Documentation and Organization for Data and Processes

Peg Burnette
Ayla Stein
Heidi Imker
Data Management Workshop Series

• **Introduction to Data Management**
  – Feb 10\textsuperscript{th} 4PM – 5PM
  – Apr 6\textsuperscript{th} 1PM – 2PM

• **Documentation and Organization for Data and Processes**
  – Feb 17\textsuperscript{th} 4PM – 5PM
  – Apr 13\textsuperscript{th} 1PM – 2PM

• **Making Research Data Public: Why, What, and How**
  – Feb 24\textsuperscript{th} 4PM – 5PM
  – Apr 20\textsuperscript{th} 1PM – 2PM
Objectives

1. Consider and adopt an overall organizational structure for project and data files
2. Learn how to visualize your research processes through a workflow diagram
3. Identify what kind of data documentation is beneficial for your research
Reality

From PhD Comics (http://www.phdcomics.com/comics/archive.php?comicid=1323)
Process Documentation
Workflow Diagrams

- **good starting point for mapping out elements of a project**
- **provide a visual of the entire process**
- **help you (and others) understand the steps needed to acquire and analyze your data**
Example Workflow Diagrams

Doesn’t have to be pretty
Example Workflow Diagrams

Can be simple
Example Workflow Diagrams

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0080278
Don’t forget project level documentation

• Context of the project
• Funding & IRB review
• Personnel
• Instruments
• Decisions made
Basic Organization
Controlled Vocabulary

- Take the guess work out of choosing between:
  - a preferred spelling
    - behavior vs behaviour
  - a scientific or popular term
    - pig vs porcine vs Sus domesticus
  - determining which synonym to use
    - record vs entry
  - determining which abbreviation to use (if you have to)
    - USA vs US
Controlled Vocabulary Examples

• MeSH – Medical Subject Headings

• WordNet – English Lexicon Database

• ISO 3166 – Codes for countries & their regions

• ISO 639.2 – Abbreviations for language names
Consistent File and Folder Naming

For quickly finding and sorting files and folders, the names should be consistent but unique. Avoid special characters.

- project name/acronym
- experiment/instrument type
- site location information (if applicable)
- researcher initials
- date (consistently formatted, i.e. YYYYMMDD)
- version number (w/ leading zeros)
ISO Date Standard

PUBLIC SERVICE ANNOUNCEMENT:

Our different ways of writing dates as numbers can lead to online confusion. That's why in 1988 ISO set a global standard numeric date format.

This is THE correct way to write numeric dates:

2013-02-27

The following formats are therefore discouraged:

02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013 II. 27 27⅔-13 2013.158904109
MMXIII-II-XXVII MMXIII \( \frac{LXXX}{VII} \) 1330300800
\((3+3)\times(111+1)-1\times3/3-1/3^3 \) \( \frac{2373}{3} \)
10/11031/1101 02/27/2013 \( \frac{01237}{5} \)
Power Naming

Bacillus anthracis methyltransferase co-expression activity experiment 01 September 04 2014.txt
Bacillus anthracis methyltransferase co-expression activity experiment 02 September 04 2014.txt
Bacillus anthracis methyltransferase co-expression activity experiment 03 September 04 2014.txt
Versioning

- **Raw data** (no manipulation)
  - HJI BAM Co-Exp Run 01 20140904 RAW.txt
  - HJI BAM Co-Exp Run 02 20140904 RAW.txt
  - HJI BAM Co-Exp Run 03 20140904 RAW.txt

- **Working data**
  - HJI BAM Co-Exp Run 01 20140904.txt
  - HJI BAM Co-Exp Run 02 20140904.txt
  - HJI BAM Co-Exp Run 03 20140904.txt

- **Versioned working data**
  - HJI BAM Co-Exp Run 01 20140904 V01.exls
  - HJI BAM Co-Exp Run 02 20140904 V01.exls
  - HJI BAM Co-Exp Run 03 20140904 V01.exls

- **Analyzed Data**
  - HJI BAM Co-Exp 2014 09 Analysis V01.enz
  - HJI BAM Co-Exp 2014 09 Analysis V10.enz
Storage & Back-Up Best Practices

• Security
• Robustness
• Number of copies
• Location of copies
• Scheduling
• Efficacy
Data Documentation
“Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information” (NISO, Understanding Metadata 2004;1).
Data Documentation Continuum

- Low-BARRIER
  - Fast
  - Easy

- Low-QUALITY
  - Irregular
  - Incomplete

- High-QUALITY
  - Standardized
  - Rich

- High-BARRIER
  - Slow
  - Skilled
Why take the time?

- find data from other researchers to support your research;
- use the data that you do find;
- help other professionals to find, understand, and use data from your research; and
- use your own data in the future when you may have forgotten details of the research.
- save time when depositing your research
ReadMe Files

Simple text file that accounts for all files and folders in a dataset

- basic information about the dataset (author, year, associated publication as appropriate)
- explanation of naming conventions
- relationship between directory structure and the data
- additional information for re-use (see extensive example)
# Codebooks

Primary source of data documentation
- Study-level information
- descriptions and annotations of the variable and data items

<table>
<thead>
<tr>
<th>#</th>
<th>Variable / Field Name</th>
<th>Field Label</th>
<th>Field Attributes (Field Type, Validation, Choices, Branching Logic, Calculations, etc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>study_id</td>
<td>Study ID</td>
<td>text</td>
</tr>
<tr>
<td>2</td>
<td>date_ensroled</td>
<td>Data subject signed consent YYYY-MM-DD</td>
<td>text (date_ymd)</td>
</tr>
<tr>
<td>3</td>
<td>patient_document</td>
<td>Upload the patient's consent form</td>
<td>file</td>
</tr>
<tr>
<td>4</td>
<td>first_name</td>
<td>First Name</td>
<td>text, Identifier</td>
</tr>
<tr>
<td>5</td>
<td>last_name</td>
<td>Last Name</td>
<td>text, Identifier</td>
</tr>
<tr>
<td>6</td>
<td>address</td>
<td>Street, City, State, ZIP</td>
<td>notes, Identifier</td>
</tr>
<tr>
<td>7</td>
<td>telephone_1</td>
<td>Phone number Include Area Code</td>
<td>text (phone), Identifier</td>
</tr>
<tr>
<td>8</td>
<td>email</td>
<td>E-mail</td>
<td>text (email), Identifier</td>
</tr>
<tr>
<td>9</td>
<td>dob</td>
<td>Date of birth</td>
<td>text (date_ymd), Identifier</td>
</tr>
<tr>
<td>10</td>
<td>age</td>
<td>Age (years)</td>
<td>odbc Calculation.rounddown(datediff([dob],today),Y)</td>
</tr>
<tr>
<td>11</td>
<td>ethnicity</td>
<td>Ethnicity</td>
<td>radio</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0 Hispanic or Latino</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 NOT Hispanic or Latino</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 Unknown / Not Reported</td>
</tr>
</tbody>
</table>

Custom alignment: LH
Data Dictionaries

provides a detailed description for each element or variable in your dataset and data model
- descriptive name
- the data type
- allowed values, units, and text description
- additional information for re-use
Metadata Schemas/Standards

• Some disciplines already have systems
  – Ecological Metadata Language (EML)
  – Data Documentation Initiative (DDI)

• Generic standards to consider
  – Dublin Core

• Sometimes multiple/no standards
Objectives

1. Consider and adopt an organizational structure
2. Visualize research processes through a workflow diagram
3. Identify what kind of data documentation is beneficial for your research
Contact

Peg Burnette
Biomedical Librarian
phburn@illinois.edu

Ayla Stein
Metadata Librarian
astein@illinois.edu

Heidi Imker
Director, Research Data Service
imker@illinois.edu

Visit:
310-312 Main Library

Call:
(217) 300-3513

Website:
researchdataserviceternal@illinois.edu
Data Management Workshop Series

• **Introduction to Data Management**
  - Feb 10\textsuperscript{th} 4PM – 5PM
  - Apr 6\textsuperscript{th} 1PM – 2PM

• **Documentation and Organization for Data and Processes**
  - Feb 17\textsuperscript{th} 4PM – 5PM
  - Apr 13\textsuperscript{th} 1PM – 2PM

• **Making Research Data Public: Why, What, and How**
  - Feb 24\textsuperscript{th} 4PM – 5PM
  - Apr 20\textsuperscript{th} 1PM – 2PM
## Scholarly Commons

http://www.library.illinois.edu/sc/index.html

## Library Home Page

http://www.library.illinois.edu/

## Subject Specialists

<table>
<thead>
<tr>
<th>Subject Area</th>
<th>Library</th>
<th>Subject Specialist</th>
<th>Phone (217)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accounting</td>
<td>Business Information Services</td>
<td>Cerissa Phillips</td>
<td>333-0169</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Becky Smith</td>
<td>244-0388</td>
</tr>
<tr>
<td>Advertising</td>
<td>Communications Library</td>
<td>Lisa Romero</td>
<td>333-6348</td>
</tr>
<tr>
<td>African Studies</td>
<td>International &amp; Area Studies Library</td>
<td>Atoma Batuma</td>
<td>333-2904</td>
</tr>
<tr>
<td>Afro-American Studies</td>
<td>African American Research Center in the History, Philosophy, &amp; Newspaper Library</td>
<td>Thomas Weissinger</td>
<td>333-3006</td>
</tr>
<tr>
<td>Aging/Gerontology</td>
<td>Social Sciences, Health, and Education Library</td>
<td>JJ Pionke</td>
<td>265-0002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Peg Burmette</td>
<td>300-5942</td>
</tr>
</tbody>
</table>

http://www.library.illinois.edu/administration/collections/subjectdirectory.html
Useful?
Tell your friends and colleagues!

Not Useful?
Tell us!