The reason we’re here is to explain what data management is, why we think it should be important to you, provide you with some things to think about in managing your data and provide some resources for managing your data.

This is a first, introductory workshop.... Invite questions and comments throughout.
As we talk with you today, be thinking about how what we are telling you works with your own research process.
Research data is:

“...the recorded factual material commonly accepted in the scientific community as necessary to validate research findings.”

Does not include, “any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This "recorded" material excludes physical objects (e.g., laboratory samples).”

You can begin to describe your own research data by characterizing the method(s) you use to gather it and where it resides.
Main goal of data curation

“Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time” –

Covers all aspects of the data lifecycle from planning digital data capture methods, whittling down, ingestion to databases, providing for access and reuse, to transformation.
Properly managed data are:

- Preserved
- Accessible
- Reusable

Accessible to other researchers
  research & knowledge base develops more rapidly
Increases the impact of your research
Allows validation of your research

It saves YOU time
  You don’t spend time looking for or recreating data
  You don’t need to worry about backup – plan is in place
  Once the data is archived – access is not your worry

Allows easier exchange with other scientists
It will be protected/preserved – important for unique and historically significant data
Increases citation of not only the data, but the scholarly work(s) associated with it

More likely to be re-visited and re-used in ways that may be different from their original intention
Available to assist in the development of multidisciplinary approaches to an issue & the development
  of combined/multi-method/across-time data sets
THE CASE FOR DATA MANAGEMENT
STEWARDSHIP
CURATION
ETC.
1. Requirements: funder, journal, etc.
2. Further your field: speed the pace of discovery by enhancing discovery, access and reuse
3. Increase your visibility & impact: share your data, get credit with data citation
4. Integrity: data (documentation will preserve integrity of data over time) & you (achieve ethics compliance [HIPAA] and enhance your reputation via transparency)
5. Efficiency: avoid digital archeology, avoid duplication of effort
6. Preserve your lifelong body of work!
RESEARCH DATA LIFECYCLE

Research Cycle

New research question posed
Research planning & design
Data collection & description
Data processing & analysis
Dissemination & publication of findings
Data archived
Data transformed / repurposed
Accessible data located
Anything in blue, we can do.
TYPES, FORMATS & STAGES OF DATA — RAW DATA

COULD BE:

SENSOR OUTPUT

AUDIO/VISUAL RECORDING

LABORATORY/FIELD NOTES

GPS COORDINATES

SURVEY RESULTS

INTERVIEW NOTES

CHARACTERISTICS:

RAW UNITS

NO EXTRA METADATA

NO CONTEXT
HOW SHOULD YOU MANAGE RAW DATA?

MULTIPLE COPIES, MULTIPLE LOCATIONS

DON’T MODIFY THEM!
TYPES, FORMATS & STAGES OF DATA

INTERMEDIATE DATA

CHARACTERISTICS:

- DATA THAT ARE *IN-PROCESS*
- ANYTHING BETWEEN ‘RAW’ AND ‘FINAL’
- MULTIPLE STAGES, MULTIPLE FORMATS, USING VARIOUS SOFTWARE
- MAY BE SHAREABLE W/COLLEAGUES, BUT NOT LIKELY W/WIDER AUDIENCE
HOW SHOULD YOU MANAGE INTERMEDIATE DATA?

MULTIPLE COPIES (OF MOST RECENT VERSIONS) IN MULTIPLE LOCATIONS

USE SENSIBLE FOLDER HIERARCHIES & FILE NAMING CONVENTIONS

ADD METADATA IMMEDIATELY & THEN ADDITIONALLY AS YOU MAKE CHANGES
TYPES, FORMATS & STAGES OF DATA

CHARACTERISTICS:

MAY BE A SUBSET OF RAW DATA, OR A MUCH LARGER DATASET

SHAREABLE WITH ANYONE

NON-PROPRIETARY FORMATS ARE IDEAL

METADATA ARE THOROUGH & COMPLETE
HOW SHOULD YOU MANAGE FINAL DATA?

MULTIPLE COPIES IN MULTIPLE LOCATIONS

REFRESH DATA EVERY 2-5 YEARS

IF YOU USE A PROPRIETARY FORMAT, MIGRATE DATA TO NEW VERSION REGULARLY

SHARE VIA AN OPEN REPOSITORY
THE BIG PICTURE

Planning
Storage & backup
Organization & naming
Documentation & metadata
Legal & ethical considerations
Sharing & reuse
Archiving & preservation
WHERE DO I START?

Make a plan! Consider:

- How much data?
- Resources needed
- Roles & responsibilities
- Metadata
- Data formats
- Data storage
- Ethics & consent
- Copyright (open data)
- Sharing

Resources: supplies and services, costs associated with them
Metadata: Project level (documentation) and data level (description)
Storage: working data vs. archive vs. shared/discoverable
Storage needs will depend on the volume and type of data in question. A plan should anticipate the particular storage needs of your research. How much data do you expect to collect and how will storage infrastructure be funded?

How will files be arranged on disk. This may seem like a trivial issue, but it’s important. If the person who has been responsible for maintaining files leaves, you need to be able to find things. You’ll also want to group files meaningfully with their related metadata.

Data are, of course, subject to loss and corruption. (Bonus joke: there are two types of computer users... those who have suffered hard drive failure and those who haven’t suffered hard drive failure yet). Will there be multiple copies—on-site and off-site? Are there checks against file corruption?

Will you use specialized software or services to manage these aspects of the process?

You should know which data you intend to retain locally and for how long. Who will be responsible for this after the main project period is over?

Will you deposit in a preservation repository? Different repositories may have different preservation policies which you should be aware of.
Offsite, redundant storage is strongly encouraged.
Data storage concerns throughout the lifecycle.
**FILE NAMING STRATEGY?**

<table>
<thead>
<tr>
<th>Filename</th>
<th>Date Modified</th>
<th>Size</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>data_2010.05.28_test.dat</td>
<td>3:37 PM 5/28/2010</td>
<td>420 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.28_re-test.dat</td>
<td>4:29 PM 5/28/2010</td>
<td>421 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.28_re-re-test.dat</td>
<td>5:43 PM 5/28/2010</td>
<td>420 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.28_calibrate.dat</td>
<td>7:17 PM 5/28/2010</td>
<td>1,256 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.28_huh?.dat</td>
<td>7:20 PM 5/28/2010</td>
<td>30 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.28_WTF.dat</td>
<td>9:58 PM 5/28/2010</td>
<td>30 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.29_saaarjgh.dat</td>
<td>12:37 AM 5/29/2010</td>
<td>30 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td><a href="mailto:data_2010.05.29_mg@kll.dat">data_2010.05.29_mg@kll.dat</a></td>
<td>2:40 AM 5/29/2010</td>
<td>0 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.29_crab.dat</td>
<td>3:22 AM 5/29/2010</td>
<td>437 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.29_notbad.dat</td>
<td>4:16 AM 5/29/2010</td>
<td>670 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.29_woo ho!dat</td>
<td>4:47 AM 5/29/2010</td>
<td>1,349 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>data_2010.05.29_USETHISONE.dat</td>
<td>5:08 AM 5/29/2010</td>
<td>2,894 KB</td>
<td>DAT file</td>
</tr>
<tr>
<td>analysis_graphs.xls</td>
<td>7:13 AM 5/29/2010</td>
<td>455 KB</td>
<td>XLS file</td>
</tr>
<tr>
<td>ThesisOutline.doc</td>
<td>7:26 AM 5/29/2010</td>
<td>38 KB</td>
<td>DOC file</td>
</tr>
<tr>
<td>Notes_Meeting_with_ProfSmith.txt</td>
<td>11:30 AM 5/29/2010</td>
<td>1,673 KB</td>
<td>TXT file</td>
</tr>
<tr>
<td>JUNK...</td>
<td>2:45 PM 5/29/2010</td>
<td>Folder</td>
<td></td>
</tr>
<tr>
<td>data_2010.05.30_startingover.dat</td>
<td>8:37 AM 5/30/2010</td>
<td>420 KB</td>
<td>DAT file</td>
</tr>
</tbody>
</table>
Within reason, include relevant information such as:

- Unique identifier (i.e., 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)
- Use application-specific codes in 3-letter file extension and lowercase: mov, tif, wrl

When using sequential numbering, make sure to use leading zeros to allow for multi-digit versions. For example, a sequence of 1-10 should be numbered 01-10; a sequence of 1-100 should be numbered 001-010-100.

No special characters: & , * % # ; ( ) ! @$ ^ ~ ' { } [ ] ? < > -

Use only one period and before the file extension (e.g., name_paper.doc
NOT name.paper.doc OR name_paper..doc)
LEGAL & ETHICAL CONSIDERATIONS

• Intellectual Property
  • Office for Commercialization & Corporate Development (OCCD)
  • Copyright
• Licensing
• Charging for data?
• Data attribution & citation

HUMAN SUBJECTS?
• Informed consent & anonymization required prior to publishing
• Resources @ OSU:
  • Office of Research Integrity, Institutional Review Board (IRB)
  • Responsible Conduct of Research (RCR) Program
COPYRIGHT — WHO OWNS THE DATA?

Copyright applies to text, images, recordings, videos etc. in a digital form, in the same way that it would to analog versions of those works. The code of computer programs (both the human readable source code and the machine readable object code) is protected by copyright as a literary work. Data compilations such as datasets and databases can be protected by copyright in the literary works category, which includes ‘tables’ or ‘compilations’. A table or compilation, consisting of words, figures or symbols (or a combination of these) is protected if it is a literary work and has the required degree of originality.
DATA SHARING & REUSE

“...digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and **publicly accessible** to search, retrieve, and analyze.”

Office of Science and Technology Policy
The White House
HOW TO SHARE?

1. Repository: discipline-specific, journal, etc.
2. OSU's digital repository, ScholarsArchive@OSU
3. Personal web site
4. By request

BEST PRACTICES:

• Well documented, standard metadata schema
• Non-proprietary file formats
There are storage options for various disciplines including discipline related repositories. Some scientific journals also have their own data repositories. OSU has an institutional repository, Scholars Archive that serves as a repository for University Scholarship. Currently we store data sets on a case-by-case basis. If you don’t know where to store your data, you may consult with the library or talk to people in your department and/or your field to see if there is a logical repository for your work.
What does this mean? Data or information that will allow another user to understand your data. Let’s take 2 very simple examples.
WHAT IS METADATA?

• Data about data
• Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.

NISO, Understanding Metadata
“The metadata accompanying your data should be written for a user 20 years into the future -- what does that person need to know to use your data properly? Prepare the metadata for a user who is unfamiliar with your project, methods, or observations.”

Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics (ORNL DAAC)
METADATA ISSUES

"Researchers prefer rich descriptive metadata supporting discovery and reuse, although they are not necessarily dedicated to allotting time required for creating good quality metadata."

“Metadata generation is inefficient, with automatic applications not being fully employed, and often the same metadata is being generated via humans in more than one setting.”
HOW IS METADATA USED?

Data management
- Automatic or semi-automatic metadata
- Interoperable
- Secure

Data quality control
- Verifiable
- Replicable
- Reproducible

Data use
- Interworkable
- Analysis-ready

Data discovery
- Findable
- Identifiable
- Visualizable
- Selectable
- Obtainable
You already use metadata whether you realize it or not. For example, take this set of data. Is this obvious what it is? What if I tell you these are temperature readings? It’s still not clear. Is this Celsius or Fahrenheit” Where and when are these temperatures taken? Okay, we need metadata or information that will help other users understand your data.
Now, if I present the same three temperatures with metadata included, what the temperatures represent is much clearer. We know it was -23 degrees Fahrenheit in Anchorage, Alaska at City Hall on February 12, 2010 at 2 pm or 1400 hours. And so on.
Describe the data and how it is being collected. Provide enough detail so that a future user can understand the data. Information about each piece of data is often kept in a spreadsheet as we saw in the previous example. You may also need information about the data set as a whole. This might be as simple as a readme text file that you include with the data, to explain the data.

**METADATA**

1. Describe each piece of data
2. Describe the data set

This data set is weather readings from Anchorage Alaska for the month of February, 2010. An ABC thermometer was used, which has an accuracy of... An XYZ barometer was used...

The data set contains part of the data for a study of oral condition of cancer patients conducted at the XYZ Center. The oral conditions of the patients were measured and recorded at the initial stage, at the end of the second week, at the end of the fourth week, and at the end of the sixth week. Patients were divided into two groups at random: One group received a placebo and the other group received aloe juice treatment. Sample size, $n = 25$ patients with neck cancer.
One of the most basic schemes is Dublin Core. It is used when the metadata required is very basic. Darwin core is a very specific scheme used in the biological sciences. Talk about discovery vs description and the collection of data.
Dublin core has very basic elements which can be used to describe a variety of items. The elements on the left could be used to describe a scholarly article, an image, a movie. There is also what is called Qualified Dublin Core which allows multiple uses of each element, for example description.abstract, description.provenance, description.

The elements on the right are all details which would be under coverage in Darwin core. There are approximately 40 geographical elements in Darwin Core. For a paining, Italy or Venice is enough information for coverage, but for a scientific experiment, specific details about location are necessary.
LET'S DESCRIBE THIS DATASET

Bright orange Garibaldi fish
_Hypsypops rubicundus_
California, USA

Ornate Butterfly fish
_Chaetodon ornatissimus_
Indo-Pacific
RESEARCH INTO SPECIFIC DETAILS OF SPECIFIC TYPES OF FISH

**Darwin Core**

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Hypsypops rubicundus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource Relationship</td>
<td>Osteichthyes</td>
</tr>
<tr>
<td>Record Number</td>
<td>0034</td>
</tr>
<tr>
<td>Recorded by</td>
<td>Jane Doe</td>
</tr>
<tr>
<td>LifeStage</td>
<td>Adult</td>
</tr>
<tr>
<td>Reproductive Condition</td>
<td>Gravid</td>
</tr>
</tbody>
</table>
RESEARCH ON WHAT FISH ARE LOCAL TO A PARTICULAR AREA. THE PHOTOS ARE THE DATA.

<table>
<thead>
<tr>
<th>ID</th>
<th>0034</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>20120212</td>
</tr>
<tr>
<td>Time</td>
<td>1324</td>
</tr>
<tr>
<td>Camera</td>
<td>Nikon-5D MarkII</td>
</tr>
<tr>
<td>Lens</td>
<td>18-55mm f/3.5-5.6G</td>
</tr>
<tr>
<td>Location.bodyOfWater</td>
<td>Southern CA Bight</td>
</tr>
<tr>
<td>Location.Area</td>
<td>Channel Islands NP</td>
</tr>
<tr>
<td>Location.LatLon</td>
<td>34.0N_119.4W</td>
</tr>
<tr>
<td>Filename</td>
<td>DSC_0034.NEF</td>
</tr>
</tbody>
</table>
RESEARCH FOR PRESCHOOLERS TO SEE IF THEY LEARN COLORS AND PATTERNS BETTER FROM REAL LIFE EXAMPLES

<table>
<thead>
<tr>
<th>ID</th>
<th>Representation</th>
<th>Description</th>
<th>Pattern/Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>Fish</td>
<td>Striped Fish</td>
<td>Stripes</td>
</tr>
<tr>
<td>002</td>
<td>Fish</td>
<td>Yellow Fish</td>
<td>Yellow</td>
</tr>
</tbody>
</table>
METADATA EXAMPLES

Santa Barbara Coastal Long Term Ecological Research (LTER)
web link

web link
Your data is being managed whether you realize it or not, but in order to preserve data and make the most of it, researchers should actively plan their data management. A data management plan provides a clear roadmap through the process of collecting, storing, sharing and preserving your data. The earlier in the process a researcher starts in the data management process, not only makes it easier for her, but many repositories have requirements about how the data has been formatted etc. . . And it may not be possible to meet those requirements retroactively.
These are the five sections for the generic NSF data management plan, but it what is discussed in any data management plan. Briefly describe each section and add that there will be a follow up workshop to write individual DMPs in the summer. We are always happy to meet with individuals who need help.

- the types of data to be authored;
- the standards that would be applied, for example format and metadata content;
- provisions for archiving and preservation;
- access policies and provisions;
- and plans for eventual transition or termination of the data collection in the long-term future.
- Create ready to use plans
- Meet agency requirements
- Step-by-step guidance
- OSU-specific guidance
Please feel free to contact Amanda Whitmire or Maura Valentino with any data management questions you might have or questions about depositing your research to ScholarsArchive@OSU digital repository, especially research data associated with published research articles and/or your thesis or dissertation.

We plan to offer an advanced data management workshop in the Fall at which we’ll discuss … Interest in this workshop? Get emails so we can contact you when those are scheduled.