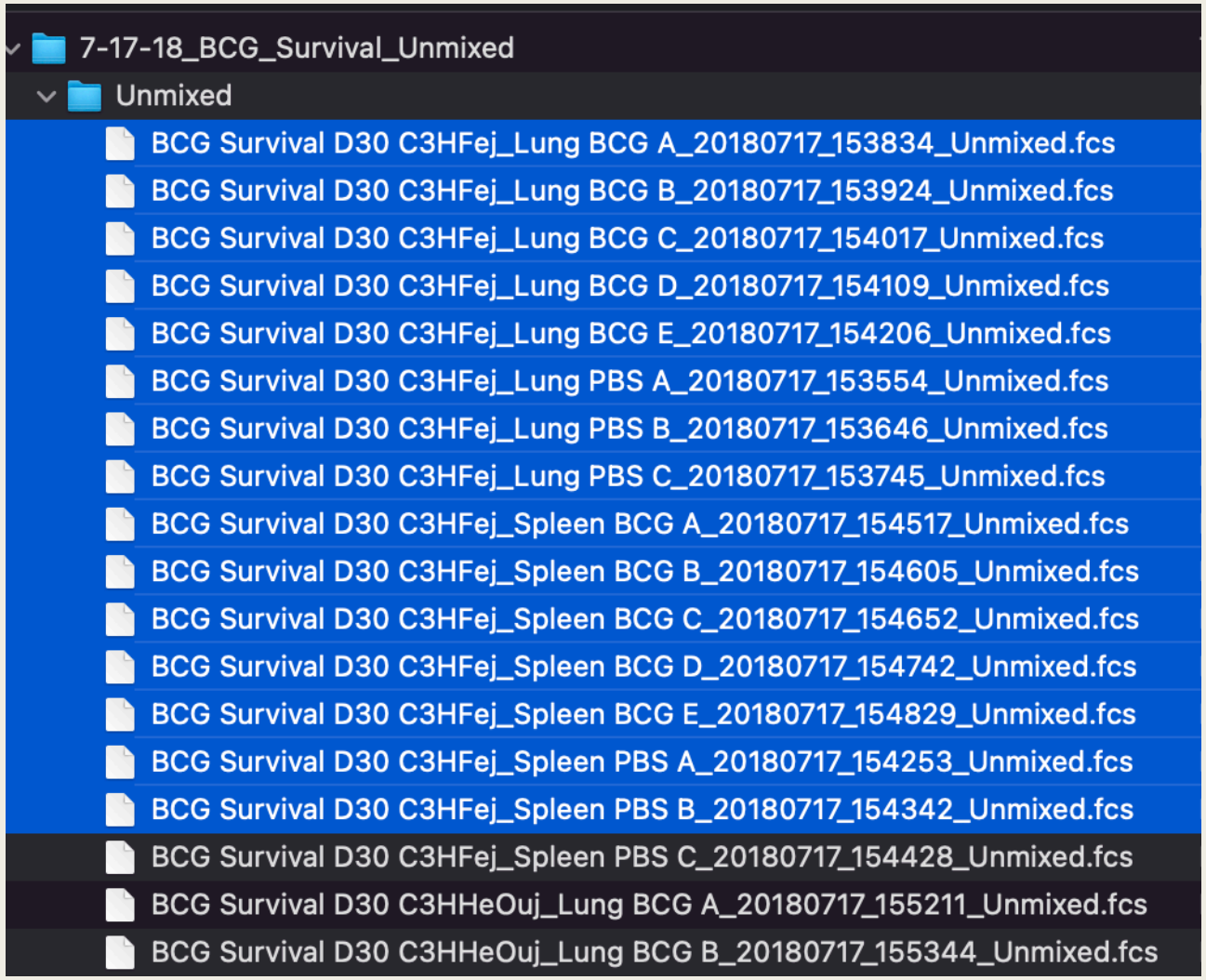




Coding to a standard directory structure

"Organizing data files into a single directory with consistent filenames prepares us to iterate over *all* of our data…

Think of it this way: remember when you discovered you could select many files with your mouse cursor? With this trick, you could move 60 files as easily as six files… By using consistent file naming and directory organization, you can do the same programmatically suing the Unix shell and other programming languages." –Buffalo, 2015

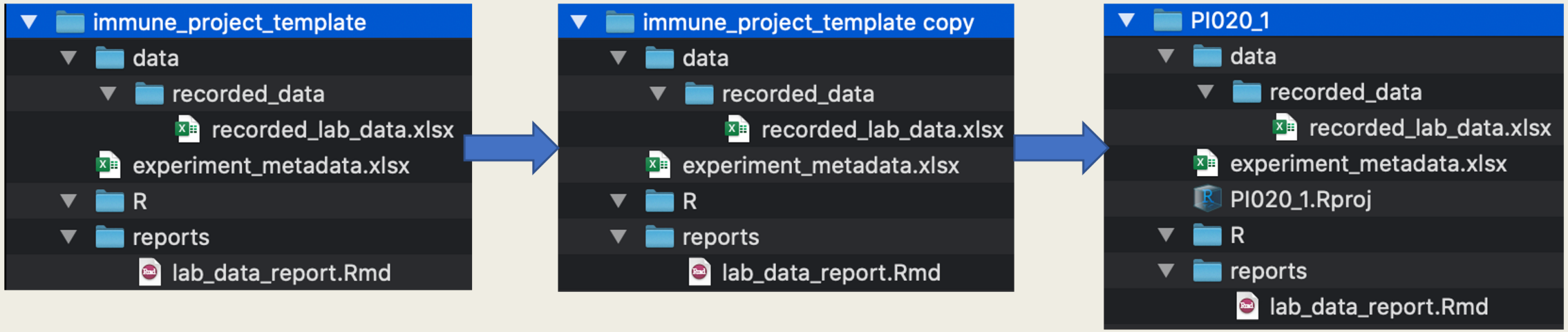


Regular expressions and filenames

- R can list all files in a directory with a system call
- Map-apply programming lets you do the same thing to a list of files
- Regular expressions help you pull information out of character strings, including filepaths

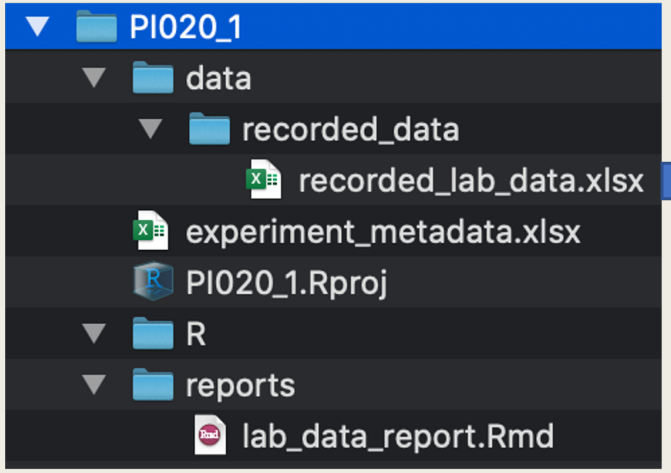
1

Find the project directory template in the file finder program on your computer. Copy the entire directory, paste the copy where you want to store the project directory for your new study, and rename the directory to the name of your new study.



2

Open data recording templates and replace the placeholder data (saved in red font to indicate that it's placeholder data) with data from the real project. Change the font color to black to show that these are data from the project, rather than placeholder data.

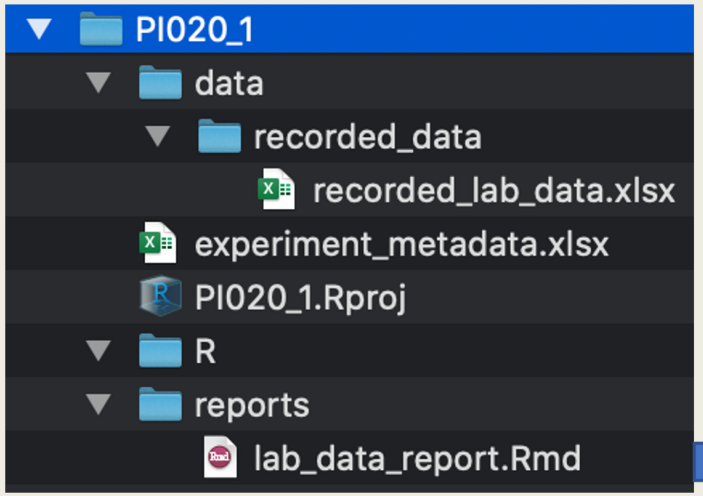


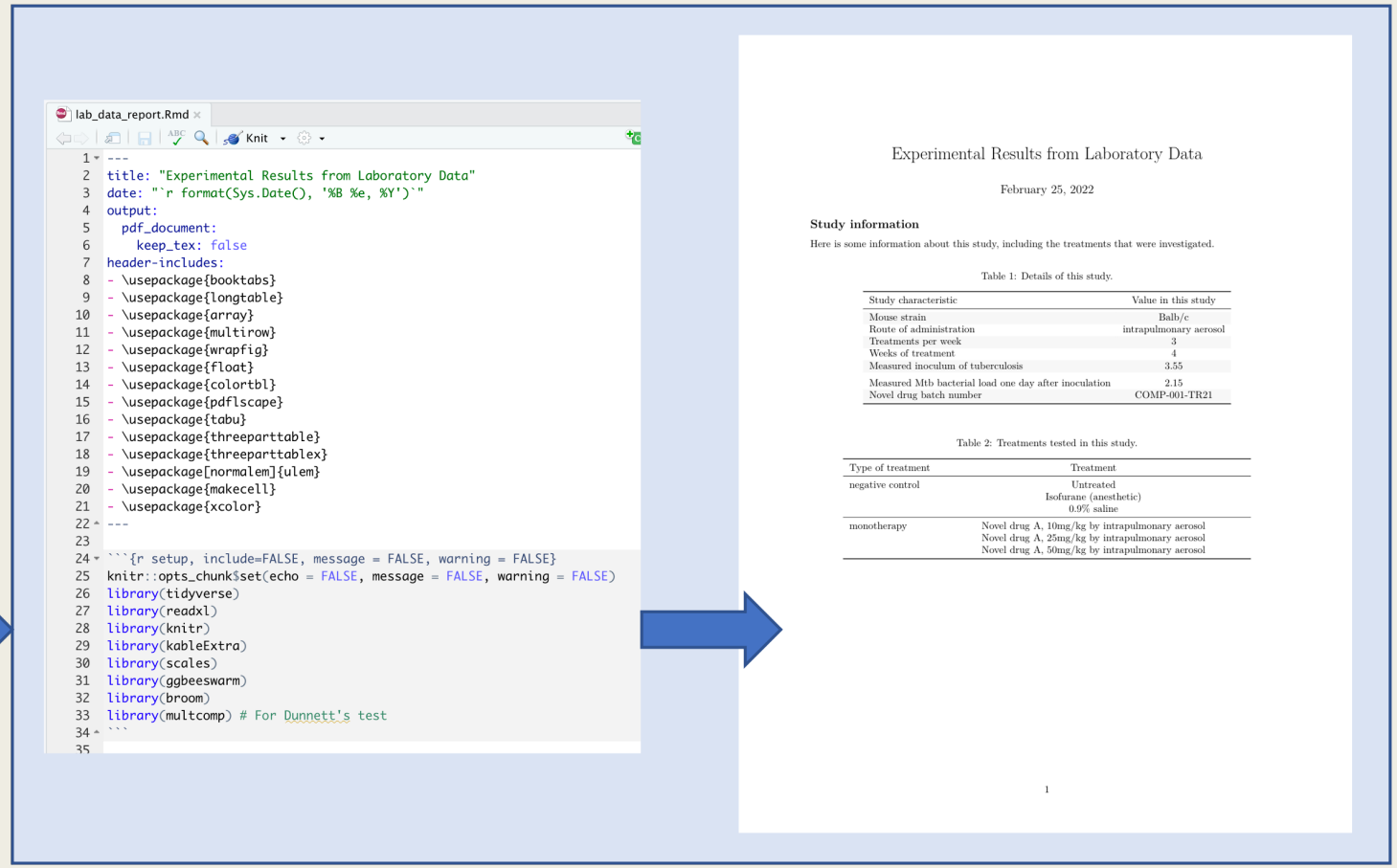
rx_group	week	weight_g
0	0	22.3
0	1	22.6
0	2	23.4
0	3	24.7
0	4	24.7
0	5	24.9
1	0	19.5
1	1	20.3
1	2	20.3
1	3	22.2
1	4	21.3
1	5	22
2	0	20.5
2	1	20.8
2	2	21.2
2	3	20.8

rx_group	week	weight_g
0	0	38.2
0	1	38.4
0	2	38.5
0	3	38.8
1	0	33.7
1	1	33
1	2	32.8
1	3	32.3
2	0	29.2
2	1	28.8
2	2	28.7
2	3	27.4
3	0	35.8
3	1	35.3
3	2	34.8
3	3	35.1
6	0	31.2

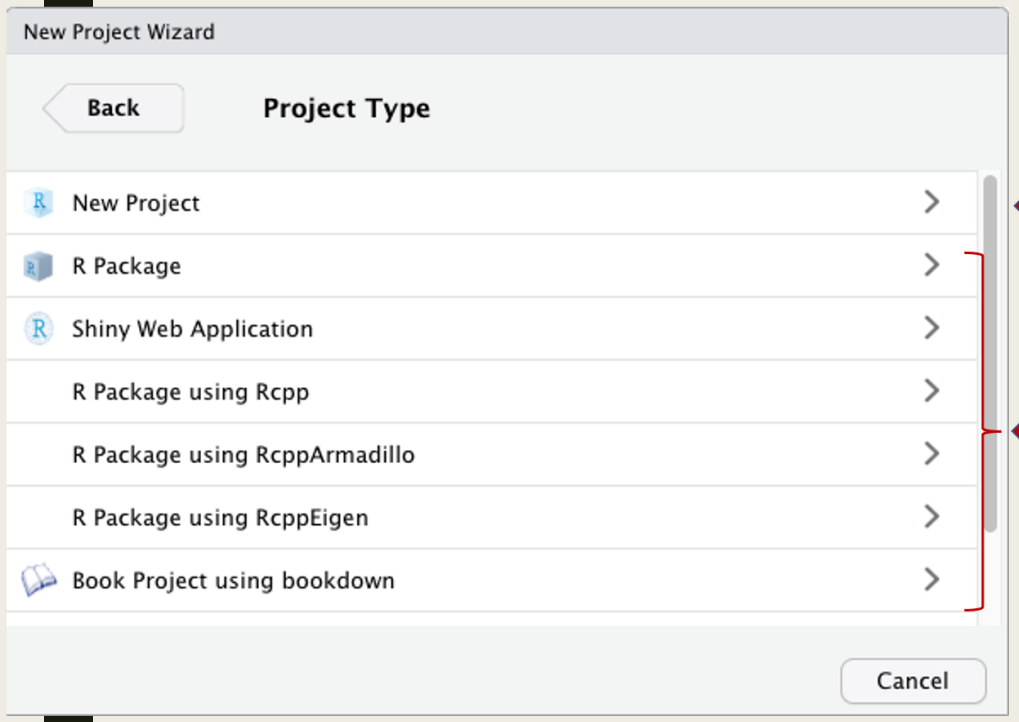
3

Open the project report template. Render it to PDF to create the report. If you'd like, you can make changes to the template Rmarkdown report file to customize it for this project.





Creating an R Project template



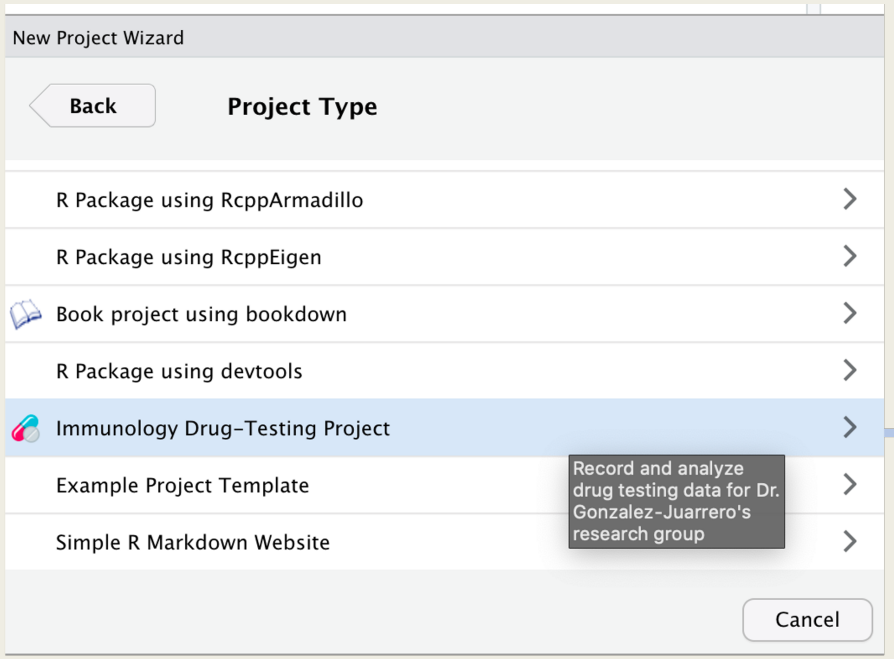
Use "New Project" to create a Project without using a template

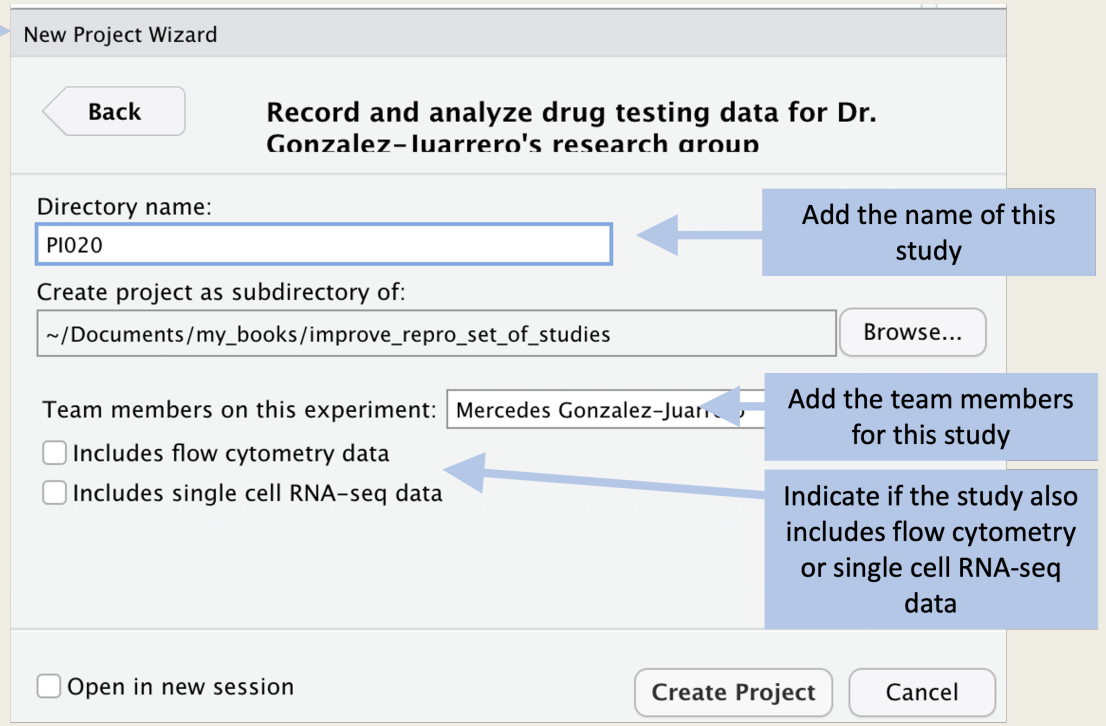
Use one of the templates in your copy of R to initialize the project

R Project templates create more customized R projects. When you begin a new project with one, it will self-populate with a directory structure and potentially template files.

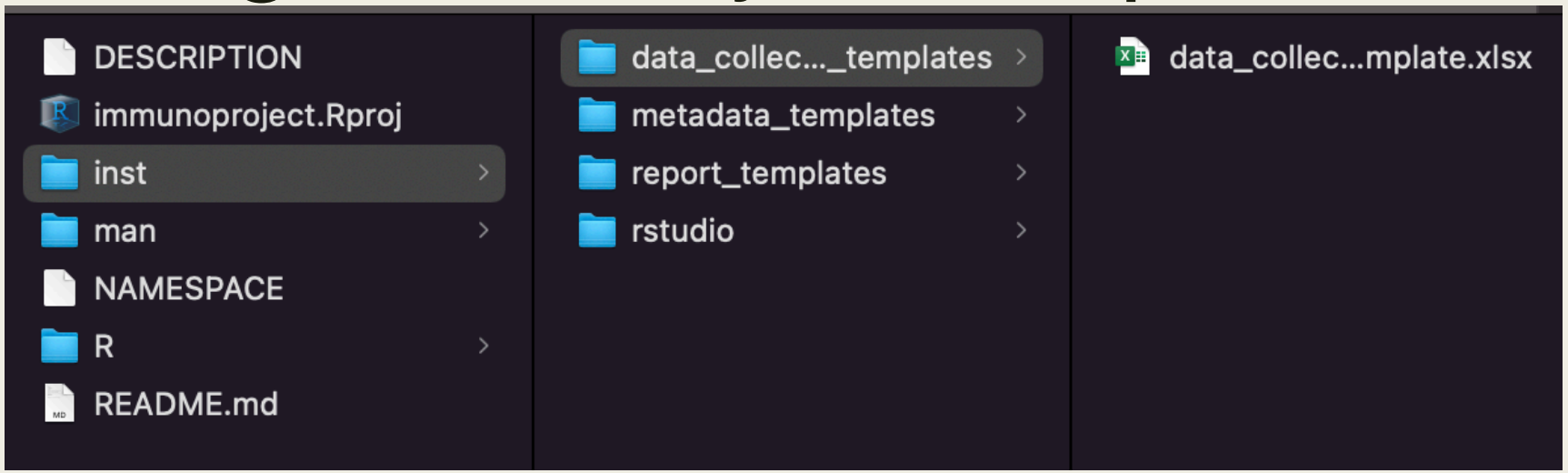
You get new templates by installing R packages.

Creating an R Project template





Creating an R Project template



- You create an R Project template as an R package, which means it can be shared and loaded in the same way as other R packages
- Most template files will be saved in the “inst” directory
- You can use code in the R directory to set up the project directory structure and copy in template files
- You can also use code in the R directory to customize the project set-up based on selections the user makes in the R Project Wizard

Git

- Git is a **version control system**.
- Git manages the evolution of a set of files (a **repository**) in a sane, highly structured way
- Git is like the “Track Changes” features from Microsoft Word on steroids.
- Git’s original purpose was to help groups of developers work collaboratively on big software projects. The data science community has re-purposed Git to use it to work on all of the different types of files that make up a typical project (data, figures, reports) in addition to source code.

Source: happygitwithr.com

Should I Git…?

- The initial installation process and change in workflow is **all worth it** for the benefit of version control and ease of communicating and collaborating with other people

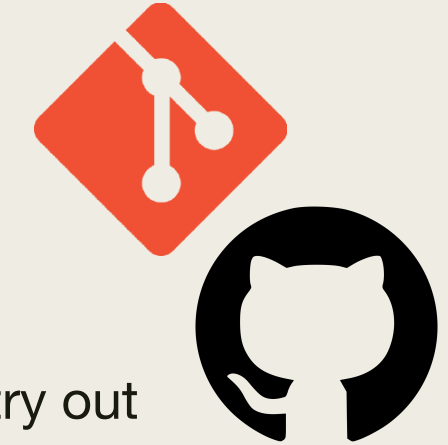
1.3 Is it going to hurt?

Yes.

You have to install Git, get local Git talking to GitHub, and make sure RStudio can talk to local Git (and, therefore, GitHub). This is one-time or once-per-computer pain.

Source: happygitwithr.com

Benefits of Git (and GitHub)



- Exposure: if someone needs to see your work or if you want people to try out your code, they can clone or fork your repository with Git, or browse your project on GitHub.
- You can track the development of projects (like R packages) on GitHub. You can modify your fork to add features or fix bugs and send them back to the owner as a proposed change.
- If you need to collaborate on analysis or code development… use Git and GitHub! This is more analogous to Google Docs than it is to the edit, save, attach workflow.

Source: happygitwithr.com

happygitwithR.com by Jenny Bryan

Happy Git provides opinionated instructions on how to:

- Install Git and get it working smoothly with GitHub, in the shell and in the RStudio IDE.
- Develop a few key workflows that cover your most common tasks.
- Integrate Git and GitHub into your daily work with R and R Markdown.

The target reader is someone who uses R for data analysis or who works on R packages, although some of the content may be useful to those working in adjacent areas.

Source: happygitwithr.com

happygitwithR.com by Jenny Bryan

1.9 What this is NOT

We aim to teach novices about Git on a strict "need to know" basis. Git was built to manage development of the Linux kernel, which is probably very different from what you do. Most people need a small subset of Git's functionality and that will be our focus. If you want a full-blown exposition of Git as a directed acyclic graph or a treatise on the Git-Flow branching strategy, you will be sad.

Source: happygitwithr.com

Further resources

- Marwick, B., Boettiger, C., and Mullen, L. (2018). Packaging data analytical work reproducibly using R (and friends). *The American Statistician*, 72(1):80–88.
- Savage, A. (2020). *Every Tool's a Hammer: Life Is What You Make It.* Atria Books.
- Making R Project templates: https://rstudio.github.io/rstudio-extensions/rstudio_project_templates.html
- Jenny Bryan's Happy Git and Github for the useR: https://happygitwithr.com/