BEST PRACTICES OF DATA ORGANIZATION

Mikhail Dozmorov

Summer 2018

Spreadsheets

- Store data in text files, delimiter-separated
- Spreadsheets may be used for data entry and storage
- Analysis and visualization should happen outside of spreadsheets

Broman, Karl W, and Kara H. Woo. "Data Organization in Spreadsheets." Accessed March 29, 2018. https://doi.org/10.7287/peerj.preprints.3183v1. - Excel spreadsheet tips and best practices

General spreadsheet tips

- Use a consistent data layout in multiple files
  - Your primary data file should contain just the data and nothing else: no calculations, no graphs
- Enter one piece of information in one cell
  - Don't use formatting, e.g., coloring, to convey additional meaning to the cell value. Instead, add another column, e.g., "outlier" with a boolean `TRUE` or `FALSE` indicator variable
- Use a consistent format for all dates, YYYY-MM-DD strongly recommended
  - Use "Text" format, where possible, to avoid autoreformatting, e.g., "SEPT9" will become "9/9/2018" displayed as "9-Sep"
  - Use "'forward tick" trick to force text format

http://www.datacarpentry.org/spreadsheet-ecology-lesson/02-common-mistakes/

General spreadsheet tips (cont.)

- Fill in all cells. No empty cells
  - Use a consistent fixed code for any missing values, e.g. “NA”. Don’t use numerical values like 999
- Use consistent file names for similar datasets
  - Never include “final” in a file name
- Save and distribute the data in plain text files
  - Use comma-delimited (CSV) format

Spreadsheet horror stories

- A public archive of spreadsheet "horror stories"
- In 13 audits of real-world spreadsheets, an average of 88% spreadsheets contained errors (Panko, Raymond R. "What we know about spreadsheet errors." Journal of Organizational and End User Computing (JOEUC) 10, no. 2 (1998): 15-21.)

http://www.eusprig.org/horror-stories.htm

Project organization principles

- One project = one folder
  - Create readable names for subfolders/code. E.g. `00_raw_data`, `01_raw_data_QC` etc.
  - Choose file names carefully. Don't put spaces and special characters in file names!
- Be sure to get and keep any/all data and meta-data possible
- Get the data in the most-raw form possible. Keep the original files, names intact. (gzipped) CSV Text format is the most preferable
- Separate data from code. Use relative paths in code. Create multiple `README.md`

Convert Excel files to CSV https://github.com/dilshod/xlsx2csv

Another project organization idea

```
project/
| data/
| | processing_scripts
| | raw/
| | proc/
| tools/
| | src/
| | bin/
| exps
| | pipeline_scripts
| | results/
| | analysis_scripts
| | figures/
```

Project organization principles

- **Script everything** - All analysis steps, including data cleaning (removal of outliers, correcting numbers, typos, renaming columns etc.) should be scripted
- **Scalability and universality** - ask yourself a question, if the data are updated (e.g., additional subjects) or you find some artifact that needs fixing, can you just "press a button" to update? If you work on a similar project, can you reuse your existing scripts with minimal modifications?
- **Document everything** - Text format, human readable. Explicitly tie files together. Have a plan to organize, store and make your work understandable by others

File naming principles

- Machine readable
- Human readable
- Plays well with default ordering

http://www.mnhs.org/preserve/records/electronicrecords/erfnaming.php

File naming principles

- Machine readable
  - Regular expression and globbing friendly
    - Avoid spaces, punctuation, accented characters, case sensitivity
    - Easy to compute on
  - Deliberate use of delimiters, e.g. "-", "_" (think `cut -d "_" -f1,3,5`, `grep Notes` commands)

```
2018-05-16_Lecture_Slides_01.pdf
2018-05-16_Lecture_Notes_01.pdf
2018-05-16_Lecture_Slides_02.pdf
2018-05-16_Lecture_Notes_02.pdf
```

File naming principles

- Human readable
  - Name contains info on content
- Easy to figure out what something is, based on its name

```
01_preprocessing.R
02_quality-control.R
helper01_rename-files.sh
helper02_merge-duplicates.py
```

File naming principles

- Plays well with default ordering
  - Put something numeric first
  - Use the ISO 8601 standard for dates (YYYY-MM-DD *or* YYMMDD)
  - Left pad other numbers with zeros

Good

```
2018-05-16_Lecture_Slides_01.pdf
2018-05-16_Lecture_Notes_01.pdf
```

Bad

```
10_final-figures.R
1_data-cleaning.R
2_quality-control.R
```

http://en.wikipedia.org/wiki/ISO_8601

Data management

- Save the raw data
- Ensure that raw data are backed up in more than one location
- Create the data you wish to see in the world
- Create analysis-friendly data
- Record all the steps used to process data
- Anticipate the need to use multiple tables, and use a unique identifier for every record
- Submit data to a reputable DOI - issuing repository so that others can access and cite it

Data management

- Data are cheap, time is expensive
  - A terabyte hard drive costs about US$50 retail, which means that 50 Gigabytes costs less than US$5
  - How much of your time is needed to generate 50Gb of code?