Development Data Boot Camp
Introduction and Preparation: General Workflow Management

Ge Sun

University of Notre Dame

May 12, 2023

Outline

Introduction

Folder Structure

Outline

Introduction

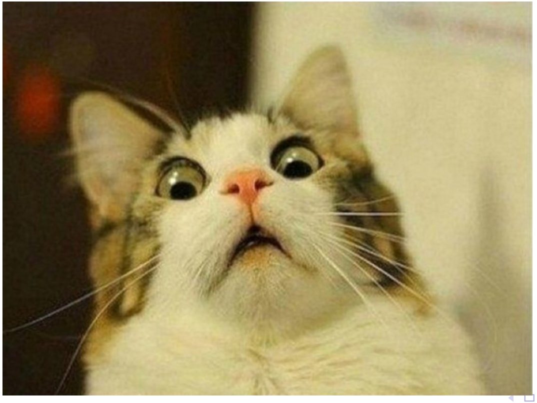
Folder Structure

Warm up

- **Case 1**:
  I spent a whole summer as a research assistant to complete a project. In the end, I generate hundreds of codes and data files (that's a lot of hard work!). Then, I put them into a folder and handed them over to my supervisor.

Warm up

- **Case 1**:
  I spent a whole summer as a research assistant to complete a project. In the end, I generate hundreds of codes and data files (that's a lot of hard work!). Then, I put them into a folder and handed them over to my supervisor.
- My supervisor:



Warm up

- **Case 1**:
  I spent a whole summer as a research assistant to complete a project. In the end, I generate hundreds of codes and data files (that's a lot of hard work!). Then, I put them into a folder and handed them over to my supervisor.
- To spare my supervisor from shock:
  - ASK YOUR ADVISOR how she/he wants them organized at the start
  - Archive subfolders are **very** useful
  - Provide a “map” (Readme files)

Warm up

- **Case 2**:
  The other two research assistants and I write code that refers to one set of data files (It is so huge that it is hard to save it on everyone's computer). We coordinate our work through a shared drive, like Google Drive and Dropbox.
- What are the things we need to take care of in this collaborative situation?

Warm up

- **Case 2**:
  The other two research assistants and I write code that refers to one set of data files (It is so huge that it is hard to save it on everyone's computer). We coordinate our work through a shared drive, like Google Drive and Dropbox.
- What are the things we need to take care of in this collaborative situation?
  - to keep the raw data as it is
  - divide the tasks, separate the "workspace"
  - MEET regularly. It is both helpful and **fun**!

Warm up

- Workflow management is crucial for effective and efficient data analysis.
- Experts from various disciplines, including computer science and data management, have dealt with the issues we've just discussed for years and have come up with solutions that can be useful in our work.
- Workflow management encompasses many aspects beyond just folder structure, which is what we will be discussing today. Throughout the course, we will delve into other important components of workflow management.

Outline

Introduction

Folder Structure

Folder Structure

- Rules of thumbs:
  1. ASK YOUR ADVISOR at the start.
  2. Organization should be done at the **first** step.
  3. Subfolders are **very** useful.
  4. Separate directories by function.
     * data, codes, results, paper, graph, etc
  5. Separate files into inputs and outputs.
     * raw data, temporary results, final results

Folder Structure

- Rules of thumbs:
  1. ASK YOUR ADVISOR at the start.
  2. Organization should be done at the **first** step.
  3. Subfolders are **very** useful.
  4. Separate directories by function.
     - data, codes, results, paper, graph, etc
  5. Separate files into inputs and outputs.
     - raw data, temporary results, final results
  6. About naming the files:
     - do not name them as "*myproject_final.do*". Use the **dates**.

Folder Structure

- Rules of thumbs:
  1. ASK YOUR ADVISOR at the start.
  2. Organization should be done at the **first** step.
  3. Subfolders are **very** useful.
  4. Separate directories by function.
     - data, codes, results, paper, graph, etc
  5. Separate files into inputs and outputs.
     - raw data, temporary results, final results
  6. About naming the files:
     - do not name them as *"myproject_final.do"*. Use the **dates**.
     - try to use **underline** *"myproject_0509.do"* instead of **space** *"myproject 0509.do"* to connect the words in the file name. (not matter that much in Stata)

Folder Structure

- Rules of thumbs:
  1. ASK YOUR ADVISOR at the start.
  2. Organization should be done at the **first** step.
  3. Subfolders are **very** useful.
  4. Separate directories by function.
     - data, codes, results, paper, graph, etc
  5. Separate files into inputs and outputs.
     - raw data, temporary results, final results
  6. About naming the files:
     - do not name them as *"myproject_final.do"*. Use the **dates**.
     - try to use **underline** *"myproject_0509.do"* instead of **space** *"myproject 0509.do"* to connect the words in the file name. (not matter that much in Stata)
  7. Using "Archives"
     - Put all the historical versions (like dofiles) in the archive subfolder and only leave the most current version on the "surface" to smooth your nerve.

Folder Structure Examples

- A Single Directory Containing Everything

```
                      ---C:/tv_and_potato/---

chips.csv     mergefiles.do         tv_potato_submission.pdf
cleandata.do  regressions_alt.do    tv_potato.tex
extractOB.xls regressions_alt.log   tv.csv
fig1.eps      regressions.do        tvdata.dta
fig2.eps      regressions.log       rundirectory.bat
figures.do    tables.txt            export_to_csv.stc
```

Re-organized Structure

```
---C:/build---             ---C:/analysis---

/input                     /input

    extract0B.xls              tvdata.dta (link to C:/build/output)


/code                      /code

    rundirectory.bat           rundirectory.bat

    export_to_csv.stc          regressions.do

    mergefiles.do              regressions_alt.do



/output                    /output

    tvdata.dta                 fig1.eps

                               fig2.eps

                               tables.txt


/temp                      /temp

    chips.csv                  regressions.log

    tv.csv                     regressions_alt.log
```

Another Folder Structure — from Taryn

01 origdata
- all the original data files stored here
- may use subfolders to distinguish between multiple data sources

02 cleandata
- all the processed data stored here

03 syntax
- all my do files are here.
- there is a master.do file that organizes all of the do files for analysis

04 output
- anything created in the do files, e.g. graphs, tables; is stored here.

05 writing
- the actual paper is stored here.
- used to be word; now tex files

Another Folder Structure — from Taryn

06 epapers

- all the literature stored for the project.
- sometimes organized into sub folders.

07 replication

- all the syntax and orig data required to replicate the project.
- a readme file is in here too

08 presentations

- all presentations for the project

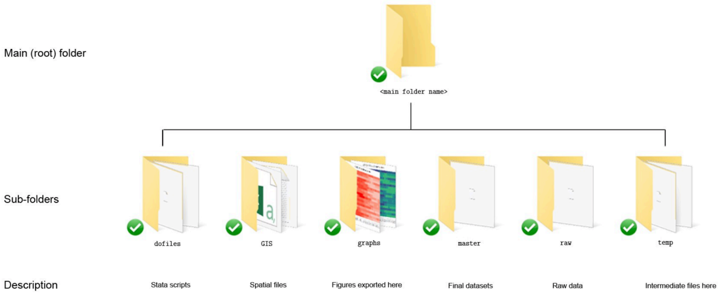
09 funding

- if the project has funding, the relevant proposal documents, grant budgets etc are in here

10 submissions

- separate sub folders for different journal submissions, referee reports, and revisions.

One more Example for your Future Reference



Main (root) folder

<main folder name>

Sub-folders

dofiles

GIS

graphs

master

raw

temp

Description

Stata scripts

Spatial files

Figures exported here

Final datasets

Raw data

Intermediate files here

- **raw**: all the raw files
- **dofiles**: the scripts to process, clean, and analyze the raw files
- **temp**: intermediate files that are generated from the raw data
- **master**: the final data that is ready for analysis
- **graphs**: the figures

Reading Materials:

- Code and Data for the Social Sciences: A Practitioner's Guide. Ch4.
- The Stata Workflow Guide