Research Data Management 101

Michelle Hudson, Kristin Bogdan

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

21 February 2014

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

Overview:

DDI lifecycle here

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?

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] 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.

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).

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. 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.

Study concept

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.

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.

Data collection & documentation

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.

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.

- 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: https://github.com/ 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: https://knb.ecoinformatics.org/morphoportal.jsp Morpho was developed for data management in ecology.
- Earthcube: http://earthcube.org/ Earthcube is a community driven data and knowledge management system that will allow for data sharing across the geosciences.
- Colectica: http://www.colectica.com/ Colectica is software that helps design, document, and publish statistical data and survey research using open data standards.

Data processing & analysis

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

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.

Guidelines for data processing & analysis:

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. Back up data in accordance with good practice.

1. CSSSI Workshops
2. High Performance Computing
3. Geographic Information Systems

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.

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.

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

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)

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`

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.