The Lifecycle of a Dataset

Ashok Mudgapalli

Director of Research IT

A Case Study (concepts from MIT)

One example of how a dataset is managed

What is data?

- Observational Data
  - Patient survey data etc.,
- Experimental Data
  - Gene sequences, images, chromatograms
- Simulation Data
- Derived/compiled Data

Researcher Goals

- Organize Data
- Store and Backup Data
- Archive the Data for future use
- Share the data with other researchers

Getting Started

- Consider your goals – what do you want to get out of managing your data?
- Figure out your criteria for keeping data
- Where you want to store your data
- Consider the metadata you want to collect to document your datasets

Is Research Data correct format?

- Research data generated from machines, computers and other ways is in variety of formats
- NOW what to do?
- Convert the files into correct format, and also name the file appropriatly

Process, Analyze & Weed

- Process the data
- Analyze the data
- Weed out duplicate or erroneous files

Don’t forget to back up your data

- You want to keep three copies of your important data
  - 1 local, e.g. on your workstation
  - 1 local external drive, e.g. on an external hard drive*
  - 1 remote, e.g, network storage or cloud
    - Verify that remote has Disaster Recovery level backup

* CDs and DVDs aren’t built to last

But my files are huge!

- You can compress your data, but make sure one copy (somewhere) is uncompressed
- Document the version of compression software you used
- Use open source compression software

Versioning

- Save a copy of every iteration of data file
- Follow a file naming convention
- Consider using version control software such as GIT, GNU RCS, Mercurial (Hg), or Apache Subversion

My Research is Top-secret

- Depending on data source, you need to use encryption (FIMS, NIST standard. and so on)
- Don’t rely on 3rd party encryption alone
- Use some thing like PGP (Pretty Good Privacy)

Add Metadata

- Why does metadata even matter?
  - Metadata, or “data about data” explains your dataset and allows you to document important information for:
    - Finding the data later
    - Knowing what the data is later
    - Sharing the data later

Metadata Standards

- There are so many to choose from
- Why use a standard?
  - So later, your dataset can be organized with other datasets
  - So you have a complete, standard set of information about each Part of you data
- Not using a standard? Document anyway
  - Write down/type up everything you know about the data. Context is very important. Don’t assume you will remember it

Some well known Metadata Standards

- DDI (Data Documentation Initiative)
- Dublin Core
- Use Dublin Core http://dublincore.org/
- A lot of repositories that store data use a variety of Dublin Core
- Darwin Core
- ABCD (Access to Biological Collections Data)

Let’s Walk Through an Example

- Start with microscopic image file
- The file is in “Sashimi Environmental Scanning Electron Microscope (ESEM)” format
- Use Bio-Formats, a stand alone free software for reading and writing life sciences image file formats
- The Bio-Formats can be downloaded from http://www.loci.wisc.edu/software/bio-formats

Example Continued..

- Original file xxxxxxxx.sam got converted to xxxxxxxx.tif
- Take xxxxxxxx.tif and name it with meaningful format…… sam_expt_08242014_am_001.tif
  - Sam = Sashimi Microscope
  - expt = project or experiment abbreviation
  - 08242014 = experiment date
  - am = my initials
  - 001 = image version

Apply Metadata to the File

Directory: sam_expt_08242014

- Metadata for this directory
  - Creator: Ashok Mudgapalli
  - Subject: experiment title
  - Description: this directory contains Sashimi ESEM images of a experiment I took after finding a …….
  - Contributor: Mike Gleason helped me with these images
  - Date: 08/24/2014
  - Type: image
  - Original Format: Sashimi Microscope format (.sam)
  - Identifier: 001 schema
  - Relation: this is a directory that will contain multiple files
  - Coverage: DRC I & II
  - Rights: NIH owns the data (grant number: 00213)

Where do I put Metadata?

- In a readme file
- In a text file
- In a spreadsheet
- In an XML file
- Into a database (when I share the data)

Sharing and Storing Datasets

- Delete obsolete data
- Decide what to keep for long term
- Share and/or archive datasets

Consider: What are Your Goals

- To store and backup your data?
- To share your data with other researchers?
- To preserve your data for the future?

Storing/Backing Up Your Data

- Ideally keep three copies of your data (local, local/remote, remote)
  - Local = local computer,
  - Local / Remote = External hard drive or another lab computer
  - Remote = Network backed Enterprise storage, Cloud

You may want to share your data…

- To further science as a whole
- To further your research
- To enable new discoveries with your data
- To comply with funder/publisher
- requirements

Publishing and Sharing

- Publish references on your personal or departmental web site after you publish in Journal
- Use Cloud storage to collaborate and share files, certain restrictions and constraints apply when sharing PHI or HIPAA data
- Can share more formally in a repository

Preserving your Data

- What happens to your data when…
  - The software you use to render it changes or becomes obsolete?
  - The platform on which you manipulate it changes?
  - The hardware you created it on becomes obsolete?

File Formats for Long Term Access

Not all file formats are created equal

- ASCII text, not Excel
- PDF/A + Word, not just Word
- MPEG-4, not Quicktime
- TIFF or JPEG2000, not JPG
- XML or RDF, not RDBMS

Try for non-proprietary, open-source, standard formats

Preserving Your Data

A place to put your data where it will be:

- Stored
- Backed up
- Discoverable
- Accessible for the future (as much as possible)
- Preservation means that it is a particular person's job to make every effort to make the data usable in the future
- Preservation=Long-term access
- Some repositories ensure preservation of data over time

Repositories and Preservation

- What you should do:
  - Keep thorough documentation
  - Keep at least one copy of your data in an opne, non-proprietary format
- What the repository may do:
  - Migrate your data to contemporary formats as popular formats change (a good archive will do this for you)

Repositories - Advantages

- Put your data in a **repository** – Domain repository such as GenBank where applicable
- Provides a metadata structure for you to fill in
- Serves as a backup vehicle for your data
- May preserve your data for the future
- Makes sharing your data easy
- Others may cite your research more
- May provide some computational/online analysis tools for people to use your data
- Publishes the data for you by giving your dataset a unique persistent identifier, e.g., DOI

Unique Identifiers

- Many repositories are equipped to issue them
- Will always direct to the correct location
- DOI: http://www.doi.org
- PURL: http://purl.org
- NCBI Accession #: http://www.ncbi.nlm.nih.gov
- InChI: http://www.iuipac.org/inchi
- URI: http://www.ietf.org/rfc/rfc2396.txt
- Handle: http://hdl.handle.net

Domain Repositories - Advantages

- Your data will be stored with similar datasets (by subject, format, or both)
- Researchers will find your data easily
- The repository will understand what your data needs in terms of storage, archiving and preservation
- Computational/online analysis tools may be available tailored to analyzing that particular kind of data (e.g. GenBank for genome data)

Intellectual Property Issues

- Data is not copyrightable, but an expression of data such as a table can be
- UNMC or the funder may own your data (consult with the Technology Licensing Office)
- You can share your data if you, in fact, own it
- You can license data to limit what others can do with it (e.g., require attribution)

- It’s incumbent upon you to police usage of your data

- You can use a CC0 Declaration to emphatically put it in the public domain

Using Other People's Data

- Perhaps you are using/reusing data you got from elsewhere
- Make sure that data doesn't have a license agreement that prevents you from sharing the data
- Most databases or registries on the UNMC campus have security access controls around them and carry restrictions on use, but many do allow for educational and research use, which allows for sharing limited portions of data
- When others data is used, cite the dataset (Genbank accession number) or cite the publication

Measure Twice, Cut Once

- As you’re working on your research, always double check over time:
  - Is the data still what I think it is? (use checksums)
  - Is the metadata still available and understandable?
  - Are the formats still usable?
  - Is the software still available?
  - Is any specialized hardware still available?
  - Is the data still in the correct location?
  - Are my backups working as I expect?

Questions?

Ashok.mudgapalli@unmc.edu

x9-9072