Data Management Best Practices

"The goal of data management is to produce self-describing data sets." DataONE Primer on Data Management. (Strasser)

What not to do.

Data Sharing and Management Snafu in 3 Short Acts: A data management horror story by Karen Hanson, Alisa Surkis and Karen Yacobucci.

http://www.youtube.com/watch?v=N2zK3sAtr-4

Don't be the brown bear.

Why manage research data?

- *You can find and understand your data when you need to use it*
- *There is continuity if project staff leave or new researchers join*
- *You can avoid unnecessary duplication e.g. re-collecting or re-working data*
- *Data underlying publications are maintained, allowing for validation of results*
- *Data sharing leads to more collaboration and advances research*
- *Research is more visible and has greater impact.*
- *Other researchers can cite your data so you gain credit (Jones)*

NEH-ODH Proposals and Awards

Application and Submission Information

7. Data Management Plan

"Prepare a data management plan for the project. The plan should describe how the project team will manage and disseminate data generated or collected by the project… Include as an attachment (not to exceed two pages) a description of the project data management…"

http://www.neh.gov/files/grants/digital-humanities-start-sept-11-2014.pdf

Data Management Best Practices

- Data Management Plan
- File Management
- Documentation
- Storage and backup
- Long term planning

Data Management Plans (DMPs)

- DMPs for grant applications are a “light touch”
- Should be considered living documents
- Can act as standard operating procedure
- Can help ensure documentation is complete
- Can save time while writing up results for publication

Data Management Plans (DMPs)

- What types of data will be created?
- Who will own, have access to, and be responsible for managing these data?
- Are there legal or ethical restrictions on the data?
- What equipment and methods will be used to capture and process data?
- Where will data be stored during and after?

Slide Credit: Modified from Module 1: Overview of Research Data Management New England Collaborative Data Management Curriculum

NEH Data Management Plans

1. Expected Data (including):

- types of data generated and shared, and under what conditions
- how data are to be managed during the project including roles and responsibilities
- limiting factors including legal and ethical restrictions
- lowest level of aggregated data that will be shared
- mechanism for sharing or making accessible to others
- other types of information that should be maintained including documentation and metadata

2. Period of data retention

3. Data formats and dissemination

4. Data storage and preservation of access

http://www.neh.gov/files/grants_data_management_plans_2014.pdf

A Successful Data Management Plan…

Data Management Plan

A unified approach to preserving cultural software objects and their development histories

1. Roles and responsibilities

This data management plan will be implemented and managed by Eric Kaltman, under the project supervision of Noah Wardrip-Fruin. Christy Caldwell will assist with transferring data to the University of California Curation Center (UC3). UC3 will have long-term responsibility for the permanent storage needs of the data. All transferred data will be made publically accessible.

2. Expected data

We are developing an approach to preserving software objects. Therefore, our data is at two levels: the objects we are preserving, and the documentation of the preservation process.

The data from preservation objects will include:

- interview transcripts from Prom Week team members
- text files of correspondence, notes, academic papers and planning documentation from the development history of Prom Week

http://www.neh.gov/files/grants/university_of_california_santa_cruzpreserving_cultural_software_objects_and_their_development.pdf

Data management plan help

- DMPTool –Specific guidance for mostly U.S. funders, customized for Princeton. http://dmptool.org
- DMPOnline –From the U.K. https://dmponline.dcc.ac.uk/
- MIT Data Planning Checklist - http://libraries.mit.edu/guides/subjects/data-management/checklist.html
- DCC Checklist for a Data Management Plan - From the U.K. http://www.dcc.ac.uk/resources/data-management-plans/checklist
- Jones, S. (2011). 'How to Develop a Data Management and Sharing Plan'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides
- DMP review and consultation services available from Princeton Library's RDMTeam, rdmteam@princeton.edu, http://library.princeton.edu/research-data-management

File and folder management

"**Figure 4:** A snapshot of data management practices. File names given by students are shown for a sampling of .1sc files, illustrating the variety of naming conventions used." (Ferguson)

Restriction Digest 5.26.11 DNA 1 & 2	1sc
awesome 2010	1sc
awesomer 2010	1sc
Deb 2010-03-09 yeast gel	1sc
7.22.10 Gel 1=60bp wt706 Tmp Tm & cycle	1sc
7.22.10 Gel 2=60bp Temp-wt001& Neal	1sc
7.22.10 Gel 2=60bp wt001 & Neal Temp	1sc
7.22.10 Gel 3=100bp wt706 Tmp, Tm &cycle	1sc
7.22.10 Gel 4=100bp wt001 & Neal Temp	1sc
ura gel 1 Mon R starred	1sc
ura gel 2 Monday Ravenclaw un-starred	1sc
Dpn Gel 5 10 WT	1sc
a	1sc
attractive	1sc
group	1sc
joe	1sc

Ferguson, Jen. "Lurking in the Lab: Analysis of Data from Molecular Biology Laboratory Instruments." *Journal of eScience Librarianship* 1, no. 3 (March 13, 2013).

File Naming Best Practices

- Files should be named consistently
- File names should be descriptive but short (<25 characters) (Briney)
- Avoid special characters in a file name.
- Use capitals or underscores instead of periods or spaces.
- Use date format ISO 8601:YYYYMMDD
- Include a version number (Creamer et al.)
- Write down naming convention in data management plan

File Naming Conventions

- How?
  - Pick what is most important for your name
    - Date
    - Work
    - Analysis
    - Sample
    - Short description
    - File Modification
    - Version

Slide Credit: Briney

Example

Date_Subject_Modification
- 20130420_tina_original
- 20130420_tina_cropped
- 20130420_tina_mustache

Work_Analysis_Version
- Ulysses_wordCloud_v1
- Ulysses_map_v1
- Ulysses_map_v2

File Organization

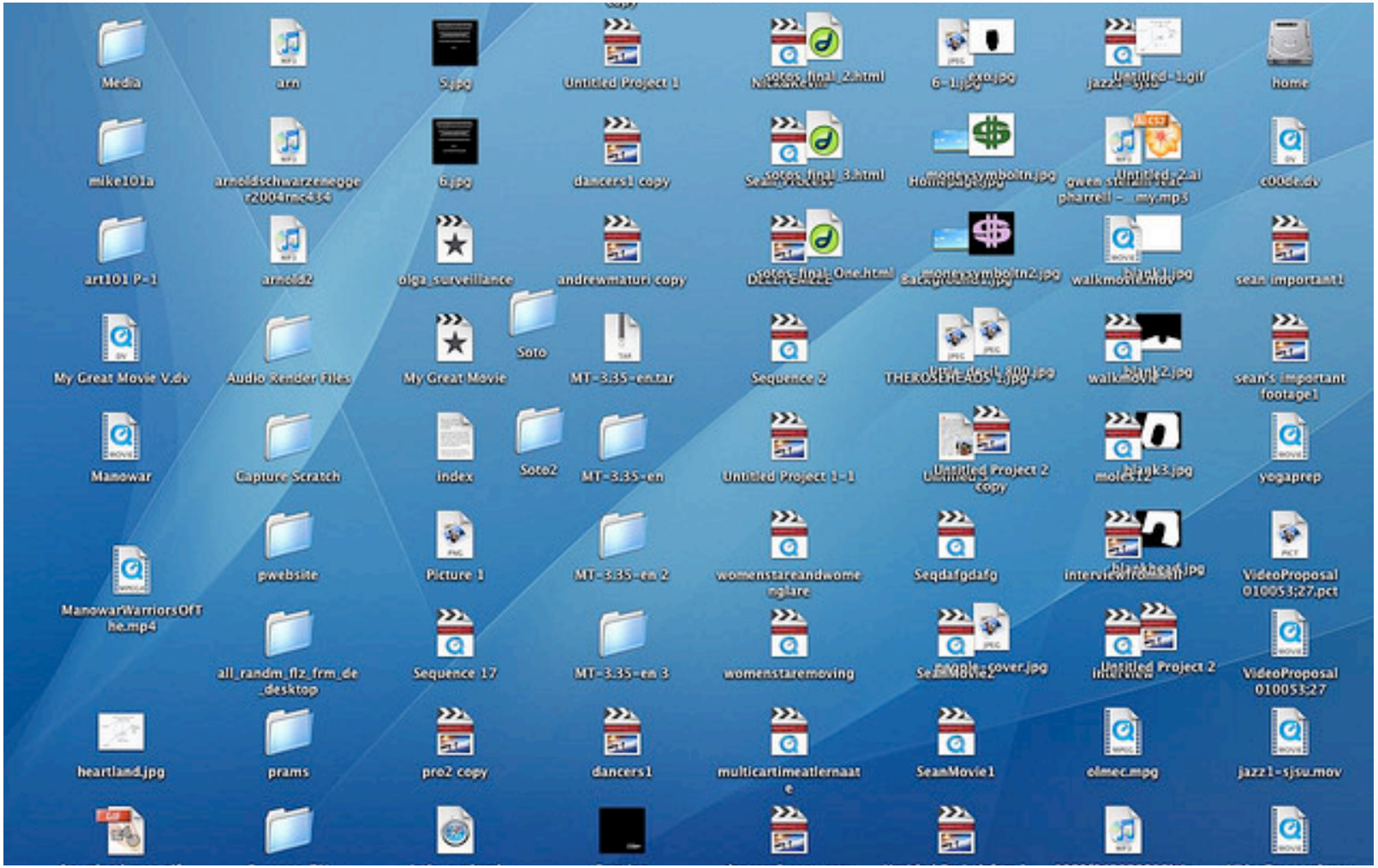


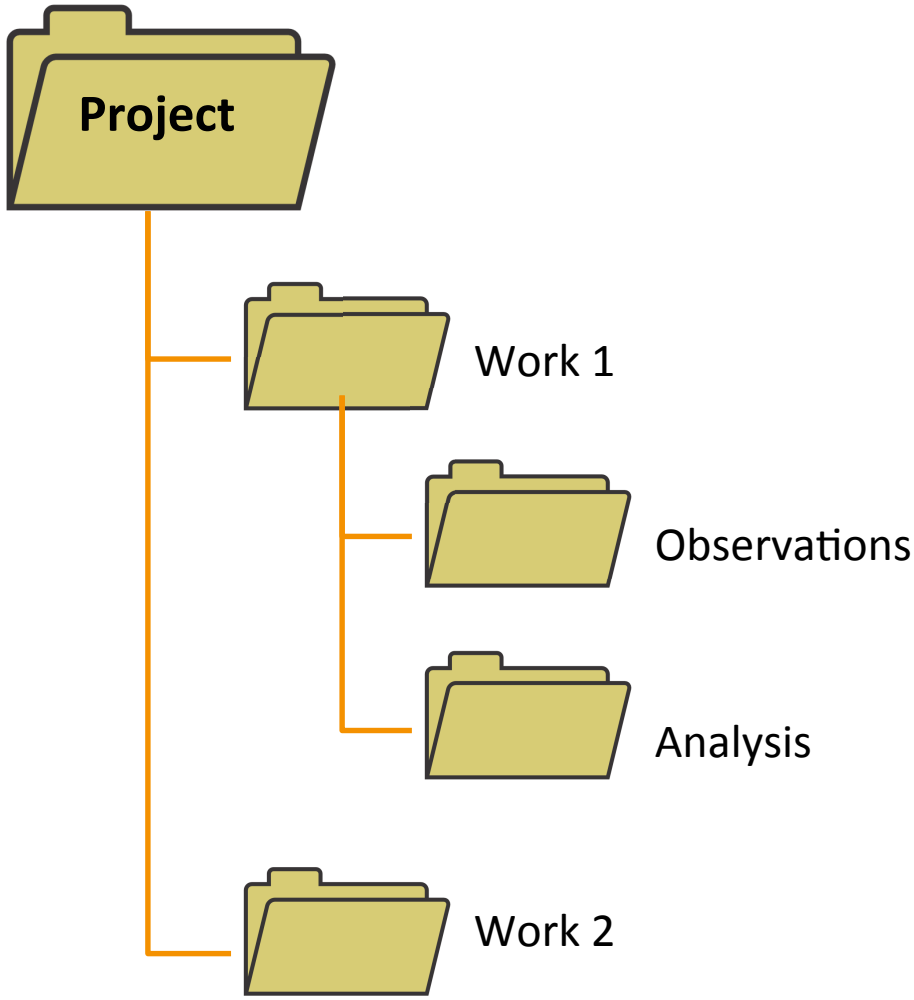
Photo by Michael Lowell on flickr, CC-BY-SA

File Organization

- How?
  - Any system is better than none
  - Make your system logical for your data
  - Possibilities
    - By project
    - By analysis type
    - By date
    - …

Slide Credit: Briney

Example



Documentation

The Who, What, When, Where and Why of Your Data

Why?

- Data without notes are unusable
- Because you won't remember everything
- For others who may need to use your files (Briney)

Study-Level Documentation
UK Data Archive

- *the context of the data collection: project history, aims, objectives and hypotheses*
- *data collection methods: data collection protocols, sampling design, instruments used, hardware and software used, data scale and resolution, temporal coverage and geographic coverage, and digitisation or transcription methods*
- *structure of data files, number of cases, records, variables and relationships between files*
- *data sources used and provenance of materials, e.g. for transcribed or derived data*
- *data validation, checking proofing, cleaning…*
- *modifications made to data over time…*
- *Information on data confidentiality, access and use conditions, where applicable (UK Data Archive)*

http://www.data-archive.ac.uk/create-manage/document/study-level

Data-Level Documentation

UK Data Archive

- *names, labels and descriptions for variables, records, and their values*
- *explanation of codes and classification schemes used*
- *codes of, and reasons for, missing values*
- *derived data created after collection, with code, algorithm or command file used to create them*
- *data list describing cases, individuals or items studied, for example for logging qualitative interviews (UK Data Archive)*

http://www.data-archive.ac.uk/create-manage/document/data-level

Documentation

- How?
  - Take good notes
  - Metadata schemas
    - http://www.dcc.ac.uk/resources/metadata-standards
  - Templates
    - Like structured metadata but easier
    - Decide on a list of information before you collect data
      - Make sure you record all necessary details
      - Takes a few minutes upfront, easy to use later
    - Put in data management plan
    - Print and post in prominent place or use as worksheet

Slide Credit: Briney

Sample documentation

- Title
- Creator
- Identifier
- Subject
- Funders
- Rights
- Access information
- Language
- Dates
- Location
- Methodology
- Data processing
- Sources
- List of file names
- File Formats
- File structure
- Variable list
- Code lists
- Versions
- Data Citation Format

Slide Credit: Module 1: Overview of Research Data Management
New England Collaborative Data Management Curriculum

Documentation

- Where?
  - README.txt
    - For digital information, address the questions
      - “What the heck am I looking at?”
      - “Where do I find X?”
    - Use for project description in main folder
    - Use to document conventions
    - Use where ever you need extra clarity

Slide Credit: Briney

Special considerations for DH (Flanders)

- "The importance of **interpretive layering."** Digital humanities data is not just the original object, but often takes its form as the interpretive framework itself.
- "The importance of information about **how the data is captured and prepared**… Crucial decisions affecting the usability and meaning of humanities data are made at each stage of data creation and management, and the documentation concerning these decisions is likely to be valuable to both future users and curators of the data."
- "The importance of capturing **responsibility, editorial voice, and debate."**

Storage and Backup

- 3-2-1 Backup Rule:
  - Keep **3 copies** of anything important, the original plus 2 backups
  - On at least **2 different media types**
  - With at least **1 off-site or in cloud storage**
- Where?
  - Personal computer hard drives. Backup available for faculty and staff.
  - External hard drives (Available at OIT Tech Depot)
  - Central File Server (H: Drive) – 5 GB,
    Departmental Storage (M: Drive) varies
  - Cloud Storage: All undergrads have a 30 GB Google Drive account. Faculty, Staff, and Graduate students can request.

Backups… what and when

- “What will you need to restore in the event of data loss?” In general only backing up data is sufficient. (UK Data Archive)
- OIT Knowledge Base: What files should I back up? http://helpdesk.princeton.edu/kb/display.plx?ID=9690
- “Backups should be made after every change of data [and]/or at regular intervals.” (UK Data Archive)
- UK Data Archive help pages on storing, backing up, data security, transmitting and encrypting data, file sharing, and data disposal. http://www.data-archive.ac.uk/create-manage/storage

Long term planning - Preservation

- At the completion of a project
- Not the same as storage during a project
- "What is Preservation and Why Does it Matter?", *Companion to Digital Humanities* http://www.digitalhumanities.org/companion/
- What are the funder or journal requirements?
  - How long does it need to be preserved?
  - Who is responsible for the data at the end of the project?
  - Does funder or journal specify a repository?

Long term planning – Repositories

- Increases discoverability
- Provide persistent unique identifiers and information to aid data citation
- Different options available
  - Many disciplinary repositories available http://www.re3data.org/
  - General repositories: Dataverse, Figshare, Zenodo
  - DataSpace at Princeton: http://dataspace.princeton.edu

Future File Usability

- Why?
  - You may want to use the data in 5 years
  - Prep for data sharing
  - May be needed to verify journal article results
  - Per U.S. Office of Management and Budget Circular A-110, must retain data at least 3 years post-project
    - Better to retain for >6 years

Slide Credit: Briney

File Format Best Practices

Is the file format open (i.e. open source) or closed (i.e. proprietary)?

Is a particular software package required to read and work with the data file? If so, the software package, version, and operating system platform should be cited in the metadata…

Do multiple files comprise the data file structure? If so, that should be specified in the metadata…

When choosing a file format, select a consistent format that can be read well into the future and is independent of changes in applications.

Non-proprietary: Open, documented standard, Unencrypted, Uncompressed, ASCII formatted files will be readable into the future.

Slide Credit: Module 1: Overview of Research Data Management
New England Collaborative Data Management Curriculum

Future File Usability

- How?
  - Convert file formats
    - Can you open digital files from 10 years ago?
    - Use open, non-proprietary formats that are in wide use
      - .docx → .txt
      - .xlsx → .csv
      - .jpg → .tif
      - .pdf → .pdf/a
      - See National Archives FAQ for more http://www.archives.gov/records-mgmt/initiatives/sustainable-faq.html
    - Save a copy in the old format, just in case
    - Preserve software if no open file format

Slide Credit: Briney

Future File Usability

- How?
  - Move to new media
    - Hardware dies and becomes obsolete
      - Floppy disks!
    - Expect average lifetime to be 3-5 years
    - Keep up with technology

Slide Credit: Briney

Campus resources

Center for Digital Humanities
https://digitalhumanities.princeton.edu/

Citation Management Tools Help
https://library.princeton.edu/help/citation-tools

Digital Maps and Geospatial Information Center
http://library.princeton.edu/collections/pumagic

Data and Statistical Services
http://dss.princeton.edu/about.html

Institutional Review Board
http://www.princeton.edu/ria/human-research-protection/committee-information/

Other Resources

Create & Manage Data, UK Data Archive.
http://www.data-archive.ac.uk/create-manage

Companion to Digital Humanities. Hardcover. Blackwell Companions to Literature and Culture. http://www.digitalhumanities.org/companion/

Data Management General Guidance, DMPTool.
https://dmptool.org/dm_guidance

Digital Humanities Data Curation, http://www.dhcuration.org/

DiRT: Digital Research Tools, http://dirtdirectory.org/

Guidelines for Effective Data Management Plans, Inter-University Consortium for Political and Social Research (ICPSR)
http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/index.html

How-to Guides, Digital Curation Centre
http://www.dcc.ac.uk/resources/how-guides

Attribution

This work is a derivative of:

Practical Data Management, ACRL DCIG Webinar. April 30, 2014.
Kristen Briney http://www.slideshare.net/kbriney
CC-BY (http://creativecommons.org/licenses/by/4.0/)

And

New England Collaborative Data Management Curriculum, Module 1: Overview of Research Data Management. Andrew Creamer et al. http://library.umassmed.edu/necdmc/modules
CC-BY-NC (http://creativecommons.org/licenses/by-nc/4.0/)

Slides used from each presentation are noted at the bottom of the slide.

Works Cited

Flanders, Julia, and Trevor Muñoz. "An Introduction to Humanities Data Curation." *DH Curation Guide*, September 22, 2011. http://guide.dhcuration.org/intro/

Jones, S. (2011). 'How to Develop a Data Management and Sharing Plan'. *DCC How-to Guides.* Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

"Documenting your data." *UK Data Archive*. Accessed December 8, 2014. http://www.data-archive.ac.uk/create-manage/document

"Back-Ups & Security, Data Management." *MIT Libraries*. Accessed October 15, 2014. http://libraries.mit.edu/data-management/store/backups/.

Strasser, Carly et al. "Primer on Data Management: What You Always Wanted to Know." http://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

About me

Willow Dressel
Plasma Physics and E-Science Librarian
wdressel@princeton.edu

Research data management services
http://library.princeton.edu/research-data-management

Data Management Plan and Documentation Exercise

Historical Architecture of Minnesota

http://mhs.iath.virginia.edu/