Research Data Management: File Organization

Ashok Mudgapalli

Director of Research IT

Courtesy of Drs. McNeill and Bailey of MIT

Research Data Management Services

- Workshops
- Our website:
- Individual assistance/consultations
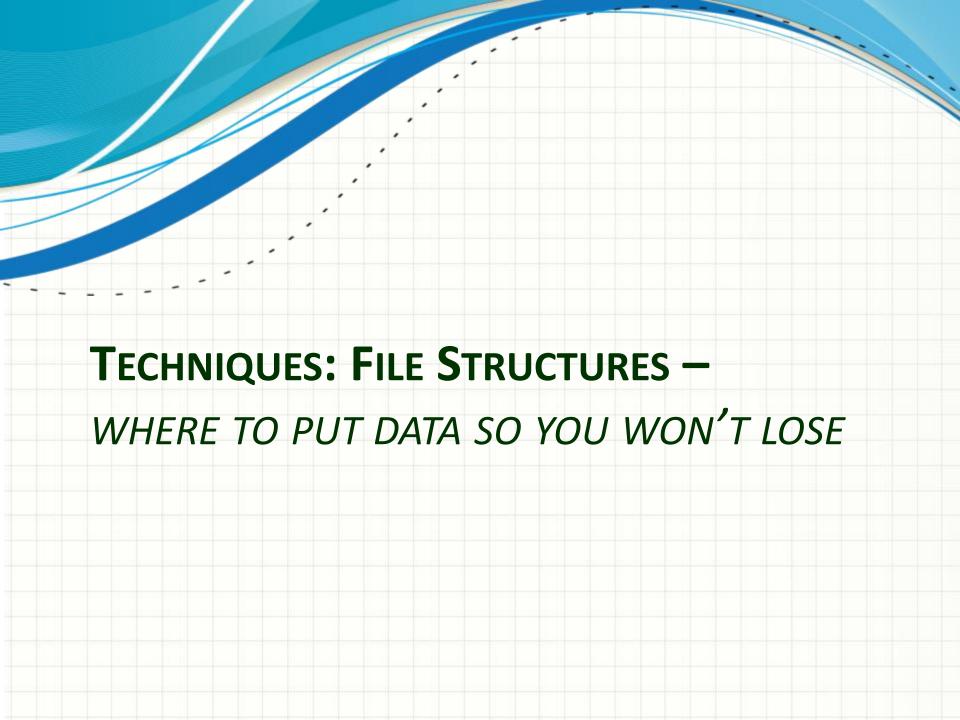- Contact: data-management@unmc.edu

What will you Learn

- Why file organization of your research data is important

- Specific techniques for organizing your research data, including developing plans for:

  – File structures - where to put data so you won't lose it (including tips on embedding metadata)

  – File naming - what to call data so you know what it is

  – A bit on version control - keeping track of data

- Will also include opportunities for:

  – Small group discussion

  – Exercise for organizing your own data

- Focuses on research data, but applies to other types of files as

Why file organization is important?

- You think you'll remember things, but over time...

- Multitude of formats and version of data and documentation

- Investment of time at the beginning in an efficient system can save time in the long run

- Good file management practices/naming protocols enable sharing with collaborators

Key principles

- Organization is a means to efficient research, not an end in itself
- Some extra work when you collect material may prevent a lot of future hassle; think of what information you need to document now so that your files make sense to you (and others) in the future
- There's no single right way to do it
- Establish and document a system that works for you
- Strike the balance between doing too much and too little: be realistic
- The 5 Cs: be Clear, Concise, Consistent, Correct, and Conformant

Techniques: File Structures –

Where to put data so you won't lose

Methods of organizing electronic material

- Hierarchical

    – Items organized in folders and sub folders

- Tag-based

    – Each item assigned one or more tags

- Remember: You can do a hybrid combination of hierarchical and tag-based

Hierarchical systems: benefits

- Familiar and widely used
- Good at representing the structure of information
  - Constructing the hierarchy can itself be a helpful exercise
- Similar items are stored together
- Sub-folders can function as task lists
- Great for location-based finding

Hierarchical systems: drawbacks

- Surprisingly hard work to set up and maintain – 'a heavyweight cognitive activity'
- Can be hard to get the right balance between breadth and depth
- Items can only go in one place
- Time consuming to reorganise if the hierarchy becomes out of date

Tag-based systems: benefits

- Items can go in more than one category
  - Moreover, multiple types of category can be used
- Many people find tagging quicker and easier than hierarchical filing
- When collaborating, can be easier to combine than hierarchical systems

Tag-based systems: drawbacks

- Not how operating systems store files
- If material isn't tagged properly when first acquired, it can be hard to find later
- There's a risk of inconsistent tagging
- And of similarly named categories getting mixed
- Less good at representing the structure of information

Tips for managing a hierarchical system

- In Windows, Windows Explorer is a good tool
- If possible, avoid overlapping categories
  - Find other ways of linking items
- Don't let your folders get too big – or your structure get too deep
- Create separate folders for older (no longer active) material

Creating a tag-based system

- Possible tools include:
- Bibliographic software
  - EndNote, Zotero, Mendeley...
- Image management programs
  - Flickr, Picasa...
- Google tools

Embedding metadata

- If feasible, try to enter basic information about the data file within its contents (e.g., author, date created/modified, project, grant, version)

    – May be able to <comment> information in a file

    – May help to identify files using your system's   full-text searching capabilities

- Embed metadata in header

- May also be able to assign this information as tags (external to your files

    – Caveat: some programs strip tags during file transfer transformation, so don't rely solely upon these

Adding searchable keywords to files in Windows

- Open up the Windows folder view and highlight (don't click to open) your file of interest
- In the pane at the bottom of the folder window, you'll see metadata about your file
- Click the property that you want to change/add (you'll see the box for tags all the way on the right), type the new property, and then click Save.
- To add >1 tag, separate each with a semicolon.
- Terms entered here will be found by the Windows search function

Adding tags on a Mac

- When you save a file, from the document menu, or in Finder

- • Spotlight Comments (and use Spotlight to search)

- http://support.apple.com/kb/HT5839

- http://www.maclife.com/article/howtos/mavericks_howto_organizing_files_and_folders_tags

- http://computers.tutsplus.com/tutorials/how-to-tag-files-and-create-spotlight-comments-on-a-mac--mac-46431

Hierarchical folder structure - Create a file structure system: why?

- Organization - important for future access and retrieval

- Simplifies your workflow in managing files

- Data files are easier to locate and browse

- Eases data sharing: clear organization is intuitive to team members and colleagues

- Data files are distinguishable from each other within and across folders

- Document your system and use iconsistently!

Good practices for organizing data files

- First: define the types of data and file formats for the research
- Be Clear, Concise, Consistent, Correct, and Conformant
- Choose a meaningful directory hierarchy/naming convention
- Includes important contextual information
- Could organize folders by primary, secondary, tertiary subject or collection method
- Document your system and use it consistently - choose a naming convention and ensure that the rules are followed systematically by always including the same information in the same order

Review the Case Study: Typical Project

- Image files (in multiple file formats)
- Data in tabular format
- Project documents (grant proposal, etc.)
- PDFs
- And more

Example file structure systems/directory hierarchy conventions:

/[Project]/[Grant Number]/[Event]/[Date]

/[Project]/[Sub-project]/[Run of an experiment]/[Person]/[Date]

/[Research area]/[Project]/[Data vs. documentation]/[Date]

/[Project]/[Type of file]/[Person]/[YYYYMMDD]

/[Instrument]/[Date]/[Sample]

FILE NAMING: TECHNIQUES

Create a file naming system: why?

- Organization - important for future access and retrieval
- Provides contextual information: a filename is a key identifier for a research data file (data files are not self-describing and you can't always embed metadata)
- Create logical structure for skimming through many files and versions; data files are distinguishable from each other within and across folders
- Eases data sharing: clear organization is intuitive to team members and colleagues

Good Practices for file naming

- Document your system and use it consistently!
- First: define the types of data and file formats for the research
- Be Clear, Concise, Consistent, Correct, and Conformant
- Context: provides content-specific or descriptive information
- Avoid using generic data file names that may conflict when moved from one location to another.
- Consistency - choose a naming convention and ensure that the rules are followed systematically
- Keep file names short but meaningful
- Reserve the 3-letter file extensions for the codes the system assigns to the file type, e.g. WRL, CSV, TIF (don't modify)
- Domains may have specific file naming recommendations – E.g., GIS datasets from the state of Massachusetts, http://www.mass.gov/mgis/dwn-name.htm
- Don't rely on file names as your sole source of documentation

Possible elements for file names

- Project/grant name and/or number
- Date of creation: useful for version control, e.g., YYYYMMDD
- Name of creator/investigator: last name first followed by (initials of) first name
- Name of research team/department associated with the data
- Description of content/subject descriptor
- Data collection method (instrument, site, etc.)
- Version number

Some specific considerations

- Capital letters or underscores (alternative: %20) can differentiate between words (avoid spaces)

- Avoid special characters such as: &-amp; * % $ £] { ! @ / as these are often used for specific tasks in a digital environment

- Number order files only if using leading zeros: e.g., 001, 002, 003, etc. will order files up to 999

- Consider how scalable your data file naming policy needs to be: e.g., don't limit your project number to two digits, or you can only have ninety nine projects.

- Capitals in file names affect ordering – be consistent.

- Note that not all systems/software are case-sensitive and recognize capitals; assume that TANGO, Tango and tango are the same

Example file naming convention systems:

[investigator]_[method]_[subject]_[YYYYMMDD]_[version].[ext], or

[project #] _[method]_[version]_[YYYYMMDD].[ext], or

[YYYYMMDD] _[version]_[subject]_[datacollector].[ext]

[type of file]_[specimen number]
 [version][collector]_[YYYYMMDD]_[geolocation].[ext]

[type of file]_[author]_[date].[ext]

Data collection equipment: file naming

- Check to see if your instrument, software, or other equipment that outputs your data files can be set with a file naming system

- Less work than retrospectively changing filenames

- But if you still have to change many file names downstream...

Batch renaming of files

- Useful for retrospectively aligning file/folder names with naming conventions
- Software tools can organize files and folders in a consistent and automated way through batch renaming (also known as mass file/bulk renaming)
- CAVEATS:
    - Take care that your bulk renaming software doesn't change the file format extension by mistake (common)
    - Given the importance of file names, ideally you'd want to keep track of the old file names along with the new ones

Batch renaming tools

Windows:
- Adobe Bridge (via any Creative Cloud products): http://ist.mit.edu/adobe-creative-cloud
- Ant Renamer: http://www.antp.be/software/renamer
- Bulk Rename Utility: http://www.bulkrenameutility.co.uk/
- ImageMagick: http://www.imagemagick.org/
- PSRenamer: http://www.powersurgepub.com/products/psrenamer.html
- RenameIT: http://sourceforge.net/prpjects/renameit

Mac:
- Adobe Bridge (via any Creative Cloud products): http://ist.mit.edu/adobe-creative-cloud
- ImageMagick: http://www.imagemagick.org/
- Name Changer: http://web.mac.com/mickeyroberson/MRR_Software/NameChanger.html
- PSRenamer: http://www.powersurgepub.com/products/psrenamer.html
- Renamer4Mac : http://renamer4mac.com/
- Name Mangler: http://manytricks.com/namemangler/

Linux:
- GNOME Commander: http://www.nongnu.org/gcmd/
- GPRename: http://gprename.sourceforge.net/
- ImageMagick: http://www.imagemagick.org/
- PSRenamer: http://www.powersurgepub.com/products/psrenamer.html

Unix
- The use of the grep command to search for regular expression

Version Control

Versioning: program vs. data files

- Ideal: keep the original version of the data file the same and save iterative versions of the analysis/program/scripts files

- If you need to modify data files: save a copy of every iteration of a data file

Version control: principles

- Document your convention and be consistent

- Record every change

- Consider: discard or delete obsolete versions (while retaining the original 'raw' copy) if appropriate

- Consider your version control needs regarding:

  – single site vs. across locations

  – single vs. multiple users

  – different versions to be stored vs. files to be synchronized

Version control: tips and resources: 1

- In the file/folder names, use ordinal numbers (1,2,3, etc.) for major version changes and the decimal for minor changes e.g v1, v1.1, v2.6

- Beware of using imprecise labels: revision, final, final2, definitive_copy as you may find that those aren't as definitive as you thought

- May put old versions in separate folder

- May create a version control table or file history w/in or alongside data file

Version control: tips and resources: 2

- Record relationships between files, e.g. data file and documentation; similar data files
- Keep track of file locations, e.g., laptop vs. PC
- Some software has built in version control facilities, e.g.:
    – control rights to file editing: read/write permissions (Windows Explorer)
    – versioning or tracking features in collaborative documents (Wikis, GoogleDocs)
    – versioning/file sharing software: check files out/in
- Consider using version control software e.g., GIT, GNU RCS, Mercurial (Hg) or Apache Subversion, TortoiseSVN

Summary

- Don't count on remembering things about your data

- Investing time at the beginning in an efficient system can save time in the long run

- Plan ahead and establish a system

- Make a system that works for you (and your collaborators)

QUESTIONS?