Data Management Basics

Thea P Atwood, University of Massachusetts Amherst
Today:

- Data management tips
  - Files, Documentation, Storage & Backup
- Post-Project Activities
- Resources
- Q & A time
But first...

Introduce yourselves!

- Name
- Department
- Most common data format
  - Text, Excel, SPSS, DICOM, Google Docs, CSV, etc.
Data Management Tips

Files | Documentation | Storage & Backup
Files

Three tips:

● File naming conventions
● File versioning
● File organization
A quick note:

Try to keep your files in as open a standard as possible. This dramatically increases your ability to share your work and maintain its integrity over time.

Some suggestions:
<table>
<thead>
<tr>
<th>File type</th>
<th>Preferred Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital Audio Data</td>
<td>Free Lossless Audio Codec (.flac)</td>
</tr>
<tr>
<td></td>
<td>Waveform Audio Format (.wav)</td>
</tr>
<tr>
<td></td>
<td>MPEG-1 Audio Layer 3 (.mp3) -- spoken word only</td>
</tr>
<tr>
<td></td>
<td>Audio Interchange File Format (.aif)</td>
</tr>
<tr>
<td>Chemistry Data</td>
<td>Convert NMR, Raman, UV, Mass Spectrometry to JCAMP</td>
</tr>
<tr>
<td>Documentation &amp; scripts</td>
<td>Open Document Text (.odt)</td>
</tr>
<tr>
<td></td>
<td>Rich Text Format (.rtf)</td>
</tr>
<tr>
<td></td>
<td>Plain Text (.txt)</td>
</tr>
<tr>
<td></td>
<td>PDF/a or PDF (.pdf)</td>
</tr>
<tr>
<td>Geospatial Data</td>
<td>ESRI Shapefile (essential: .shp, .shx, .dbf; optional: .prj, .sbx, .sbn)</td>
</tr>
<tr>
<td></td>
<td>geo-referenced TIFF (.tif, .tfw)</td>
</tr>
<tr>
<td></td>
<td>CAD data (.dwg)</td>
</tr>
<tr>
<td>Digital Image Data</td>
<td>TIFF version 6 uncompressed (.tif)</td>
</tr>
<tr>
<td></td>
<td>JPEG 2000 (.jp2)</td>
</tr>
<tr>
<td>Qualitative Data</td>
<td>eXtensible Mark-up Language (SML)</td>
</tr>
<tr>
<td></td>
<td>Rich Text Format (.rtf)</td>
</tr>
<tr>
<td>Quantitative Tabular Data</td>
<td>Comma-Separated Values (.csv)</td>
</tr>
<tr>
<td></td>
<td>Tab-delimited file (.tab)</td>
</tr>
</tbody>
</table>

Some suggestions based on file format.
A file naming convention is a standard, documented, human and machine understandable way that you (and ideally, your lab and your collaborators) name your files.

Benefits:

- Easier to find files
- Avoid duplicates or ambiguity
- Easier to wrap up a project -- you know what belongs where, and with what project
Files: Naming Convention

Under your naming convention:

- Files should be named consistently
- No more than 25 characters long
- useCamelCapsToDistinguishBetweenWords
- Use_underscores_instead_of_spaces (though CamelCaps is better)
- Avoid special characters
  - e.g., “/ \ : * ? ` < > [ ] & $
- Use internationally accepted standards, where you can & where appropriate
  - For dates: YYYY-MM-DD (follows International Organization for Standardization (ISO) 8601)
File versioning is a system that allows multiple copies of a file to exist at one time.

- Automatically via specialized programs
  - Apache Subversion
  - Concurrent Versions System
- Or! DIY: save your file with a new suffix
  - e.g., change the date, update the version number
Ideally, in your file versioning system:

- Keep the **original version** of the data file the same (in its raw format, where possible)
- Save iterative versions of the file following a naming convention
- Save iterative versions with information on the type of manipulation performed  
  - e.g., data cleaning, analysis, script, etc.
- For code, try GIT
Files: Versioning

Example of file versioning:

subj01_raw_20141009
subj01_analysis_20141015
subj02_raw_20141010
subj02_analysis_20141020
subj02_analysis_20141110

LetterLearningAnalysisV01
LetterLearningAnalysisV02
LetterLearningAnalysisV03
...
LetterLearningAnalysisV10

Are these file names descriptive enough? How would you make them better?
File organization is the practice of organizing the files in a project around a central theme.

This can mean:

- One project, one folder (for small projects)
- Separate folders for data, or data types
- Separate folders for project stages
- Date-based folders (great if you have a lab notebook to map to!)

Any system is better than none!
Consider your goals when applying your organizational system.

- What do you want to get out of managing your data?
- Where do you want to data to end up?
- Will you separate your current versions from your previous versions?
- etc.

Avoid:

- One person with all the data
- Data scattered across several machines in your lab
  - Exception: backups!
Develop a naming convention for your most common data type.
Three tips:

- Document
- Document
- Document!
Documentation

No point in having a system without documentation!

**Documentation** is the practice of including pertinent information on the process and products of your research. It provides anyone* with an idea about the parameters of your data -- the who, what, when, where, why, and how of your data.

*This includes you!
One main project-wide README.txt

- Include basic project information
  - Title
  - Contributors
  - Grant or funding info
  - etc.

- Contact information for at least one person
- All data locations, including backups
- Information about the files and how they are organized
  - Describe your naming convention
  - Filing method, etc.
Two granular components: metadata and methods
Metadata is a set of data that describes and gives information about other data. It’s a way to contextualize a piece of information (e.g., a file) that otherwise lacks descriptive documentation.

In other words, metadata is another way to talk about documentation! Sometimes referred to as “data about data” or “information about information,” among others...
# Documentation

## Methods
- How the data were gathered
- How the data should be interpreted
- What you did
- Limitations of the study
- Tools used
  - e.g., code, programs
- Anything to help others replicate your processes

## Metadata
- What your study entails
- Who made what & when
- How it got there
- What it means
- What you can do with it
- Units used
- Definitions of acronyms & unusual or abbreviated terms
- etc.

**Help others understand your work before looking at the file**

*Build trust in your data*
(Honest) Documentation

You can be overly honest, too...

Dominique @microhapa
buffer was thawed quickly by placing the tube in my jeans pocket #overlyhonestmethods
12:47 PM - 6 Nov 2014
4 RETWEETS 3 FAVORITES

Eleanor Senior @EleanorSnr
@deveena152 @EmilyJAngus 'we positioned the fungal plugs in this manner as it was the way the fell onto the agar' #overlyhonestmethods
11:16 AM - 14 Nov 2014
1 RETWEET

Dan Drodge @drd1983
The dead-load applied was 53.8 newtons because that's what the ASTM handbook weighs #overlyhonestmethods
11:15 AM - 8 Jan 2013
6 RETWEETS 13 FAVORITES
You can include **informal or formal** descriptions of your data (aka, your **metadata**):

- **Informal:**
  - Can fit your unique research
  - No special standards to consider -- can include as much or as little information

- **Formal:**
  - Completeness -- may include information you wouldn’t have initially thought to include
  - Aids in sharing (commonalities between your peers)
  - Often required for deposit into a repository
    - May be required by your funder
**Documentation**

**Example Formal Standard: Dublin Core**

<table>
<thead>
<tr>
<th>Contributor</th>
<th>Publisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td>Relation</td>
</tr>
<tr>
<td>Creator</td>
<td>Rights</td>
</tr>
<tr>
<td>Date</td>
<td>Source</td>
</tr>
<tr>
<td>Description</td>
<td>Subject</td>
</tr>
<tr>
<td>Format</td>
<td>Title</td>
</tr>
<tr>
<td>Identifier</td>
<td>Type</td>
</tr>
<tr>
<td>Language</td>
<td></td>
</tr>
</tbody>
</table>
Contributor
  Karin Harman Collaborator

Creator
  Thea Atwood

Date
  2014-11-13

Description
  The three experimental runs and one anatomical run for Subject 01 for research project Letter Learning in Adults. Parameters for data collection in the readme.txt file associated with this project.

Format
  DICOM

Identifier
  Subj01_20141113

Relation
  Functional scans
    Subj01_20141113_run01
    Subj01_20141113_run02
    Subj01_20141113_run03
  Anatomical scan
    Subj01_20141113_anat

Subject
  fMRI; experimental scans

Title
  Letter Learning in Adults, Subject 01
Lots of standards! Some examples:

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Discipline</th>
<th>Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>---</td>
<td>Social &amp; Behavioral Sciences</td>
<td>Data Documentation Initiative</td>
</tr>
<tr>
<td>---</td>
<td>Ecology</td>
<td>Ecological Metadata Language</td>
</tr>
<tr>
<td>Spatial</td>
<td>---</td>
<td>Content Standard for Digital Geospatial Metadata (CSDGM) /FGDC/ISO 19115</td>
</tr>
<tr>
<td>Biodiversity</td>
<td>Life Sciences</td>
<td>Darwin Core</td>
</tr>
</tbody>
</table>

But wait, there’s more! A comprehensive list is available from the [Digital Curation Centre](https://www.dcc.ac.uk).
Overwhelmed with metadata standards? Can’t figure out what to choose?

Consult:
- Disciplinary repository (list from databib.org)
- Your peers
- Subject librarian
- Data Working Group
Decide on your standard before you collect the data!

Easier to record your metadata when collecting data than to try to remember (or convert to your pre-selected standard) later.

Regardless of what you choose, keep metadata consistently whenever you can.
Okay, great, but where do I put this stuff?

- Great question!
  - In a readme file
    - Make sure you use a non-proprietary format, like .txt!
  - In a spreadsheet
  - In an XML file
    - Some metadata programs will do this for you
  - Into a database
    - When you share the data.
    - Databases often have forms for you to fill out when you deposit the data
For your most common data type, make a list of the most important information to record for each dataset.
Storage & Backup

Three tips:
- Security
- Storage
- Backup
Security

Does your data fall under:

- HIPAA
- FERPA
- FISMA
- Human subject research, working with endangered or at-risk species, etc.

Ask for help! You’ll likely have special considerations:

- Secure storage, encryption
- Controlled access
- De-identification of personal information
- Security training
Security

Resources for Security:

● UMass Security Center from IT:
  ○ [https://www.oit.umass.edu/security](https://www.oit.umass.edu/security)

● Certificate in Information Security
  ○ Some discipline specific recommendations

● HIPAA

● FERPA

● More @ UMass’s Research & Engagement site: [http://www.umass.edu/research/](http://www.umass.edu/research/)
**Storage**
What you have on your hard drive -- what you’re actively working with and the space available on that device.

Items in storage are easily accessed and manipulated, but losing your storage can mean losing everything or just your most current work.

**Backup**
The frequent and regular process of copying your most important data and information to a secure place.

Backup is an often overlooked but invaluable part of our digital lives, and we only know its value when catastrophe strikes!
Storage & Backup

Storage & backup are related -- you need both for any project.

Follow the **Rule of 3:**

- **2 onsite**
  - Your working lab copy
  - Backed up periodically in the area (lab, university)

- **1 offsite**
  - Far enough away that it wouldn’t be hit by the same catastrophe that your lab would (e.g, flooding from major storms in Massachusetts not likely to occur in Arizona)
Ideally, your backup is:

- Low effort
- High reliability
- As secure as necessary
- As open as possible to collaborators
- Well organized
Storage & Backup

Consider your goals...

Do you need to:
● Store and backup your data, in case of machine error, theft, or natural disaster?
● Share your data with other researchers?
● Preserve your data for the future?

…and your **realities**.

● How much data?
● How frequently?
● What storage options are available?
There are many options & they all have pros and cons.

No one perfect option - but with layers (and following the rule of 3) you can get pretty close!
Storage & Backup: Your Computer

Pros
- You’re using it -- most up to date version of everything
- Often built-in backup (but you have to set it up and give it time to complete!)

Cons
- Stolen,
- dropped,
- lost,
- & broken easily (especially if you use a laptop)
- Little security

Verdict: Not enough by itself
Storage & Backup: USB/Flash Drive

Pros
● Small, convenient package
● Big enough for many datasets

Cons
● Stolen,
● dropped,
● lost,
● & washed easily
● Easy to perpetuate out-of-date copies

Verdict: Great for (moderately-sized) data transport, bad for backup.
Storage & Backup: CDs/DVDs

Pros
- Portable - okay for transport
- You can write on them

Cons
- Failure rate is high -- at best 5 - 10 yrs
- Damaged easily
- Slow to write to
- Challenge to keep up to date
- Kind of a pain

Verdict: Not good for quick backup, okay for periodic backup, but have to plan to transfer to new media
Storage & Backup: External Hard Drive

Pros
- More affordable
- Moore’s Law in play: Getting more space for less every year.
- Easy to establish periodic backups

Cons
- Failure rate moderate to high -- higher for HDs with more space & dependent on brand
- Must maintain & audit it yourself

Verdict: Coupled with automatic backups, okay choice for onsite backup. Still need a third - offsite - backup.
Storage & Backup: Shared Drives/Servers

Pros
- Keeps data out of lab and off of laptop
- Option for active data
- Not your problem to manage
- Often built-in redundancies

Cons
- Need to do homework on who is managing, what they’re using
- Can have storage quotas
- May require authentication from on-campus IP

Verdict: If well managed, a great choice for regular use, onsite, or offsite backup.
Storage & Backup: Tape

Pros
● Highly reliable
● Tolerably secure - not always on a network

Cons
● Can take a while to get data back
● Not always audited as often as they should be

Verdict: Great for backup & archiving onsite or offsite - just make sure they’re regularly audited!
Storage & Backup: The Cloud

Pros

● Syncing
● Often cheap
● Decently secure if client-side encryption occurs

Cons

● Required network connection (possible security risk)
● Ongoing costs
● Still a business
● Reliability, security, privacy not guaranteed

Verdict: Can be fine for offsite backup -- but don’t use it as your only backup.
Storage & Backup: Exercise

Do a quick inventory of your data:
  ○ What do you have?
  ○ How much space does it take up?

Inventory where your files are currently stored, including backups.

How safe are your data?
Final notes:

Any backup is better than none
Automatic is better than manual
Your research is only as safe as your backup plan

Test your backups periodically!
Post-project Activities
Post-project

Publication

- Licensing your work for clarity, ease of use, & reuse
  - Creative Commons helps make this clear to others what they can and can’t do
- Copyright & the impact of handing over your copyright to publishers
  - Ask Scholarly Communications for more help & guidance on publishing Open Access
Take-home exercise

See what others in your field are up to in terms of data management & data sharing!

If you haven’t done so yet, go out and try to get your hands on data in your field. Then ask yourself:

● Was this data easy to locate?
● Is this data easily understandable?
   ○ Terms defined, units provided, instrumentation outlined?
   ○ What other documentation is provided?
● If this data comes from a repository, what does the repository require for data deposit?
● What is missing?
● What would you do differently?
Okay, just one more thing...

Disambiguate yourself! Register for your unique ORCID identifier at orcid.org!

(Check out Thea’s ORCID [0000-0001-8348-1097] for an example)
In closing...

These are just a few tips to help you get started -- not a comprehensive overview of all the guidance ever.

If you get lost, have more questions, or want some advanced tips, remember - librarians can help!
Resources

Data Working Group
Data-minded librarians (datamgt@library.umass.edu)

Data management resources & tips from UMass
Data management LibGuide (http://guides.library.umass.edu/datamanagement)
Data Management (http://www.library.umass.edu/services/services-for-faculty/data-management/)

More training
MANTRA Research Data Mgmt Training (http://datalib.edina.ac.uk/mantra/)

Discipline specific help
Faculty | Mentors | Professional associations | Industry partners
Thank you & Let’s chat!

More questions? Get in touch: datamgt@library.umass.edu
This presentation available under a Creative Commons Attribution (CC-BY) license.

References: