Managing Ecological Data for Effective Use and Re-Use

Amber Budden, DataONE
Bill Michener, DataONE
Stephanie Hampton, NCEAS
Carly Strasser, CDL
Viv Hutchison, USGS
Tammy Beaty, ORNL DAAC

Is this for you?

The target audience is early-career scientists but is open to any researcher who would benefit from developing better data management skills. Faculty members who would like to include best practices for preparing data in curricula are encouraged to attend.

Who are you?

Name
Title / Day Job / Career Stage
Institution / Organization
Experience with Data Management
Favorite Food
Name that Puppy

Outcomes of the workshop

The purpose of this workshop is to give attendees a set of practical tools for organizing and sharing their data through all parts of the research cycle.

Attendees will gain
- a solid understanding of data management planning, data structure, quality control and data documentation;
- experience using tools that aid in data management;
- an appreciation for good data management practices for data sharing, collaboration, and data re-use.

Agenda

12:00 Welcome and Introductions
12:10 Data Life Cycle / Data Management Planning
12:50 Data Collection and Management
13:30 Break
13:45 Quality Assurance / Quality Control
14:25 Describing your data (Metadata)
15:05 Data Preservation
15:45 Break
16:00 Q&A, problem solving etc

Who are you?
Other information

Dropbox location:

Other DataONE activities at the ESA:
- Booth #206, Exhibit Hall
- SS3 – Data Management Planning; Mon 10:15 am Bldm 253
- WK29 – Data Sharing Culture; Tue 11:30 E141
- OOS 33 – Ecology and the 21st Century; Wed 1:30 pm A106
- OOS 33-2 –Big Data; Wed 1:50 pm A106
- WK47 – Accessing Data in R; Thu 11:30 am F150

Tweets:
@DataONEorg

Data Life Cycle / Planning
Bill Michener
DataONE and University of New Mexico
Data Management Planning (Part 1)
Failing to plan is planning to fail.
Alan Lakein

Data Management Planning Topics
1. What is a data management plan (DMP)?
2. Why prepare a DMP?
3. Components of a DMP
4. Example of a NSF DMP
5. Exercise

Why Prepare a DMP?
• Saves time
• Increases research efficiency
  • Ensures you and others will be able to understand and use data in future
  • Prevents duplication of effort
• Satisfies funding agency requirements

NSF DMP Requirements
From Grant Proposal Guidelines:
Plans for data management and sharing of the products of research. Proposals must include a supplementary document of no more than two pages labeled "Data Management Plan". This supplement should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results (in AAG), and may include:
1. the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project
2. the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies)
3. policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements
4. policies and provisions for re-use, re-distribution, and the production of derivatives
5. plans for archiving data, samples, and other research products, and for preservation of access to them

What is a DMP?
• A document that describes what you will do with your data during and after you complete your research

Credits:
Taung3npanda, Anita363, Stonebird, NeilsPhotography, Rick Smit, Jschinker
Components of a DMP

1. Information about data & data format
2. Metadata content and format
3. Policies for access, sharing and re-use
4. Long-term storage and data management
5. Budget

1. Information About Data & Data Format

1.1 Description of data to be produced
   - Experimental, Observational, Raw or derived, Physical collections, Models, images, etc.
1.2 How data will be acquired
   - When? Where? Methods?
1.3 How data will be processed
   - Software used, Algorithms, Workflows
1.4 File formats
   - csv, tab-delimited, naming conventions
1.5 Quality assurance & quality control
1.6 Existing data
   - If existing data are used, what are its origins?
   - What is the relationship between your data and existing data?
1.7 How data will be managed in short-term
   - Version control, Back-up, Security & protection,
   - Who will be responsible?

2. Metadata Content & Format

   Metadata is the documentation describing all aspects of the data (e.g., who, why, what, when, where)

2.1 What metadata are needed
   - Any details that make data understandable and usable
2.2 How metadata will be created and/or captured
   - Lab notebooks? GPS units?
   - Auto-saved on instrument? Manually entered?
2.3 What format will be used for the metadata
   - Standards for community (EML, ISO 19115, etc.)
   - Justification for format chosen

3. Policies for Access, Sharing, Reuse

3.1 Obligations for sharing
   - Funding agency, institution
3.2 Details of data sharing
   - How long?
   - When?
   - How access can be gained?
3.3 Ethical/privacy issues with data sharing
3.4 Intellectual property & copyright issues
   - Institutional policies
   - Funding agency policies
   - Embargos for political/commercial reasons
3.5 Intended future uses/users for data
3.6 Citation
   - How should data be cited when used?
   - Persistent citation?

4. Long-term Storage & Data Management

4.1 What data will be preserved
4.2 Where will it be preserved
   - Most appropriate archive for data
   - Community standards
4.3 Data transformations/formats needed
   - Consider archive policies
4.4 Who will be responsible
   - Contact person for archive

5. Budget

5.1 Anticipated costs
   - Time for data preparation & documentation
   - Hardware/software for data preparation & documentation
   - Personnel
   - Archive costs

5.2 How costs will be paid

DataONE
5.

Budget

We will first document our metadata by taking careful notes in the laboratory for reuse and storage will require approximately one month of salary for one technician. The technician will be responsible for data entry, quality control and assurance, and metadata generation. These costs are included in the budget in lines 12-16.
What a researcher can do:
- View sample plans
- Preview funder requirements
- Create, save, edit, publish plan
- View, use, past plans
- Use help (generic and institution specific)
- View news and latest changes

A goal without a plan is just a wish.
Larry Elder
Data Management Planning Topics
1. When do you plan?
2. How do you develop a plan?
3. Who does the planning?

When do you plan?
At the beginning ... and throughout the research life cycle

How do you plan?
- View DM planning as a part of the research process
- “The better is the enemy of the ‘good enough’”
  - Don’t hesitate because you don’t know all the answers
    - Use metadata from your proposal, publications, etc.
- Review and revise routinely

Who plans?

DMP Resources
http://dmp.cdlib.org/
http://dataone.org

Thanks to Carly Strasser, Amber Budden, Viv Hutchison, and DMP Tool contributors!
Data collection, entry & manipulation
Stephanie Hampton
National Center for Ecological Analysis and Synthesis
University of California, Santa Barbara

See also: http://www.dataone.org/education-modules

Goals in data entry
Data are:
• Valid
• Organized
• Easy to subset

Capture as close to source as possible

A personal example of bad data practices...

Random notes

2 tables
Goals in data entry

Data are:
- Valid
- Organized
- Easy to subset

Standardize protocols

Personal data management problems are magnified over time, and in collaboration
- Data organization – standardize
- Data documentation – standardize descriptions of data (metadata)
Standardize protocols

- Descriptive column names without spaces or special characters
  - Soil_T30 → Soil_Temp_30cm
  - Species-Code → Species_Code (avoid using -, +, ^ in column names – software may interpret as an operator)

Missing data

- In numeric fields, use a distinct value such as 9999 to indicate a missing value
- In text fields, use NULL or NA (“Not Applicable”)
- Use Data flags in a separate column to qualify empty cells

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>NCO_1_Corr</th>
<th>NCO_1_Corr_Flag</th>
</tr>
</thead>
<tbody>
<tr>
<td>20080211</td>
<td>1200</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td>20080211</td>
<td>1330</td>
<td>0.016</td>
<td>M1</td>
</tr>
<tr>
<td>20080211</td>
<td>1400</td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td>20080211</td>
<td>1430</td>
<td>0.001</td>
<td>E1</td>
</tr>
</tbody>
</table>

Reduce possibility of manual error

- Constrain data entry choices as much as possible

Goals in data entry

- Data are:
  - Valid
  - Organized
  - Easy to subset
- Capture as close to source as possible
- Protect raw data
  - Manipulate outside of raw data sheets

Example – using R for data manipulation

```R
# Simple Linear Regression - 0+ age Trout in Hoh River, WA against Temp Celsius
# Load data
HohTrout <- read.csv("Hoh_Trout0_Temp.csv")
# http://knb.ecoinformatics.org/knb/style/skins/nceas/
# Look at the data
HohTrout
# Plot TROUT against TEMPC
plot(TROUT ~ TEMPC, data=HohTrout)
# Log Transform the independent variable (x+1) - this method for transform creates a new column in the data frame
HohTrout$LNtrout <- log(HohTrout$TROUT + 1)
# Plot the log-transformed y against x
windows()
plot(LNtrout ~ TEMPC, data=HohTrout)
```

Modified from K. Vanderbilt
Example – using R for data manipulation

### Simple Linear Regression - age Trout in Hoh River, WA against Temp Celsius

#### Load data

HohTrout <- read.csv("Hoh_Trout_Age_Temp.csv")


#### http://knb.ecoinformatics.org/doi:10.5063/AKINS0001

#### Look at the data

plot(HohTrout$Age ~ HohTrout$Temp)

#### Log Transform the independent variable (x+1) - this method for transform creates a new column in the data frame

HohTrout$LNage <- log(HohTrout$Age + 1)

#### Plot the log-transformed y against x

First I'll ask R to open new windows for subsequent graphs with the windows command

windows()

plot(HohTrout$LNage ~ HohTrout$Temp)

---

Spreadsheets & relational databases

### Spreadsheets: primordial data entry tool

- No shame in using CSV (comma-separated value) to archive data tables – just enforce best practices
  - Every computer can read it
  - It's been around forever
  - It will be around for a long while more

- Databases typically require a specific application to read their proprietary binary formats: DBF (DB IV), MDB (MS Access), or e.g. even XLS

---

Spreadsheets & relational databases

### Relational databases: Relationships are defined between tables

<table>
<thead>
<tr>
<th>Date</th>
<th>Site</th>
<th>Species</th>
<th>Flowering?</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/13/2010</td>
<td>A</td>
<td>BOG2</td>
<td>Y</td>
<td>34.1</td>
<td>-109.3</td>
</tr>
<tr>
<td>2/13/2010</td>
<td>B</td>
<td>HOCR</td>
<td>Y</td>
<td>35.2</td>
<td>-108.6</td>
</tr>
<tr>
<td>4/15/2010</td>
<td>B</td>
<td>BOER4</td>
<td>Y</td>
<td>35.2</td>
<td>-108.6</td>
</tr>
<tr>
<td>4/15/2010</td>
<td>C</td>
<td>PLUA</td>
<td>n</td>
<td>32.6</td>
<td>-107.5</td>
</tr>
</tbody>
</table>

---

You can use Spreadsheets like a Database e.g., using R, or SAS, MATLAB, etc. to define relationships
Handy References for Better Data Structures


Exercise

Look at your sample data
These sheets are all part of the same study
Make suggestions for data organization

15 Minute Break

Assuring Quality Data
Carly Strasser
DataONE, California Digital Library

Defining Quality Assurance & Control

Mechanisms for preventing errors from entering a data set

Quality assurance
• Activities to ensure data quality before collection

Quality control
• Monitoring & maintaining data quality during the study

The Data Life Cycle

Plan
Collect
Assure
Integrate
Describe
Preserve
Discover

Plan
Collect
Assure
Integrate
Describe
Preserve
Discover

Error Definitions

Data contamination
- Process or phenomenon, other than the one of interest, that affects the variable’s value

Errors of Commission
- Incorrect or inaccurate data entered
- Examples: malfunctioning instrument, mistyped data

Errors of Omission
- Data or metadata not recorded
- Examples: inadequate documentation, human error, anomalies in the field

Prevent Errors Before Collection

- Define & enforce standards
  - Formats
  - Codes
  - Measurement units
  - Metadata
- Assign responsibility for data quality
  - Be sure assigned person is educated in QA/QC & error detection

Minimize Errors During Collection

- Double entry
  - Data keyed in by two independent people
  - Check for agreement with computer verification
- Use text-to-speech program to read data back
- Design data storage well
  - Minimize number of times items must be entered repeatedly
  - Use consistent terminology
  - Atomize data: one cell per piece of information
  - Consider using a database

Minimize Collection Errors: Tools

Databases
- FileMaker Pro (Mac)
- Access (PC)

Google Forms
- web-based

Excel
- lists
- data validation

Minimize Errors After Collection

- Ensure data line up in columns
- Check for missing, impossible, or anomalous values
  - sort by fields to highlight discrepancies
- Perform statistical summaries
  - Compare summaries before and after transformations

Detect Errors After Collection

- Look for outliers
  - Goal is not to eliminate outliers but to identify potential data contamination
- Graphical methods
  - Normal probability plots
  - Regression
  - Scatter plots
  - Maps
**Lesson Topics**

Definition of metadata
Examine information included in a metadata record
Examples of metadata standards and how to choose
Illustrate the value of metadata to data users, data providers, and organizations
Tips on how to write quality metadata records

**Handle Errors**

- Case-by-case decision
  - Flag them?
  - Remove them?
  - Fix them?
- Document all changes (readme.txt, scripts)
- Keep original data separate
  - Analyze & clean with scripts

**Exercise**

Using the raw data provided...
Check the data for different types of error using one or more of the software tools introduced earlier
Make appropriate error handling decisions
Document the error check workflow

**Metadata**

Viv Hutchison
USGS

**The Data Life Cycle**

Plan → Collect → Integrate → Describe → Preserve

Discover → Analyze → Assure
### Data Collection

![Image](https://example.com/data_collection_image)

### From Field Notes to Datasets

#### Average Temperature of Observation for Each Species

<table>
<thead>
<tr>
<th>Species</th>
<th>Average Temp</th>
<th>Temp Deviation</th>
<th>Number of Observations</th>
<th>Minimum Temp</th>
<th>Maximum Temp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern Red-legged Frog</td>
<td>4.4</td>
<td>0.0</td>
<td>1</td>
<td>4.4</td>
<td>4.4</td>
</tr>
<tr>
<td>Tailed Frog</td>
<td>7.0</td>
<td>3.0</td>
<td>2</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Arizona Toad</td>
<td>10.0</td>
<td>---</td>
<td>1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Doctor's Choice Frog</td>
<td>10.5</td>
<td>2.0</td>
<td>11</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>Oregon Spotted Frog</td>
<td>11.0</td>
<td>15.5</td>
<td>2</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>New Jersey Chorus Frog</td>
<td>11.5</td>
<td>4.5</td>
<td>17</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>United Frog</td>
<td>12.5</td>
<td>3.5</td>
<td>99</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>Tailing Frog</td>
<td>15.2</td>
<td>5.6</td>
<td>500</td>
<td>5</td>
<td>32</td>
</tr>
<tr>
<td>Northern Red-legged Frog</td>
<td>15.3</td>
<td>5.9</td>
<td>10</td>
<td>4</td>
<td>27</td>
</tr>
</tbody>
</table>

### From Datasets to Published Papers

![Image](https://example.com/datasets_to_papers_image)

### Working with Data

#### Providing data:
- Why were the data created?
- What limitations, if any, do the data have?
- What does the data mean?
- How should the data be cited if it is re-used in a new study?

#### Receiving data:
- What are the data gaps?
- What processes were used for creating the data?
- Are there any fees associated with the data?
- In what scale were the data created?
- What do the values in the tables mean?
- What software do I need in order to read the data?
- What projection are the data in?
- Can I give these data to someone else?

### What is Metadata?

**Metadata is: Data ‘reporting’**

**WHO** created the data?

**WHAT** is the content of the data?

**WHEN** were the data created?

**WHERE** is it geographically?

**HOW** were the data developed?

**WHY** were the data developed?

### Metadata in Real Life

**Metadata is all around...**

![Image](https://example.com/metadata_real_life_image)
**What is a Metadata Standard?**

A standard provides a structure to describe data with:

- Common terms to allow consistency between records
- Common definitions for easier interpretation
- Common language for ease of communication
- Common structure to quickly locate information

In search and retrieval, standards provide:

- Documentation structure in a reliable and predictable format for computer interpretation
- A uniform summary description of the dataset

---

**What Does a Metadata Record Look Like?**

*Southwest Non-native Invasive Plant Database (SWNNIPD)*

**Metadata**

- **Title:** Southwest Non-native Invasive Plant Database (SWNNIPD)
- **Abstract:** The Southwest Non-native Invasive Plant Database (SWNNIPD) is a collaborative effort to compile data on non-native invasive plants in the Southwest. The database contains information on plant species, distribution ranges, and associated management strategies.

---

**The Value of Metadata**

*Data developers* help...

---

**What is the Value to Data Users?**

Metadata gives a user the ability to:

- Search, retrieve, and evaluate data set information from both inside and outside an organization
- Find data: Determine what data exists for a geographic location and/or topic
- Determine applicability: Decide if a data set meets a particular need
- Discover how to acquire the dataset you identified; process and use the dataset

---

**What is the Value to Data Developers?**

Metadata allows data developers to:

- Avoid data duplication
- Share reliable information
- Publicize efforts – promote the work of a scientist and his/her contributions to a field of study

---

**What is the Value to Organizations?**

Metadata helps ensure an organization's investment in data:

- Documentation of data processing steps, quality control, definitions, data uses, and restrictions
- Ability to use data after initial intended purpose

Transcends people and time:

- Offers data permanence
- Creates institutional memory

Advertises an organization's research:

- Creates possible new partnerships and collaborations through data sharing
Concerns About Creating Metadata

Even if the value of data documentation is recognized, concerns remain as to the effort required to create metadata that effectively describe the data.

Concerns About Creating Metadata

<table>
<thead>
<tr>
<th>Concern</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workload required to capture accurate robust metadata</td>
<td>Incorporate metadata creation into data development process – distribute the effort</td>
</tr>
<tr>
<td>Time and resources to create, manage, and maintain metadata</td>
<td>Include in grant budget and schedule</td>
</tr>
<tr>
<td>Readability/usability of metadata</td>
<td>Use a standardized metadata format</td>
</tr>
<tr>
<td>Discipline specific information and ontologies</td>
<td>Use ‘profile’ standard to require specific information and use specific values</td>
</tr>
</tbody>
</table>

Informative Entropy

Sound information management, including metadata development, can arrest the loss of dataset detail.

Writing Quality Metadata

Multiple Metadata Standards Exist: Examples

**Dublin Core Element Set**
- Emphasis on web resources, publications

**FGDC Content Standard for Digital Geospatial Metadata (CSDGM)**
- Emphasis on geospatial data
- **Biological Data Profile (BDP) of the CSDGM**
  - Profile to the CSDGM emphasis on biological data (and geospatial)
Multiple Metadata Standards Exist:
Examples

Ecological Metadata Language (EML)
- Focus on ecological data

Darwin Core
- Emphasis on museum specimens
- http://rs.tdwg.org/dwc/index.htm

ISO 19115/19139 Geographic information: Metadata
- Emphasis on geospatial data and services
- http://www.fgdc.gov/metadata/geospatial-metadata-standards#fgdcendorsedisostandards

Choosing a Metadata Standard

Many standards collect similar information
Factors to consider:
- Your data type:
  - Are you working mainly with GIS data? Raster/Vector or point data? Do you have biological or shoreline information in your dataset?
    - Consider the FGDC Content Standard for Digital Geospatial Metadata with one of its profiles: the Biological Data Profile or the Shoreline Data Profile.
  - Are you working with data retrieved from instruments such as monitoring stations or satellites?
    - Are you using geospatial data services such as applications for web-mapping applications or data modeling?
    - If so, then consider using the ISO 19115-2 standard
  - Are you mainly working with ecological data?
    - Consider Ecological Metadata Language (EML)

Steps to Create Quality Metadata

Organize your information
- Did you write a project abstract to obtain funding for your proposal? Re-use it in your metadata!
- Did you use a lab notebook or other notes during the data development process that define measurements and other parameters?
- Do you have the contact information for colleagues you worked with?
- What about citations for other data sources you used in your project?

Tips for Writing Quality Metadata

- Do not use jargon
- Define technical terms and acronyms:
  - CA, LA, GPS, GIS: what do these mean?
- Clearly state data limitations
  - E.g., data set omissions, completeness of data
- Express considerations for appropriate re-use of the data
- Use “none” or “unknown” meaningfully
  - None usually means that you knew about data and nothing existed (e.g., a “0” cubic feet per second discharge value)
  - Unknown means that you don’t know whether that data existed or not (e.g., a null value)
Tips for Writing Quality Metadata

- Titles are critical in helping readers find your data
  - While individuals are searching for the most appropriate data sets, they are most likely going to use the title as the first criteria to determine if a dataset meets their needs.
  - Treat the title as the opportunity to sell your dataset.
- A complete title includes: What, Where, When, Who, and Scale
- An informative title includes: topic, timeliness of the data, specific information about place and geography

Tips for Writing Quality Metadata

- Be specific and quantify when you can! The goal of a metadata record is to give the user enough information to know if they can use the data without contacting the dataset owner.
  - Vague: We checked our work and it looks complete.
  - Specific: We checked our work using a random sample of 5 monitoring sites reviewed by 2 different people. We determined our work to be 95% complete based on these visual inspections.

Tips for Writing Quality Metadata

- Remember: a computer will read your metadata
  - Do not use symbols that could be misinterpreted: Examples: ! @ # % { } / \ < > ~
  - Don’t use tabs, indents, or line feeds/carrige returns
  - When copying from other sources, use a text editor to eliminate hidden characters

Tips for Writing Quality Metadata

- Select keywords wisely
  - Use descriptive and clear writing
  - Fully qualify geographic locations
  - Use thesauri for keywords whenever possible
  - Example: USGS Biocomplexity Thesaurus (over 9,500 terms)

Share Your Metadata: Distribution

Share your metadata with other researchers
Examples of metadata search portals:
  - Data.gov
    Federal e-gov geospatial data portal
    http://www.geo.data.gov
  - Metacat
    Repository for data and metadata
    http://knb.ecoinformatics.org/index.jsp
  - US Geological Survey
    USGS Core Science Metadata Clearinghouse:
    http://mercury.ornl.gov/clearinghouse
  - ArcGIS Online
    ESRI sponsored national geospatial data portal
    http://www.arcgisnetwork.com
**DataONE Search**

**Summary**

- Metadata is documentation of data
- A metadata record captures critical information about the content of a dataset
- Metadata allows data to be discovered, accessed, and re-used
- A metadata standard provides structure and consistency to data documentation
- Standards and tools vary – select according to defined criteria such as data type, organizational guidance, and available resources
- Metadata is of critical importance to data developers, data users, and organizations
- Writing quality metadata is important because records are expected to last with the data over decades
- Metadata completes a dataset.

*Creating robust metadata is in your OWN best interest!*

---

**Data Preservation: Data Stewardship and Reuse**

Tammy Walker Beaty
Environmental Sciences Division
Oak Ridge National Laboratory
beatytw@ornl.gov

**Topics:**
- Introduction
- Protection
- Archiving
- Sharing & Reuse
- Classroom Exercise

**Introduction: Cycles of Research**

Research Cycle = Repetitive process
Research Product = many iterations of Data = Final Data Product

**Introduction: Data Stewardship & Reuse**

**Stewardship**

“An ethic that embodies responsible planning and management of resources” - Wikipedia

“The careful and responsible management of something entrusted to one’s care” - Webster
Introduction: The 20-Year Rule (NRC 1991)

The metadata accompanying a data set should be written for a user 20 years into the future — what does that investigator need to know to use the data?

Prepare the data and metadata/documentation for a user who is unfamiliar with the details of your project, methods, and observations.

Data Protection: Backups

Create back-up copies
- At least three copies – original, one one-site (external), one off-site (e.g., dropbox, Carbonite, etc.)
- Frequency based on need/risk

Version your data
- As you develop a data product, keep previous versions – be able to go back to yesterday if you make a critical mistake today

Know that you can recover from a data loss
- Periodically test your ability to restore information

Data Protection: File Transfers

Ensure that file transfers are without error
- Compare checksums before and after transfers (calculated summary of a data portion, fingerprint)
- Example tools to generate checksums
  - hdp://checksumchecker.sourceforge.net
  - hdp://www.openoffice.org/dev_docs/using_md5sums.html

Ensure that files are transferred across platforms properly
- Correct ftp transfer mode (e.g., dos2unix)
- Proper line feeds and control characters

Data Archiving

Data Center: Stewardship and Archive Functions

- Acquisition
  - Identify how best to serve the scientific community
  - Establish how and when to receive data
- Ingest
  - Perform QA checks
  - Complete purchase and metadata
  - Convert to archival file format
- Enhance (as required)
  - Convert to standard formats & units, aggregate files
- Metadata / Documentation
  - Prepare final metadata record and documentation

Archive / Publish
- Generate citation

Exploration and Distribution
- Provide tools to explore, access, and extract data for users worldwide

Post-Project Data Support
- Serve as a buffer between end-users and researchers
- Provide usage statistics

Stewardship
- Provide long-term secure archiving of the data
- Security, disaster recovery
- Migration to new computer systems

Choosing a Data Archive – Sharing & Reuse

Institution vs. science discipline archive
- Keep discipline data together
- Resources ($$)
- Data center and project requirements
- Discovery and access requirements
- Specialized data types (geospatial data, genetic sequences, etc.)

Data sharing and reuse policies
- Institutions’ Policy - Intellectual property?
- Restrictions – protection policies and procedures for legitimate needs based on data type (privacy, confidentiality, intellectual property, or other security needs)
- US Funding agencies
  - Open Access (NASA, NSF)
  - Short period of exclusive use for QA/QC (NOAA)
Data Sharing & Reuse: Citation

Practice analogous to journal article citations

Enable readers to find data products themselves
• Reproduce results
• Use data for new hypothesis

Add to data authors' CV – citation indices for the data publication

Data authors get credit for the data publication and subsequent citations

Can be used to show funders the impact of their research programs on the advancement of science

Shows the scientific impact of data centers' data holdings

Classroom exercise, Background:

You became the data manager for FLUXNET in April 2008. FLUXNET is a network of regional networks that coordinates regional and global analysis of observations from micrometeorological tower sites using eddy covariance methods to measure the exchanges of carbon dioxide, water vapor, and energy between terrestrial ecosystems and the atmosphere. The FLUXNET database contains information about the tower location, site characteristics, and data availability. It does not contain the actual data collected at each site, only pointers to that data. http://daac.ornl.gov/FLUXNET/fluxnet.shtml


This was the suggested citation:


You will need to update and re-publish this map every six months or so. What do you need to reproduce this map?

Summary

Plan to manage iterative versions of your data throughout your research – document along the way

Protect your data – backups, versioning, file transfer verification procedures, plan for archival and dissemination

Archival – identify an appropriate final destination for your results, factor destination’s requirements into your research

Sharing and Reuse – Citation, Sponsor requirements, document for layperson

Classroom exercise, Background:

Software: ESRI ARCMAP, map_name.mxd (esri map file)

Source Data itself source_data_file.csv (forget the database itself for now, and focus only on reproducing the map).

Modified source file, with displayed networks, map_name_data.csv

Documentation file (*.readme)

World polygon layer (layername.dbf, layername.shp, layername.sbn, layername.sbx, layername.shx)

Logos – file names and sources.

*.readme – file to self documenting location of data sources, process, etc.

This map will be updated and re-published about every six months. What is your plan for managing this data?

What do you need to reproduce this map?

Classroom Exercise

Document data used to produce the FLUXNET map at:


• This map was published in April 2008, Today is August 5, 2012.

On March 30, 2008, there were 543 sites http://en.wikipedia.org/wiki/File:Fluxnet_Growth.jpg. Today there are 544 sites.

• What issues do you see?

The networks are different – there were xx networks displayed in 2008, and yy networks displayed in changed, the networks are different.

• FLUXNET is a global data base; of you are asked to your site data and produce a global map of networks to be published.
15 Minute Break
Followed by Question & Answer and Troubleshooting Session