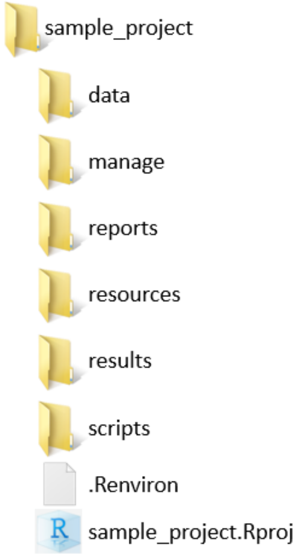
Principles for a reproducible workflow

497 / 597 Reproducible Research



sample_project

data

manage

reports

resources

results

scripts

.Renviron

sample_project.Rproj

Richard Layton

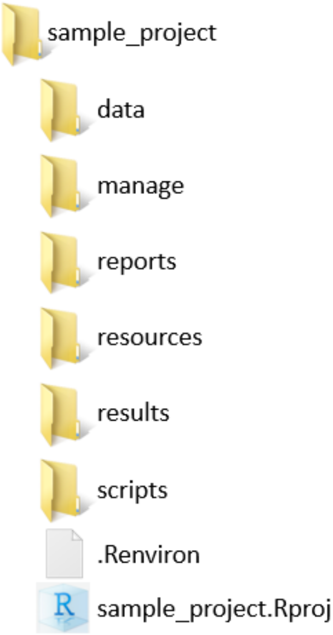
Rose-Hulman Institute of Technology

Fall 2018

Organize for reproducibility from the beginning

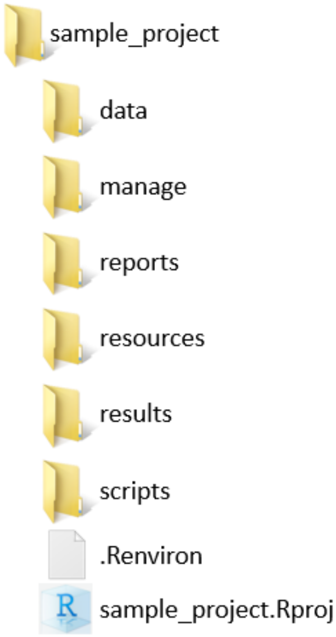
- Plan your directory structure
- Script everything — point/click/copy/paste is not reproducible
- Strive for simplicity & readability
- Link files explicitly
- Adopt a file naming scheme
- Use version control

From the beginning — plan your directory structure



```
sample_project
    data
    manage
    reports
    resources
    results
    scripts
    .Renviron
    sample_project.Rproj
```

From the beginning — plan your directory structure



sample_project

data

manage

reports

resources

results

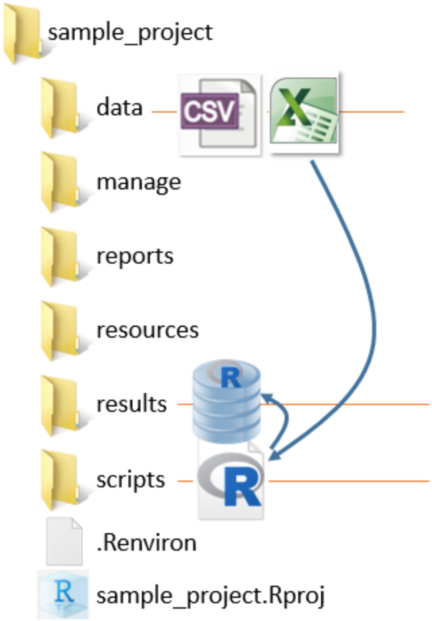
scripts

.Renviron

sample_project.Rproj

- 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

Script everything

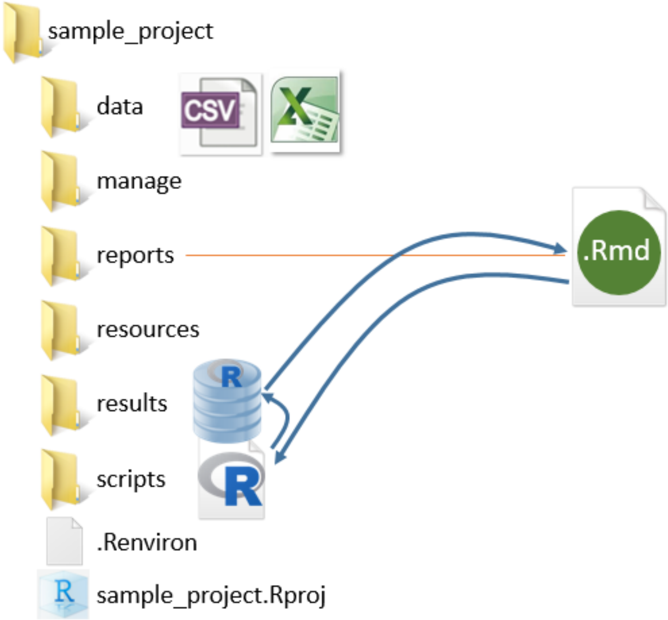


Use an R script to

- read a raw data file
- produce tidy data saved to results

Raw data files are stored unaltered.

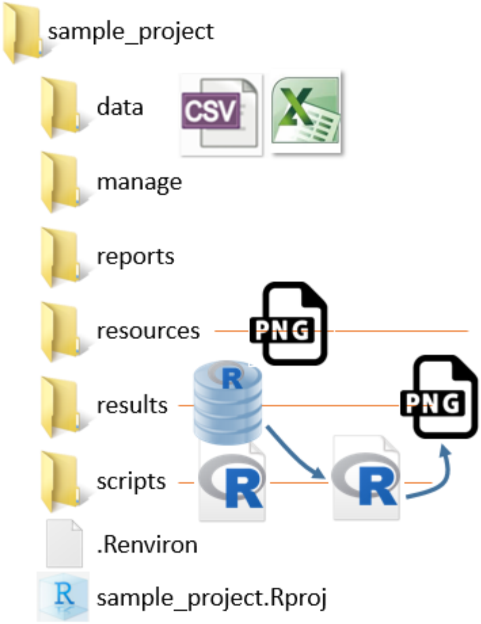
Link files explicitly



Start the Rmd script

- write prose to explain the work
- write R code chunks to execute the scripts
- import data from results to create data tables

Script everything

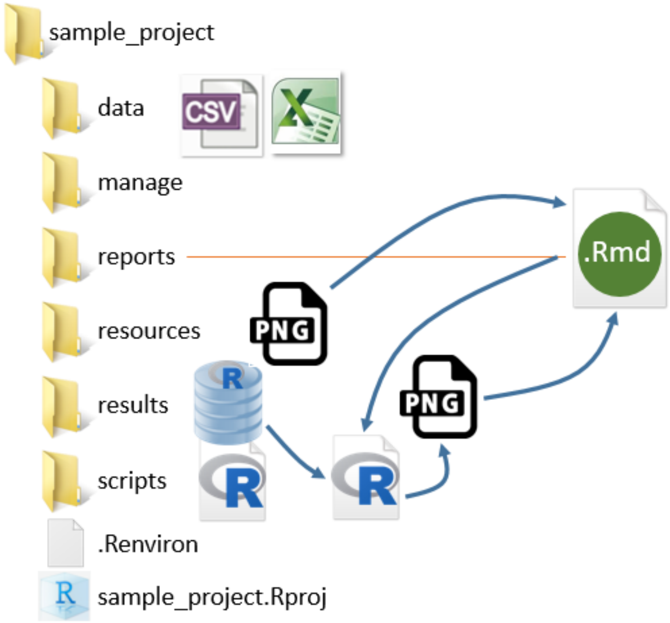


Use an R script to

- read tidy data from results
- produce a graph saved to results

Non-reproducible images stored in resources

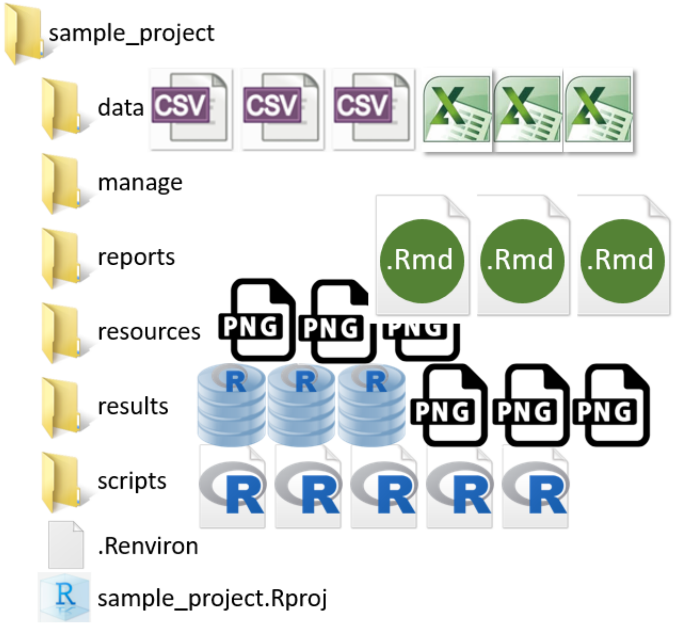
Link files explicitly



Continue the report

- write prose to explain the work
- write R code chunks to execute the scripts
- import images

Strive for simplicity & readability



sample_project
data
manage
reports
resources
results
scripts
.Renviron
sample_project.Rproj

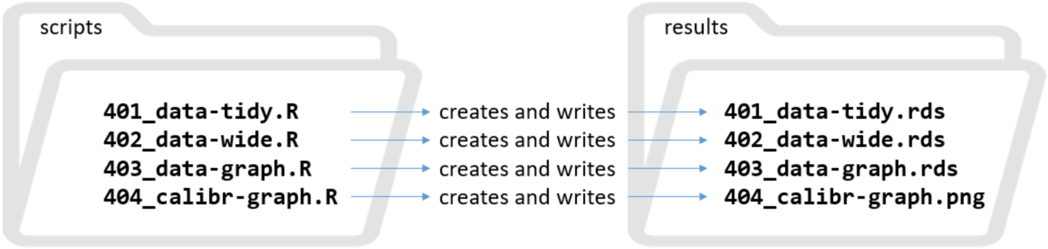
One Rmd script for each project milestone

- proposal
- progress report
- final report

Strive for simplicity & readability

R scripts are generally short, between 60–100 lines, to

- produce one object written to file, e.g., CSV, PNG
- simplify editing, testing, & debugging
- improve readability



scripts

401_data-tidy.R → creates and writes → 401_data-tidy.rds

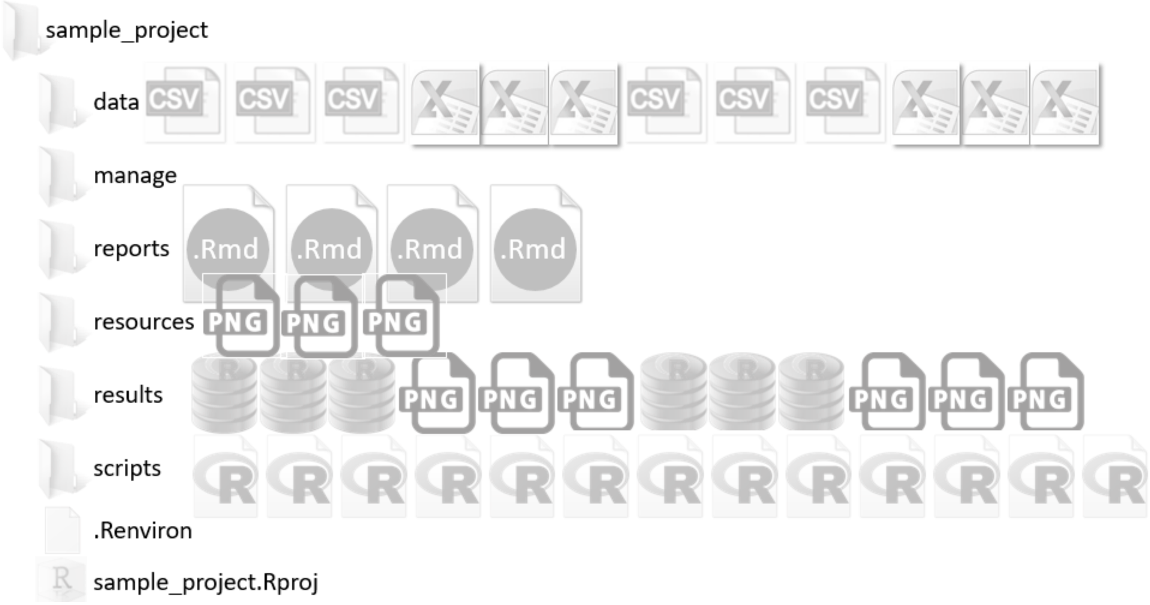
402_data-wide.R → creates and writes → 402_data-wide.rds

403_data-graph.R → creates and writes → 403_data-graph.rds

404_calibr-graph.R → creates and writes → 404_calibr-graph.png

results

From the beginning — adopt a file naming scheme



In this scheme, every file name starts with 3 digits

Use “slugs“ to facilitate file searches, for example **`_report_`**

000-series manage

```
001_RFP_2018-05-25.pdf
002_contract_2018-06-05.pdf
invoice_201801.pdf
invoice_201802.pdf
```

100-series data

```
101_raw-data_2018-07-25.xlsx
102_raw-data_2018-08-01.xlsx
```

200-series resources

```
201_apparatus_2018-08-12.png
202_load-cell_2018-08-12.png
```

300-series reports

```
301_proposal_2018-07-05.Rmd
302_progress_2018-08-12.Rmd
303_report_2018-09-03.Rmd
```

400-series scripts

```
401_data-tidy.R
402_data-wide.R
403_data-graph.R
404_calibr-graph.R
```

400-series also used for results

```
401_data-tidy.rds
402_data-wide.rds
403_data-graph.rds
404_calibr-graph.png
```

Use version control

See the website for instructions

obtain a free account for asynchronous collaboration

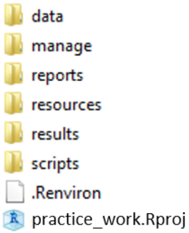
create an online repository for each project

link each repository to a local RStudio Project

commit and push your changes to the repository

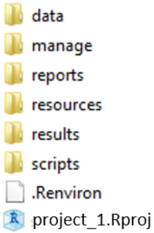
Create the folders after version control is set up

`practice_work/`



data
manage
reports
resources
results
scripts
.Renviron
practice_work.Rproj

`project_1/`



data
manage
reports
resources
results
scripts
.Renviron
project_1.Rproj

See the website for instructions

- the **`.Rproj`** file denotes the R Project working directory level
- copy the **`.Renviron`** file to the top level of every project

Organize for reproducibility from the beginning

- Plan your directory structure
- Script everything — point/click/copy/paste is not reproducible
- Strive for simplicity & readability
- Link files explicitly
- Adopt a file naming scheme
- Use version control