Research Data Management 101

Michelle Hudson, Kristin Bogdan

21 February 2014

https://github.com/michellehudson/datamanagement/

michelle.hudson@yale.edu,
kristin.bogdan@yale.edu

Overview:

Using the DDI data lifecycle model as a guide, this workshop will 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?

Helpful guides:

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

More resources:

CSSSI Workshops:
http://statlab.stat.yale.edu/workshops/

High Performance Computing:
http://its.yale.edu/services/research-technologies/high-performance-computing

Geographic Information Systems:
http://guides.library.yale.edu/gis

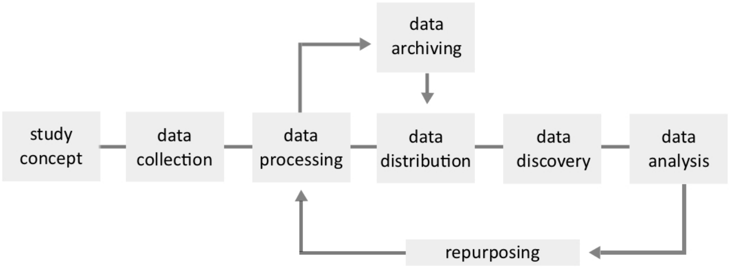

Figure 1: Data lifecycle model based on DDI.

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." [1]

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.

[1] OMB Circular A-110.

Study concept

DMPTool https://dmp.cdlib.org/

Data Management Planning Consultation Group http://csssi.yale.edu/dmp

StatLab consultants http://csssi.yale.edu/csssi-statistical-consultants-schedule

Data collection & documentation

Yale-supported resources:

- Box
- LabArchives
- EliApps
- Qualtrics

Data processing & analysis

- Stata, SAS, MatLab, R, OpenRefine, Python
- DataONE software tools catalog
- Kepler: https://kepler-project.org : open source scientific workflow application.
- VisTrails: http://www.vistrails.org : open source scientific workflow application, emphasis on visualization.

Data archiving, preservation, distribution, and citation:

- DataCite https://www.datacite.org/
- re3data http://www.re3data.org
- DataBib http://databib.org/

Additional services & software:

- GitHub: https://github.com/
- Morpho https://knb.ecoinformatics.org/morphoportal.jsp
- Earthcube http://earthcube.org/
- Colectica http://www.colectica.com/

Guidelines:

1. Visit the StatLab before you start your project.
2. Consider making a data management plan even if you aren't seeking a grant.

Guidelines for data collection & documentation:

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.

Guidelines for data processing & analysis:

1. Visit the StatLab before you start your project.
2. Keep track of everything you do and always keep versions of your data sets.
3. Best practices for working with data during analysis – folder structures, naming conventions, statistical package considerations.
4. Back up data in accordance with good practice.

Guidelines for data archiving & preservation:

1. Backup is not sufficient for preservation.
2. Doing preservation yourself requires format migration and ensuring integrity of files.
3. Handing over your data to a repository like ICPSR is possible, and will ensure the data is usable over the long-term.

Guidelines for data distribution & citation:

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.
3. Look for domain-appropriate distribution channels.