Research Data Management 101

Michelle Hudson

CSSSI StatLab workshop Oct 14 2016

https://github.com/michellehudson/datamanagement

Description:

This two-hour workshop will provide an overview of the data lifecycle and the critical steps within it that need to be addressed to ensure integrity of research data. It is appropriate for students and faculty in all disciplines, however, the constrained time-frame and high level overview of the issues only warrant a few in-depth examples of tools and resources for specific disciplines. The workshop will focus on general good practices for data management that span disciplines. There will be Q&A time for specific questions, and attendees are always welcome to follow up with instructors or other specialist for more tailored data management instruction or assistance.

Data lifecycle model

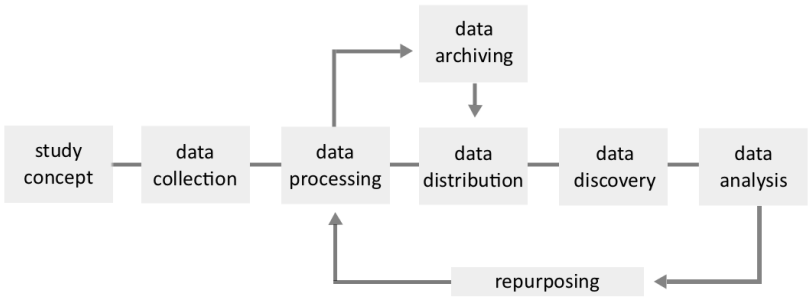

Figure 1: DDI lifecycle model

Outline:

1. Research Data Consultation Group
2. What is data?
3. Why manage it?
4. Data management checklist
5. Data Management Planning
6. Metadata and data description
7. Organizing and working with data files
8. Re-using data and making your data reusable
9. Storage, computing, and analysis
10. Data preservation and archiving data
11. Sharing and publishing data
12. Resources list

Research Data Consultation Group

http://researchdata.yale.edu

RDCG has 13 consultants from across Yale to help consult on finding and using data, analyzing data, storage options, preserving and archiving data, and more. Use the contact form at http://researchdata.yale.edu to get assistance with research data issues at any point in your project.

What is research data?

Research data is defined as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings." - OMB Circular A-110. More loosely, it's defined as *information collected, observed, or created for purposes of analysis to produce original research.*

There are four general types of research data:

1. Observational: captured in real time, usually irreplaceable (sensor readings, telescope images, sample data, surveys).
2. Experimental: data from lab equipment, can be reproducible but may be expensive (gene sequences).
3. Simulation: data generated from test models (climate models).
4. Derived or compiled: reproducible but expensive (data mining, compiled databases).

Research data comes in many formats of information: documents, spreadsheets, field notebooks, survey responses, audio and video recordings, images, film,

specimens, software code, and can be structured and stored in a variety of file formats.

Why manage research data?

There are many reasons why good data management is important for your research career, ranging from long-term effects on the future of science to personal productivity and accomplishment.

Transparency, integrity, and reproducibility:

Managing data and making it accessible by peers decreases the chances of an article being retracted because of falsified or missing data sets. Reproducibility is a fundamental part of scientific research, and failing to make all the components of a research study available makes reproducibility impossible.

Compliance:

Data management plans are required by funding agencies, and there is increased expectation that the products of federal funding will be required to be accessible to the public. See more information about this on the whitehouse.gov post on the OSTP Memo on Expanding Public Access to the Results of Federally Funded Research. In addition, many journals are requiring data deposit before an article may be published.

Personal & professional benefits:

If data is managed within your lab, research group, or simply well-organized for your own use, you will save time, energy, and resources. All members of the team will have an understanding of the well-documented processing and analysis of the project's data, and be able to carry out their research components more effectively. Sharing research data is now regarded as an integral and valuable part of the research process, and archiving your data in a repository will allow other researchers to build upon your work and cite you in the process.

Data management checklist

1. always keep original data
2. back up regularly

3. document your data thoroughly
4. name and organize files according to a schema
5. use version control
6. secure the data appropriately
7. cite any secondary data you use
8. consider your long-term plan: What will you keep, for how long, where, and who will pay for it? What kinds of reuse or sharing will be allowed? In what timeframe?

Data Management Planning

What is a DMP?

Some funding agencies require a short document called a Data Management Plan that supplements a grant proposal and explains how data will be managed throughout a project, who will be responsible for managing it, and how it will be shared when the project ends.

Which agencies require a DMP?

All NSF directorates, the NIH, and several other smaller funding agencies require a formal data management plan to be submitted with proposals. In addition, every federal agency that receives more than 100 million in R&D expenditures has been instructed to develop plans to require researchers to "better account for and manage the digital data resulting from federally funded scientific research." - OSTP. Many of the agencies have responded by requiring DMPs. You can find a list of all the agencies and their guidelines at SPARC.

What tools and resources are available to help me write a DMP?

DMPTool: https://dmptool.org:

Yale is a DMPTool partner. Logging in with your Yale ID and password will give you access to the DMPTool, which will give you an overview of funder requirements (for various NSF, NIH, and other directorates and divisions), and walk you through building a data management plan, asking the right questions along the way.

Research Data Consultation Group: http://researchdata.yale.edu/contact:

If you have to submit a DMP as part of a grant proposal and have trouble using the DMPTool or answering questions you think are critical to the good management of data, you can contact the Research Data Consultation Group for help. This group can review written plans and offer feedback, or connect you with more resources at Yale you might be able to cite or consider including in your plan to make a stronger proposal.

StatLab consultants: http://csssi.yale.edu/data/csssi-statistical-consulting:

Even if you aren't submitting a grant proposal, it's a good idea to come to the StatLab at the beginning of your project. If you know what analyses you want to do on your data, the StatLab can make sure you set out to collect your data correctly. If you anticipate using StatLab services near the end of your project, it's much easier for them if you connect in the beginning of the project, as well.

Things to consider when writing a DMP

- What data are generated by your research?
  - Describe intended file formats, instruments and software used, collection notes, collection materials (surveys or tests), metadata standards applied
- What is your plan for managing the data?
  - How will the data be stored while the project is in progress?
  - What security measures are necessary, if any?
  - Who will be responsible for managing the data generated for the project, during and after?
  - How will data be disseminated or shared after the project, and in what timeframe, under what conditions?
  - Are there any limitations that you need to address re: personally identifiable or health information?
  - How will necessary data be archived and preserved?

Metadata and data description

Metadata

Metadata is the "data about data" that is needed to make numeric data usable. Without proper metadata and documentation of the research methods, analysis, variables, units, codes, and locations relevant to the numeric information, digital data is unusable.

Types of metadata

Descriptive

Title, author, abstract, keywords, geographic coordinates, species...

Structural

Schematic relationships between files, relational information...

Administrative

Formats of data files, intellectual property information, preservation metadata...

Levels of metadata

Study-level description

1. Context of the data collection (project history, aim, objectives, and hypotheses)
2. Data collection methods (sampling, data collection process, instruments used, hardware and software used to collect data, scale and resolution, temporal and geographic coverage, secondary data sources used, if any)
3. Data set structure – of files, study cases, and relationships between files
4. Changes made to data over time
5. Information on access and use conditions or data confidentiality

File-level description

1. Names, labels, and descriptions for variables, records, and their values
2. Definition of codes & classification schemes used
3. Codes of and reasons for missing values

Variable-level description

1. What does each variable mean in the context of the research?

Examples of metadata standards

Data Documentation Initiative: http://www.ddialliance.org

A freely available, international standard for describing statistical and social science data.

Ecological Markup Language: https://knb.ecoinformatics.org/#external//emlparser/docs/index.html

The EML project is an open source, community oriented project dedicated to providing a high-quality metadata specification for describing data relevant to the ecological discipline.

Darwin Core: http://rs.tdwg.org/dwc/

The Darwin Core is body of standards. It includes a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries.

ISO Geospatial Metadata Standards: https://www.fgdc.gov/metadata

Guides from the Federal Geographic Data Committee on applying appropriate metadata to geospatial information.

More: list of metadata standards from the Research Data Alliance

What tools and resources are available?

Morpho: https://knb.ecoinformatics.org/morphoportal.jsp

Morpho was developed for data management in ecology.

Colectica: http://www.colectica.com

Colectica is software that helps design, document, and publish statistical data and survey research using open data standards.

ISO geospatial metadata editors: https://www.fgdc.gov/iso-metadata-editors-registry/editors

A comparison of tools available for editing geographic metadata.

More: list of metadata tools from the Research Data Alliance

Reminder!

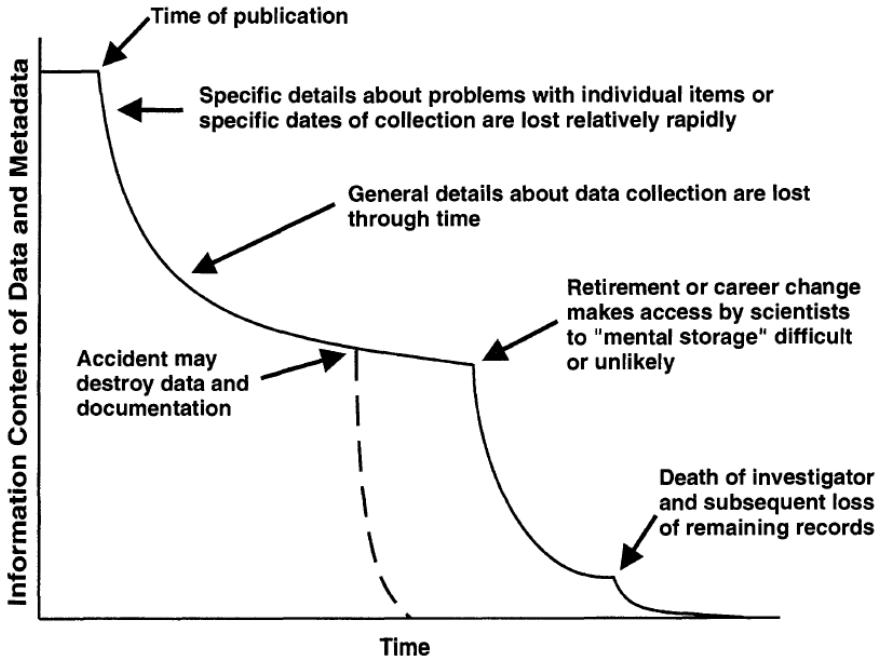

Figure 2: Bill Michener's description of data completeness over time

Organizing and working with data files

Main takeaways

1. Keep a codebook for all data
2. Keep a log of all transforms and analyses (syntax)
3. Save data often and back up files
4. Use a versioning system
5. Organize files

Codebooks

Codebooks are used in social science research to serve as a companion to the numeric data – a human-readable manual containing all the metadata needed to understand and use data related to a project. You should have a codebook for your project that explains each variable and its values, any variables you've added or computed, etc.

Example: General Social Survey codebook, 1972 - 2014

In addition to creating a codebook, you can also create a readme.txt file that lives in the home directory for your project and explains the latest notes, updates, and reminders for your project's data files.

Working with syntax files

Syntax files are separate text files used to enter commands that are then performed on the data. Keeping a log of your analyses this way makes your research more reliable (you can share your syntax code along with your data sets for replication purposes). With syntax files, you can add comments to commands so you remember why you performed which actions.

The ISPS Data Archive keeps data and syntax files for all studies. Example: Butler, Daniel M. et al. (2015) Replication Materials for 'Ideology, Learning and Policy Diffusion: Experimental Evidence.' http://hdl.handle.net/10079/1zcrjs8." ISPS Data Archive.

Backup and versioning

1. Implement a system for backing up all your project-related files (Box, USB drives, network shares).
2. If you can't automate a system, back up manually according to a schedule and stick to it.
3. Explore options for using version tracking for data files, especially if more than one person is working on the same project.

Organize files and folders

Before you begin a project, decide (in cooperation with your lab, PI, or others if necessary) on a folder structure and naming convention. There are few best practices around this during the active stage of working with data, and researchers do it differently according to the needs of their lab and their data. The best advice is to decide on something and stick with it.

Folder structure

- Use a hierarchical structure.
- Keep original data and working files separate.
- Keep a readme.txt.
- Keep any geospatial data together in its folders.

Folder structure example

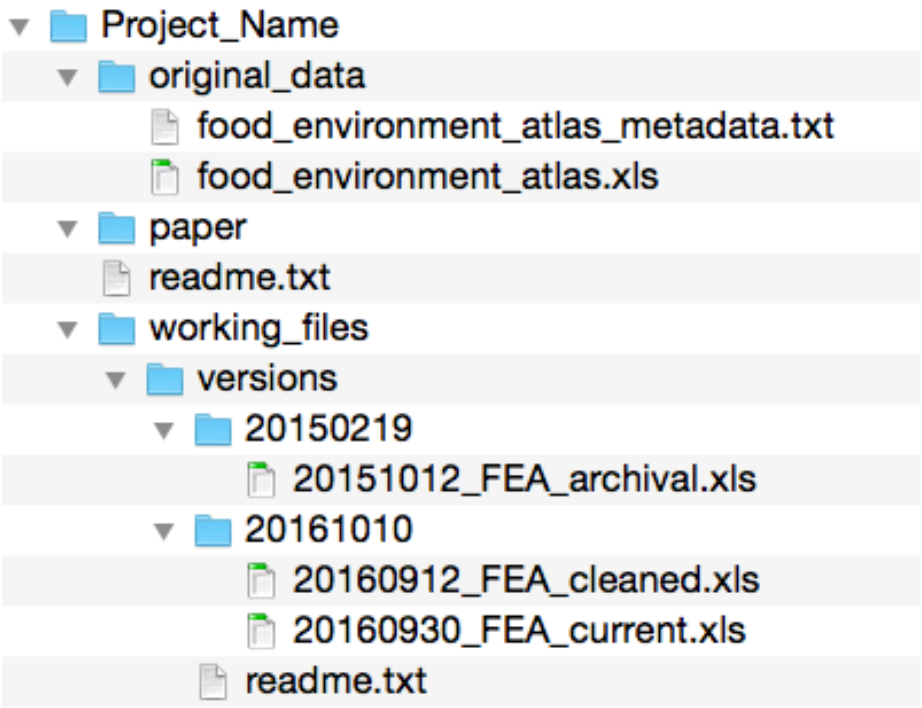

Figure 3: Example folder structure

https://github.com/michellehudson/datamanagement/tree/master/research_data_management_101/2016_fall/ex_folder_structure

File naming

- Use descriptive filenames, but not too long.
- Do not use special characters.
- Include date information at the beginning or end of the file and be consistent.
- Use underscores, not spaces.

- Considering including: project name, researcher initials, version number.

File naming example

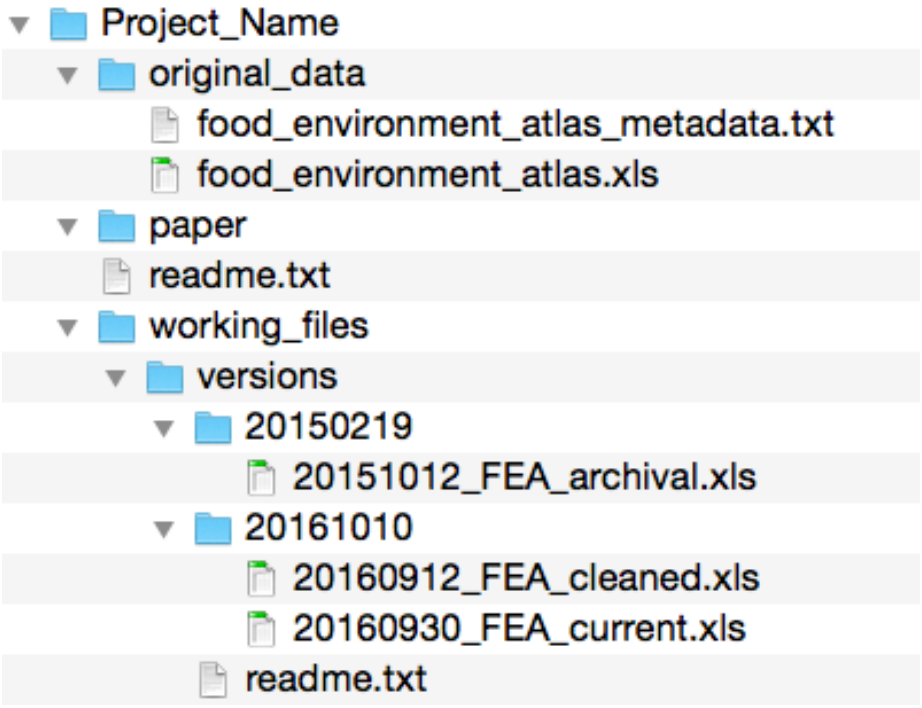

Figure 4: Example folder structure

https://github.com/michellehudson/datamanagement/tree/master/research_data_management_101/2016_fall/ex_folder_structure

Re-using data and making your data re-usable

If you're re-using data from another source (downloaded from another institute, a data archive, from another researcher, etc.), you want to get to know it as completely as possible. If you're providing your data to an archive or to another researcher so they can re-use it, think about what information you'd need if you were using someone else's data, and make sure that information is included for your own.

1. How was the data collected? What instruments (survey or scientific) were used? If it was a survey, what was the wording of the questions? Who coded the questions?

2. How is the data coded? What are the codes for missing values, and what do they mean?
3. Is the data pre-processed or cleaned? Is it weighted? Are any values interpolated?

Finding data for re-use

Librarians have created guides to assist with finding data in different disciplines.

- Social science data & statistics
- Science data

Or, feel free to schedule a meeting with the data librarian to discuss finding data for a project.

Storage, computing, and analysis

Data storage

- Box
- Storage@Yale

Computing

- Yale Center for Research Computing for HPC support and training
- StatLab windows server for smaller jobs (talk to Themba Flowers for access)

Data analysis

- StatLab consultants
- StatLab workshops

Preservation and archiving

Archiving and preserving research data is different from distributing it or backing it up regularly. Preservation ensures long-term retention of the data and the necessary migration from format to format that will be required to keep the data usable over a time period. How long you retain your data is often up to what

your funding dictates – some grants say three years, others five. In some cases, your data may have value for an indefinite period of time.

Available repositories:

The Registry of Research Data Repositories http://www.re3data.org aims to list all the data repositories available for submission or for finding research data to reuse, and you can search or browse by subject.

Guidelines:

1. Doing preservation yourself requires format migration and ensuring integrity of files.
2. Handing over your data to a repository like ICPSR is possible, and will ensure the data is usable over the long-term.

Examples:

Institution for Social & Policy Studies: http://isps.yale.edu/research/data

ISPS is a Yale department that maintains a data archive of research that has been conducted by their affiliates.

ICPSR: http://icpsr.umich.edu

The Inter-university Consortium for Political and Social research is a domain archive that has been curating and maintaining access to data sets for over 50 years.

Data sharing and publishing

Platforms

There are many platforms for data distribution that are easy, free, and meet many researcher needs. These solutions do not necessarily guarantee preservation-level archiving for research data, but they make data available and citable.

openICPSR: https://www.openicpsr.org

openICPSR is a branch of the ICPSR and is free for Yale researchers depositing social science and behavioral health related datasets.

Open Science Framework: https://osf.io

The Open Science Framework is funded by federal agencies, private foundations, and commercial entities, and offers a free platform for data management and publication.

Dataverse: https://dataverse.harvard.edu

Dataverse is repository software that institutions can set up and host, but it's also a network of these repository nodes. The Harvard instance of Dataverse is open to all researchers for data submission and publication through a personal account.

Data citation

It's important for your data to be citable, and it's important to cite any data you use in your analyses thoroughly. Look for a data sharing platform that will give you a permanent identifier (like a DOI or a handle) for your project.

DataCite https://www.datacite.org/index.html is an international organization that provides permanent identifiers for data, and they provide a helpful citation formatter for data.

Resources

Data Management Research Guide: http://guides.library.yale.edu/datamanagement

CSSSI's data management guide.

MANTRA: http://datalib.edina.ac.uk/mantra

Mantra is series of useful research data management training modules you can complete online.

Contact info

Michelle Hudson

- Science and Social Science Data Librarian
- michelle.hudson@yale.edu

Joshua Dull

- Research Data Support Specialist
- joshua.dull@yale.edu

StatLab Consultants

- Schedule: http://csssi.yale.edu/csssi-statistical-consultants-schedule
- 203.432.3277
- Contact the StatLab