

# Basic elements of file management

ME 447/547 Visualizing Data

---

Richard Layton

December 2018

Rose-Hulman Institute of Technology

# Effective file management begins when the project begins

Plan your directory  
structure

- carpentry
- data
- data-raw
- design
- figures
- reports
- README.Rmd
- portfolio.Rproj

Adopt a scheme to consistently  
name your files

Explicitly link files using  
relative file paths



## Planning the structure

---

# The portfolio project has a mandatory structure

 carpentry 

 data 

 data-raw 

 design 

 figures 

 manage 

 practice 

 reports 

 resources 



























 .gitignore 

 .Renviron 



























 README.Rmd 

 portfolio.Rproj 














# Open **portfolio.Rproj** to start every work session

|   |                 |   |
|---|-----------------|---|
|  | carpentry       |    |
|  | data            |    |
|  | data-raw        |    |
|  | design          |    |
|  | figures         |    |
|  | manage          |    |
|  | practice        |    |
|  | reports         |    |
|  | resources       |    |
|  | .gitignore      |    |
|  | .Renviron       |    |
|  | README.Rmd      |    |
|  | portfolio.Rproj |  Sets the project directory as the working directory |














# README introduces your portfolio to the reader

|   |                 |   |
|---|-----------------|---|
|  | carpentry       |    |
|  | data            |    |
|  | data-raw        |    |
|  | design          |    |
|  | figures         |    |
|  | manage          |    |
|  | practice        |    |
|  | reports         |    |
|  | resources       |    |
|  | .gitignore      |    |
|  | .Renviron       |    |
|  | README.Rmd      |  Creates the main page of your portfolio website |
|  | portfolio.Rproj |    |

## Other top-level files perform administrative duties














|   |   |
|---|---|
|  carpentry       | ◀   |
|  data            | ◀   |
|  data-raw        | ◀   |
|  design          | ◀   |
|  figures         | ◀   |
|  manage          | ◀   |
|  practice        | ◀   |
|  reports         | ◀   |
|  resources       | ◀   |
|  .gitignore      | ◀ Directs Git to ignore specific files              |
|  .Renviron       | ◀ Stores packages in a library separate from base R |
|  README.Rmd      | ◀   |
|  portfolio.Rproj | ◀   |

## Raw data are never edited manually














|   |                 |                             |
|---|-----------------|-----------------------------|
|  | carpentry       | <                           |
|  | data            | <                           |
|  | data-raw        | < Data in its original form |
|  | design          | <                           |
|  | figures         | <                           |
|  | manage          | <                           |
|  | practice        | <                           |
|  | reports         | <                           |
|  | resources       | <                           |
|  | .gitignore      | <                           |
|  | .Renviron       | <                           |
|  | README.Rmd      | <                           |
|  | portfolio.Rproj | <                           |





























# Data carpentry converts raw data to tidy data

|   |   |
|---|---|
|  carpentry       | ◁ R scripts that turn raw data into tidy data |
|  data            | ◁   |
|  data-raw        | ◁ Data in its original form                   |
|  design          | ◁   |
|  figures         | ◁   |
|  manage          | ◁   |
|  practice        | ◁   |
|  reports         | ◁   |
|  resources       | ◁   |
|  .gitignore      | ◁   |
|  .Renviron       | ◁   |
|  README.Rmd      | ◁   |
|  portfolio.Rproj | ◁   |














# Data carpentry converts raw data to tidy data

|   |  |
|---|--|
|  carpentry       | ◁ R scripts that turn raw data into tidy data  |
|  data            | ◁ Tidy data saved here, read by design scripts |
|  data-raw        | ◁  |
|  design          | ◁  |
|  figures         | ◁  |
|  manage          | ◁  |
|  practice        | ◁  |
|  reports         | ◁  |
|  resources       | ◁  |
|  .gitignore      | ◁  |
|  .Renviron       | ◁  |
|  README.Rmd      | ◁  |
|  portfolio.Rproj | ◁  |



























# Graph design converts to tidy data to graphs

|   |  |
|---|--|
|  carpentry       |   |
|  data            |  Tidy data saved here, read by design scripts |
|  data-raw        |   |
|  design          |  R scripts that create and save graphs        |
|  figures         |   |
|  manage          |   |
|  practice        |   |
|  reports         |   |
|  resources       |   |
|  .gitignore      |   |
|  .Renviron       |   |
|  README.Rmd      |   |
|  portfolio.Rproj |   |






















# Graph design converts to tidy data to graphs

|   |   |
|---|---|
|  carpentry       | <   |
|  data            | <   |
|  data-raw        | <   |
|  design          | < R scripts that create and save graphs         |
|  figures         | < Graphs saved here, imported by report scripts |
|  manage          | <   |
|  practice        | <   |
|  reports         | <   |
|  resources       | <   |
|  .gitignore      | <   |
|  .Renviron       | <   |
|  README.Rmd      | <   |
|  portfolio.Rproj | <   |



























# Reports commingle data, scripts, graphs, prose, and references

|   |   |
|---|---|
|  carpentry       |    |
|  data            |    |
|  data-raw        |    |
|  design          |    |
|  figures         |  Graphs saved here, imported by report scripts |
|  manage          |    |
|  practice        |    |
|  reports         |  Reports draw from data, graphs and resources  |
|  resources       |  Image downloads and bibliography files        |
|  .gitignore      |    |
|  .Renviron       |    |
|  README.Rmd      |    |
|  portfolio.Rproj |    |














# README creates the main page of your portfolio website

|   |  |
|---|--|
|  carpentry       |   |
|  data            |   |
|  data-raw        |   |
|  design          |   |
|  figures         |   |
|  manage          |   |
|  practice        |   |
|  reports         |  Reports draw from data, graphs and resources |
|  resources       |   |
|  .gitignore      |   |
|  .Renviron       |   |
|  README.Rmd      |  Provides explicit links to every report      |
|  portfolio.Rproj |   |

# Reduce clutter by excusing some resources from version control

|   |   |
|---|---|
|  carpentry       |                                        |
|  data            |                                        |
|  data-raw        |  Version control optional              |
|  design          |                                        |
|  figures         |                                        |
|  manage          |  Correspondence and project management |
|  practice        |  Scripts for practicing and learning R |
|  reports         |                                        |
|  resources       |                                        |
|  .gitignore      |  Directs Git to ignore specific files  |
|  .Renviron       |                                        |
|  README.Rmd      |                                        |
|  portfolio.Rproj |                                        |

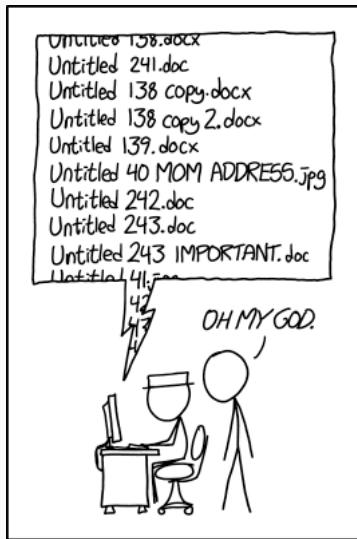
# Summary: Use the given directory structure for the portfolio

|   |  |
|---|--|
|  carpentry       | ◁ R scripts that create and save tidy data             |
|  data            | ◁ Tidy data saved here, read by design scripts         |
|  data-raw        | ◁ Data in its original form (version control optional) |
|  design          | ◁ R scripts that create and save graphs                |
|  figures         | ◁ Graphs saved here, imported by report scripts        |
|  manage          | ◁ Correspondence and project management                |
|  practice        | ◁ Scripts for practicing and learning R                |
|  reports         | ◁ One report per display type                          |
|  resources       | ◁ Image downloads and bibliography files               |
|  .gitignore      | ◁ Directs Git to ignore specific files                 |
|  .Renvi          | ◁ Stores packages in a library separate from base R    |
|  README.Rmd      | ◁ Creates the main page of your portfolio website      |
|  portfolio.Rproj | ◁ Sets the project directory as the working directory  |



## Naming files

---



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

Source: <https://xkcd.com/1459/>

# Three basic principles should guide your choice of filenames

Filenames should be **machine readable**

- use delimiters “\_” and “-” instead of spaces
- avoid symbols, punctuation marks, and case-sensitivity

Filenames should be **human readable**

- include information about the file content

Filenames should be **friendly to default ordering**

- start filenames with a numeric ID
- use leading zeros, e.g., 001, 002, ..., 999

# A sample set of portfolio file names illustrates the principles

Numeric display ID starts every file name: **d1, d2, ..., d7**

Hyphenated **content-information** supports human readability

```
carpentry/ d7_extract-and-tidy.R
data/      d7_survey-data.csv
data-raw/  d7_survey-data-raw.csv
design/     d7_div-stack-bar.R
figures/   d7_div-stack-bar.png
reports/   d7_report.Rmd
```

All lowercase, no special symbols, no spaces

Underscores support machine readability

## Add a sequence number 01, 02, etc., for related files

When related files are run in order


















```
carpentry/ d7_01-extract-financials.R  
           d7_02-extract-mortality.R  
           d7_03-tidy-inequality-data.R
```

When content is saved in different forms

```
data/      d7_01-survey-data.csv  
           d7_02-survey-data-wide.csv
```

Not version numbers. For version control, use version control (git).

# With a plan for managing files, we can start writing them

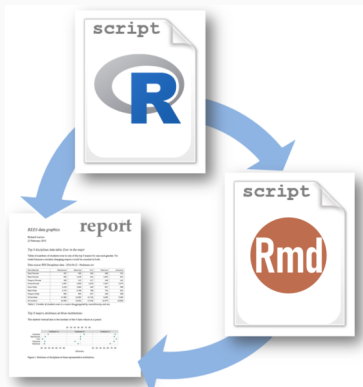
|   |  |
|---|--|
|  carpentry |  d7_extract-and-tidy.R                |
|  data      |  d7_survey-data.csv                   |
|  data-raw  |  d7_survey-data-raw.csv               |
|  design    |  d7_div-stack-bar.R                   |
|  figures   |  d7_div-stack-bar.png                 |
|  manage    |  |
|  practice  |  d7_practice-div-stack-bar.R          |
|  reports   |  d7_report.Rmd                        |
|  resources |  d7_Heiberger-Robbins-2014-Likert.pdf |

Icons for csv, pdf, and png by [Freepik](#) from [Flaticon](#) licensed [CC BY 3.0](#).

## Linking files

---

# Explicitly linking files supports reproducibility



Remove all ambiguity about what files are used to create a report

portfolio.Rproj sets the working directory and supports relative file paths



## Relative file paths document the **data tidying** workflow

Write an R script for data tidying

 **carpentry/d7\_extract-and-tidy.R**

that reads the raw data, prepares it for graphing,

 **read\_csv("data-raw/d7\_survey-data-raw.csv")**

and writes the tidy dataframe to the data directory.

 **write\_csv(dataframe, "data/d7\_survey-data.csv")**

## Relative file paths document the **graph design** workflow

Write an R script for graph design

 **design/d7\_div-stack-bar.R**

that reads the tidy data, creates the graph,

 **read\_csv("data/d7\_survey-data.csv")**

and writes the image to the figures directory.

 **ggsave("figures/d7\_div-stack-bar.png")**

## The **report script** runs all the files in order

Write an Rmd report script containing the report text

 **reports/d7\_report.Rmd**

interleaved with Rmd code chunks that run every R script,

 **source("carpentry/d7\_extract-and-tidy.R")**

 **source("design/d7\_div-stack-bar.R")**

import data to print a data table,

 **read\_csv("data/d7\_survey-data.csv")**

and import figures.

 **include\_graphics("figures/d7\_div-stack-bar.png")**

# The **README script** includes links to each report



README.Rmd

creates the portfolio main webpage

`## Displays and critiques`

Your prose as needed.

`[D1 Title](reports/d1_report.md)`

`[D2 Title](reports/d2_report.md)`

`[D3 Title](reports/d3_report.md)`

`[D4 Title](reports/d4_report.md)`

`[D5 Title](reports/d5_report.md)`

`[D6 Title](reports/d6_report.md)`

`[D7 Title](reports/d7_report.md)`

## Portfolio of data displays

Your name

2018-12-04

(Place an image to illustrate your best work.)

### Introduction

Your prose.

### Displays and critiques

Your prose as needed.

[D1 Title](#) (graph type)

[D2 Title](#) (graph type)

[D3 Title](#) (graph type)

[D4 Title](#) (graph type)

[D5 Title](#) (graph type)

[D6 Title](#) (graph type)

[D7 Title](#) (graph type)

### Discussion notes

Your prose as needed.

[Reading prompts](#)

[Presentation prompts](#)

# Effective file management begins when the project begins

Plan your directory  
structure

- carpentry
- data
- data-raw
- design
- figures
- reports
- README.Rmd
- portfolio.Rproj

Adopt a scheme to consistently  
name your files

Explicitly link files using  
relative file paths



# References

Bryan J (2015) Naming things. <https://speakerdeck.com/jennybc/how-to-name-files>

Bryan J (2018) Excuse me, do you have a moment to talk about version control? *The American Statistician* 72(1), 20–27 (doi:10.1080/00031305.2017.1399928)

Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L and Teal TK (2017) Good enough practices in scientific computing. *PLoS Computational Biology* 13(6)  
<https://doi.org/10.1371/journal.pcbi.1005510>