Introduction to LTER Data Management

Shamelessly scavenged from presentations by John Porter, Kristin Vanderbilt, Hook et. al, and DataONE
Scientific Use of Data

- The traditional model of using data

Diagram:
- Data Collection (sensors, people) → Data Entry → Raw Data → Quality Assurance and Quality Control → Verified Data → New research questions → Publications → Visualization & Analysis
Challenge: Understanding Data

Michener et al. 1997. Ecological Applications

Without Metadata, the usable information content of data declines over time.

- Time of publication
- Specific details
- General details
- Accident
- Retirement or career change
- Death
Scientific Use of Data

- A new model incorporates sharing and archiving

Michiner et. al. 2011, Ecological Informatics
Increasing value of data over time

Serendipitous Discovery

Inter-site Synthesis

Gradual Increase In Data Equity

Methodological Flaws, Instrumentation Obsolescence

Non-scientific Monitoring

Data Value

Time

Slide from James Brunt
Data Sharing and Archiving
Data Sharing

- The LTER Network Data Policy dictates that almost all data should be made available within 2-years.
- Exceptions must be justified.
- NSF now requires Data Management Plans for LTER.
- A better plan increases your chance of funding.
Additional Incentives

- NSF now requires Data Management Plans for non-LTER data as well
  - A better plan increases your chance of funding
- Journals are increasingly requiring data submission as a condition of publication for papers (e.g., evolution, genomics journals)
- Increasingly data is citable
  - Allows you to tally the citations of your data as well as citations of your publications
- Data can even be published: e.g., Ecological Archives publishes “data papers” that are peer-reviewed
Value of Data Sharing

- Open science and new research
- Data longevity
- Data reusability
- Greater exposure to data
- Potential increased citation of source papers (Piwowar, 2007)
- Confirmation of results from publications (Thornton et al., 2005)
- Generation of value added products
- Possibility for future research collaborations
- More value for the sponsor’s research investment

Benefits Pay Off when Data is Well Documented and Well Structured...
But First... Backup Your Data

- Have a working version and at least two backup copies
- Have your data in three separate physical places: Here, Near, and There [Brunt's Axiom]
  - Computer, a local physically separate backup, and far, far away (e.g., in the Cloud--Dropbox, SugarSync, Mozy, Carbonite)
  - Some agencies have restrictions on use of the Cloud. Follow your local guidelines.
Preparing Environmental Data Sets: 7 Best Practices

- Use Consistent Data Organization
- Use Consistent File Structure
- Use Descriptive File Names
- Assign Descriptive Data Set Titles
- Perform Basic Quality Assurance
- Provide Documentation (JRN LTER Project & Dataset forms)
- Define the Contents of Your Data Files
Organization by Data Type

- If collecting observations of different types of measurements at a site (e.g., leaf area index and above- and belowground biomass), place each type of measurement in a separate data file.
- For each data file, use similar data organization, parameter formats, and common site names, so that users understand the interrelationships between data files.
Data types collected on different time bases (e.g., per hour, per day, per year) are handled more efficiently in separate files.
Use Consistent File Structure

- Use the same structure throughout the file - don't have a different number of columns or re-arrange the columns within the file.
- Use a consistent structure across all data files prepared for a study or project.
- Figures and analyses should be reported in companion documents - don't place figures or summary statistics in the data file.
Assign Descriptive File Names

- File names should reflect the contents of the file and include enough information to uniquely identify the data file.
- Clear, descriptive, and unique file names may be important later when your data file is combined in a directory or FTP site with your own data files or with the data files of other investigators. Avoid using file names such as mydata.dat or 1998.dat.
Assign Descriptive Data Set Titles

- Be aware that your data sets may be accessed many years in the future by people who will be unaware of the details of the project.
- Limit length of the title to 85 characters (spaces included) to be compatible with clearinghouses of ecological and global change data collections.
Define Contents of Your Data Files

In order for others to use your data, they must fully understand the contents of the data set, including the parameter names, units of measure, formats, and definitions of coded values.
Use Consistent Data Organization

- Each row in a file represents a complete record, and the columns represent all the parameters that make up the record. This arrangement is similar to a spreadsheet or matrix.
- In general, keep a set of similar measurements together (e.g., same investigator, methods, time basis, and instruments) in one data file. Please do not break up your data into many small files, e.g., by month or by site.
Data Structure

- The ways researchers typically use data are frequently not compatible with best practices for archiving [the Confetti Toss—one time usage]

<table>
<thead>
<tr>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>My Ugly Data Spreadsheet</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Date: 5/23/2005</td>
<td></td>
<td>cond_bot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Meter Type</td>
<td>YSI_Model_30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Tide State</td>
<td>slack-high</td>
<td>conductivity_top</td>
<td>&lt;30</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Salinity</td>
<td>Bottom</td>
<td>0.6 Top</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Date: 10/2/2005</td>
<td></td>
<td>cond_bot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>TIDESTATE</td>
<td>-0.34</td>
<td>MeterType</td>
<td>YSI Model 30</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Salinity_bottom</td>
<td>0.3</td>
<td>Conductivity Top</td>
<td>349</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Salinity_Top</td>
<td>None</td>
<td>cond_bot</td>
<td>39%</td>
<td></td>
</tr>
</tbody>
</table>
LTER Solutions

- Data manager can help vet data
- Help communicate best practices to students and investigators
- Use of improved tools that encourage good practices
QA/QC

- Mechanisms designed to prevent introduction of errors into a dataset
Perform Basic Quality Assurance

- QA for Tabular Data
  - Ensure metadata accurately reflects data file names and the actual data file’s format, column headers, and content
  - Check that data are in the proper columns
  - Check for & document missing values
  - Check validity of values
  - Verify data (from field notebooks, dataloggers, or instruments)
Loss of data quality can occur at many stages:

- At the time of collection
  - Incorrect or inaccurate data are entered into a dataset
    - Misidentifying quad
    - Malfunctioning instrumentation
      - Sensor drift
      - Low batteries
      - Damage

Credit: http://sev.lternet.edu
Loss of data quality can occur at many stages:

- **During digitization**
  - Mistyping code
  - Transposing numbers
Loss of data quality can occur at many stages:

- **During documentation**
  - Omission: Data or metadata are not recorded
  - Inadequate documentation of experimental design, sampling methods
  - Inadequate documentation of anomalies in the field
  - Forgetting to take measurement in the field
- Datasheet Design
- Training technicians
- Documentation of procedures
- Data Entry Constraints

- Graphics
- Statistics

Cost of error correction increases
Datasheets facilitate data collection

- Date
- Time
- Habitat Type
- Macroplot #
- Tree Species
- DBH
- Condition
- Microplot #
- Shrub Species
- Shrub Cover%
- Daubenmire Plot Species
- Cover %
Emulate data file structure...

<table>
<thead>
<tr>
<th>Type</th>
<th>Study#</th>
<th>StudyName</th>
<th>Year</th>
<th>Qtr</th>
<th>Date</th>
<th>Site</th>
<th>Time</th>
<th>Tower</th>
<th>Height (cm)</th>
<th>Avg_Bag_Wt (g)</th>
<th>Bag+Sample (g)</th>
<th>Sample_Wt (g)</th>
<th>QWt</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSNE</td>
<td>122</td>
<td>ScrapeSite</td>
<td>2013</td>
<td>1</td>
<td>2/26/2013</td>
<td>Scrape Site</td>
<td>944</td>
<td>East</td>
<td>5</td>
<td>4.31</td>
<td>5.34</td>
<td>0.91</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>BSNE</td>
<td>122</td>
<td>ScrapeSite</td>
<td>2013</td>
<td>1</td>
<td>2/26/2013</td>
<td>Scrape Site</td>
<td>944</td>
<td>East</td>
<td>10</td>
<td>4.31</td>
<td>4.93</td>
<td>0.50</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>BSNE</td>
<td>122</td>
<td>ScrapeSite</td>
<td>2013</td>
<td>1</td>
<td>2/26/2013</td>
<td>Scrape Site</td>
<td>944</td>
<td>East</td>
<td>20</td>
<td>4.31</td>
<td>5.73</td>
<td>1.30</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>BSNE</td>
<td>122</td>
<td>ScrapeSite</td>
<td>2013</td>
<td>1</td>
<td>2/26/2013</td>
<td>Scrape Site</td>
<td>944</td>
<td>East</td>
<td>50</td>
<td>4.31</td>
<td>5.34</td>
<td>0.91</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>BSNE</td>
<td>122</td>
<td>ScrapeSite</td>
<td>2013</td>
<td>1</td>
<td>2/26/2013</td>
<td>Scrape Site</td>
<td>944</td>
<td>East</td>
<td>100</td>
<td>4.31</td>
<td>5.49</td>
<td>1.06</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>BSNE</td>
<td>122</td>
<td>ScrapeSite</td>
<td>2013</td>
<td>1</td>
<td>2/26/2013</td>
<td>Scrape Site</td>
<td>947</td>
<td>Middle</td>
<td>5</td>
<td>4.31</td>
<td>5.19</td>
<td>0.75</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>BSNE</td>
<td>122</td>
<td>ScrapeSite</td>
<td>2013</td>
<td>1</td>
<td>2/26/2013</td>
<td>Scrape Site</td>
<td>947</td>
<td>Middle</td>
<td>10</td>
<td>4.31</td>
<td>4.85</td>
<td>0.42</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>BSNE</td>
<td>122</td>
<td>ScrapeSite</td>
<td>2013</td>
<td>1</td>
<td>2/26/2013</td>
<td>Scrape Site</td>
<td>947</td>
<td>Middle</td>
<td>20</td>
<td>4.31</td>
<td>5.31</td>
<td>1.10</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>BSNE</td>
<td>122</td>
<td>ScrapeSite</td>
<td>2013</td>
<td>1</td>
<td>2/26/2013</td>
<td>Scrape Site</td>
<td>947</td>
<td>Middle</td>
<td>50</td>
<td>4.31</td>
<td>5.62</td>
<td>1.19</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>BSNE</td>
<td>122</td>
<td>ScrapeSite</td>
<td>2013</td>
<td>1</td>
<td>2/26/2013</td>
<td>Scrape Site</td>
<td>947</td>
<td>Middle</td>
<td>100</td>
<td>4.31</td>
<td>6.06</td>
<td>1.63</td>
<td>G</td>
<td></td>
</tr>
</tbody>
</table>
**Study: BNSE dust collectors - NPP sites, Burn site, Geomet grid data**

**Project:** Study #288 - BSNE dust collection at NPP, JER Pasture 13, Geomet sites

**Responsible investigator:** Greg Okin

**Attribute description:**
- **Type** = sampling type (core=soil core; BSNE=BSNE dust collector)
- **Study#** = LTER Study number (122 (SS), 228, 288, 308, ?)
- **StudyName** = short name for study (ConnPilot, Geomet, LTER-I_TranProbeSoil, NPP, P11aWindErosion, P13Burn, ScrapeSite)
- **Year** = year of field collection (yyyy)
- **Qtr** = quarterly collection period (1,2,3,4)
- **Date** = date of field collection (mm/dd/yyyy)
- **Site** = Scrape Site
- **Time** = time of field collection (24-hour clock)
- **Tower** = Tower (East, Middle, West)
- **Height** = BSNE collector height (5,10,20,50,100 cm above ground when opening oriented to southwest)
- **Avg_Bag_Wt** = average bag weight (in grams recorded to 3 decimal places)
- **Bag+Sample** = bag + sample weight (in grams recorded to 3 decimal places)
- **Sample_Wt** = air dry weight of dust (grams) calculated from (Bag+Sample) minus (Avg_Bag_Wt)
- **QWt** = data quality flag for Sample Wt
  - **G** = Good; no problems; normal collection
  - **TL** = Not collected; **TOO LITTLE** to collect (less than a "pinch"); Any dust (if any) in trap was emptied
  - **PC** = Partial collection; **some sample lost during collection**
- **Comments** = Explain QWt entry or other observations
- **missing value** = .
Provide Data Set Documentation and Metadata

- Project and Dataset documentation/metadata forms available online.
- Includes “who, what, where, when, why, and how” of the data.
- Write it for a user 20 years into the future.
Submit and Archive Data with Metadata Early in Your Study

- Submit Project & Dataset documentation within 90 days of initiation of data collection
- First time data submission must be submitted with updated Jornada metadata forms previously submitted.
- First time data submission should be no later than 3 months prior to graduation
Open vs Restricted Access Data

- Open Access Data is made available online.
- Restricted Access Data is not made available to others without the release authorization of the responsible investigator.
  - Data backed up or archived with Jornada less than 2 years after submission or with exemption by LTER Exec Committee.
  - Students have opportunity to publish data following graduation
Summary

• If data are:
  o Well-organized
  o Documented
  o Preserved
  o Accessible
  o Verified as to Accuracy and validity

• Result is:
  o High quality data
  o Easy to share and re-use in science
  o Citation and credibility to the researcher
Acknowledgments:

LTER slides downloaded from http://im.lternet.edu/node/937 from presentations by John Porter, Kristin Vanderbilt, and DataONE.

Downloaded from http://im.lternet.edu/node/948