Tips & Resources for Data Management Planning

LIZA BRAZIL, CUAHSI PRODUCT MANAGER
SEPTEMBER 19, 2018
Goals for Today

1. Introduce important Data Management concepts and Best Practices

2. Give you some tips for impressing reviewers with your Data Management Plan

3. Describe CUAHSI Data Management Plan resources available to you
Why Data Management?

Individual Efficiency
Easier to Collaborate
Study Reproducibility
Data can be re-used and re-purposed - *in other disciplines too!*

WWW.CUAHSI.ORG
The Research Data Management Lifecycle

**Brainstorming Exercise**

Think about or draw a flowchart to describe what you do with your data from the very beginning of your project to the very end.
The Research Data Management Lifecycle

1. Research Question
2. Collection
3. Data Description
4. Data Storage
5. Analysis
6. Re-collection
7. Data Management Plan
8. Data Search/Reuse
9. Publication
10. Archive
Best Practices for Data Management
1. Don’t Mess with the Raw Data

Always store uncorrected data with all of its “bumps and warts”

- Do not make any corrections
  - You could change something that was actually correct
  - You could make mistakes while correcting other mistakes
- Script QA/QC procedures and write results to a new file/copy

Removal of a calibration shift
2. Use Descriptive File Names

- Use only plain ASCII characters
- Brief, but descriptive of content
- Generally avoid spaces in file names, use underscores instead
- For new versions of files, use date format ISO 8601: YYYYMMDD
- Include a “readme” file when using many files in a directory
3. Use Descriptive Headers in Files and Tables

Standard convention for many software applications
- Excel understands that the first line in a file is the header line
- Subsequent lines are interpreted as data

Encapsulate data and descriptive metadata together
4. Do Not Mix Data Types in Table Columns

- Numeric, strings, date/time, boolean
- Different software packages will handle mixed data types inconsistently
- Can be more difficult to detect errors in the data
- Can cause errors in results
5. Data Collection Includes Data Entry

- **Examples include:**
  - Recording observations and notes in a field notebook
  - Transcribing field notebooks and sheets into digital forms
  - Automating processing of sensor data streams into a database
  - Adding a DEM layer to your geodatabase

- **Tips for data entry:**
  - Use pre-designed forms or templates
  - Use lists of valid values rather than free form text entry
  - Use validation checks (e.g., range checks)
  - Example: pH must be between 0 and 14. If it’s not – there is a problem!
6. Archive Data in Non-Proprietary Data Formats

- Microsoft Excel is widely available and used now, but what about in 10 years? 20 years?
- How many other software programs can open your data?
- Will your data disappear if the file format/software become obsolete?
7. Preservation of Data...To the Cloud!

Prevent data loss and preserve your data in the cloud!

- Convenience
- Cross platform
- Enhanced sharing
- Low cost

But...

- Privacy
- File formats and semantics still matter
8. Automate Your Analyses

- Code creates reproducible results
- Code is a record of the steps involved in processing and analyzing data
- Code can be shared
- Code can be reexecuted at any time

Above: Reproducible visualization in Python
9. Maintain Metadata

- WHO created the data?
- WHAT is the content of the data?
- WHEN were the data created?
- WHERE is it geographically?
- WHY were the data developed?
- HOW were the data developed?

Details about the cereal you ate for breakfast this morning.

*Borer et al.: “Do not underestimate your ability to forget details about a study!”*
9. Maintain Metadata Part II

<table>
<thead>
<tr>
<th>Level</th>
<th>Planned use</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>III</td>
<td>Publishable &amp; auditable</td>
<td>Inadequate</td>
<td>Minimal</td>
<td>Good Practice</td>
</tr>
<tr>
<td>II</td>
<td>Searchable &amp; third party reuse</td>
<td>Minimal</td>
<td>Good Practice</td>
<td>Excessive</td>
</tr>
<tr>
<td>I</td>
<td>Exchange with expert colleague</td>
<td>Good Practice</td>
<td>Excessive</td>
<td>Excessive</td>
</tr>
</tbody>
</table>

The degree of metadata format and structure necessary for different levels of projected secondary data utilization (adapted from Michener et al., 1997).
Data Management

Best Practices

Summary

- Don’t mess with the raw data
- Use descriptive file names
- Use descriptive file headers
- Do not mix data types in table columns
- Archive data in non-proprietary data formats
- Consider leveraging cloud providers
- Automate analyses
- Maintain metadata structure

**When you’re a data provider:**
- Why were the data created?
- What limitations do the data have?
- What does the data mean?
- How should the data be cited if it is re-used in a new study?

**When you’re a data user:**
- What are the data gaps?
- What processes were used for creating the data?
- Are there any fees associated with the data?
- In what scale were the data created?
- What software do I need in order to read the data?
- What projection are the data in?
- Can I give these data to someone else?

*You don’t know why your data will be used in the future!*

E.g., someone in public health or sociology may want your groundwater chemistry data
Resources for Data Sharing & Preserving

- Emerging data repositories have functionality for collaboration and preservation
- Still very much domain specific
- Impact is higher when you choose appropriately!
Tips for impressing your reviewers with your Data Management Plan

Include:

- Type(s) of data being collected and how it will be collected
- Standards that will be adhered to
- Provisions for archiving and preservation
- Expected reuse (research, education, general public?)
- Plans for eventual transition or termination of the data collection in the long-term future
### CUAHSI Resources for Data Management, Preservation, & Reuse

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>Staff Support: <a href="mailto:help@cuahsi.org">help@cuahsi.org</a></td>
</tr>
</tbody>
</table>
Let’s go find some data in HydroShare, a CUAHSI-supported repository.
https://toolbox.google.com/datasetsearch
To Review

Use Best Practices for Data Management Planning

For this part of your proposal, you know what your reviewer is looking for!

CUAHSI resources and tools are here to help
Q&A