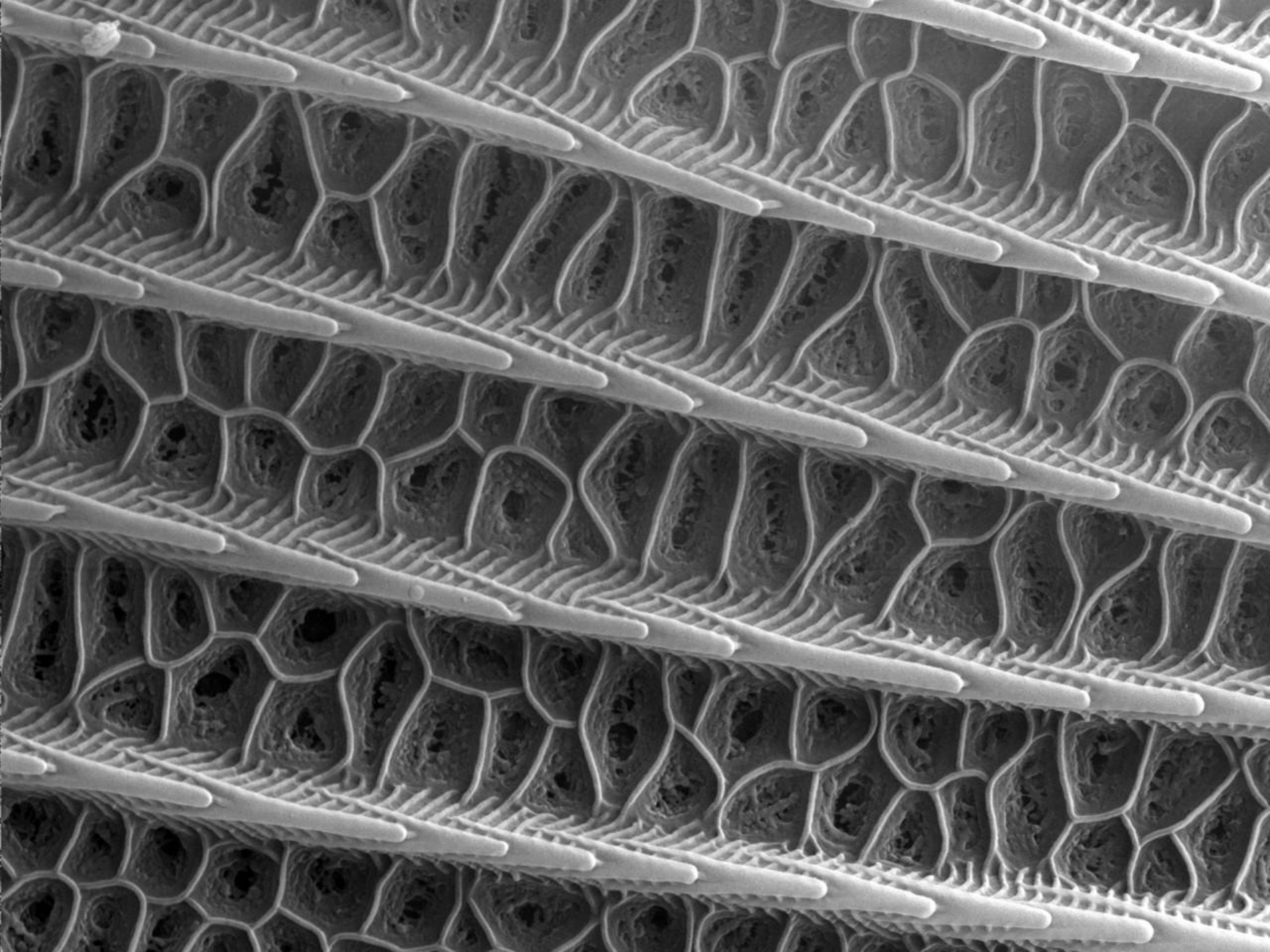
The Lifecycle of a Dataset

A Case Study

One example of how a dataset is managed


2009/12/19



What is data?

- Observational data
  - Sensor readings, images of the world as it is, survey data, telemetry (irreplaceable)
- Experimental data
  - Gene sequences, images of water flowing through a flume, chromatograms (reproducible, but expensive)
- Simulation data
  - Climate models (model the most important thing)
- Derived/compiled data
  - Compiled database (reproducible but expensive)

My Goals as a Researcher

- To organize my data
- To store and backup my data
- To preserve my data for the future
- To share my data with my colleagues

Getting Started

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

Oh great…

- My data is in “Sashimi Environmental Scanning Electron Microscope (ESEM)” format!
- NOW what do I do?

Bio-Formats to the rescue!

- Bio-Formats is a standalone Java library for reading and writing life sciences image file formats
- Will convert from Sashimi Format to OME-TIFF file format, an OPEN SOURCE FORMAT

First, Convert the Data (I found the conversion software here: http://www.loci.wisc.edu/software/bio-formats)

- I went from having a file named:
  - abcdefghijklmnopqrstuvwxyz.sam
- To having a file named:
  - abcdefghijklmnopqrstuvwxyz.tif

Process, Analyze & Weed

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

Let's clean up those file names

- abcdefghijklmnopqrstuvwxyz.tif doesn't make much sense, does it?
- How about:
  - sam_monarch_wing_24052011_ag_001.tif
  - (initials because working in a group)
- And put it in a project file called:
  - sam_monarch_wing_24052011

Why this structure?

- Oh, I just made it up! But I'm going to be *consistent.* **sam_monarch_wing_24052011_ag_001.tif**
  - sam = Sashimi Microscope format
  - monarch_wing
  - 24052011 = the day I did the experiment
  - ag = my initials
  - 001 = an accession number (I made that up, too, but I'll continue to use that schema)

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/remote, e.g. on an external hard drive*
  - 1 remote, e.g. on MIT’s TSM service or the “cloud”

  * CDs and DVDs aren’t built to last

But my images 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

sam_monarch_wing_24052011_ag_001.tif

- Save a copy of every iteration of a 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

- Then you can use encryption
- Don’t rely on 3rd party encryption alone
- Use something like PGP (Pretty Good Privacy)
  - Write the keys down on two pieces of paper
  - Store each piece of paper securely in separate locations

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

- The great thing about standards is that 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 your 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 Famous Metadata Standards

- FGDC (Federal Geographic Data Committee)
- DDI (Data Documentation Initiative)
- Dublin Core
- Darwin Core
- ABCD (Access to Biological Collections Data)
- AVMS (Astronomy Visualization Metadata Standard)
- CSDGM (Content Standard for Digital Geospatial Metadata)

Dublin Core

- When all else fails, use Dublin Core http://dublincore.org/
- A lot of repositories that store data use a variety of Dublin Core
- Why not Darwin Core, for biological diversity?
  - I'm not studying biological diversity
  - Darwin Core emphasizes taxonomy, which I don't care about
  - I'm not putting my data in a biodiversity database

Metadata for my directory could look like this…

- Directory: sam_monarch_wing_05242011
- Metadata for this directory
  - Creator: Anne Graham
  - Subject: monarch butterfly wing
  - Description: this directory contains Sashimi ESEM images of a monarch butterfly wing I took after finding a butterfly floating by the Charles River near MIT
  - Contributor: Kate McNeill helped me with these images
  - Date: 05/24/2012
  - Type: image
  - Original Format: Sashimi Microscope format (.sam)
  - Identifier: 000 schema
  - Relation: this is a directory that will contain multiple files
  - Coverage: By the Charles River in Cambridge, MA, MIT side
  - Rights: Monarch Butterfly Research Foundation (funder) owns the data (grant number: 00213)

Metadata for this Image

- Title: sam_monarch_wing_05242012_ag_001.tif
- Source: abcdefghijklmnopqrstuvwxyz.sam
- Relation: is a file in the directory: sam_monarch_wing_05242012

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

- Weed out obsolete data
- Decide what to keep for the 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

Reminder: ideally keep three copies of your data (local, local/remote, remote)

For storing your data:

- TSM (MIT IS&T)
- cloud storage is offered through private companies

Not for publishing or sharing.

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

Can be as simple as:

- Posting on a web site
- Sending via email upon request

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 open, 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)
- *Not all repositories are created equal! Some provide more full-service long-term preservation than do others.

Ideal Storage

- Put your data in a repository
  - Domain repository (such as GenBank)*
  - Institutional repository (such as DSpace@MIT)*

Advantages of a Repository

- 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

Advantages of a Domain Repository*

- 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), ICPSR (for numeric social science data)

Advantages of an Institutional Repository, e.g., DSpace@MIT

- Linked to your institution
- You can put all your datasets together
- University guarantees support of Institutional Repositories
  - Some Domain Repositories may “go out of business” once their funding ends

Putting our Monarch Wing Scan Data in DSpace@MIT

- Will be co-located with our other materials
- TIFF format is supported: http://libraries.mit.edu/dspace-mit/build/policies/format.html
- No alternative domain repository
- Must have faculty sponsorship
- Each file cannot exceed 2.5 GB
- Contact data-management@mit.edu for more information

Intellectual Property Issues

- Data is not copyrightable, but an expression of data such as a table can be.
- MIT 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 to which the MIT Libraries subscribe are licensed and carry restrictions on use, but many do allow for educational and research use, which allows for sharing limited portions of data.

Citing Data (Yours or Others')

- You can cite the dataset
  - Subject archive entry (e.g. Genbank accession number, e.g. NP_002700)
  - Rujirapisit,P, (Eleocharisdulcis Trin.),Chemical Composition and Physico-Chemical Properties of Chinese Water Chestnut (Eleocharisdulcis Trin.) Flour and Starch, http://hdl.handle.net/10527/10471
- You can cite the publication that describes the data
  - Wilson MD (1988) The MRC Psycholinguistic Database. Behavioural Research Methods. 20 (1) 6-11.

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?

Want to learn more?

- In the Libraries' IAP workshops on Managing Your Information:
- **Research Data Management: File Organization**
  Tue Jan 29, 10-11:00am, 14N-132
- **Research Data Management: Versioning**
  Tue Jan 29, 11:00am-12:00pm, 14N-132

Questions?

http://libraries.mit.edu/data-management

data-management@mit.edu