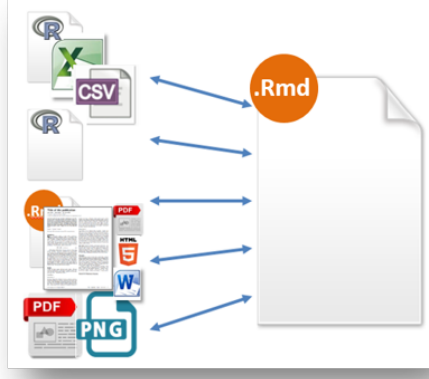
Workflow, managing files, and naming things

497 / 597 Reproducible Research



Richard Layton

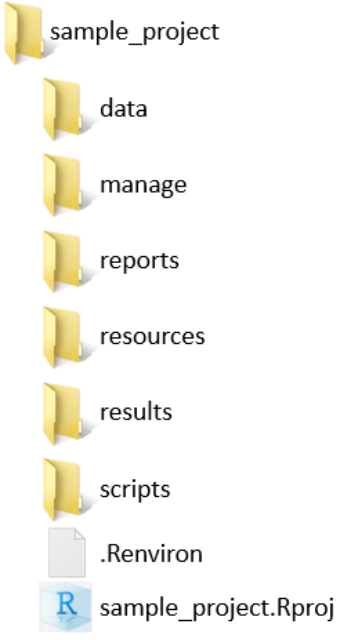
Rose-Hulman Institute of Technology
Fall 2018

Principles for reproducible research

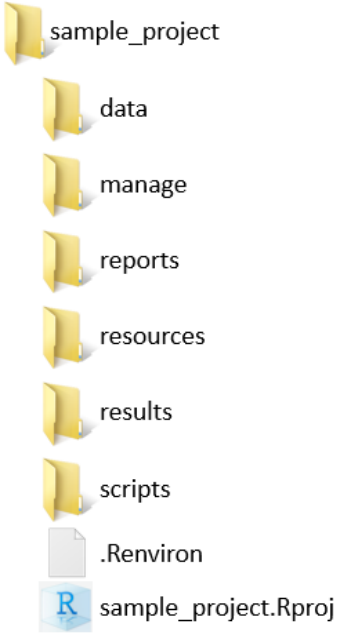
- Organize for reproducibility from the beginning
  - Adopt a directory naming scheme
  - Adopt a file naming scheme
- Everything you do (convert data files, clean data, analyze data) should be accomplished via code. Point /click /copy /paste are not reproducible.
- Explicitly link files

  `source(filename)`, `save(filename)`, `read(filename)`, `includegraphics(filename)`
- Don't repeat yourself (DRY)
  - create functions if you find yourself copying/pasting code
  - execute a task from one location only

Organize for reproducibility from the beginning

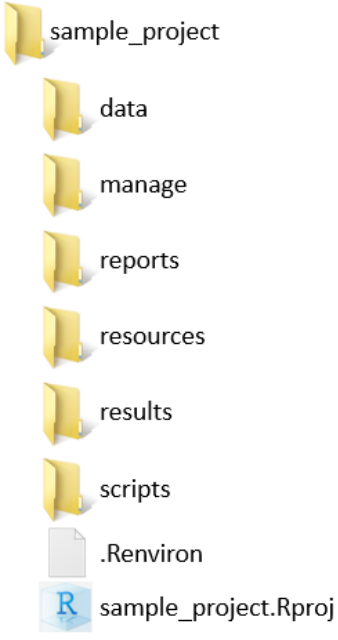


Organize for reproducibility from the beginning



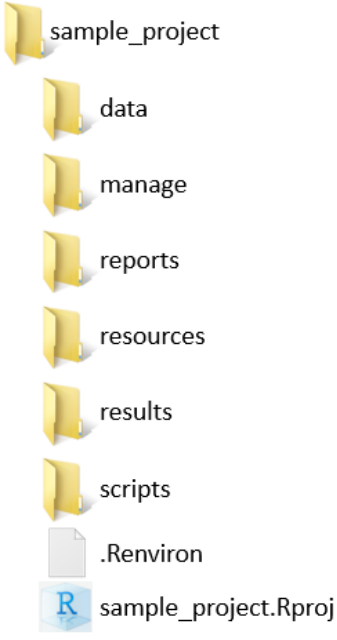
- working directory (relative file paths start here)

Organize for reproducibility from the beginning



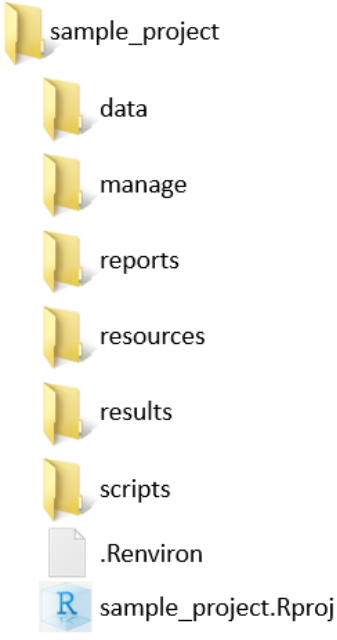
- working directory (relative file paths start here)
- unaltered raw data

Organize for reproducibility from the beginning



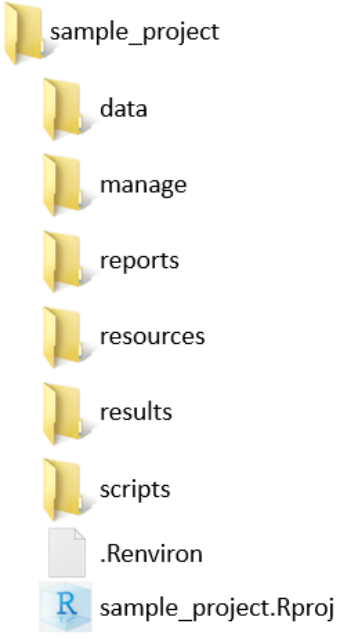
- working directory (relative file paths start here)
- unaltered raw data
- administrative files, not version controlled

Organize for reproducibility from the beginning



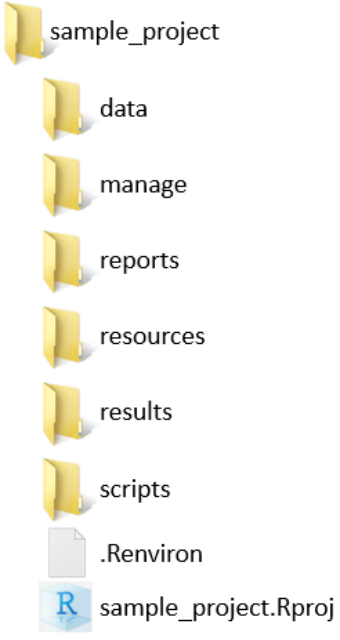
- working directory (relative file paths start here)
- unaltered raw data
- administrative files, not version controlled
- Rmd file(s) of the project report(s)

Organize for reproducibility from the beginning



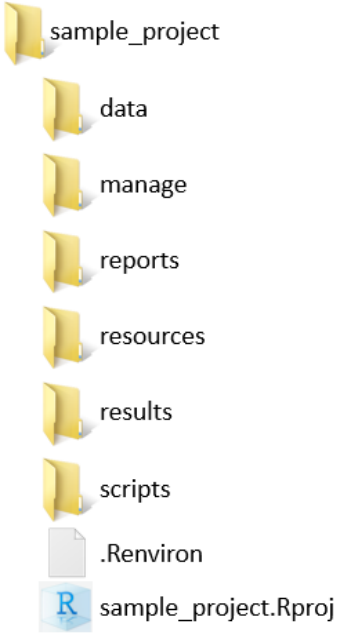
- working directory (relative file paths start here)
- unaltered raw data
- administrative files, not version controlled
- Rmd file(s) of the project report(s)
- images and pdfs from other sources

Organize for reproducibility from the beginning



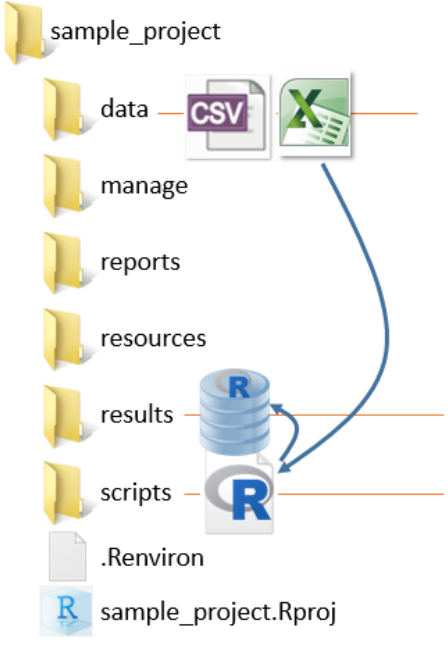
- working directory (relative file paths start here)
- unaltered raw data
- administrative files, not version controlled
- Rmd file(s) of the project report(s)
- images and pdfs from other sources
- save script output (tidy data and graphs) here

Organize for reproducibility from the beginning



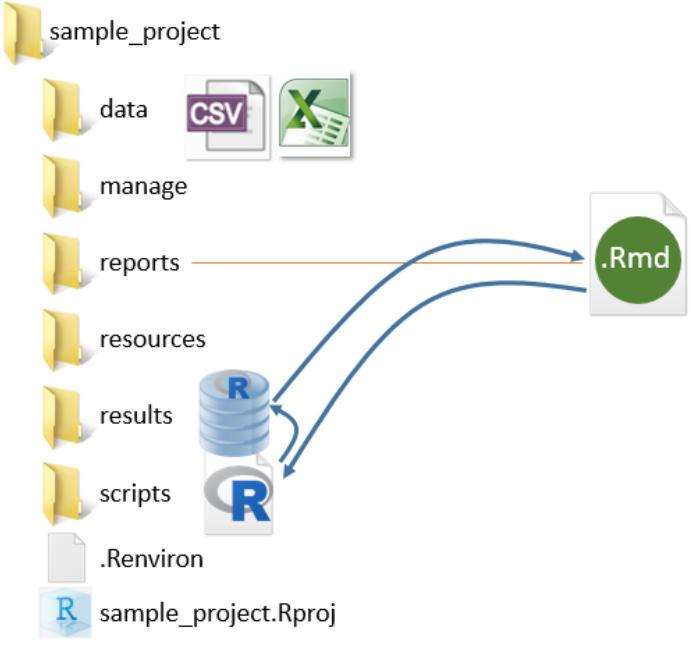
- working directory (relative file paths start here)
- unaltered raw data
- administrative files, not version controlled
- Rmd file(s) of the project report(s)
- images and pdfs from other sources
- save script output (tidy data and graphs) here
- R files to tidy data, do analysis, & create graphs

Do everything with a script: data



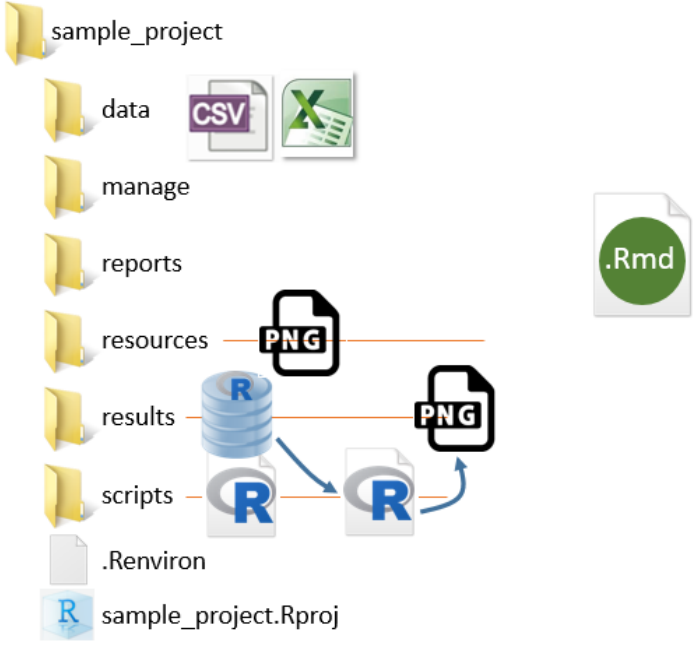
- raw data files
- read by R scripts
- produce tidy data saved in results

Explicitly link files: start the report



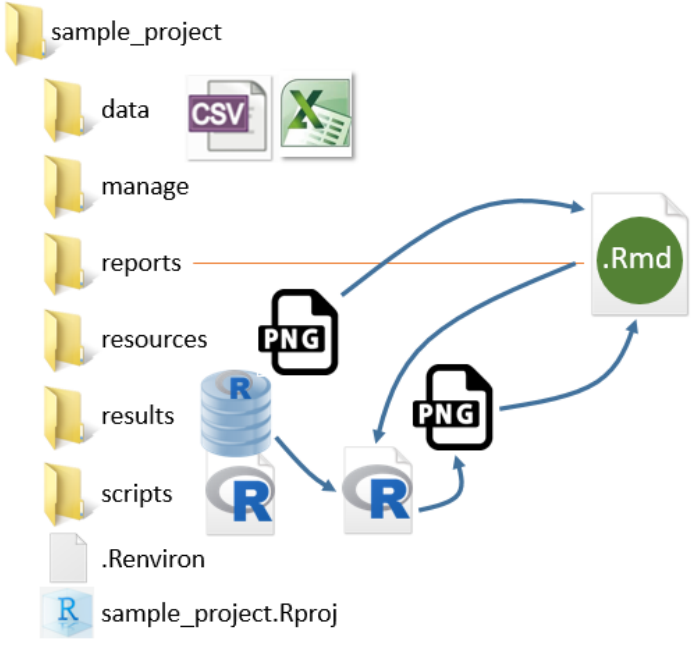
- prose that explains the work
- R code chunks to execute the scripts
- import data to create data tables

Do everything with a script: analysis and graphs



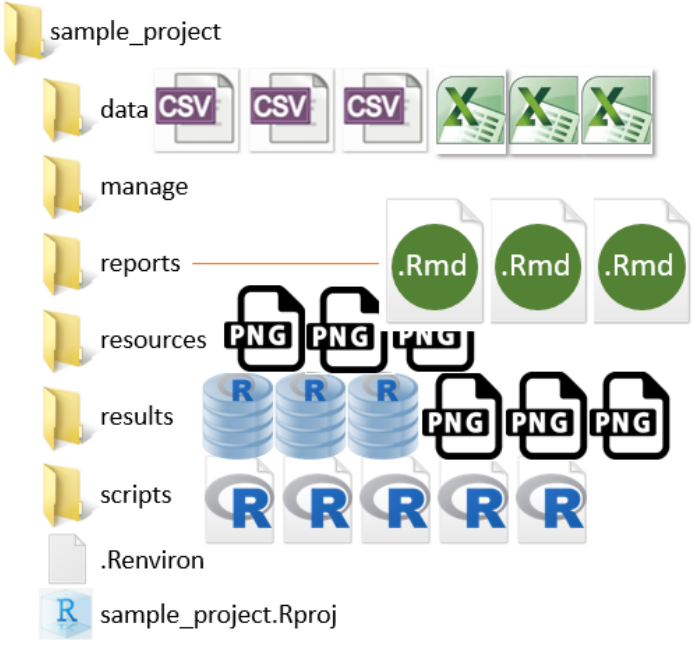
- R scripts read tidy data from results
- producing graphs saved in results
- other images stored in resources

Explicitly link files: continue the report



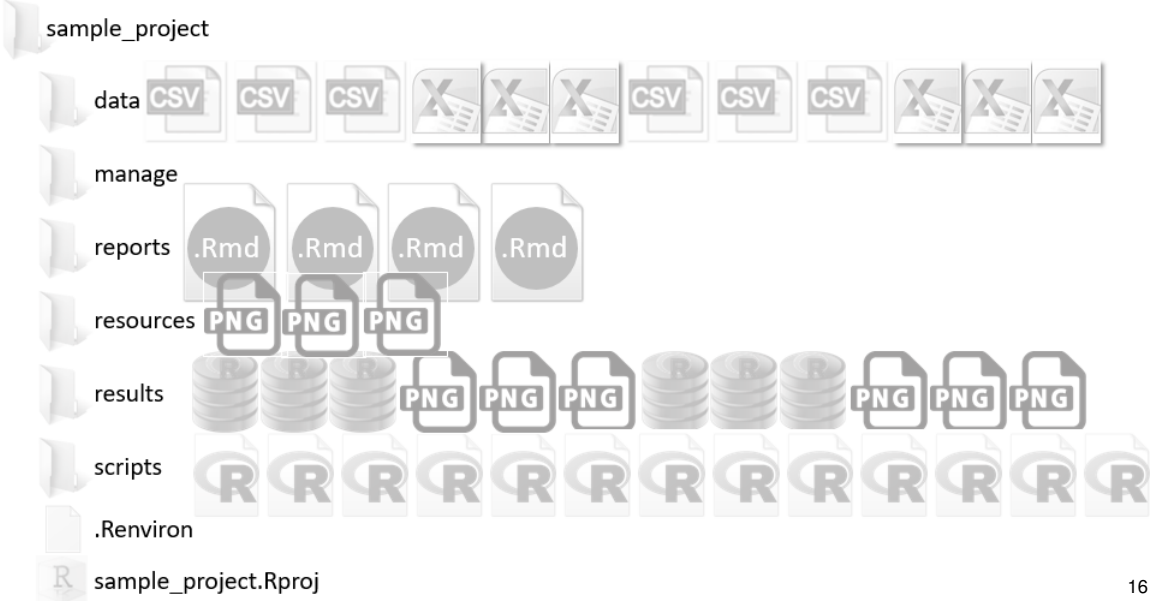
- prose that explains the work
- R code chunks to execute the scripts
- import images

One R Markdown script for each project report



- proposal
- progress report
- final report

By the end of a project, you will have a lot of files! Adopt a file naming scheme at the beginning.



Example file names for a calibration project

Project directory

```
load-cell-calibration\
  |-- data\
  |-- manage\
  |-- reports\
  |-- resources\
  |-- results\
  |-- scripts\
  |-- load-cell-calibration.Rproj
  |-- .Renviron
  |-- .gitignore
  `-- README.md
```

Sub-directories

```
data\
  |-- 001_raw-data_2018-07-25.xlsx
  `-- 002_raw-data_2018-08-01.xlsx

resources\
  `-- 901_load-cell-setup-786x989px.png
```

```
reports\
  |-- 101_proposal_2018-07-05.Rmd
  |-- 102_progress_2018-08-12.Rmd
  |-- 103_report_2018-09-03.Rmd
  |-- 101_proposal_2018-07-05.docx
  |-- 102_progress_2018-08-12.docx
  `-- 103_report_2018-09-03.docx

results\
  |-- 401_data-wide.csv
  |-- 402_data-tidy.csv
  |-- 403_graph-draft.png
  |-- 404_calibr_regression.png
  `-- 405_calibr_graph.png

scripts\
  |-- 401_data-wide.R
  |-- 402_data-tidy.R
  |-- 403_graph-draft.R
  |-- 404_calibr_regression.R
  `-- 405_calibr_graph.R
```

A script does one thing

one script	produces	one result
`401_data-wide.R`	→	`401_data-wide.csv`
`403_graph-draft.R`	→	`403_graph-draft.png`

- short, between 60–100 lines
- generally produces one object written to file, e.g., CSV, PNG
- simplifies editing, testing, readability, debugging

The full workflow is documented by the prose and the R code chunks in the Rmd file

````
---
output: word_document
---

```{r echo=FALSE}
library(tidyverse)
library(knitr)
```

# Introduction
Prose to explain the context
of the report

# Data
Prose to explain the data

```{r echo=FALSE}
# create tabulated data
source("scripts/401_data-wide.R")
```

```{r echo=FALSE}
# import and print tabulated data
df <- read_csv("results/401_data-wide.csv")
kable(df)
```

# Results
Prose to explain the results

```{r echo=FALSE}
# create the graph
source("scripts/405_calibr_graph.R")

# import the graph
includegraphics("results/405_calibr_graph.png")
```

# Conclusion
Prose to explain the conclusions
````