2-FileManagement

July 9, 2018

1 Module 2: File Management (1.5 hours)

From “Old Files”

1.1 Organizing Projects

This section draws from Karl Broman’s steps toward reproducible research

- Encapsulate everything in one directory
- Separate raw data from derived data
- Separate the data from the code
- Use relative paths
- Choose file names carefully – more on that in the following section
- Avoid using “final” in a file name
- Write README files

1.2 File Naming

This section draws from Stanford Library’s data best practices

1.2.1 Basic Information

- Project or experiment name or acronym
- Location/spatial coordinates (e.g. multiple sites), if applicable
- Researcher name/initials
- Date or date range of experiment
- Type of data
- Conditions, if applicable
- Version number of file
- Three-letter file extension for application-specific files

1.2.2 Tips

- For dates, use YYYYMMDD or YYMMDD.
- Keep the file names short!
- Avoid special characters.
- When numbering, use leading zeros.
- Do not use spaces.
- Do use underscores, dashes, or CamelCase.

1.2.3 Example

http://bit.ly/naming_exemplar

1.3 File Formats

This section draws from Stanford Library's data best practices

1.3.1 Discussion

- What formats are you saving your files as?
- Are they proprietary or open formats?
- How does this affect your ability to read the file in the future?
- Can you think of an example when proprietary is the better option?

1.3.2 Some preferred file formats

- Containers: TAR, GZIP, ZIP
- Databases: XML, CSV
- Geospatial: SHP, DBF, GeoTIFF, NetCDF
- Moving images: MOV, MPEG, AVI, MXF
- Sounds: WAVE, AIFF, MP3, MXF
- Statistics: ASCII, DTA, POR, SAS, SAV
- Still images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- Tabular data: CSV
- Text: XML, PDF/A, HTML, ASCII, UTF-8
- Web archive: WARC

1.4 Data Provenance and Version Control

1.4.1 What is provenance?

This section draws from the Wikipedia article Provenance. > **Provenance** is the chronology of the ownership, custody or location of a historical object. The primary purpose of tracing the provenance of an object or entity is normally to provide contextual and circumstantial evidence for its original production or discovery, by establishing, as far as practicable, its later history, especially the sequences of its formal ownership, custody, and places of storage. [E]stablishing provenance is essentially a matter of documentation.

1.4.2 What is data provenance?

This section draws from the Wikipedia article Data lineage. > Data lineage includes the data's origins, what happens to it and where it moves over time. **Data provenance** documents the inputs, entities, systems, and processes that influence data of interest, in effect providing a historical record of the data and its origins. The generated evidence supports essential forensic activities such as data-dependency analysis, error/compromise detection and recovery, and auditing and compliance analysis. "*Lineage* is a simple type of *why provenance*."

1.4.3 How is version control related to provenance?

It's part of documentation!

From GitHub Guides: Hello World
"Git Commit" from xkcd by Randall Munroe

1.5 Storage/Backup/Archive/Preservation

1.5.1 Definitions

This section draws from DataONE's Education Module Lesson 6.

- **Storage**: the medium you're using to keep your active files, backups, and archives (e.g. hard drive, flash drive, CD/DVD, magnetic tape, cloud)
- **Backup**: periodic snapshots of current version; stored for short or near-long-term; often done on a somewhat frequent schedule

- **Archive**: final version for historical reference or disasters; stored for long-term; created at end of project or at major milestone
- **Preservation**: Includes backups and archiving as well as processes such as data conversion, reformatting, and rescue.

1.5.2 Cloud Storage Options at The University of Utah

Resource	Security	Collab1	Backup2	Max file size	Max allocated space	File type	Cost
REDCap	Restricted	yes	Yes	na	unlimited	any	free
LabArchives	Public	yes	Yes	4GB	unlimited	any	free
CHPC Group Space	Sensitive	??	Upon request	na	na	??	$150/TB
CHPC Archive	Sensitive	??	Yes	na	na	??	$120/TB
UBox	Restricted	yes	Yes	15GB	1TB	See Link	free
Google Drive	Public	yes	yes	See FAQ	unlimited	See FAQ	free

The CHPC also provides tools for secure large file transfer such as Globus. The table above was adapted from a handout created by Daureen Nesdill (Data Management Librarian at Marriott Library).

1.5.3 The 3-2-1 Rule

The section draws from Peter Krogh's Backup Overview.