

# The Lifecycle of a Dataset

# A Case Study

One example of how a  
dataset evolves

# What is data?

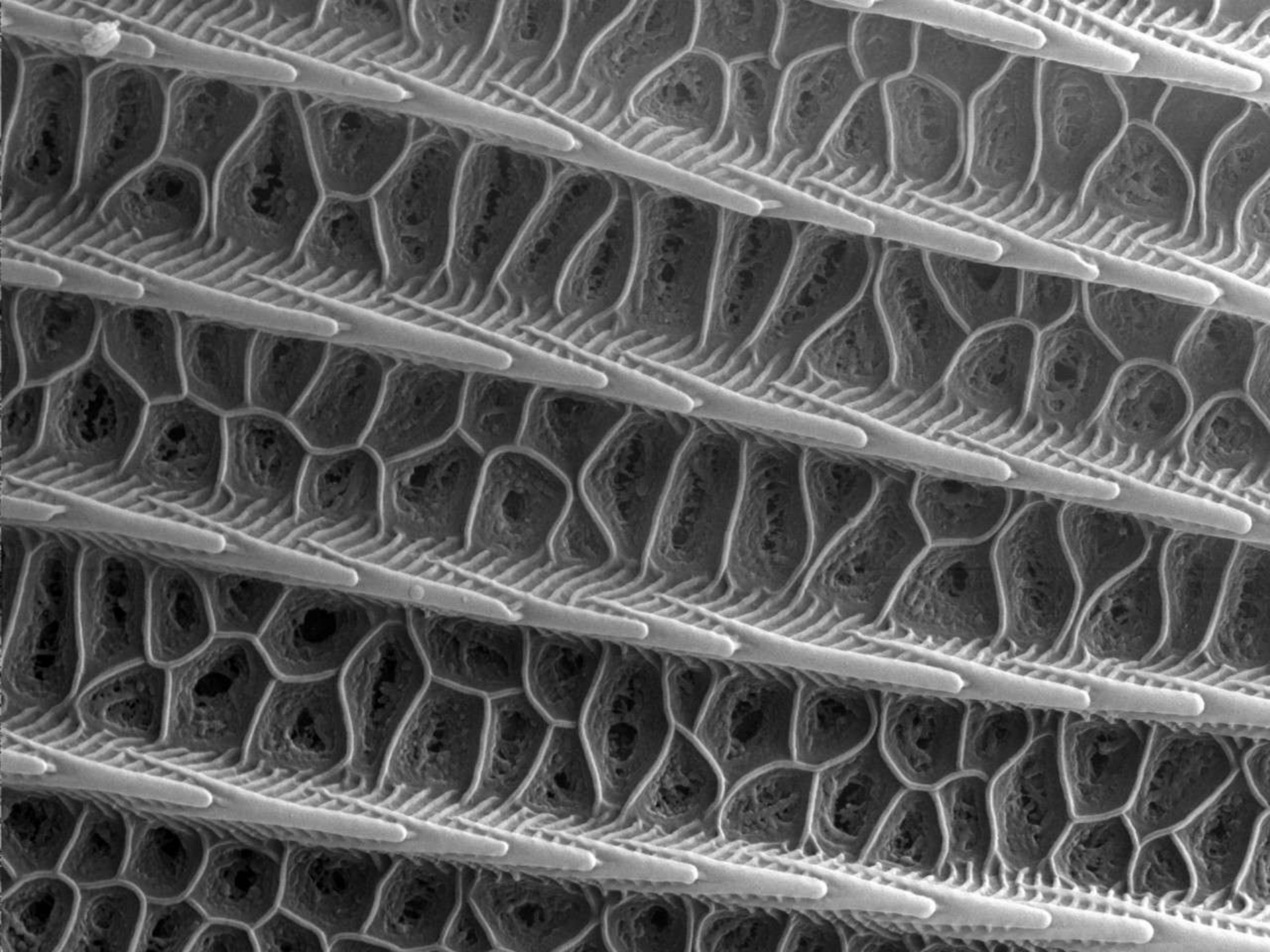
- Observational data
  - Sensor readings, telemetry (usually irreplaceable)
- **Experimental data**
  - Gene sequences, chromatograms (reproducible, but expensive)
- Simulation data
  - Climate models (model the most important thing)
- Derived/compiled data
  - Compiled database (reproducible but expensive)

# 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 to end up
- 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
- <http://www.loci.wisc.edu/software/bio-formats>

# First, Convert the Data (this will require software)

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

# Process, Analyze & Weed

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

# Let's clean up those file names

- abcdefghijklmnopqrstuvwxyz.tiff doesn't make much sense, does it?
- How about:
  - sam\_monarch\_wing\_24052011\_as\_001.tiff
  - (I put my initials because I am working in a group)
- And put it in a directory called:
  - sam\_monarch\_wing\_24052011

# Why this structure?

- Oh, I just made it up! But I'm going to be *consistent*
  - sam = Sashimi Microscope format
  - monarch\_wing
  - 24052011 = the day I did the experiment
  - as = 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
- Use open source encryption software
- Document the version of compression software you used

# Versioning

- Save a copy of every iteration of a data file
- Follow a file naming convention
- Consider using software such as Apache Subversion

# My Research is Top-secret

- Then you can use encryption
- Don't rely on 3<sup>rd</sup> 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

# Decide What your Final Datasets Are

- Once your project is over, weed out obsolete data and decide what you want to keep for the long-term

# Add Metadata

- Why does metadata even matter?
  - Metadata, or “data about data” is a surrogate for your dataset that 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

# 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
- 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 this directory could look like this...

- Directory: sam\_monarch\_wing\_05242011
- Metadata for this directory
  - Creator: Amy Stout
  - 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: Anne Graham helped me with these images
  - Date: 05/24/2011
  - Type: image
  - 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\_05242011\_as\_001.tiff
- Source: abcdefghijklmnopqrstuvwxyz.sam
- Relation: is a file in the directory:  
sam\_monarch\_wing\_05242011

# Where do I put metadata?

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

# Where can I archive my data?

- What do you mean by “archive?”
  - Store?
  - Preserve?
  - Keep in perpetuity?
  - Migrate to newer formats?
  - Provide an emulator?
  - Back up?

# By Archive I mean...

A place to put your data where it will be:

- Stored
- Backed up
- Discoverable
- Preserved for the future (as much as possible)

# Archiving/Publishing/Sharing your Data

- TSM or cloud storage is for storing your data, not publishing or sharing your data
- That may be sufficient for you
- However...

# 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:
  - Putting on a web site
  - Sending via email upon request

# Ideal Storage

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

\*Not all repositories are created equal!

# Advantages of a Repository

- Provides a metadata structure for you to fill in
- Publishes the data for you by giving your dataset a unique identifier, e.g. DOI
- 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 tools for people to use with your data

# Advantages of a Domain Repository, e.g. GenBank

- Your data will be stored with similar datasets
- Researchers will find your data easily
- The repository will understand what your data needs in terms of storage, archiving and preservation
- Computational tools may be developed to crunch a critical mass of data of a certain kind

# 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

# The Unique Identifier

- PURL: <http://purl.org>
- DOI: <http://www.doi.org>
- Accession: <http://www.ncbi.nlm.nih.gov>
- InChI: <http://www.iuipac.org/inchi>
- URI: <http://www.ietf.org/rfc/rfc2396.txt>
- Handle: <http://dspace.mit.edu>

# Digital Object Identifier (DOI)

- In common use
- Easy to get
- Many repositories are equipped to issue them
- Will always resolve to the correct location
  - As long as you keep the repository up-to-date

*Hazard/Risk Assessment***USE OF BUTTERFLIES AS NONTARGET INSECT TEST SPECIES AND THE ACUTE TOXICITY AND HAZARD OF MOSQUITO CONTROL INSECTICIDES**THAM C. HOANG,<sup>†</sup> RACHEL L. PRYOR,<sup>†</sup> GARY M. RAND,<sup>\*†</sup> and ROBERT A. FRAKES<sup>‡</sup><sup>†</sup>Ecotoxicology and Risk Assessment Laboratory, Florida International University, North Miami, Florida, USA<sup>‡</sup>U.S. Fish and Wildlife Service, Vero Beach, Florida*(Submitted 25 March 2010; Returned for Revision 3 June 2010; Accepted 5 November 2010)*

**Abstract**—Honeybees are the standard insect test species used for toxicity testing of pesticides on nontarget insects for the U.S. Environmental Protection Agency (U.S. EPA) under the Federal Insecticide Fungicide and Rodenticide Act (FIFRA). Butterflies are another important insect order and a valued ecological resource in pollination. The current study conducted acute toxicity tests with naled, permethrin, and dichlorvos on fifth larval instar (caterpillars) and adults of different native Florida, USA, butterfly species to determine median lethal doses (24-h LD<sub>50</sub>), because limited acute toxicity data are available with this major insect group. Thorax- and wing-only applications of each insecticide were conducted. Based on LD<sub>50</sub>s, thorax and wing application exposures were acutely toxic to both caterpillars and adults. Permethrin was the most acutely toxic insecticide after thorax exposure to fifth instars and adult butterflies. However, no generalization on acute toxicity (sensitivity) of the insecticides could be concluded based on exposures to fifth instars versus adult butterflies or on thorax versus wing exposures of adult butterflies. A comparison of LD<sub>50</sub>s of the butterflies from this study (caterpillars and adults) with honeybee LD<sub>50</sub>s for the adult mosquito insecticides on a  $\mu\text{g}/\text{organism}$  or  $\mu\text{g}/\text{g}$  basis indicates that

# Crossref – DOI issuer

## DOI Resolver

If you encounter a DOI string (e.g., [10.1037/0003-066X.59.1.29](http://dx.doi.org/10.1037/0003-066X.59.1.29)) that is not hyperlinked, you can enter it in the box below:

TIP: You can turn a DOI string into a URL by appending the DOI string to <http://dx.doi.org/>

### **Want to look up a DOI? Visit our [Guest Query](#) form.**

CrossRef is an independent membership association, founded and directed by publishers. CrossRef's mandate is to connect users to primary research content, by enabling publishers to work collectively. CrossRef is also the official DOI® link registration agency for scholarly and professional publications. Our citation-linking network today covers tens of millions of articles and other content items from several hundred scholarly and professional publishers.

# Preserving your Data

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

# What does preservation mean?

Usually means that every effort  
will be made to make the data  
usable in the future

# Things to consider when preserving your data...

- Provide thorough documentation
- Preserve a copy of the software used to create, manipulate and render your data
- Migrate your data to contemporary formats as popular formats change (a good archive will do this for you)
- Update the software you store to contemporary formats

# 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 Open Source, Standard formats**

# Putting my Monarch Wing Scan Data in DSpace@MIT

- Must have faculty sponsorship
- Each file cannot exceed 2.5 GB
- Contact [data-management@mit.edu](mailto:data-management@mit.edu) for more information

# Citing Data

- You can cite the publication that describes the data
  - Wilson MD (1988) The MRC Psycholinguistic Database. Behavioural Research Methods. 20 (1) 6-11.
- You can cite the dataset
  - Subject archive entry (e.g. Genbank accession number, e.g. NP\_002700)
  - Kessler, Ronald C. Detroit Area Study, 1985: Life Events in Everyday Experience [Computer file]. ICPSR06414-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1995.  
doi:10.3886/ICPSR06414

# Measure Twice, Cut Once

- Always double check over time:
  - Is the data still what I think it is? (use checksums)
  - Is the metadata still available?
  - Are the formats still usable?
  - Is the software still available?
  - Is there anything left to run it on?
  - Is the data still in the correct location?

# Intellectual Property Issues

- MIT or the funder may own your data (consult with the Technology Licensing Office)
- Data is not copyrightable, but an incarnation of data such as a table is copyrightable.
- You can license data to limit what others can do with it
  - It's incumbent upon you to police usage of your data
- You can share your data if you, in fact, own it
- 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

[http://libraries.mit.edu/data-  
management](http://libraries.mit.edu/data-management)

[data-management@mit.edu](mailto:data-management@mit.edu)

