# Research Data Management 101: The Lifecycle of a Dataset

MIT Libraries

# Data Management Services @ MIT Libraries



- Workshops
- Web guide: http://libraries.mit.edu/data-management
- Individual consultations
  - includes help with creating data management plans
  - guide you to related services across MIT

MITLibraries

# Today's session: topics



Data management planning

File organization

Funder and publisher requirements

Versioning

File formats

Metadata

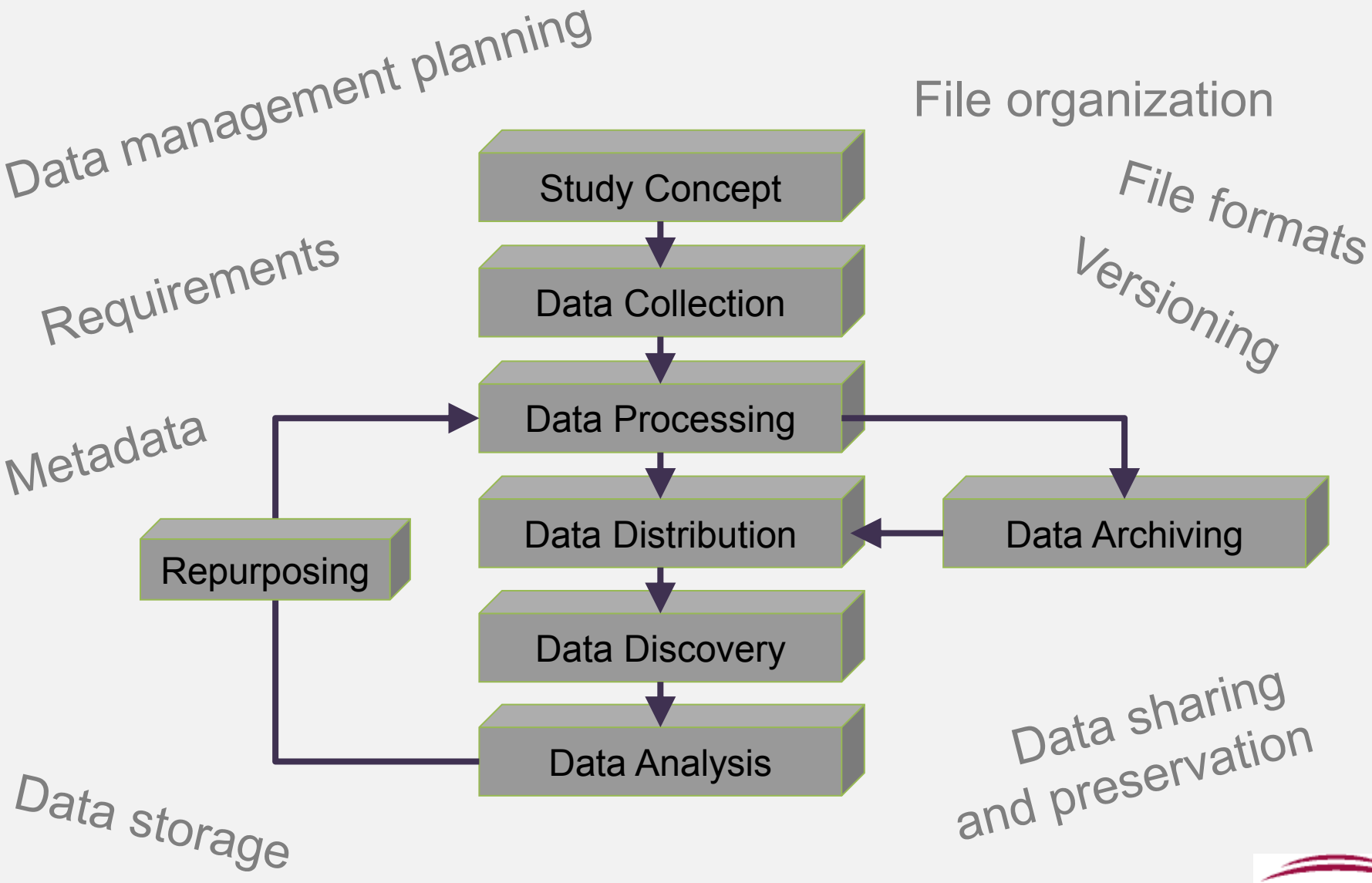
Data sharing and preservation

Data storage

MITLibraries

# Today's session: the research data lifecycle



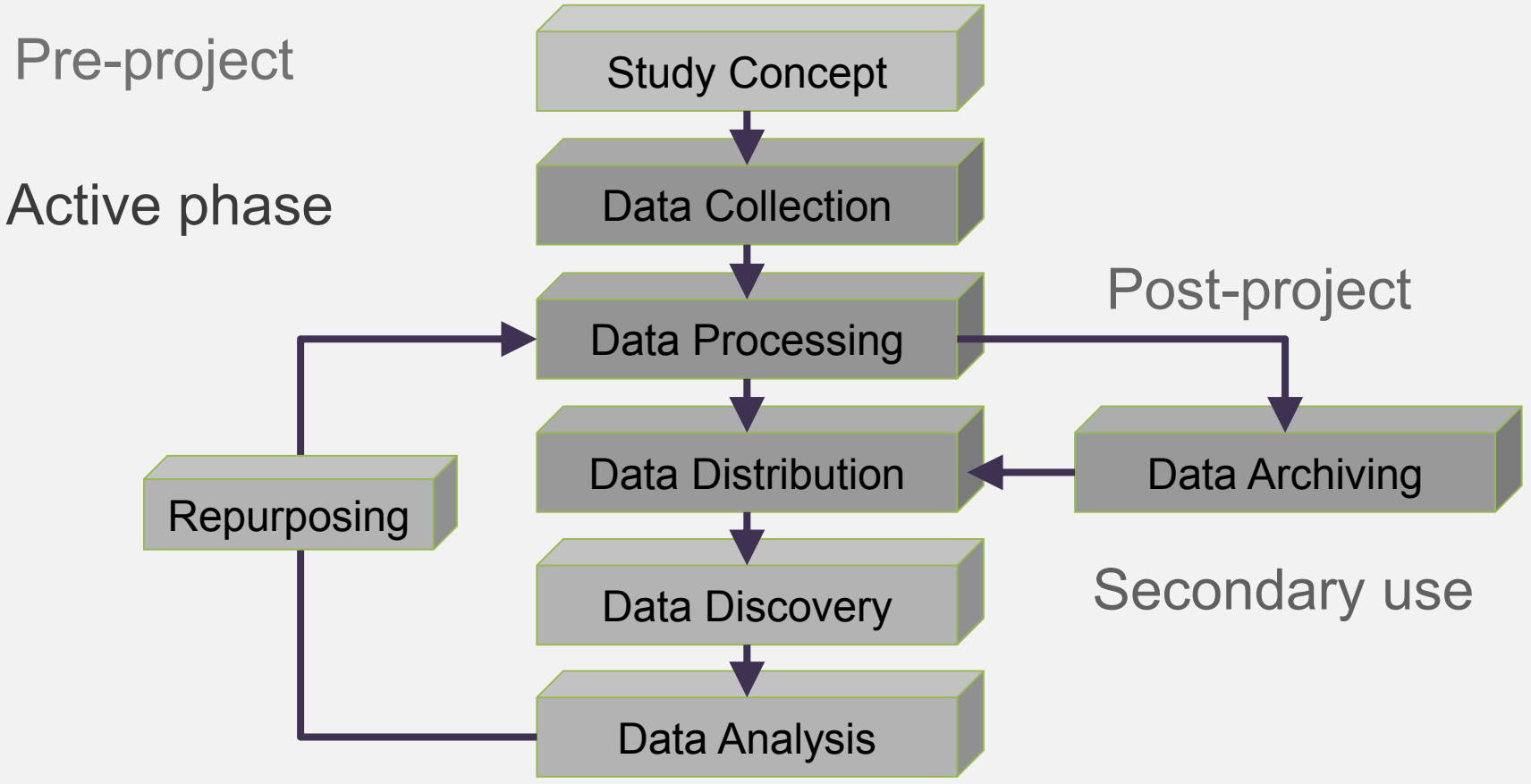



Lifecycle adapted from DDI version 3.0 Combined Life Cycle Model

MITLibraries

# Research data lifecycle phases






Lifecycle adapted from DDI version 3.0 Combined Life Cycle Model

# What do we mean by ***data***?



| General | Social Sciences | Natural/Physical Sciences |
|---|---|---|
| • images<br>• video<br>• mapping/GIS data<br>• numerical measurements | • survey responses<br>• focus group and individual interview transcripts<br>• economic indicators<br>• demographics<br>• opinion polling | • measurements generated by sensors/laboratory instruments<br>• computer modeling<br>• simulations<br>• observations and/or field studies<br>• specimen |

# A case study



Photo Courtesy of William Warby on Flickr. License CC-BY.

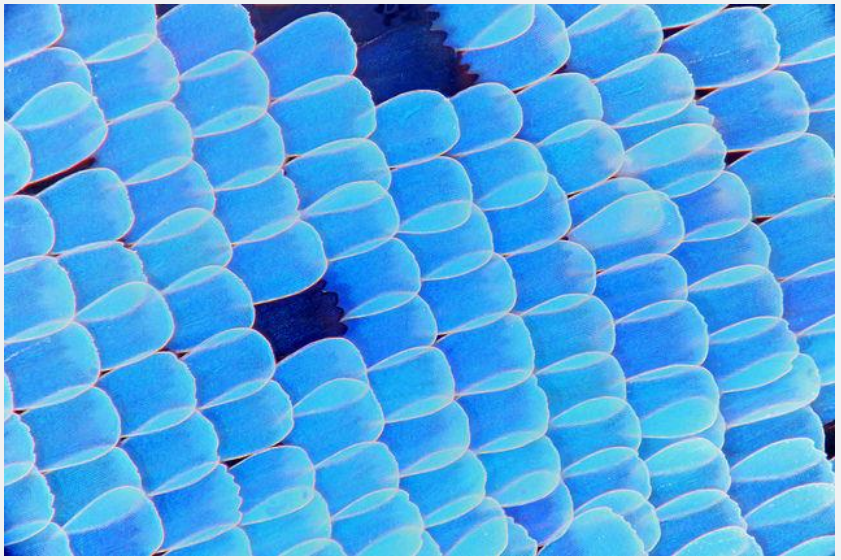
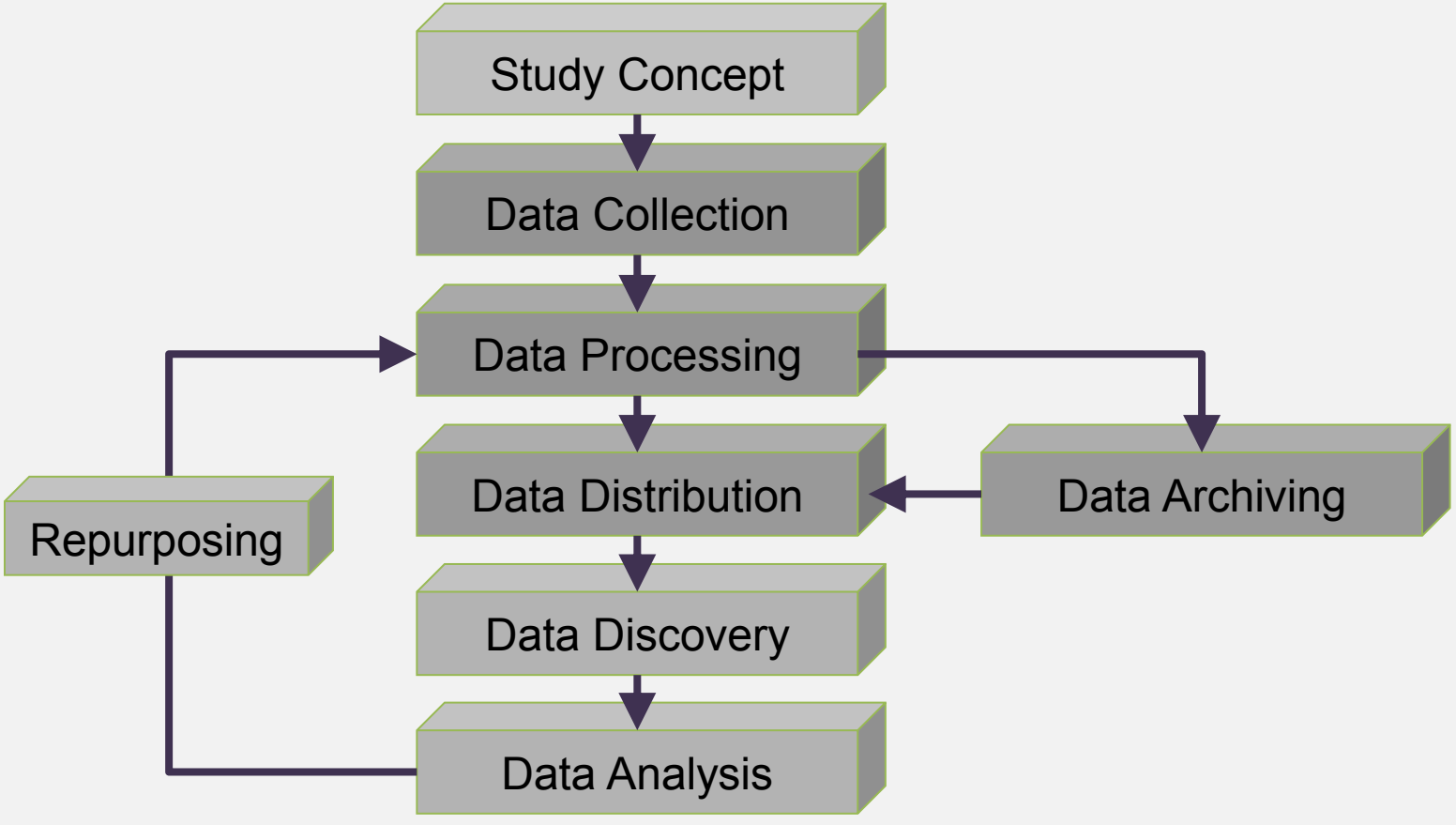


Photo Courtesy of Macroscopic Solutions on Flickr. License CC-BY.

# Pre-project: planning






Lifecycle adapted from DDI version 3.0 Combined Life Cycle Model

# Where do you start?



## 1. Consider your goals.

What do you want to get out of managing your data?

- Part of ongoing research
- Compliance with funder or publisher requirements:
- Dissemination for others' use

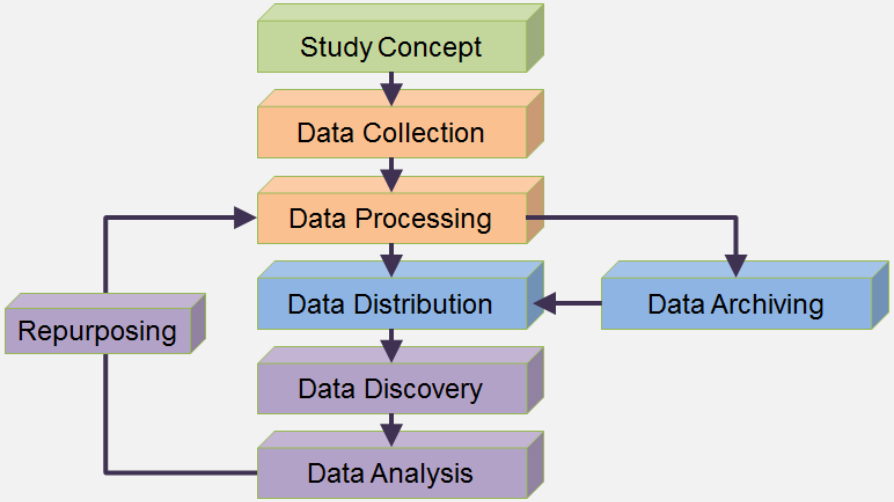
# Where do you start?



1. Consider your goals.

2. What are you collecting?

3. What are you keeping?



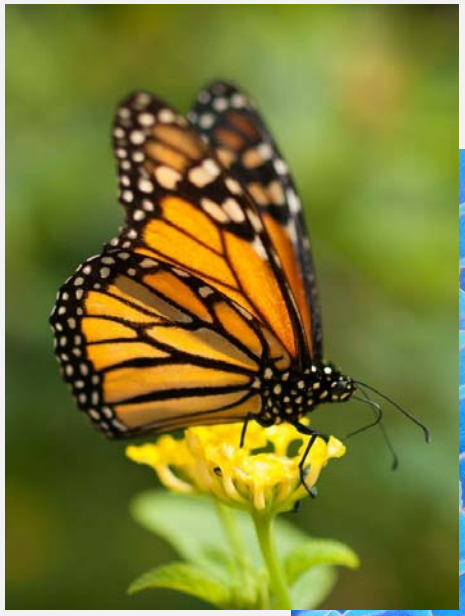



Photo Courtesy of William Warby on Flickr. License CC- BY.

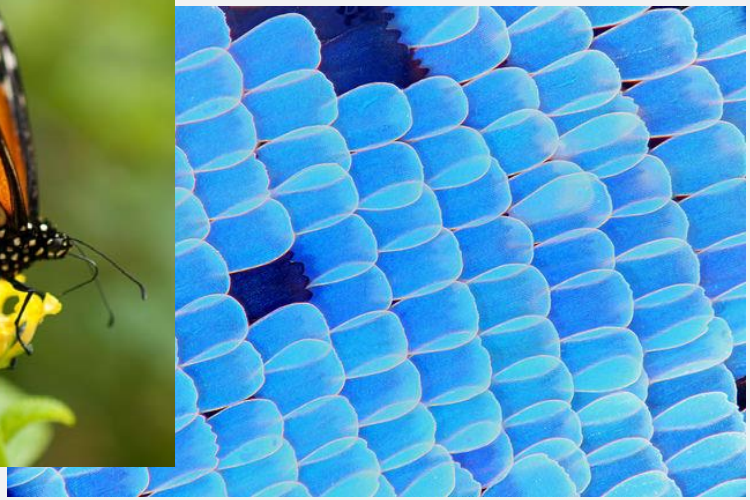


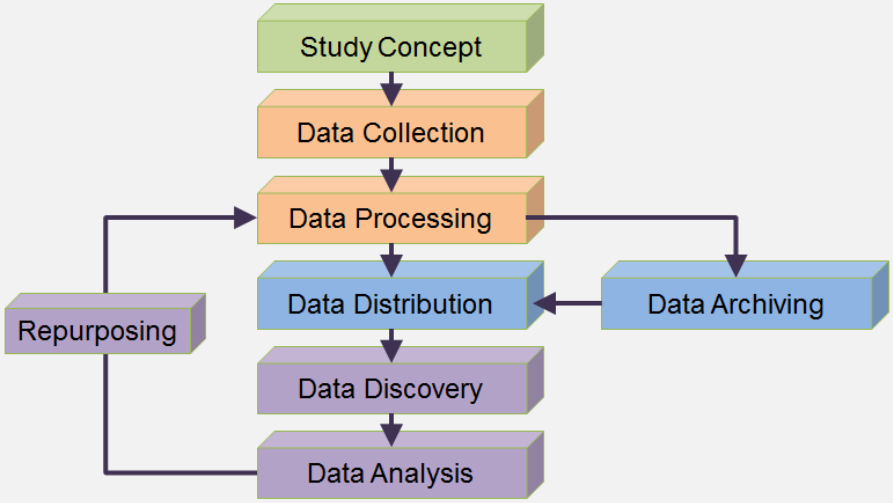
Photo Courtesy of Macroscopic Solutions on Flickr. License CC- BY.

MITLibraries

# Where do you start?



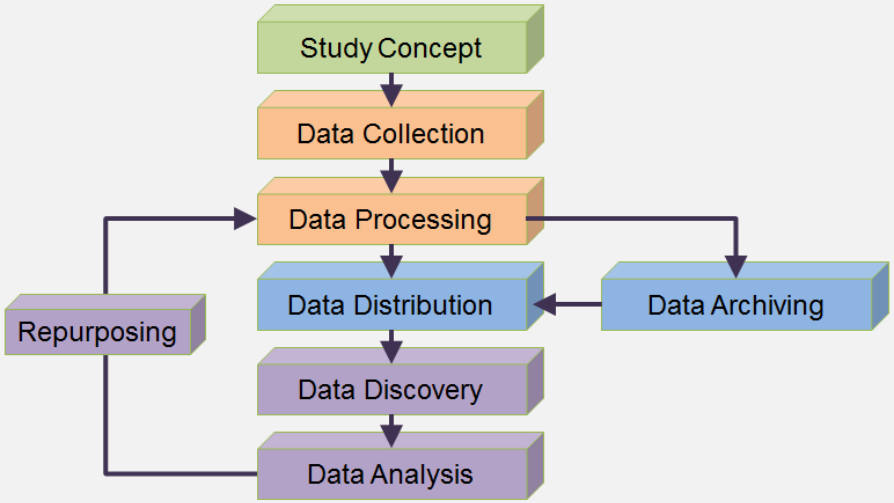
1. Consider your goals.
2. What are you collecting?
3. What are you keeping?
4. **Where do you want to keep it?**



# Where do you start?



1. Consider your goals.
2. What are you collecting?
3. What are you keeping?
4. Where do you want to keep it?
5. **What do you need to be able to use it & share it later?**



# Funder requirements



- Funding agencies increasingly requiring people to share outputs of research, including data (and publications)
- Their motivation: extend the impact of their research farther
- What funders require data sharing?
  - Most federal agencies and many private funders as well
- What do funders require?
  - Data management plan (DMP) submitted with grant application
  - Can request funds for data management in the grant
  - Sharing of final data produced
- See: http://libraries.mit.edu/oarequirements
- In our case study, my research is funded by NSF; I have to:
  - Share *all* of the data generated in my project
  - Include a DMP in my grant application

# What can help me write my DMP? DMPTool!



- *New* tool to help you create your data management plan
- Guides you through writing a DMP specific to a funder's requirements using templates


MITLibraries

# Data Documentation (AKA Metadata)

# Metadata: why does it matter?



Data is not self-describing.

Metadata, or **"data about data"** explains your dataset and allows you to document important information for:

- **Finding** the data later
- **Understanding** what the data is later
- **Sharing** the data (both with collaborators and future secondary data users)
- Consider it an **investment** of time that will save you trouble later several-fold

# Metadata standards



**Examples:**

- 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)

**Advantages:**

- Ensure you have a complete, standard set of information about each part of your data
- Enable your dataset to be organized with other datasets

# Metadata



**Do what works for you!**

Document and describe your data
in whatever way works for you.

Better “**good enough**” than doing nothing

# Metadata: our case study



## Possible metadata options:

1. **Dublin Core**
   - General metadata standard
   - Widely applicable
   - Used in many different repositories

2. **Darwin Core**
   - For biological diversity
   - Emphasizes taxonomy, *which I don't care about*
   - Frequently used in biodiversity databases

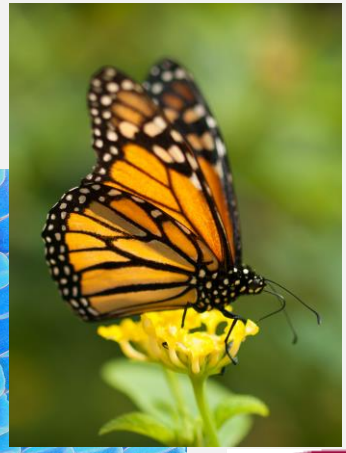


Photo Courtesy of William Warby on Flickr. License CC-BY.

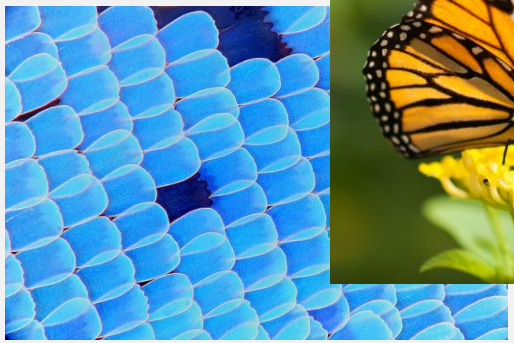


Photo Courtesy of Macroscopic Solutions on Flickr. License CC- BY.

MITLibraries

# Metadata: our case study



**Our directory:** sam_monarch_wing_20150415

**Metadata for this directory:**

- *Creator:* Katherine McNeill
- *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:* Mark Clemente helped me with these images
- *Date:* 20151015
- *Original Format:* Sashimi Microscope format (.sam)
- *Relation:* this is a directory that will contain multiple files
- *Type:* image
- *Coverage:* By the Charles River in Cambridge, MA, MIT side
- *Rights:* Monarch Butterfly Research Foundation (funder) owns the data (grant number: 00213)

# Metadata: our case study



**Metadata for this image:**

- *Title:*

  sam_monarch_wing_20150415_CM_001.tif

- *Source:*

  abcdefghijklmnopqrstuvwxyz.sam

- *Relation:*

  is a file in the directory: sam_monarch_wing_20150415

# Metadata: capturing it



In a filename

In a readme file

In a spreadsheet

In an XML file

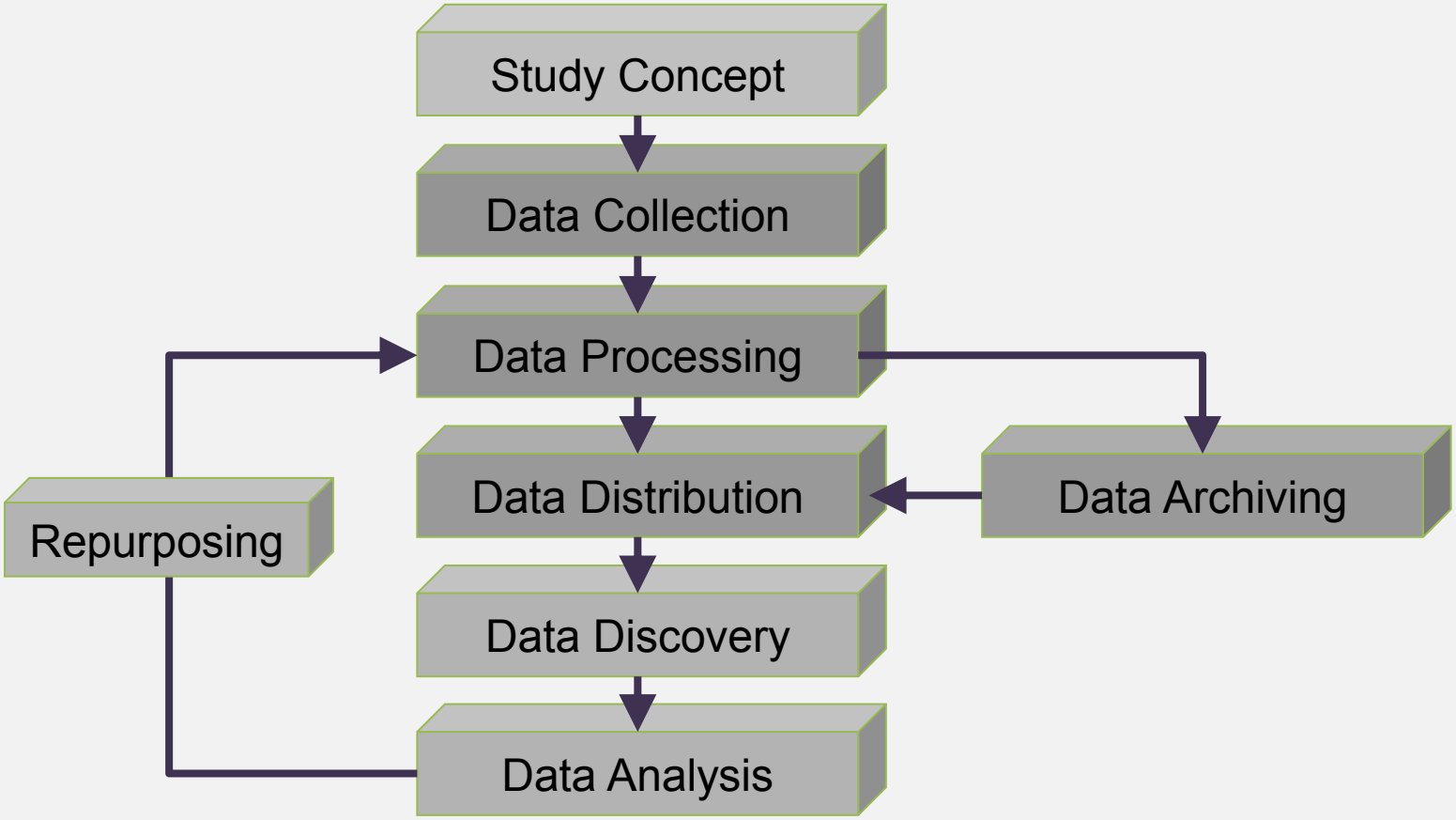
Into a database

When choosing how to capture metadata, consider:

- Expertise at your disposal
- Complexity of your project
- Collaborators
- Your own comfort level

# Active phase of research






Lifecycle adapted from DDI version 3.0 Combined Life Cycle Model

# Storing data



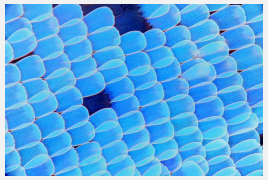
## What are your goals?

- To find a place to store your data?
- To appropriately back up your data so that it's not lost?

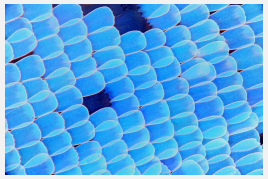
# Backing up data: storing during active phase



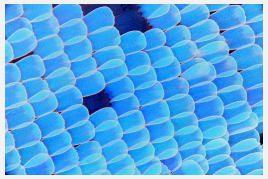
Ideally, **keep three copies** of your data:



1. local/working copy
   e.g. on your workstation or in shared workspace



2. remote copy
   e.g. on a managed backup system



3. other copy in another remote location, or local/external
   e.g., external hard drive* (CDs and DVDs not built to last)



Test file recovery at setup and on a regular schedule
This is not for publishing or sharing (storage ≠ archiving)

# Backing up data: storing during active phase



## Additional considerations:

- **What to keep?**
  Consider weeding obsolete data as you go

- **Huge file size!**
  Images can often be very large files
  - Image compression may be done by your image software automatically
  - If not, you can compress your data.
    - But always keep one uncompressed copy somewhere
    - Use open source compression software
    - Document the version of compression software used

# Backing up data: storing during active phase



## Additional considerations:

- **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

# File Organization
# &
# File Formats

# File organization: naming conventions



Naming conventions make life easier!

Naming conventions should be:

- **Descriptive**
- Consistent

Consider including:

- Unique identifier (ie. Project Name or Grant # in folder name)
- Project or research data name
- Conditions (Lab instrument, Solvent, Temperature, etc.)
- Run of experiment (sequential)
- Date (in file properties too)
- Version #

# File organization: naming conventions



Naming conventions make life easier!

Naming conventions should be:

- Descriptive
- **Consistent**

YYYYMMDD
MMDDYYYY
YYMMDD
MMDDYY
MMDD
DDMM

**Maintain order**

TimeDate
DateProjectID
TimeProjectID

Sample001234
Sample01234
Sample1234

**Include the same information**

# File organization: naming conventions



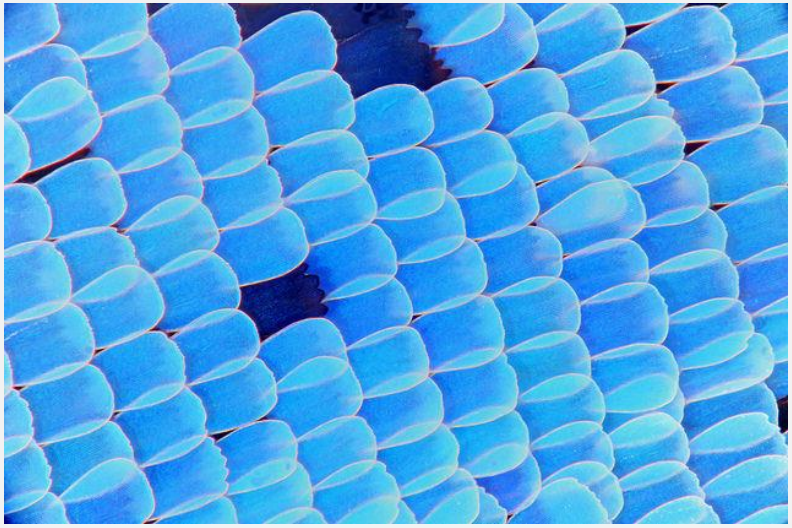
| Best Practice | Example |
|---|---|
| **Limit the file name to 32 characters** (preferably less!) | 32CharactersLooksExactlyLikeThis.csv |
| When using sequential numbering, **use leading zeros** to allow for multi-digit versions<br>For a sequence of 1-10: 01-10<br>For a sequence of 1-100: 001-010-100 | **NO** ProjID_1.csv ProjID_12.csv<br>**YES** ProjID_01.csv ProjID_12.csv |
| **Don't use special characters**<br>& , * % # ; * ( ) ! @$ ^ ~ ' { } [ ] ? < > - | **NO** name&date@location.doc |
| **Use only one period** and use it before the file extension | **NO** name.date.doc<br>**NO** name_date..doc<br>**YES** name_date.doc |
| **Avoid using generic data file names** that may conflict when moved from one location to another | **NO** MyData.csv<br>**YES** ProjID_date.csv |

MITLibraries

# File organization: naming conventions



For our case study:





Maybe started with:

**abcdefghijklmnopqrstuvwxyz.sam**

Sashimi Microscope format

Date as YYYYMMDD

Ascension # because part of a series

**sam_monarch_wing_20150415_CM_001.tif**

File format

Descriptive element

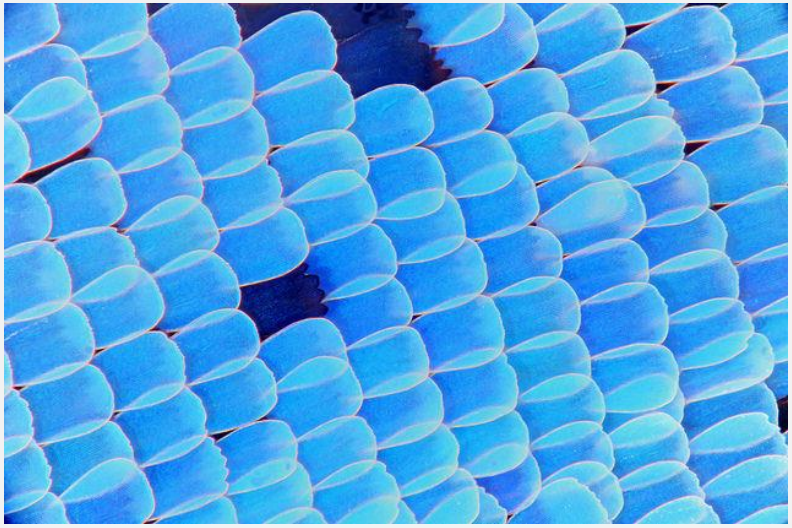
Initials because working in a group

MITLibraries

# File organization: file structure



For our case study:





**sam_monarch_wing_20150415_CM_001.tif**

↓

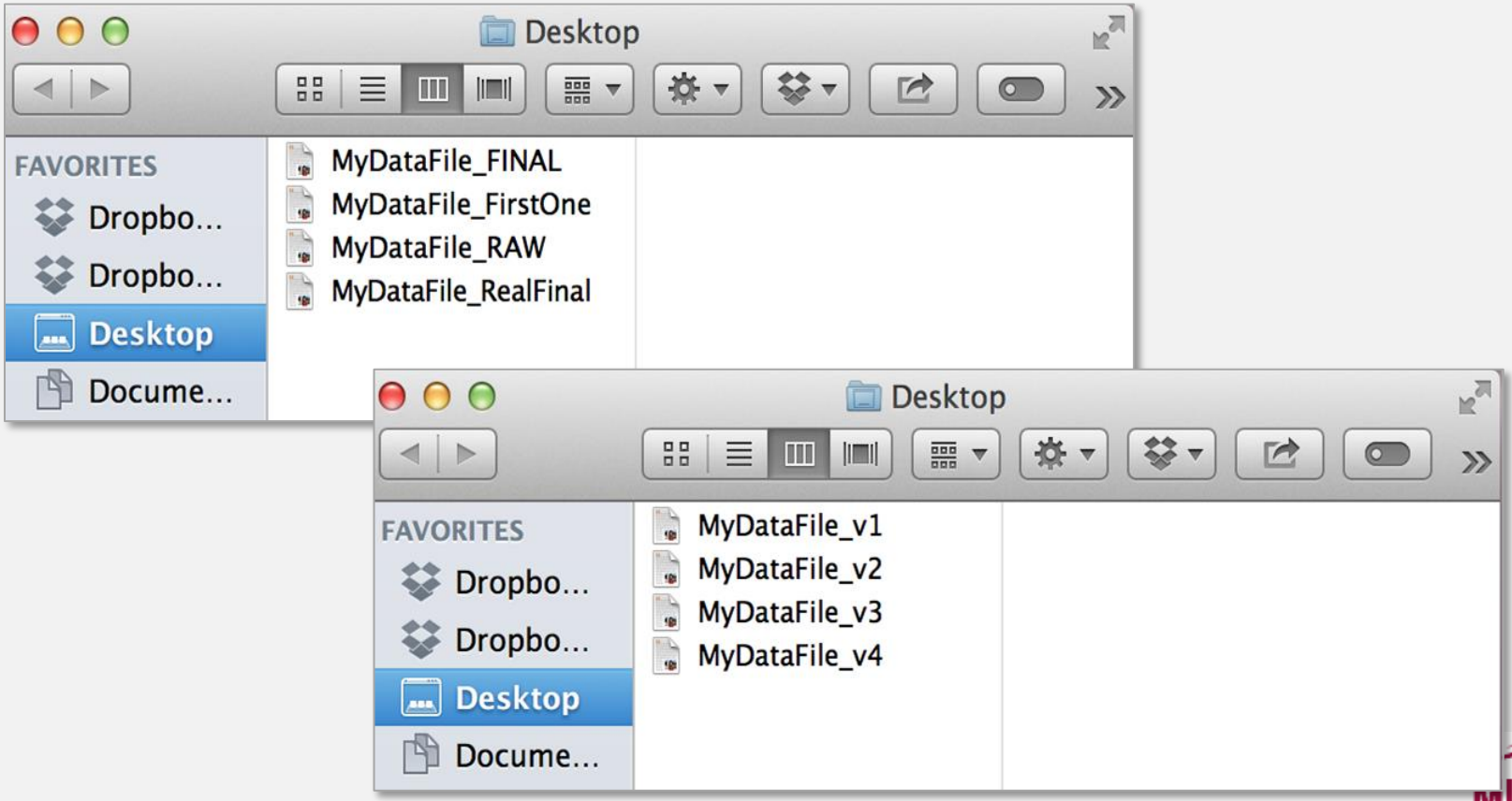
**sam_monarch_wing_20150415**

# File organization: versioning



Track versions of either:

- Analysis/program/script files, while keep the original version of the data file the same; OR
- Data files themselves



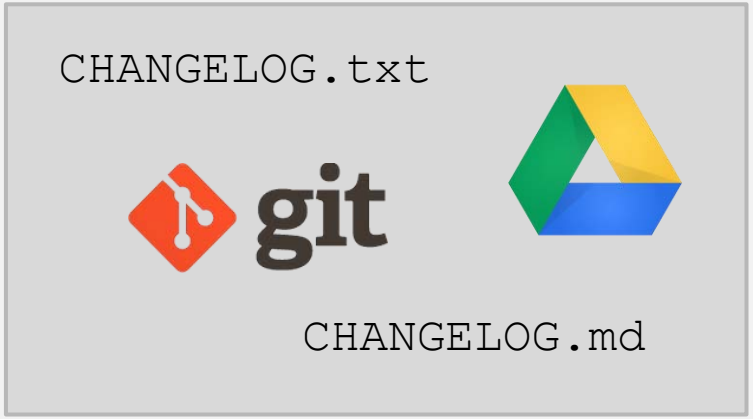
# File organization: versioning



In some cases, it may make sense to log the changes so that you can quickly assess and access the versions.

It's good to document:

- What was changed?
- Who is responsible?
- When did it happen?
- Why?



# File formats



In the best case, your data files are both:

- **Non-proprietary (also known as *open*)**, and
- Unencrypted and uncompressed

# File formats: preferred examples

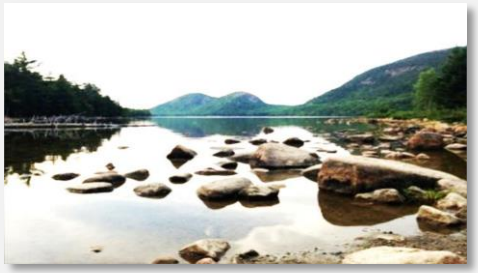


| Proprietary Format | Alternative/Preferred Format |
|---|---|
| Excel (.xls, .xlsx) | Comma Separated Values (.csv)<br>ASCII |
| Word (.doc, .docx) | plain text (.txt), XML, PDF/A, HTML, ODF<br><br>or if formatting is needed, PDF/A (.pdf) |
| PowerPoint (.ppt, .pptx) | PDF/A (.pdf), ODP, JPEG 2000, PDF, PNG |
| Photoshop (.psd) | TIFF (.tif, .tiff), |
| Quicktime (.mov) | MPEG-4 (.mp4), MOV, AVI, MXF |
| Sounds | WAVE, AIFF |
| Containers | TAR, GZIP, ZIP |
| Databases | XML, CSV |

# File formats: converting considerations







Photos courtesy of Christine Malinowski, used with permission.

**Information can be lost when converting file formats.**

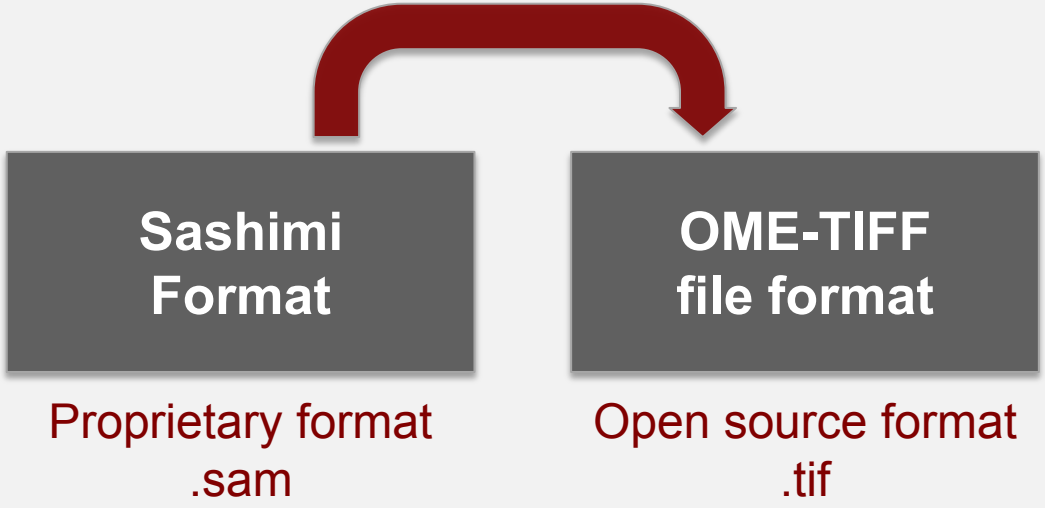
To mitigate the risk of lost information when converting:

- Note the conversion steps you take
- If possible, keep the original file as well as the converted ones

# File formats: our case study



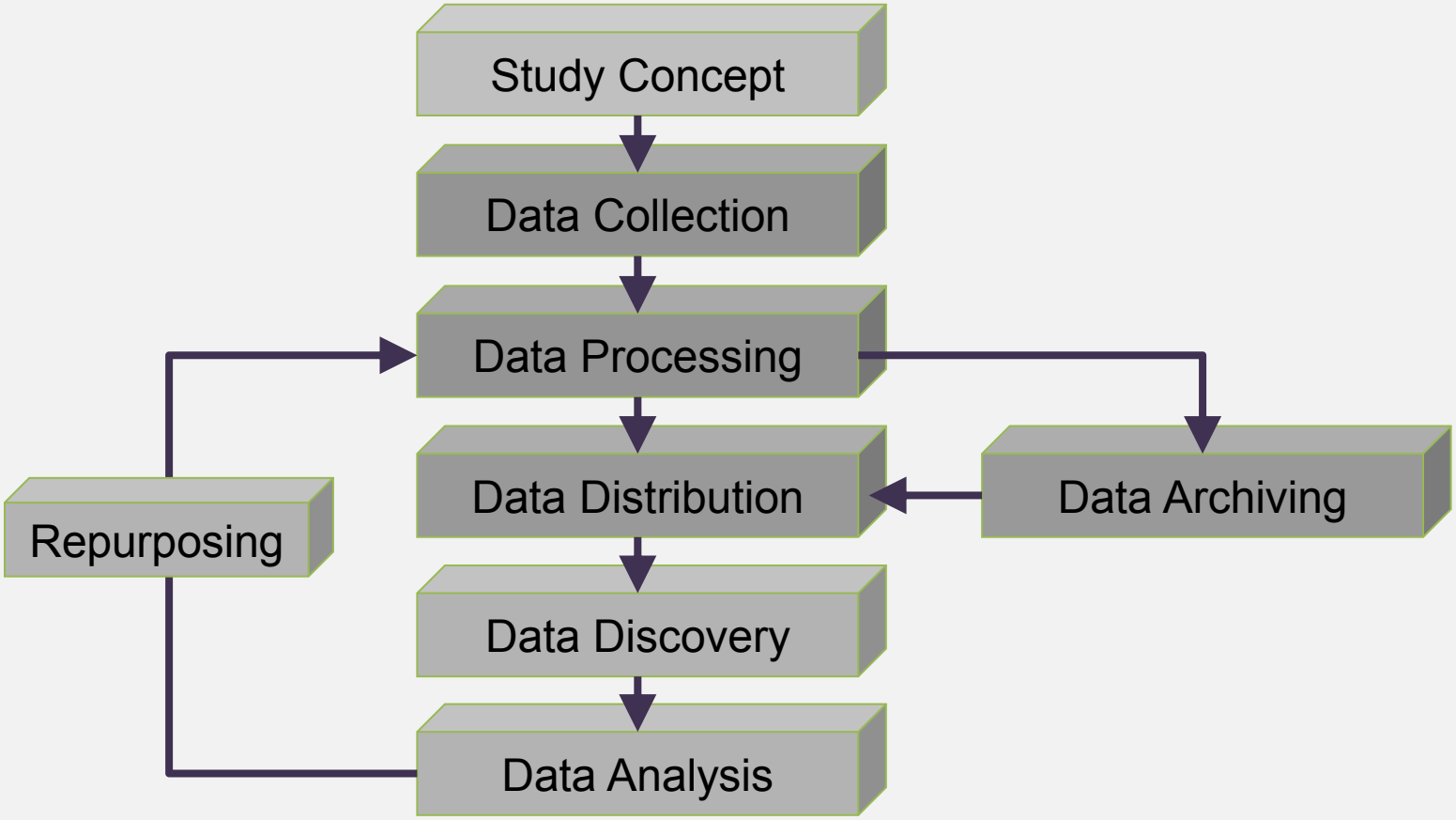
- My data is in “Sashimi Environmental Scanning Electron Microscope (ESEM)” format!
  - **Problem:** This is a proprietary format. How can I convert it to a more open-non proprietary format that will be usable in the future?
  - **Solution:** I found Bio-Formats, a standalone Java library for reading and writing life sciences image file formats into alternative open formats.



# Sharing
# &
# Preserving Data

# Sharing and preserving data






Lifecycle adapted from DDI version 3.0 Combined Life Cycle Model

# Sharing data: why?



- Further science as a whole
- Further your research/reputation
- Enable new discoveries with your data
- Comply with funder/publisher data sharing requirements

**Share Informally:**
Posting on a web site, sending via email upon request

**Share Formally:**
Via a repository, which may also provide preservation and makes your data more accessible and citable

# 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?

# Preserving your data



Preservation means storing your data in a place 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

# Storing data: repositories



**Overall advantages:**

- May preserve your data for the future
- Provides a metadata structure
- Serves as a backup vehicle for your data
- Makes sharing your data accessible and citable by others
- May provide some computational/online analysis tools for people to use your data
- Gives your dataset a unique persistent identifier, e.g., DOI

**As a researcher, you still need to:**

- Keep thorough documentation
- Keep at least one copy of your data in an open, non-proprietary format

# Storing data: repository options



## 1. Domain repositories

**Advantages:**

- Data is stored with similar datasets (by subject, format, or both)
- Other 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

**Examples:**

- GenBank (for genome data)
- ICPSR (for numeric social science data)

**re3data.org | Registry of Research Data Repositories**

# Storing data: repository options



## 2. Institutional repositories

**Advantages:**

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

**Example:**

- DSpace@MIT

# Storing data: our case study



**Repository decision:**

Put our Monarch Wing Scan Data in DSpace@MIT

**Reasoning:**

- Will be co-located with our other materials
- TIFF format is supported
- No alternative domain repository

**Additional considerations:**

- Must have faculty sponsorship
- Each file cannot exceed 2.5 GB

# Sharing Data: additional considerations



## Intellectual Property & Confidentiality

- Data is not copyrightable, but an expression of data 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


- Look at Creative Commons Licenses, including the CC0 Declaration to emphatically put it in the public domain
- Confidential and sensitive data requires additional consideration and planning

# Sharing Data: additional considerations



## Journal requirements

- Many journals require that underlying data accompany published articles, usually found in “instructions for authors”
- More resources at http://libraries.mit.edu/data-management/share/journal-requirements/

## Using other people’s data

- 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.

# Measure Twice, Cut Once



Research data management is an ongoing, but beneficial activity.

**Things you want to check and re-check over time:**

- Is the data still stable and retrievable?
- 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? Comments? Tips?



Check out our web site:
http://libraries.mit.edu/data-management

MITLibraries

RES.STR-002 Data Management
Spring 2016