Intro to Research Data Management for PoliSci

Michelle Hudson

michelle.hudson@yale.edu

Fall 2014

http://bit.ly/intro-rdm-polisci

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?

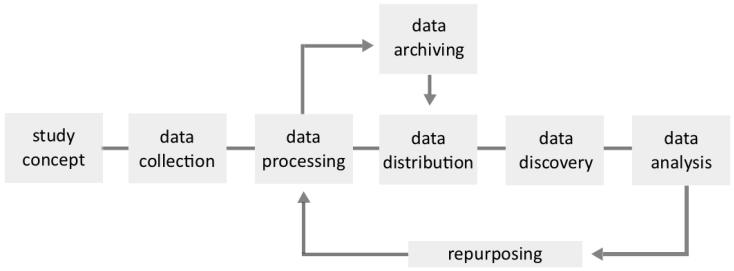

Figure 1: Data lifecycle model based on DDI.

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).

[1] OMB Circular A-110.

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.

Helpful guides:

http://guides.library.yale.edu/datamanagement
http://guides.library.yale.edu/data-statistics
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

Study concept

DMPTool https://dmptool.org Log into the DMPTool with your Yale NetID and password and get access to great tools for building a data management plan for the agency of your choice.

Data Management Consultation Group http://csssi.yale.edu/dmp The consultation group can review plans, help write plans, or help refer to needed services.

Data consultants http://csssi.yale.edu/csssi-statistical-consultants-schedule CSSSI's data consultants can help you figure out how to collect, analyze, or use your data, formulate research methodology, or just help you think through concepts.

Guidelines:

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

Data collection & documentation

Yale-supported resources:

- Box - 50G of cloud-based storage space
- EliApps - Yale's version of Google Drive, etc.
- Qualtrics - Sophisticated survey building software.
- GitHub - Code/data/file repositories with version control.

Data collection & documentation:

1. Look at great examples of documentation, like the General Social Survey.
2. Consistency: whatever you do, stick with it.
3. Level of detail: What would someone need to know to re-use your data or replicate your findings?

Data processing & analysis

- Stata, SAS, MatLab, R, OpenRefine, Python
- DataONE software tools catalog
- Tech at CSSSI http://csssi.yale.edu/tech

Data processing & analysis:

1. Visit the CSSSI 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, distribution, and citation:

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

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.

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.