Research Data Management 101

Michelle Hudson & Kristin Bogdan

CSSSI workshop Feb 21 2014

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.

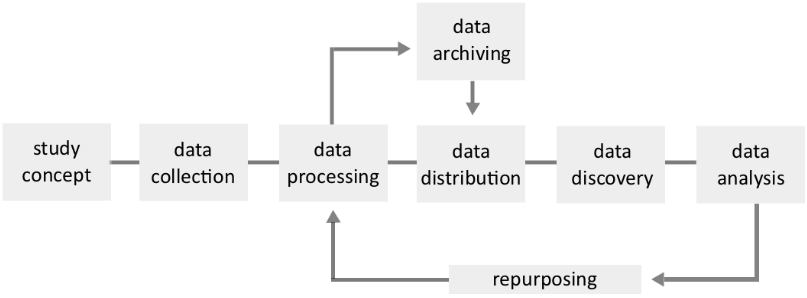

Figure 1: DDI lifecycle model

General format:

Using the DDI data lifecycle model as a guide, we'll cover the following questions:

1. What does this stage of the data lifecycle involve?
2. What resources are available for doing it well at Yale (& elsewhere)?
3. What are guidelines for managing data at this stage?

Outline:

1. What is data?
2. Why manage it?
3. Study concept
4. Data collection
5. Data processing
6. Data archiving
7. Data distribution
8. More resources
9. Q&A

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.

There are four 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. 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.

Study concept

What does this stage involve?

This is the planning process for a study. It involves formulating a research question and deciding on the methods you want to use to execute your study. It may include submitting a grant to get funding. Some grants require data management plans to be submitted as part of the proposal.

What tools and resources are available?

DMPTool: 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. In the next iteration of the tool, we'll be able to further customize it with Yale-specific resources.

DMP Consultation Group: 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 DMP 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: 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.

Data collection & documentation

What does this stage involve?

This stage involves all the collection and subsequent documentation of your data, and may involve collaboration with other people. There aren't a lot of collaborative online spaces for data collection, but we can discuss a few, and some general guidelines for documenting your research well.

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

What tools and resources are available?

Yale-supported:

Box: Box is a document-sharing cloud service available to everyone at Yale and is supported by Yale ITS. See the link for questions about security and size limits.

LabArchives: LabArchives is an electronic lab notebook solution available to everyone at Yale and is supported by Yale ITS.

EliApps: EliApps may be an appropriate place to collaborate for simple spreadsheets and forms for non-sensitive data.

Qualtrics: Qualtrics is robust survey building software, is available to everyone at Yale, and is supported by Yale ITS.

Additional services & software:

GitHub: GitHub is a free or paid service, popular for writing and sharing software code, and can be used to track changes to files and work with multiple collaborators. GitHub is not supported by Yale ITS.

Morpho: Morpho was developed for data management in ecology.

Earthcube: Earthcube is a community driven data and knowledge management system that will allow for data sharing across the geosciences.

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

Guidelines:

1. Spreadsheets vs. databases: see the upcoming workshop on database design: 4/18/2014, 1:30 - 3:30 CSSSI.
2. Consistency: whatever you do, stick with it.
3. Level of detail: decide how much detail you'll need now and in the future.

Example:

The codebook for the General Social Survey is an enormous document that helps researchers use the data effectively and ensures that every variable is described.

Reminder!

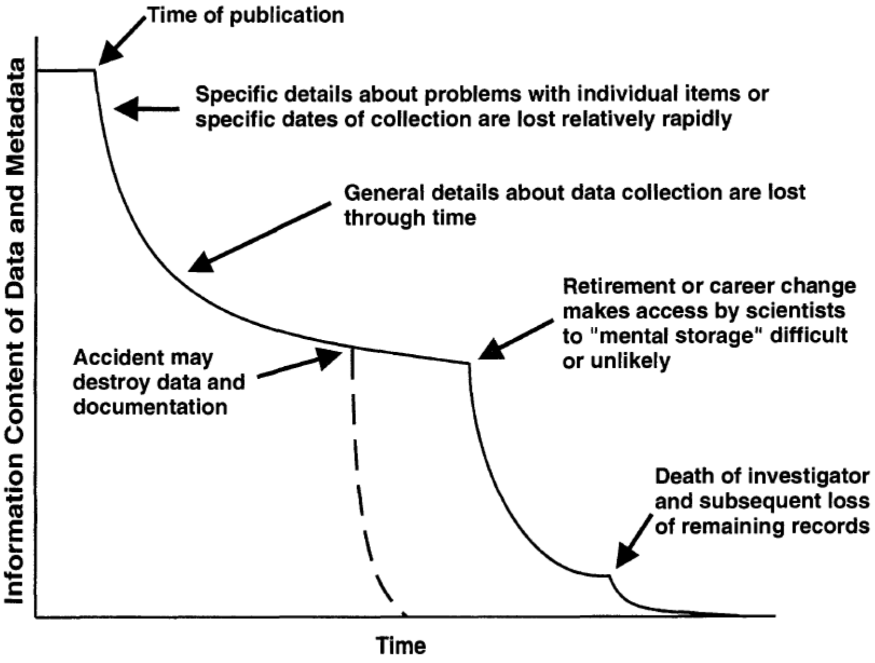

Figure 2: Bill Michener's description of data completeness over time

Data processing & analysis

What does this stage involve?

These stages are in separate boxes on the lifecycle model, and they may indeed be different steps, but not always. You usually process data in order to get to an analyzable form of it. The stages have the same considerations. This stage includes any data cleaning, refinement, integration, and organizing (combining variables, weighting variables) that you might do, as well as any computation necessary for analysis.

What tools and resources are available?

Software:

- Stata
- SAS
- MatLab
- R
- OpenRefine
- Python
- DataONE software tools catalog

CSSSI Workshops

High Performance Computing

Geographic Information Systems

Workflow tools

- Kepler: https://kepler-project.org/ : open source scientific workflow application.
- VisTrails: http://www.vistrails.org : open source scientific workflow application, emphasis on visualization.

People:

Steve Weston, HPC specialist Steve has office hours in the CSSSI from 9:30 - 1:00 on Wednesdays.

Stace Maples, GIS specialist Stace has office hours in the CSSSI, the med library, and HGS. Find out more here: http://guides.library.yale.edu/gis

StatLab consultants: StatLab consultants staff a desk in the CSSSI. Their schedules are: http://csssi.yale.edu/csssi-statistical-consultants-schedule

Kristin Bogdan & Michelle Hudson, Data Librarians Kristin & Michelle have offices in CSSSI, and you can see their offsite office hours at: http://bit.ly/datalibofficehours

Guidelines:

1. Keep track of everything you do and always keep versions of your data sets.
2. Best practices for working with data during analysis – folder structures, naming conventions, statistical package considerations.
3. How to back up data.

Data archiving & preservation

What does this stage involve?

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.

What tools and resources are available?

Lists of repositories: A few projects aim to list all the data repositories available for submission or for finding research data to reuse, and you can search or browse by subject: + DataBib + re3data

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

ICPSR 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 distribution & citation

What does this stage involve?

This is the stage (usually after archiving) where you can make your data, or a link to your data available, so that others know they can get your raw materials and use them in their own research, or check your studies for replication.

What tools and resources are available?

DataCite

Repositories (listed above)

Guidelines:

1. Give your data set a title and make it easy to credit you.
2. Always cite data that you use as if it were as important as the journal articles you cite.

Examples:

1. ICPSR data citation
2. DataCite data citation

References & other resources:

NECDMC Some material from this presentation came from the New England Collaborative Data Management Curriculum.

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

Guides & links These guides may be useful as you work on your projects: + http://guides.library.yale.edu/datamanagement + http://guides.library.yale.edu/data-statistics + http://guides.library.yale.edu/sciencedata + http://guides.library.yale.edu/eln + http://csssi.yale.edu/datamanagement

Contact info

Michelle Hudson

- Science and Social Science Data Librarian
- 203.432.4587
- michelle.hudson@yale.edu
- office hours: http://bit.ly/datalibofficehours

Kristin Bogdan

- Science and Social Science Data Librarian
- 203.436.5907
- kristin.bogdan@yale.edu
- office hours: http://bit.ly/datalibofficehours

StatLab Consultants

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