Ecoinformatics: Managing Data, Rescuing Data and Changing the Scientific Culture

William Michener
LTER Network Office
Department of Biology
University of New Mexico
Today’s Road Map

- Science Challenges
- Ecoinformatics
- Data Rescue
- Changing the Culture of Science
Today’s Road Map

• Science Challenges
  • Ecoinformatics
  • Data Rescue
  • Changing the Culture of Science
<table>
<thead>
<tr>
<th>Date</th>
<th>Site</th>
<th>Species</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/1/1993</td>
<td>N654</td>
<td>Picea rubens</td>
<td>13</td>
</tr>
<tr>
<td>10/1/1993</td>
<td>N654</td>
<td>Betula papyifera</td>
<td>3</td>
</tr>
<tr>
<td>10/31/1993</td>
<td>1</td>
<td>Picea rubens</td>
<td>13.5</td>
</tr>
<tr>
<td>10/31/1993</td>
<td>1</td>
<td>Betula papyifera</td>
<td>1.6</td>
</tr>
<tr>
<td>11/14/1994</td>
<td>1</td>
<td>Picea rubens</td>
<td>8.4</td>
</tr>
<tr>
<td>11/14/1994</td>
<td>1</td>
<td>Betula papyifera</td>
<td>1.8</td>
</tr>
</tbody>
</table>

(Michener, 2000)
Characteristics of Ecological Data

Data Volume (per dataset) vs. Complexity/Metadata Requirements

- High Data Volume: Satellite Images, Weather Stations, GIS, Most Ecological Data
- Low Data Volume: Gene Sequences, Primary Productivity, Biodiversity Surveys, Population Data, Soil Cores

- High Complexity/Metadata Requirements: Business Data
- Low Complexity/Metadata Requirements: Most Software

Most Ecological Data

Most Software
Data Integration

- Syntax and Schema transformations
- Semantic conversion

### METADATA
(from EML)

<table>
<thead>
<tr>
<th>Date</th>
<th>Site</th>
<th>Species</th>
<th>Area</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/1/1993</td>
<td>N654</td>
<td>PIRU</td>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>10/3/1994</td>
<td>N654</td>
<td>PIRU</td>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>10/1/1993</td>
<td>N654</td>
<td>BEPA</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

**Study A = White Mountains**
PIRU=Picea rubens
BEPA=Betula papyifera
Area column units = square meter

**Study B = Green Mountains**
picrub=Picea rubens
betpap=Betula papyifera
Area sampled = 1 square meter

<table>
<thead>
<tr>
<th>Date</th>
<th>Site</th>
<th>Species</th>
<th>Count</th>
<th>picrub</th>
<th>betpap</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 Oct 1993</td>
<td>1</td>
<td>13.5</td>
<td>1.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 Nov 1994</td>
<td>1</td>
<td>8.4</td>
<td>1.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### DATA

<table>
<thead>
<tr>
<th>Study Date</th>
<th>Site</th>
<th>Species</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 10/1/1993</td>
<td>N654</td>
<td>Picea rubens</td>
<td>13</td>
</tr>
<tr>
<td>A 10/1/1993</td>
<td>N654</td>
<td>Betula papyifera</td>
<td>3</td>
</tr>
<tr>
<td>B 10/31/1993</td>
<td>1</td>
<td>Picea rubens</td>
<td>13.5</td>
</tr>
<tr>
<td>B 10/31/1993</td>
<td>1</td>
<td>Betula papyifera</td>
<td>1.6</td>
</tr>
<tr>
<td>B 11/14/1994</td>
<td>1</td>
<td>Picea rubens</td>
<td>8.4</td>
</tr>
<tr>
<td>B 11/14/1994</td>
<td>1</td>
<td>Betula papyifera</td>
<td>1.8</td>
</tr>
</tbody>
</table>

### INTEGRATED DATA PRODUCT
Semantics—Linking Taxonomic Semantics to Ecological Data

- Taxon concepts change over time (and space)
- Multiple competing concepts coexist
- Names are re-used for multiple concepts

### Table

<table>
<thead>
<tr>
<th>Date</th>
<th>Species</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>1830</td>
<td>R.plumosa</td>
<td>39</td>
</tr>
<tr>
<td>1840</td>
<td>R.plumosa</td>
<td>49</td>
</tr>
<tr>
<td>1900</td>
<td>R.plumosa</td>
<td>42</td>
</tr>
<tr>
<td>1985</td>
<td>R.plumosa</td>
<td>48</td>
</tr>
<tr>
<td>1995</td>
<td>R.plumosa</td>
<td>22</td>
</tr>
<tr>
<td>2000</td>
<td>R.plumosa</td>
<td>19</td>
</tr>
</tbody>
</table>

### Graphical Representation

**Rhynchospora plumosa s.l.**

- **A**: R. plumosa v. interrupta
- **B**: R. intermedia
- **C**: R. plumosa

**From R. Peet**
Data Entropy

Time of publication
Specific details
General details
Retirement or career change
Accident
Death

(Michener et al. 1997)
Today’s Road Map

- Science Challenges
- Ecoinformatics
- Data Rescue
- Changing the Culture of Science
Ecoinformatics

• “a broad interdisciplinary science that incorporates conceptual approaches and practical tools for the generation, processing, understanding and dissemination of ecological data and information.”
Data Design

• Conceptualize and implement a logical structure within and among data sets that will facilitate data acquisition, entry, storage, retrieval and manipulation.
7 Habits of Highly Effective Data Set Design

1. Assign descriptive file names
2. Use consistent and stable file formats
3. Define the parameters
4. Use consistent data organization
5. Perform basic quality assurance
6. Assign descriptive data set titles
7. Provide comprehensive documentation (metadata)

Adapted from Cook et al. 2000
1. Assign descriptive file names

- **File names should be unique and reflect the file contents**
  - Bad file names
    - Mydata
    - 2001_data
  - A better file name
    - Sevilleta_LTER_NM_2001_NPP.asc
      - Sevilleta_LTER is the project name
      - NM is the state abbreviation
      - 2001 is the calendar year
      - NPP represents Net Primary Productivity data
      - asc stands for the file type--ASCII
2. Use consistent and stable file formats

- **Use ASCII file formats** – avoid proprietary formats
- **Be consistent in formatting**
  - don’t change or re-arrange columns
  - **include header rows (first row should contain file name, data set title, author, date, and companion file names)**
  - column headings should describe content of each column, including one row for parameter names and one for parameter units
  - **within the ASCII file, delimit fields using commas**, pipes (|), tabs, or semicolons (in order of preference)
3. Define the parameters

- Use commonly accepted parameter names that describe the contents (e.g., precip for precipitation)
- Use consistent capitalization (e.g., not temp, Temp, and TEMP in same file)
- **Explicitly state units** of reported parameters in the data file and the metadata (SI units are recommended)
- Choose a format for each parameter, explain the format in the metadata, and use that format throughout the file
  - e.g., use yyyyymmdd; January 2, 1999 is 19990102
4. Use consistent data organization (one good approach)

<table>
<thead>
<tr>
<th>Station</th>
<th>Date</th>
<th>Temp</th>
<th>Precip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Units</td>
<td>YYYYMMDD</td>
<td>C</td>
<td>mm</td>
</tr>
<tr>
<td>HOGI</td>
<td>19961001</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>HOGI</td>
<td>19961002</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>HOGI</td>
<td>19961003</td>
<td>19</td>
<td>-9999</td>
</tr>
</tbody>
</table>

Note: -9999 is a missing value code for the data set
4. Use consistent data organization (a second good approach)

<table>
<thead>
<tr>
<th>Station</th>
<th>Date</th>
<th>Parameter</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOGI</td>
<td>19961001</td>
<td>Temp</td>
<td>12</td>
<td>C</td>
</tr>
<tr>
<td>HOGI</td>
<td>19961002</td>
<td>Temp</td>
<td>14</td>
<td>C</td>
</tr>
<tr>
<td>HOGI</td>
<td>19961001</td>
<td>Precip</td>
<td>0</td>
<td>mm</td>
</tr>
<tr>
<td>HOGI</td>
<td>19961002</td>
<td>Precip</td>
<td>3</td>
<td>mm</td>
</tr>
</tbody>
</table>
5. Perform basic quality assurance

- Assure that data are delimited and line up in proper columns
- Check that there no missing values for key parameters
- Scan for impossible and anomalous values
- Perform and review graphical & statistical summaries
- Map location data (lat/long) and assess errors
- Verify automated data transfers
- For manual data transfers, consider double keying data and comparing 2 data sets
6. Assign descriptive data set titles

- Data set titles should ideally describe the type of data, time period, location, and instruments used (e.g., Landsat 7).
- Titles should be restricted to 80 characters.
- Data set title should be similar to names of data files
  - Good: “Shrub Net Primary Productivity at the Sevilleta LTER, New Mexico, 2000-2001”
  - Bad: “Productivity Data”
7. Provide comprehensive documentation (metadata)

"I think you should be more explicit here in step two."
What are metadata?

“Data about data”

or, more appropriately,

“the information necessary to understand and effectively use the data”
Metadata helps you decide which can you would like to eat!
Metadata Content Specifications

- Dublin Core
- NBII Biological Data Profile / CSDGM
- ISO CD 19115, Geographic information - metadata
- Darwin Core
- Ecological Metadata Language (EML)
ESRI ArcCatalog

Metadata creation/import selections

Metadata Tab

Metadata Sections

Catalog list

Metadata Parts

Description Spatial Attributes

Keywords
Theme: topography
Place: Oregon, Cascades

Description

Abstract
Polygon layer of subdivision of landscape based on topographic position: crest, valley, north, south, east, or west slope. Vectorized from cascomp.

Purpose
Used in a PhD dissertation for Oregon State University Department of Geosciences, comparing human-impacted and natural landscapes in the western Cascades of Oregon

Status of the data

Time period for which the data is relevant

Publication Information

Data storage and access information

Details about this document
An Example EML Document

```xml
<?xml version="1.0"?>
<eml:eml packageId="piscoUCSB.5.20" system="knob" xmlns:eml="eml://ecoinformatics.org/eml-2.0.0">
  <dataset>
    <shortName>Alegria Temperatures</shortName>
    <creator id="C.Blanchette">
      <individualName>
        <givenName>Carol</givenName>
        <surName>Blanchette</surName>
      </individualName>
      <organizationName>PISCO</organizationName>
      <address>
        <deliveryPoint>UCSB Marine Science Institute</deliveryPoint>
        <city>Santa Barbara</city>
        <administrativeArea>CA</administrativeArea>
        <postalCode>93106</postalCode>
      </address>
    </creator>
    <abstract>
      <para>These temperature data were collected at Alegria Beach, California, and were ...</para>
    </abstract>
    <keywordSet>
      <keyword>OceanographicSensorData</keyword>
      <keyword>Thermistor</keyword>
      <keywordThesaurus>PISCOCategories</keywordThesaurus>
    </keywordSet>
    <intellectualRights>
      <para>Please contact the authors for permission to use these data. Please also acknowledge the authors in any publications.</para>
    </intellectualRights>
    <contact>
      <references>C.Blanchette</references>
    </contact>
  </dataset>
</eml:eml>
```

Ecological Metadata Management Software
How much work is this going to be???
Rules of Thumb

(Michener 2000)

• the more comprehensive the metadata, the greater the longevity (and value) of the data

• structured metadata can greatly facilitate data discovery, encourage “best metadata practices” and support data and metadata use by others

• metadata implementation takes time!!!

• start implementing metadata for new data collection efforts and then prioritize “legacy” and ongoing data sets that are of greatest benefit to the broadest user community
Make metadata implementation a team effort! Include:

• Team Leader
• GIS Specialist
• Field Personnel
• Database Manager
• Laboratory Specialist
• Voucher/Repository Specialist
• And others as appropriate….
An Idealized Information Environment

“Value-Added” or “Integrated” Infobases

National/Regional Systems

Project or Site-Based Systems

Individual datasets
Individual Datasets

Metadata

Data

**Individual Datasets**

**Metadata**

**Data**

---

**Dataset Title:** Fish in North Temperate Lakes Primary Study Lakes: Corrected Catch

**Investigator:** John J. Magnuson, Timothy K. Krats

**Contact:** John J. Magnuson, Center for Limnology, University of Wisconsin-Madison, 680 North Park Street, Madison, WI 53726

**Temporal and Spatial Resolution:**

- Study duration: 07/29/81
- Sampling frequency: annually

**Number of sites:** Multiple sites (year dependent) on 7 lakes

**Algorithim used to synthesize data:** Gill net data have been standardized to a 24-hour sampling period. Assumptions used in the standardization are available from the investigators. Key species were only used in 1991 and have been eliminated from this data set to make sampling effort across years comparable. Number caught for each species is summed over repetitions of a gear within a lake and over depth.

**Abstract:**

This data set is a derived data set based on fish abundance data. Data are collected annually to enable us to track the fish assemblages of seven primary lakes (Allegheny, Big Musky, Crystal, Sparkling, Trout, and log lakes) and seven secondary lakes. Parameters measured include species identification and lengths for all fish caught, and weight and scale samples from a subset.

**Dataset includes:**

- Catch per unit effort, growth, and age distribution by species, lake, and year.
- Dominant species vary from lake to lake. Perch, crappies, and bluegill are common, with walleye, large and smallmouth bass, northern pik and muskellunge as major piscivores.

**Species Listed:**

- BSEINE, 9, BIGHOUNTHUFFALO, 2
- BSEINE, 9, BLACKCRAPPIE, 112
- BSEINE, 9, BLUEGILL, 8
- BSEINE, 9, BLUSTNEOMINNOW, 6
- BSEINE, 9, BROOKSILBERSIDE, 227
- BSEINE, 9, COMMOCARP, 1
- BSEINE, 9, CRAPPIE, 216
- BSEINE, 9, FATHEAD, 1
- BSEINE, 9, LOGPERCH, 1
- BSEINE, 9, ROCKBASS, 1
- BSEINE, 9, SMALLMOUTH, 8
- BSEINE, 9, WHITEBASS, 12
- BSEINE, 9, WHITECRAPPIE, 2
- BSEINE, 9, YELLOWPERCH, 35
- CRAYTR, 20, BLACKCRAPPIE, 2
- CRAYTR, 20, BLUEGILL, 5
- CRAYTR, 20, CRAPPIE, 2
- CRAYTR, 20, YELLOWPERCH, 1
- ESHOCK, 2, BIGHOUNTHUFFALO, 3
- ESHOCK, 2, BROOKSILBERSIDE, 1
- ESHOCK, 2, COMMOCARP, 16
- ESHOCK, 2, CRAPPIE, 20
- ESHOCK, 2, FRESHWATERDRUM, 18
- ESHOCK, 2, LARGMOUTH, 5
- ESHOCK, 2, LONGNOSEGAR, 3
Climate database integrates data from a number of sites
Key Elements Needed at each level

• Site/Project
  – Metadata – ideally, standards-based

• National or Network
  – Consistent keyword vocabularies
  – Standards for metadata content

• “Value Added”
  – Domain Expertise
  – Need for structured metadata
  – Standards for data products
Lessons Learned

1. **Keep data close to home**
   - Experts on a data set are those who collected the data
Lessons Learned

2. **Build a modular system**

- Traditional approach of building a single, large system is unwieldy and inflexible
- Each module should be able to take advantage of new technologies
3. **Build a database partnership**

- Successful design and implementation depends on combining expertise of:
  - Scientists (what is needed)
  - IM’s (how to make it work)
  - Computer scientists (software tools)

In that order!
4. Make extensive use of prototypes

- Describing a system as “user friendly” does not make it so—testing is required
- Prototypes lend themselves to the exploitation of new technologies and ideas; prototypes are inexpensive so you can have lots of them.
Lessons Learned

5. **Data should be expected to far outlive the database**
   - Databases are ephemeral
   - Have a clear exit strategy
Lessons Learned

6. **Give something back**

- If you build it, **it does not mean** that they will come.
- **Data contributors should receive something of value back**
  - “*If we do this, we both benefit.*”
Today’s Road Map

- Science Challenges
- Ecoinformatics
- Data Rescue
- Changing the Culture of Science
What is responsible for the apparent rise/decrease in _______???
What relevant data exist?
Where are those data?
What do the data mean?
(How) Can I use the data?
(How) Can I integrate the various data sources?

*Weeks > Months > Years*
Finding Data & Metadata

- Colleagues
- Scientific literature
- WWW searches
- Data and metadata registries
  - Global Change Master Directory
- Metadata Clearinghouses
  - National Biological Information Infrastructure
  - NODC
NBII Clearinghouse Nodes to Query

Select one or more NBII Clearinghouse nodes in the list below. You may query multiple nodes to query simultaneously. Hold your shift or control key down while selecting multiple nodes.

See Descriptions of NBII Clearinghouse Nodes

- NBII Metadata Clearinghouse
- Eastern Sierra Geospatial Data Clearinghouse
- EMAN Data Set Library (Environment Canada Server)
- National Wetlands Research Center (NWRC) Clearinghouse Node
- ONRC Clearinghouse for the Olympic Peninsula

Maximum Number of Responses to View: 20

Submit Query  Reset  Reset the form to default values.

When you are done searching, please give us your comments, and then select the LOGOUT button.
Frequently-anticipated questions:

- What does this data set describe?
  1. How should this data set be cited?
  2. What geographic area does the data set cover?
  3. What does it look like?
  4. Does the data set describe conditions during a particular time period?
  5. What is the general form of this data set?
  6. How does the data set represent geographic features?
  7. How does the data set describe geographic features?

- Who produced the data set?
  1. Who are the originators of the data set?
  2. Who also contributed to the data set?
  3. To whom should users address questions about the data?

- Why was the data set created?

- How was the data set created?
  1. Where did the data come from?
  2. What changes have been made?

- How reliable are the data; what problems remain in the data set?
  1. How well have the observations been checked?
  2. How accurate are the geographic locations?
  3. How accurate are the heights or depths?
  4. Where are the gaps in the data? What is missing?
  5. How consistent are the relationships among the data, including topology?
Today’s Road Map

- Science Challenges
- Ecoinformatics
- Data Rescue
- Changing the Culture of Science
Cycles of Research
“A Conventional View”
Cycles of Research

“A New View”

Planning

Analysis and modeling

Selection and extraction

Publication

Archive of Data

Collection

Original Observations

Problem Definition (Research Objectives)

Secondary Observations

Planning
Reasons to Not Share Data:

My data will be misinterpreted ...

I will get to publishing on it later ...

I will be scooped ...

Someone will find errors ...
Benefits of Data Sharing

- Publicity, accolades, media attention
- Renewed or increased funding
- Teaching:
  - long-term data sets adapted for teaching & texts
- Archival: back-up copy of critical data sets
- Research:
  - new synthetic studies
  - peer-reviewed publications
- Document global and regional change
- Conservation and resource management:
  - species and natural areas protection
  - new environmental laws
Culture Change

- Scientific societies
- Funding agencies and programs
- Universities and science/resource management agencies
Increasing value of data over time

- Serendipitous Discovery
- Inter-site Synthesis
- Gradual Increase in Data Equity
- Methodological Flaws, Instrumentation Obsolescence
- Non-scientific Monitoring
Thanks !!!
References


Porter (2000) Ch. 3 in Michener and Brunt (2000)


Michener (2000) Ch. 7 in Michener and Brunt (2000)

Metadata Resources

• http://knb.ecoinformatics.org and
• http://seek.ecoinformatics.org -- for ecologists
• http://www.w3.org/DesignIssues -- for technologists