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

Additional help:

CSSSI Workshops
High Performance Computing
Geographic Information Systems

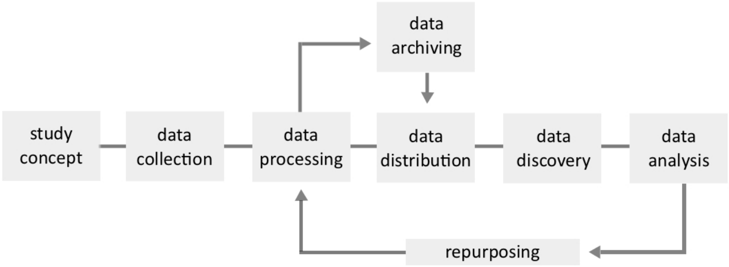

Figure 1: Data lifecycle model based on DDI.

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]

There are four types of research data: [2]

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).

[1] OMB Circular A-110.

[2] 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?

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.

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.

Study concept

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.

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.

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.

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.

Data collection & documentation

Yale-supported resources:

- Box

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.

- LabArchives
- EliApps
- Qualtrics

Additional services & software:

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

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.

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.

Data archiving & preservation

What does this stage involve?

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.
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 and re3data

Guidelines for data archiving & preservation:

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.

Data distribution & citation

Tools & resources

DataCite

re3data

DataBib

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.