Data Management: Why is it important?

Money

Time

Sharing

Integrity

Improve your analysis

Writing a DMP
What do we mean by *data*?

<table>
<thead>
<tr>
<th>General</th>
<th>Social Sciences</th>
<th>Natural/Physical Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>• images</td>
<td>• survey responses</td>
<td>• measurements generated by sensors/laboratory</td>
</tr>
<tr>
<td>• video</td>
<td>• focus group and individual interview transcripts</td>
<td>instruments</td>
</tr>
<tr>
<td>• mapping/GIS data</td>
<td>• economic indicators</td>
<td>• computer modeling</td>
</tr>
<tr>
<td>• numerical measurements</td>
<td>• demographics</td>
<td>• simulations</td>
</tr>
<tr>
<td></td>
<td>• opinion polling</td>
<td>• observations and/or field studies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• specimen</td>
</tr>
</tbody>
</table>
Research data lifecycle phases

Pre-project
- Study Concept

Active phase
- Data Collection
- Data Processing
- Data Distribution
- Data Discovery
- Data Analysis
- Repurposing

Post-project
- Data Archiving

Secondary use

Lifecycle adapted from DDI version 3.0 Combined Life Cycle Model
A case study
Data Documentation

AKA Metadata
Metadata: why does it matter?

Data is not self-describing.

Metadata, or “data about data” explains your dataset and allows you to document important information for:

- Finding the data later
- Understanding what the data is later
- Sharing the data (both with collaborators and future secondary data users)
- Consider it an investment of time that will save you trouble later several-fold
Documentation

Project name

Project summary

Funding info

Primary contact info

Other people working on the project

Location of data and supporting info (lab notebooks, procedures, etc.) for project

Organization and naming conventions used for data
Capturing your Metadata

In a filename

In a README.txt

In a spreadsheet

In an XML file

Into a database
Data Organization
Why file organization is important

Once your research gets underway, there may be multiple files in various formats, multiple versions, methodologies, etc., all relating to your research.
# File organization: naming conventions

<table>
<thead>
<tr>
<th>Best Practice</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limit the file name to 32 characters (preferably less!)</td>
<td>32CharactersLooksExactlyLikeThis.csv</td>
</tr>
</tbody>
</table>
| When using sequential numbering, **use** leading zeros to allow for multi-digit versions | NO   ProjID_1.csv ProjID_12.csv  
| For a sequence of 1-10: 01-10                                               | YES  ProjID_01.csv ProjID_12.csv  
| For a sequence of 1-100: 001-010-100                                       |                              |
| Don't use special characters & , * % # ; * ( ) ! @$ ^ ~ '{ } [ ] ? < > -      | NO   name&date@location.doc   |
| Use only one period and use it before the file extension                     | NO   name.date.doc            |
|                                                                             | NO   name_date..doc           |
|                                                                             | YES  name_date.doc            |
| Avoid using generic data file names that may conflict when moved from one location to another | NO   MyData.csv              |
|                                                                             | YES  ProjID_date.csv          |
File Organization: naming conventions

Example for our case study:

Picture_a2Xc38_butterfly.sam

How could this be better named?

Identifiers?
File organization: naming conventions

Maybe started with:

*abcdefgghijklmnopqrstuvwxyz.sam*

Sashimi Microscope format

Descriptive element

Date as YYYYMMDD

Initials because working in a group

Ascension # because part of a series

File format
Data Storage
Data Storage: during active phase

Ideally keep 3 copies of your data!

- Local/working copy
- Local external copy (e.g. external hard drive)
- Remote copy offsite (e.g. cloud storage)
## File formats: preferred examples

<table>
<thead>
<tr>
<th>Proprietary Format</th>
<th>Alternative/Preferred Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excel (.xls, .xlsx)</td>
<td>Comma Separated Values (.csv) ASCII</td>
</tr>
<tr>
<td>Word (.doc, .docx)</td>
<td>plain text (.txt), XML, PDF/A, HTML, ODF</td>
</tr>
<tr>
<td></td>
<td>or if formatting is needed, PDF/A (.pdf)</td>
</tr>
<tr>
<td>PowerPoint (.ppt, .pptx)</td>
<td>PDF/A (.pdf), ODP, JPEG 2000, PDF, PNG</td>
</tr>
<tr>
<td>Photoshop (.psd)</td>
<td>TIFF (.tif, .tiff),</td>
</tr>
<tr>
<td>Quicktime (.mov)</td>
<td>MPEG-4 (.mp4), MOV, AVI, MXF</td>
</tr>
<tr>
<td>Sounds</td>
<td>WAVE, AIFF</td>
</tr>
<tr>
<td>Containers</td>
<td>TAR, GZIP, ZIP</td>
</tr>
<tr>
<td>Databases</td>
<td>XML, CSV</td>
</tr>
</tbody>
</table>

- **Sashimi Format**: Proprietary format .sam
- **OME-TIFF file format**: Open source format .tif
Data Security

Software

Passwords

Safe Storage Environment

Encryption
Preservation and Sharing
Preserving and Sharing

Preservation
Repositories
Individual

Sharing
Website
Email upon request
Repository
Preserving and Sharing

Our Case Study chose…

Email upon request

Why?

informal/easy

Keep control over who sees it

Control over managing and preserving
Scenario: What’s good and bad?

A project on ground water changes using GRACE satellite data.
Quick Practices

Check backups
Organization system
Take better notes
Review security plan
Check to see if you can access old files