Research Data Management 101

Michelle Hudson & Kristin Bogdan

CSSSI workshop Feb 21 2014

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

General format:

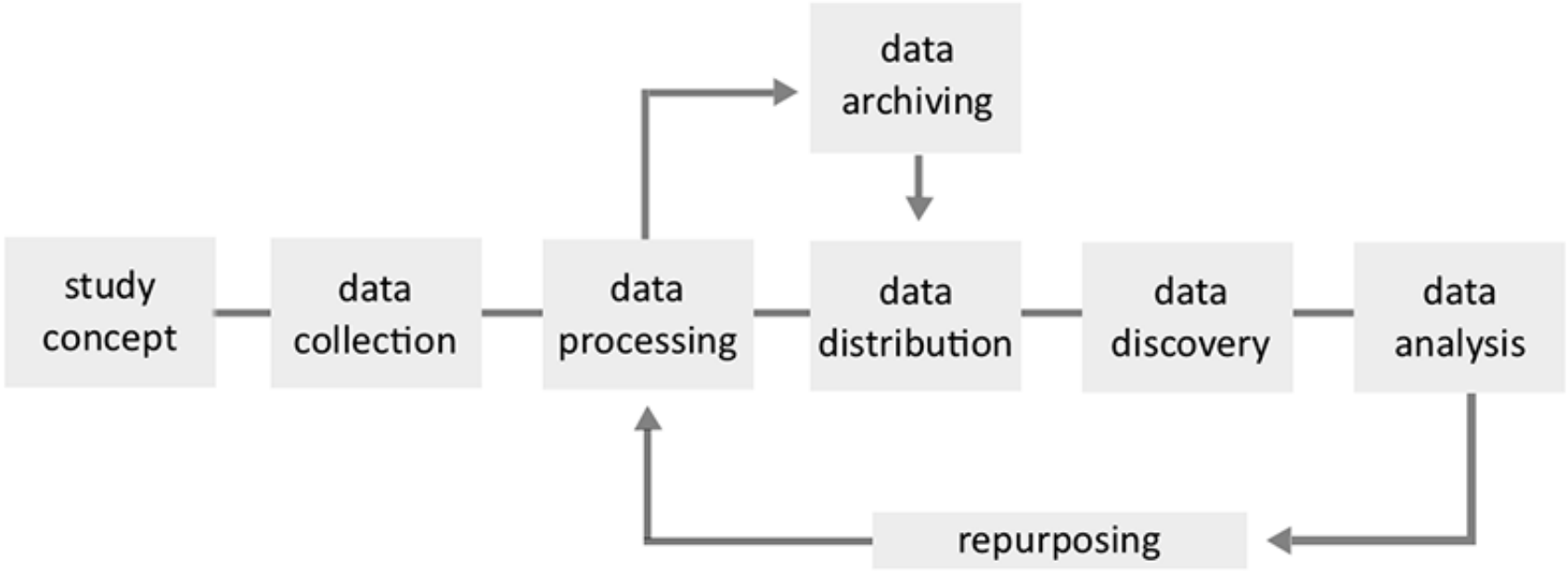
Using the DDI data lifecycle model as a guide, we'll cover the following questions:

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

DDI Lifecycle model



What is research data?

"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?

Transparency, integrity, and reproducibility

Compliance

Personal & professional benefits

Study concept

What does this stage involve?

Formulating a research question; deciding on methods; grant submission; data management planning.

Study concept

Tools & resources:

DMPTool

DMP Consultation Group

StatLab consultants

Data collection & documentation

What does this stage involve?

Collection & documentation of data; collaborations with colleagues.

Data documentation

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

Data documentation

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

Data documentation

Tools & resources:

Yale-supported

Box

LabArchives

EliApps

Qualtrics

Data documentation

Additional resources

GitHub

Morpho

Earthcube

Colectica

Data documentation

Guidelines:

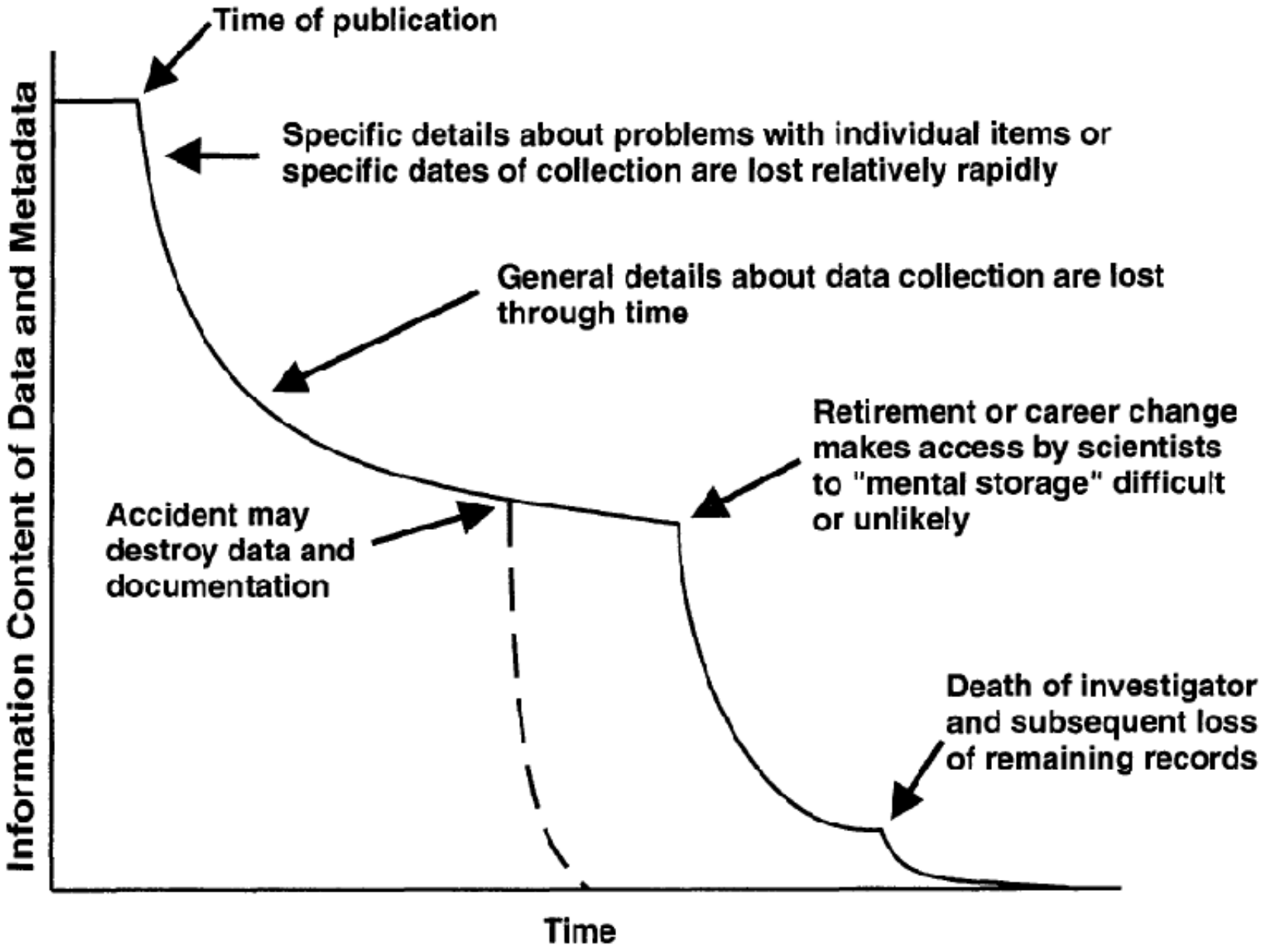
1. Spreadsheets vs. databases: see the upcoming workshop on database design: 4/18/2014, 1:30 - 3:30 CSSSI.
2. Consistency: whatever you do, stick with it.
3. Level of detail: decide how much detail you'll need now and in the future.

Data documentation

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.

Bill Michener's description of data completeness over time:



Data processing & analysis

What does this stage involve?

processing data: cleaning, refining, integrating, organizing, analyzing.

Data processing & analysis

Tools & resources:

Software:

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

Data processing & analysis

Tools & resources

CSSSI Workshops

High Performance Computing

Geographic Information Systems

Data processing & analysis

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.

Data processing & analysis

People!

Steve Weston, HPC specialist

Steve has office hours in the CSSSI, TBD

Stace Maples, GIS specialist

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

Data processing & analysis

People!

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

Data processing & analysis

Guidelines:

1. Keep track of everything you do.
2. Best practices for working with data during analysis -- folder structures, naming conventions, statistical package considerations.
3. How to back up data

Data processing & analysis

Tools & resources:

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 processing & analysis

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.

Data processing & analysis

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.

Data distribution & citation

Tools & resources:

DataCite

DataBib

re3data

Data distribution & citation

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.

Data distribution & citation

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

Contact info

StatLab Consultants

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

Thank you!

Q&A