Reproducible Research Workflows

Oregon Data Science Collaborative

Spring 2022

Oregon State University

University of Oregon

Portland State University

Reproducible research… what is it good for?

A published data analysis is reproducible if the analytic data sets and the computer code used to create the data analysis are made available to others for independent study and analysis

This definition is sufficiently vague that it ultimately raises more questions than it answers. What is an “analytic data set“? What does it mean to be “available”? What is included with the “computer code”?

Peng and Hicks, 2021, *Reproducible Research: A retrospective*

Open science: a new framework for research

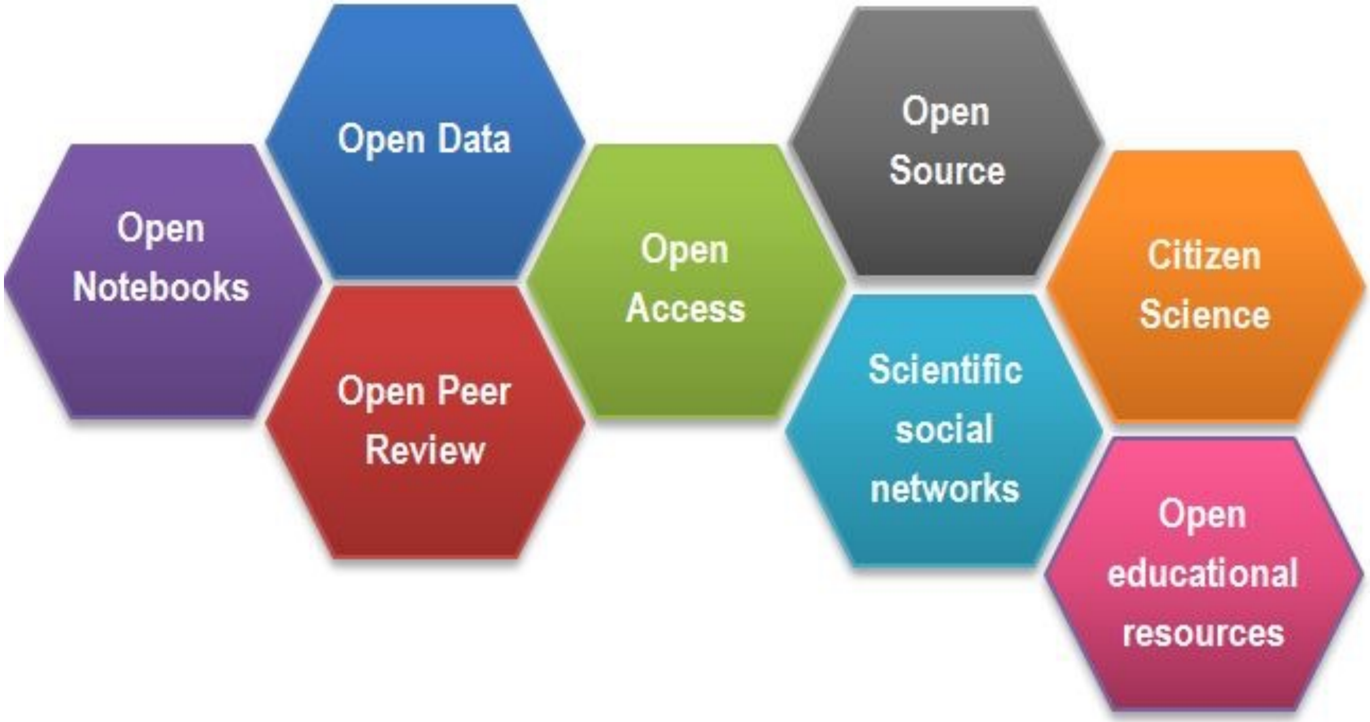
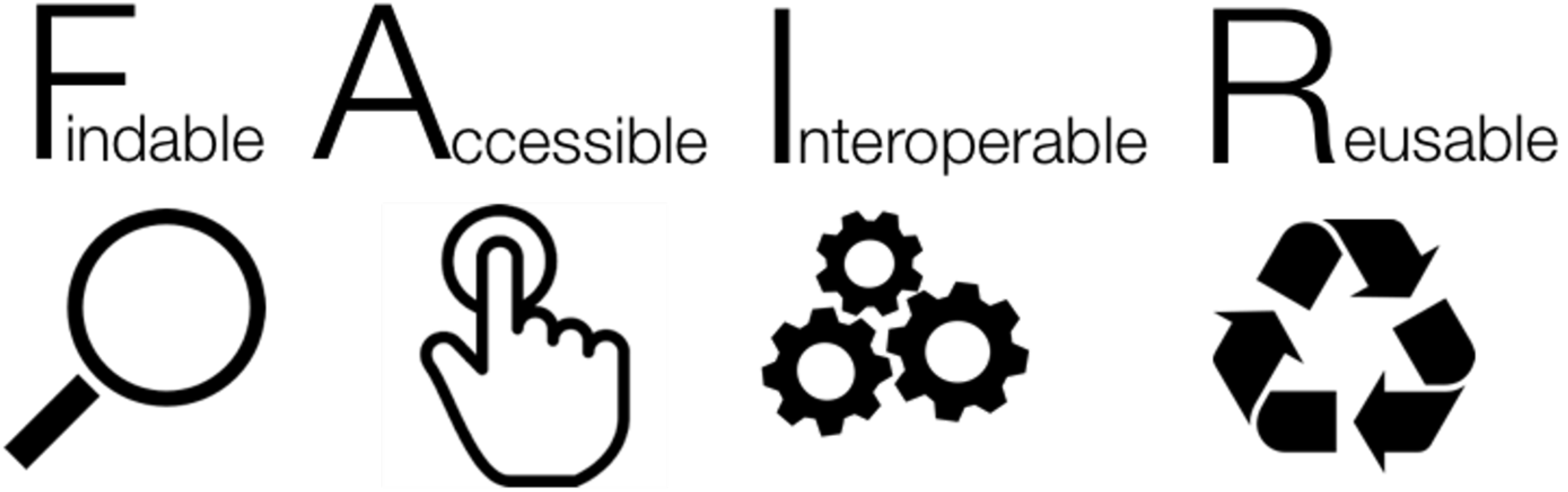
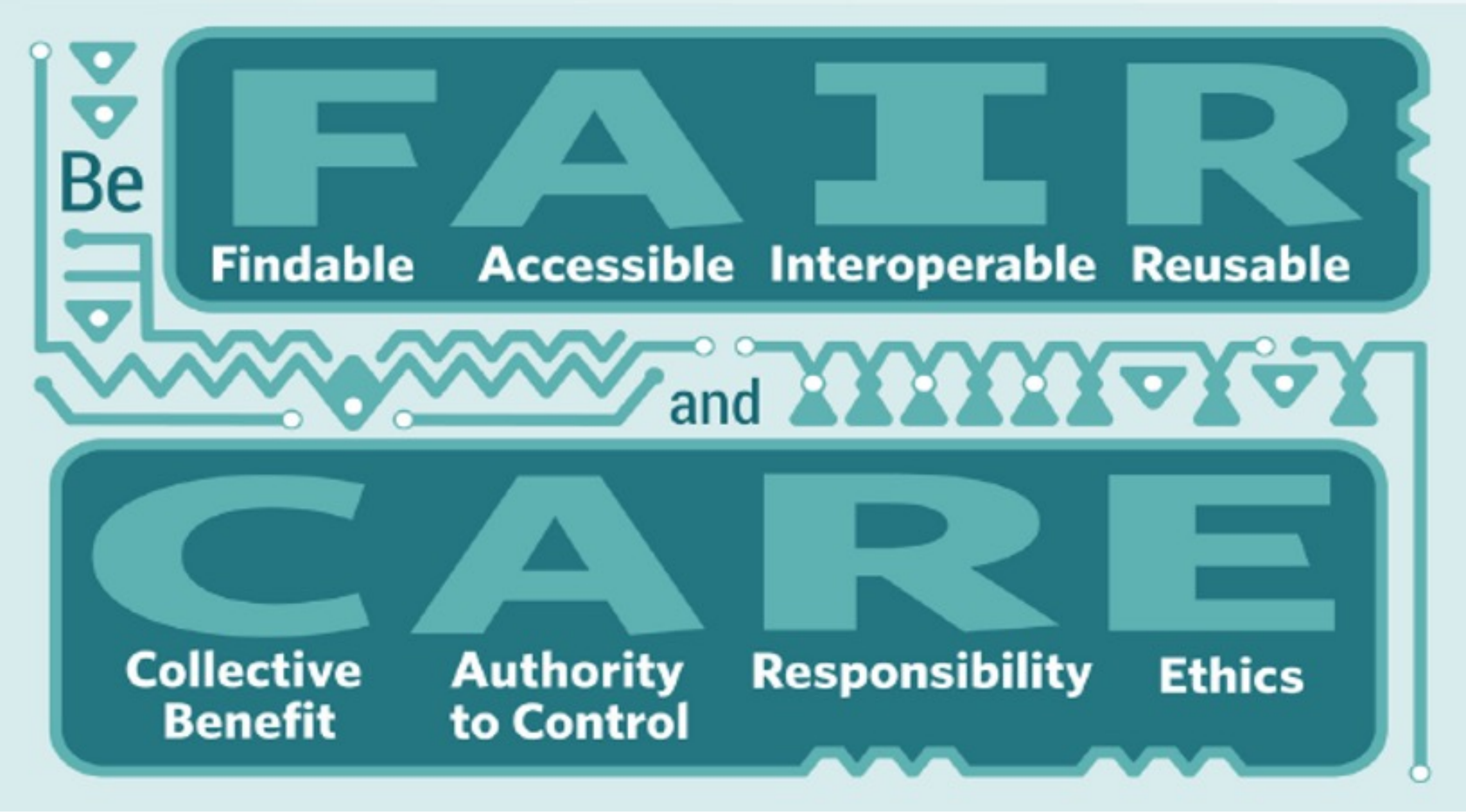


Image credit G. de la Fuente



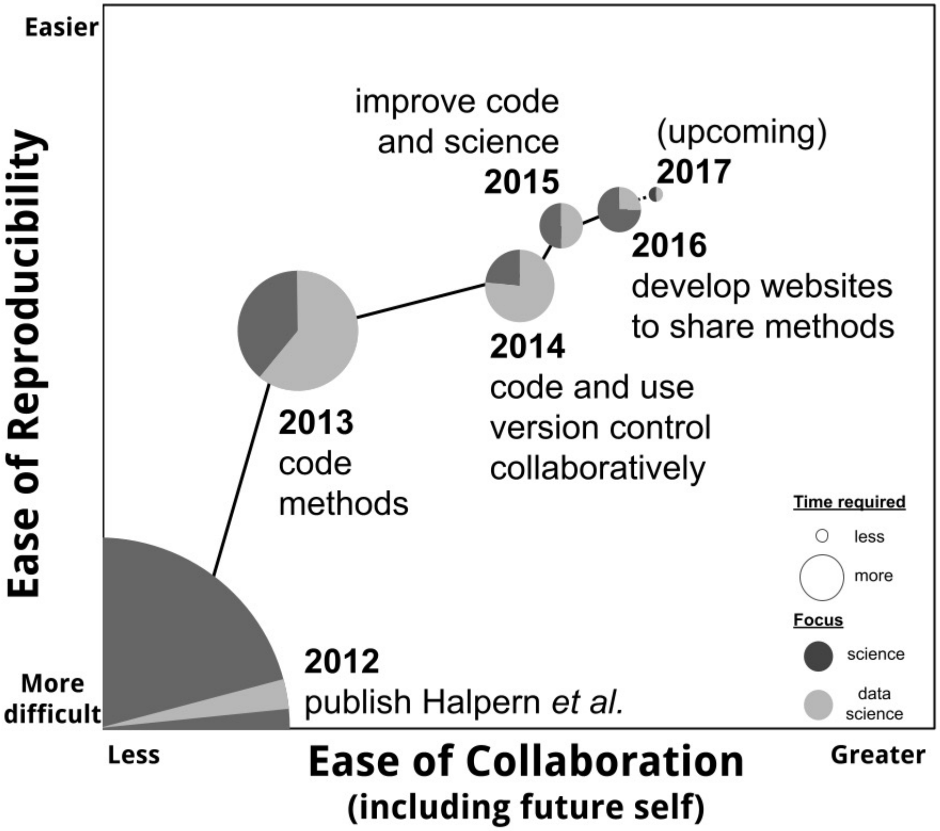
https://www.go-fair.org/fair-principles/

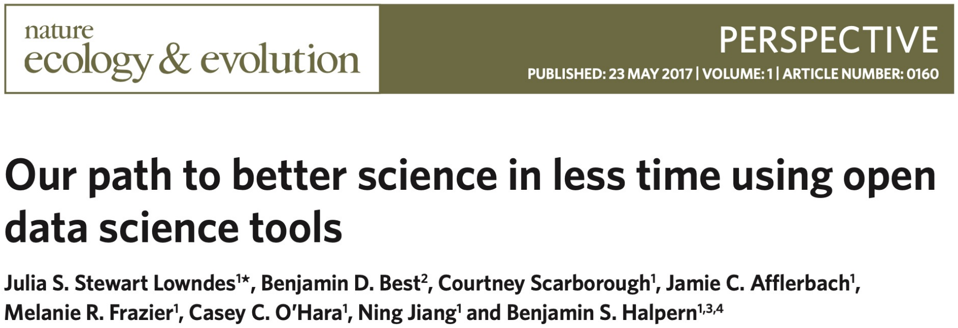


Global Indigenous Data Alliance

https://www.gida-global.org/

Reproducibility enhances collaboration





nature ecology & evolution

PERSPECTIVE

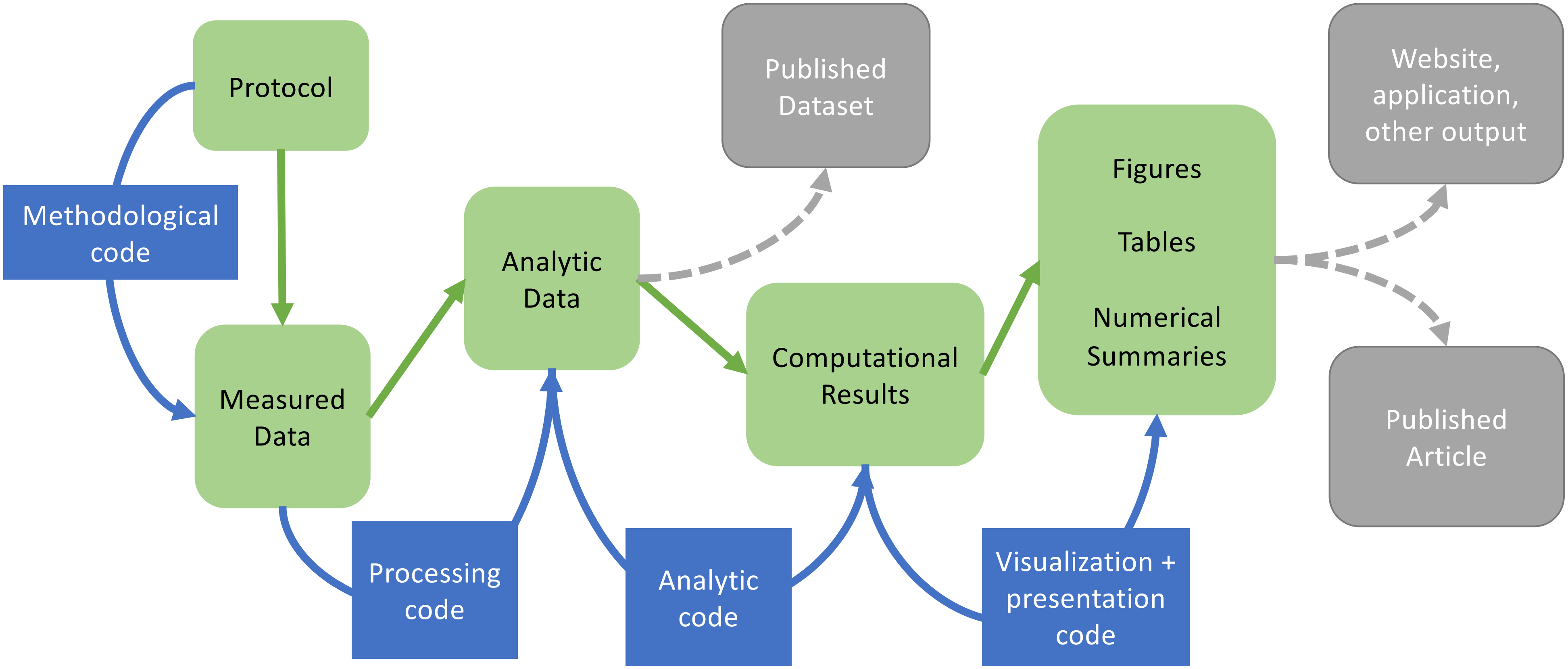
PUBLISHED: 23 MAY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0160

Our path to better science in less time using open data science tools

Julia S. Stewart Lowndes[1]*, Benjamin D. Best[2], Courtney Scarborough[1], Jamie C. Afflerbach[1], Melanie R. Frazier[1], Casey C. O'Hara[1], Ning Jiang[1] and Benjamin S. Halpern[1,3,4]

http://ohi-science.org/betterscienceinlesstime/

Research workflow



Adapted from R. Peng

How can we build a reproducible workflow?

Tools

- Version control (e.g. Git)
- Transparent collaboration (e.g. GitHub)
- Documentation
- Data repositories

Practices

- Think about the whole workflow
- Avoid doing things by hand
- Use best practices for coding
- Don’t save output

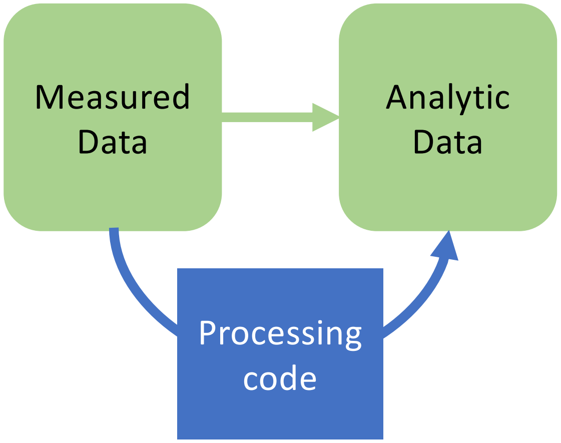
Collaboration exercise

1. Diagram (part of) your research workflow
2. Identify collaborators who contribute at different steps
3. Pick one step (e.g. moving from raw to processed data)
    1. What access do your collaborators need to data, analysis, or products at this step? How do they contribute?
    2. How do you maintain reproducibility with these collaborators at this step?

Don’t forget to include your future self as a collaborator!

Project-oriented workflows

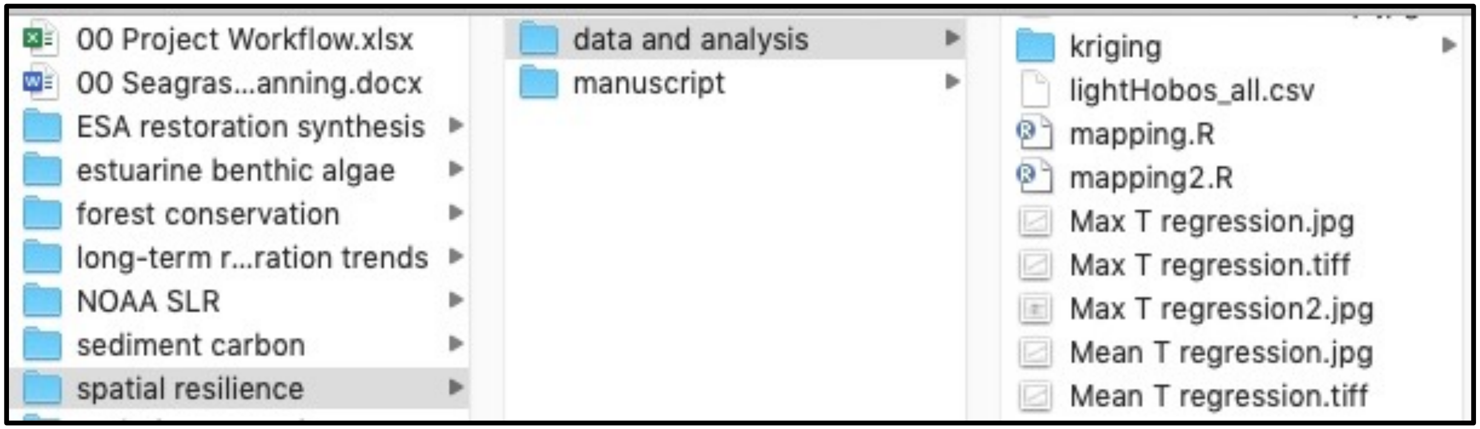
- Separate ‘workflow’ from ‘product’
  - Workflow = personal choices
  - Product = elements you want to reproduce
- Avoid hard-wiring your workflow into your product
- Organize work into ‘projects’



https://www.tidyverse.org/blog/2017/12/workflow-vs-script/
https://rstats.wtf/project-oriented-workflow.html

File and project organization

How can file organization enhance your research workflow?

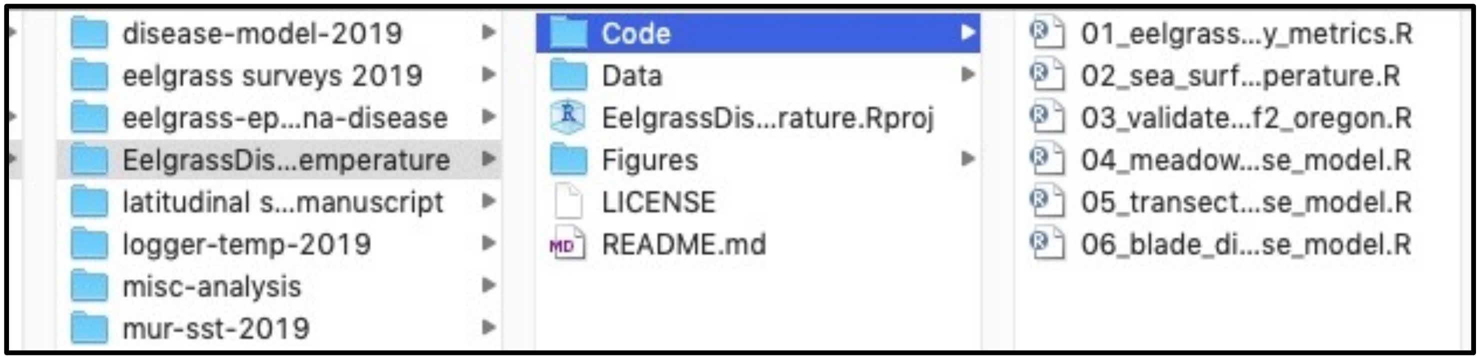


00 Project Workflow.xlsx
00 Seagras...anning.docx
ESA restoration synthesis
estuarine benthic algae
forest conservation
long-term r...ration trends
NOAA SLR
sediment carbon
spatial resilience

data and analysis
manuscript

kriging
lightHobos_all.csv
mapping.R
mapping2.R
Max T regression.jpg
Max T regression.tiff
Max T regression2.jpg
Mean T regression.jpg
Mean T regression.tiff

VS

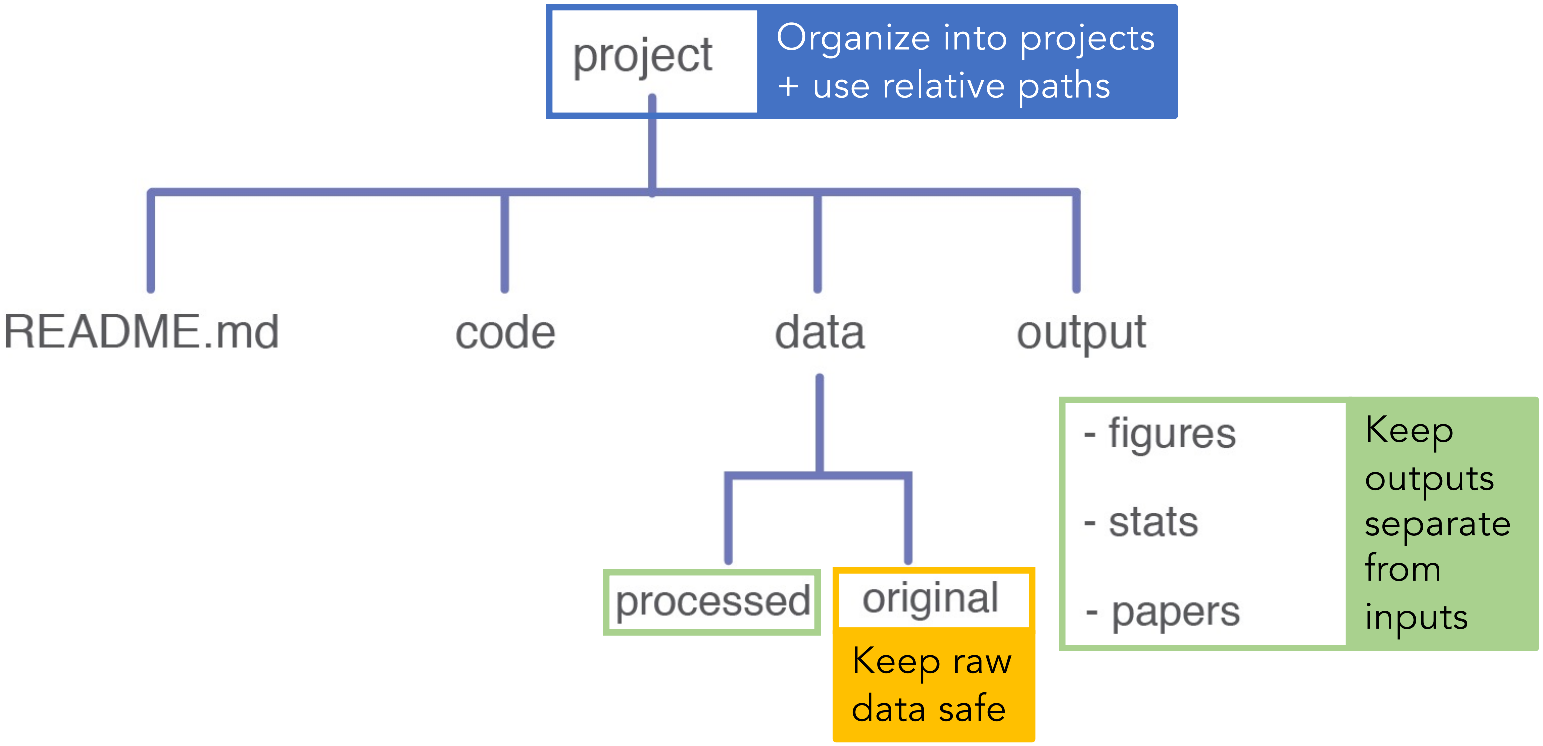


disease-model-2019
eelgrass surveys 2019
eelgrass-ep...na-disease
EelgrassDis...emperature
latitudinal s...manuscript
logger-temp-2019
misc-analysis
mur-sst-2019

Code
Data
EelgrassDis...rature.Rproj
Figures
LICENSE
README.md

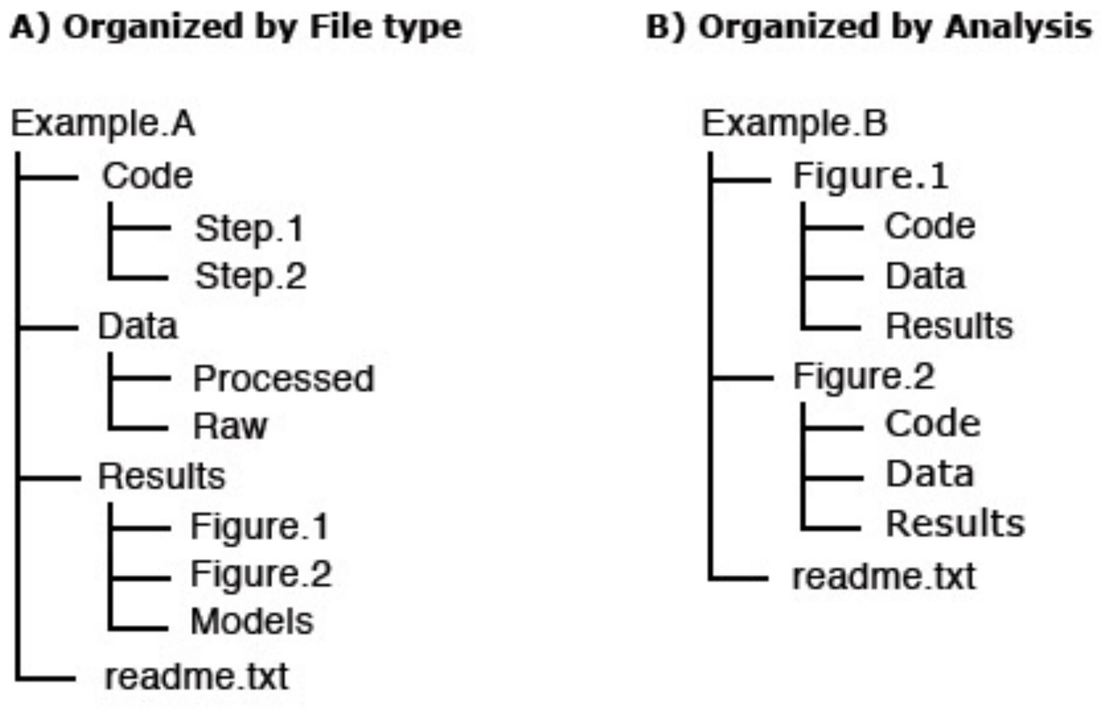
01_eelgrass...y_metrics.R
02_sea_surf...perature.R
03_validate...f2_oregon.R
04_meadow...se_model.R
05_transect...se_model.R
06_blade_di...se_model.R

Best practices for project structure



Adapted from [*https://reproducible-science-curriculum.github.io/*](https://reproducible-science-curriculum.github.io/)

No one way to organize your research



A) Organized by File type

```
Example.A
├── Code
│   ├── Step.1
│   └── Step.2
├── Data
│   ├── Processed
│   └── Raw
├── Results
│   ├── Figure.1
│   ├── Figure.2
│   └── Models
└── readme.txt
```

B) Organized by Analysis

```
Example.B
├── Figure.1
│   ├── Code
│   ├── Data
│   └── Results
├── Figure.2
│   ├── Code
│   ├── Data
│   └── Results
└── readme.txt
```

- Decide what works for you!
- Aim for consistency
- Automate?

Adapted from Helsinki University Library

Best practices for file and folder naming

- Machine readable
  - Avoid spaces, special characters
  - Deliberate_delimiters
- Human readable
  - CamelCase
  - more_deliberate-delimiters
- Works well with default ordering
  - 01_first_script
  - 10_tenth_script
  - 2002-09-06_data.csv
  - 2004-06-09_data.csv

Adapted from J. Bryan

File Organization Exercise

1. Consider the files for (one of) your research projects. Diagram or screenshot your directory structure.

What works and what doesn’t work about this structure?

Who else might need access to these files?

2. Assess your naming scheme for the files related to this project.

What kinds of files do you create and in what formats?

What are the unique characteristics of these files? E.g. date created, experiment number, investigator, location

Use the unique identifiers to draft file names

3. Create a systemic folder hierarchy

How can you group the individual files into folders?

Can you improve the directory structure to address the needs you identified in (1)?

Adapted from MIT Libraries