Get Your Data Organized

HARVARD LONGWOOD MEDICAL AREA
RESEARCH DATA MANAGEMENT WORKING GROUP

Learning Objectives

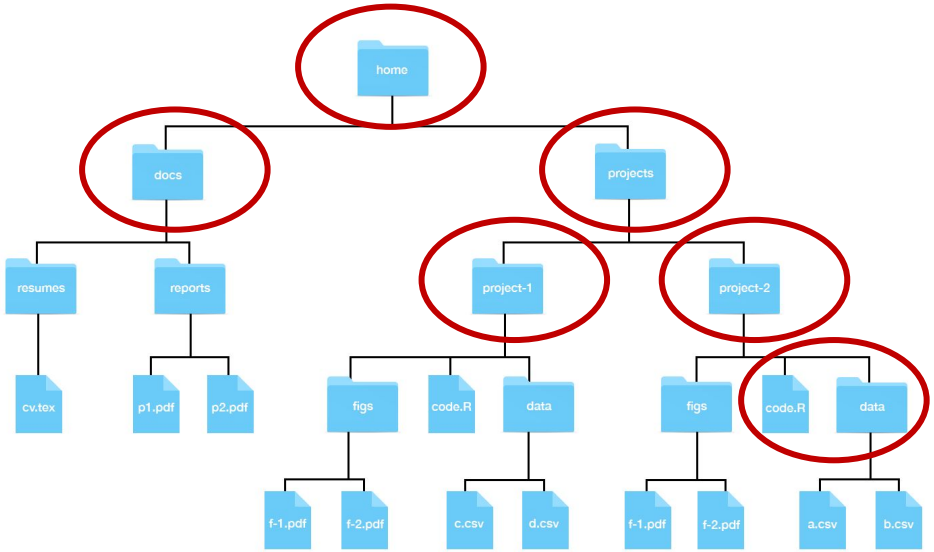
- Understand why project organization is essential for data management
- Learn best practices for organizing folders and files
- Learn best practices for file naming and versioning
- Find out who to contact for assistance

Get Organized!

- Establish systems and use them consistently
- Any system is better than none
- Separate folders for data or project stages
- One project, one folder
- Date-based folders utilizing ISO 8601 standards

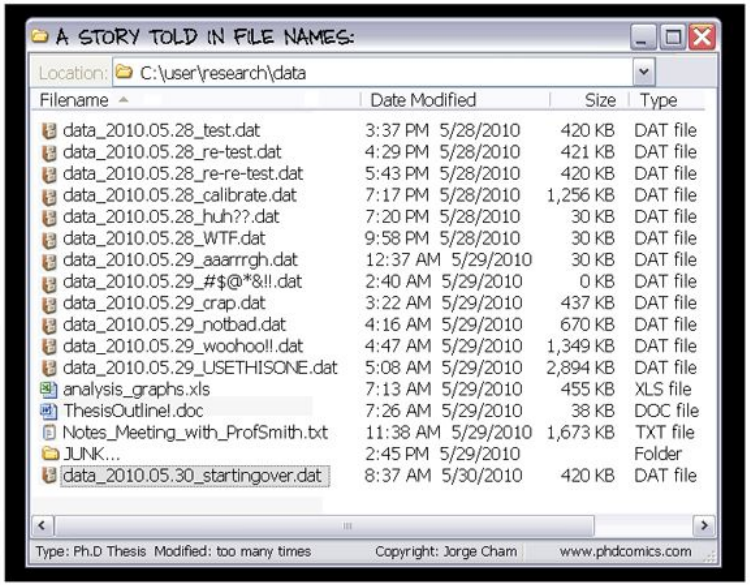
Folder System

- Organize your data hierarchically
- Identify ways to divide your data into categories (Attributes)
  - Project
  - Time
  - Location
  - File type
- Top level organization is the most important attribute



Naming Conventions

- Should be consistent and used consistently
  - Use ISO 8601 standard for dates (YYYYMMDD)
  - Sequential numbering (e.g. 001, 010)
- Should be descriptive and provide context
  - Project or experiment name or acronym
  - Lab name/location or researcher name/initials
  - Date or date range of experiment
  - Type of data
  - Experiment conditions



http://phdcomics.com/comics.php?f=1323

https://datamanagement.hms.harvard.edu/file-naming-conventions

File Naming

USE DESCRIPTIVE NAMES

Bad name: file.txt

Ok name:
02-07-2020-mouse-data.txt

Good name:
2020-02-07-mouse-weight.csv

Human readability: name contains information about content

GO FROM GENERAL TO SPECIFIC

Bad name:
rep1-2-7-2020-gene-expression.csv

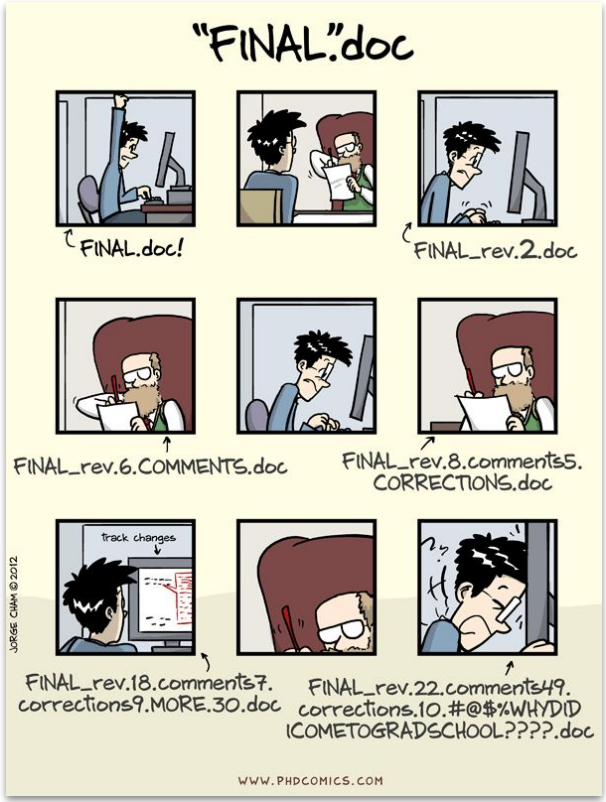
Ok name:
2020-02-07-rep1-gene-expression.csv

Good name:
2020-05-07-gene-expression-rep01.csv

Machine readability: can be sorted meaningfully

Version Control

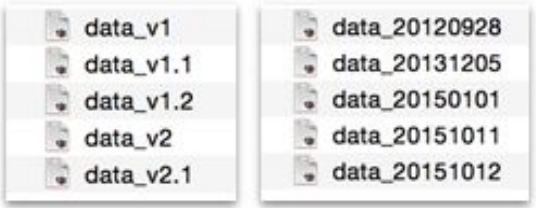
- Version control captures a snapshot of a file at any moment in time
- Provides a way to document and track changes
- Allows you to revert to previous (working) versions
- Enables collaboration so multiple people can work on a file at once
- Important to always record all changes:
  - Who is responsible?
  - What was changed?
  - When did it happen?



http://phdcomics.com/comics/archive.php?comicid=1531

Versioning Strategies

Basic – file names (e.g. V1; 01, 1.1, date)



Intermediate – built-in software capabilities

Advanced – version control software

https://datamanagement.hms.harvard.edu/versioning

Project Organization

- Create a directory structure for output files before running the analysis workflow
- Create README.txt files in higher level directories briefly describing their contents

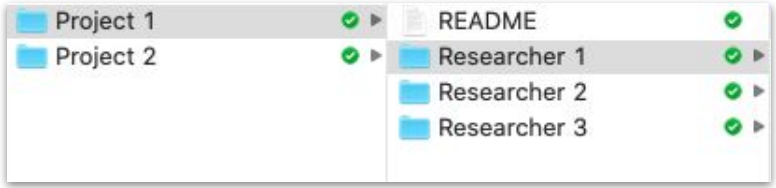
https://datamanagement.hms.harvard.edu/directory-structure

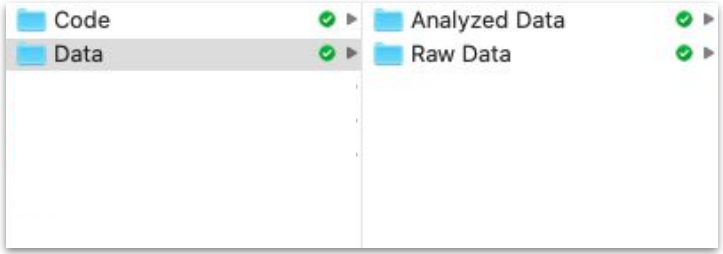
Project Organization

1. Put each project in its own directory, which is named after the project. Create a "README" file that outlines basic information about the project.

2. Create folders that will separate your code and data.

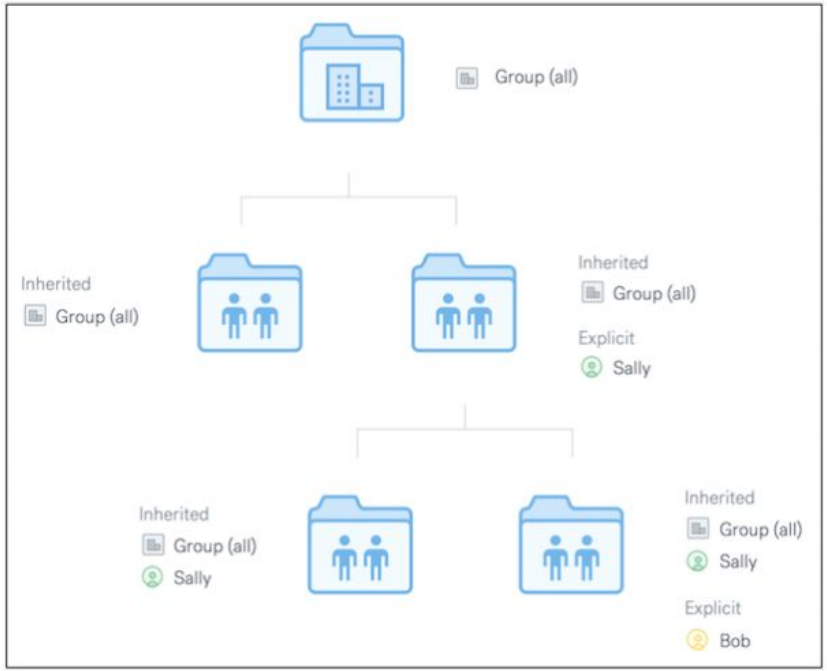
3. In your data folder, ensure that your **raw** data are separated from any data you have **processed** (i.e. your clean datasets).





Collaborative Folders

- Collaborative folders can help you share your research files with your group
- Creating a clear plan for file sharing will help you simplify the process and limit data risks
- **Examples**: Microsoft OneDrive, Google Drive, Dropbox



Example of team folder and shared subfolder structure
Example from Dropbox

Takeaways

- Data organization refers to the method of classifying and organizing data sets to make them more useful. Organization is a key aspect of data management and will help keep the project on track by saving time, storage, and data loss.
- Before you even start collecting or working with data, you should decide how you will structure and name files and folders. This will allow for standardized data collecting and analysis by many team members.
- Structure project folders hierarchical and divide data into categories. Apply a standard and descriptive file naming system and version control method.