Smart and Simple Data Management
Data is defined, consistent with OMB circular A-110, as:

the digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications

but generally does not include ...

lab notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, physical objects, such as lab specimens
National Data Policy - 2013

OSTP MEMO: INCREASING ACCESS TO THE RESULTS OF FEDERALLY FUNDED SCIENTIFIC RESEARCH

“requiring researchers to better account for and manage the digital data resulting from federally funded scientific research”

• Data management plans will be compulsory
• Providing public access to data will become more routine

http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research
Data Management Practices

- complete records
- find
- understand
- back-up
- migrate
Data Management Reality

- I’ve lost two patent claims because we didn’t have **complete records** in lab notebooks.
- I panic every time I try to **find** my students’ data.
- I know I won’t be able to publish anything after my grad student leave because it’s too hard to **understand** what he/she did.
- We don’t have the assay results with substrate X, because my student didn’t **back-up** the computer. And we’re out of substrate X.
- The only reason I was able to do the analysis was because on a whim I decided to **migrate** some of my data off of floppy disks to CDs back in the 90s. I wish I had done all of it.
What are the elements of good data management?
Plan
Data Management Plans

• Usually short (2 page) documents submitted as part of a grant application
• Required by many federal funding agencies (more expected) and some funding announcements require them already regardless of agency policy
• When required, generally subject to panel review
• Currently DMP expectations are variable
Where to start with a DMP?

• On a tight deadline for a grant?
  – RDS Review via researchdata@library.illinois.edu or DMPtool.org
  – Revisit post-submission to refine

• Have more time?
  – Craft a DMP that you can implement and reuse
  – We’re still happy to review! researchdata@library.illinois.edu
Organize
Organization

- Distinguish Raw Data
- Be Consistent in File Naming and Directory Structures
- Define and Exercise Version Control

Used with permission from “Piled Higher and Deeper” by Jorge Cham
Copyright 2010 Jorge Cham.
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. YYYY-MM-DD)
- version number (w/ leading zeros)
Then Make it Clear and Known

**Pat’s draconian and non-negotiable naming system**

All experiment data, stats, and graph files shall be named with the following attributes:

<table>
<thead>
<tr>
<th>PROJECT NAME</th>
<th>experiment group</th>
<th>file type</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>choose one:</td>
<td>refer to master list of experiment groups and numbers e.g. GP129</td>
<td>choose one:</td>
<td>Use international standard format: YYYY-MM-DD 2015-09-23</td>
</tr>
<tr>
<td>BSUB6755</td>
<td></td>
<td>data (for raw data)</td>
<td></td>
</tr>
<tr>
<td>UICCC</td>
<td></td>
<td>stats (for input to R)</td>
<td></td>
</tr>
<tr>
<td>TRT</td>
<td></td>
<td>stats (for output of models)</td>
<td></td>
</tr>
<tr>
<td>LAS56</td>
<td></td>
<td>graphi (for input to ggplot in R)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>grapho (for graphs created in ggplot)</td>
<td></td>
</tr>
</tbody>
</table>

Example: UICCC_GP129_grapho_2012-09-23.jpg is a graph created for the data from the 129 group of the UICCC project on Sept 23 2012.
Date Tip

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
2013-02-27 2013.02.27 27-02-13
27.2.13 2013.2.27 27-II-13 2013.158904109
MMXIII-II-XXVII MMXIII 2013
X3 0300800
(3+3)×(11+1)−1×3−1×3 2013
10/1101/1101 02/27/1013

BAM Co-Exp Run 01 2014-09-04.txt
BAM Co-Exp Run 02 2014-09-04.txt
BAM Co-Exp Run 03 2014-09-04.txt

VS.

Run 1 B anth meth Sept 4 .txt
BAM Rxn 2 2014_09_04.txt
20140904_meth_3.txt
Describing Data

- Enables someone unfamiliar with your data to find, evaluate, understand, and reuse your data
- Helps YOU to remember details about your data after time has passed

Examples

- Read Me File: [http://hdl.handle.net/2142/28689](http://hdl.handle.net/2142/28689)
- Standardized Metadata: [https://www.icpsr.umich.edu/icpsrweb/neutral/ddi25/studies/35383](https://www.icpsr.umich.edu/icpsrweb/neutral/ddi25/studies/35383)
- Codebook: [https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/35383](https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/35383)
Basic Documentation

• Types of Documentation:
  – Descriptive (e.g., creator, title, keywords)
  – Structural (e.g., relation to other files)
  – Administrative (e.g., software & hardware requirements, rights information)

• If you know the data will be deposited in a repository, understand the documentation requirements early in the process
Data Documentation Continuum

Informal ReadMe

- Low-Barrier
- Fast
- Easy

- Low-Quality
- Irregular
- Incomplete

Formal Schema

- High-Quality
- Standardized
- Rich

- High-Barrier
- Slow
- Skilled

Note that the http://example.org/myapp/schema.xsd XML schema does not exist - this is a fictional example.
Controlled Vocabulary

- Take the guess work out of choosing between:
  - a preferred spelling
    - behavior vs behaviour
Storage & Backup
Storage & Backup

• Number of copies
• Location of copies
• Schedule
• Robustness
• Efficacy

Photograph used under a CC-By license. Photography taken by Gail Steinhart and available at http://www.flickr.com/photos/gailst/7824341752/
The 3-2-1 Rule

3: copies of data
2: kinds of media
1: copy remote

“There are two kinds of people in the world - those who have had a hard drive failure, and those who will.”

Example
– One: Laptop
– Two: External hard drive
– Three: Cloud storage
Data Security

• All research data is a theft target (not just military)

• Risk extends beyond just “getting hacked”
  – Confidentiality (premature disclosure or violation of law/NDA)
  – Integrity (data tampered with, not trustworthy)
  – Availability (data/system disappears, cannot work for days or weeks)

• For human subject research, IRB can offer advice on security

• Tech Services Security Office welcomes questions
  (securitysupport@illinois.edu)

• http://go.illinois.edu/uoficloud for a guide to selecting a cloud platform
Workflow mapping activity

- We’ve given you a ton of information on approaches to organization
- Getting started can be overwhelming
- Workflow mapping can help
  - At the start...to have a plan
  - In the middle...to double check
  - At the end...to know what you’ve done
- Think of it as retracing your steps
Workflow mapping activity

• Take the post-it notes
• Lay out five in a row
• Think of a specific project, or generalize
• Write out your activities for that project, from start to finish
  – Go high level – try to just use five
Workflow mapping activity

• Once you’ve written out the activities...
• Think of any data, scripts, or other files that you touch, both items you **use** and items you **make**.
  – Write down what they are
  – Their file names and/or location
  – And if/where they’re backed up
Workflow mapping activity

• And now you have a pretty solid piece of documentation!
• Number your stickies
• Take a cell phone picture, redraw this in PowerPoint, or just stick them in your handout
Archiving and Sharing
Anticipate the Long Term

• What happens to your data in...
  • 2 years?
  • 5 years?
  • 20 years?
• Will the formats be valid?
• Will the hardware be defunct?
• Will anyone know where it is?
Anticipate the Long Term

- More Control
- More Responsibility
- Less Control
- Higher Potential Impact

Keep Private  Make Public

re3data.org

Registry of Research Data Repositories
Data “Publication”

Making datasets (in and of themselves) publically accessible

• improves transparency and reproducibility of research
• save time by reducing duplication of effort (yours or theirs)
• makes the data itself independently discoverable
• another way to expose to your work
• maybe you’ll need that data again some day
Limitations for Data Publication

- protecting confidentiality and personal privacy
- recognizing proprietary interests, business confidential information, and intellectual property rights
- preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden
Illinois Data Bank (databank.illinois.edu)

A self-serve publishing platform that centralizes, preserves, and provides persistent and reliable access to Illinois research.

- can be linked to related materials, such as articles, theses, code, and other datasets
- can include files of any format and self-deposit sizes up to 15 GB/file via Box.com, larger sizes welcome by arrangement
- can be deposited for immediate release or temporarily embargoed
- receive a stable, unique identifier (DOI) for persistent access and ease of citation
- are registered in a central, world-wide catalog for better discovery
- are professionally managed and curated by the Research Data Service staff at the University Library
- are preserved for a minimum of 5 years
Contact

visit:
310-312 Main Library

call:
(217) 300-3513

website:
researchdataservice.illinois.edu
Data Management Workshop Series

1. Intro to Data Management
   - Hand-on activities: data survey, inventory, and catalog
2. Data Documentation
   - Hands-on activities: using documentation and documentation building
3. Data Sharing
   - Hands-on activities: exploring repositories and considerations for sharing
4. Data Workflows
   - More detailed workflow mapping activity, great for large and small project teams

*Check your handout for the schedule and registration info!*
Useful?
Tell your friends and colleagues!

Our goal is for every graduate student and faculty member on campus to see this presentation.

Not Useful?
Tell us!