Data Management & Intro to R

Figure 1. The data life cycle includes the following steps: (i) plan; (ii) collect; (iii) assure (i.e. quality assurance and quality control); (iv) describe (i.e. ascribe metadata); (v) preserve (i.e. deposit data in a secure data repository); (vi) discover (i.e. identify data that might be needed to answer a question); (vii) integrate (e.g. merge data from multiple data sources); and (viii) analyze (e.g. statistical analysis, visualization). Modified after Figure 1 in [22] with the permission of C. Strasser.
Figure 1. The data life cycle includes the following steps: (i) plan; (ii) collect; (iii) assure (i.e. quality assurance and quality control); (iv) describe (i.e. ascribe metadata); (v) preserve (i.e. deposit data in a secure data repository); (vi) discover (i.e. identify data that might be needed to answer a question); (vii) integrate (e.g. merge data from multiple data sources); and (viii) analyze (e.g. statistical analysis, visualization). Modified after Figure 1 in [22] with the permission of C. Strasser.

Michener and Jones 2011
Barriers to Synthesis

• Data not preserved
  – Tiny proportion of ecological and evolutionary data are readily available

• Dispersed, isolated repositories
  – Each community has its own; disconnected; underutilized

• Lack of software interoperability

• Heterogeneous data
  – Many data formats, metadata formats, and varying semantics
Dispersed data

Global Biodiversity Information Facility (GBIF) Downloaded 01/26/2015
Data diversity

• Biological
  – e.g., Gene, Organism, Population, Species, Community, Biome, Ecosystem

• Environmental
  – e.g., Atmospheric, Chemical, Ecological, Hydrological, Oceanographic, Physical

• Social
  – e.g., Land use, human population

• Economic
  – e.g., trade, ecosystem services, resource extraction
# Biodiversity data heterogeneity

**Space**

<table>
<thead>
<tr>
<th>Site</th>
<th>Community</th>
<th>Treatment</th>
<th>Replicate</th>
<th>Species_Code</th>
<th>ITIS_TSN</th>
<th>Plant_Mass</th>
<th>Plant_Mass_m2</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCE6</td>
<td>BorJun</td>
<td>N</td>
<td>1</td>
<td>JROM</td>
<td>39238</td>
<td>75.13</td>
<td>300.5</td>
</tr>
<tr>
<td>GCE6</td>
<td>BorJun</td>
<td>N</td>
<td>2</td>
<td>JROM</td>
<td>39238</td>
<td>179.81</td>
<td>719.2</td>
</tr>
<tr>
<td>GCE6</td>
<td>BorJun</td>
<td>N</td>
<td>3</td>
<td>JROM</td>
<td>39238</td>
<td>443.20</td>
<td>1772.8</td>
</tr>
</tbody>
</table>

**Time**

**Taxa**

<table>
<thead>
<tr>
<th>Code</th>
<th>Collector</th>
<th>Year</th>
<th>Long</th>
<th>Lat</th>
<th>Cat</th>
<th>Sci_Name</th>
<th>Institution</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>4815</td>
<td>DOWNEY, J.</td>
<td>1947</td>
<td>-109.34</td>
<td>37.86</td>
<td>null</td>
<td>Mephitis mephitis</td>
<td>ESTOR</td>
<td>mephitis</td>
</tr>
<tr>
<td>6767</td>
<td>NEILSON, L.</td>
<td>1948</td>
<td>-109.52</td>
<td>40.41</td>
<td>null</td>
<td>Mephitis mephitis</td>
<td>ESTOR</td>
<td>mephitis</td>
</tr>
<tr>
<td>6768</td>
<td>NEILSON, L.</td>
<td>1948</td>
<td>-109.52</td>
<td>40.41</td>
<td>null</td>
<td>Mephitis mephitis</td>
<td>ESTOR</td>
<td>mephitis</td>
</tr>
</tbody>
</table>

**NOAA Ocean Buoy Data**

Station 46069 - South Santa Rosa Island, CA

<table>
<thead>
<tr>
<th>YYYY</th>
<th>MM</th>
<th>DD</th>
<th>hh</th>
<th>mm</th>
<th>WD</th>
<th>WSPD</th>
<th>GST</th>
<th>WVHT</th>
<th>DPD</th>
<th>APD</th>
<th>MWD</th>
<th>BAR</th>
<th>ATMP</th>
<th>WTMP</th>
<th>DEWP</th>
<th>VIS</th>
<th>TIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>01</td>
<td>01</td>
<td>00</td>
<td>00</td>
<td>284</td>
<td>6.9</td>
<td>8.2</td>
<td>3.50</td>
<td>13.79</td>
<td>8.11</td>
<td>275</td>
<td>1013.2</td>
<td>15.5</td>
<td>14.3</td>
<td>999.0</td>
<td>99.0</td>
<td>99.00</td>
</tr>
<tr>
<td>2006</td>
<td>01</td>
<td>01</td>
<td>01</td>
<td>00</td>
<td>287</td>
<td>6.0</td>
<td>7.6</td>
<td>3.13</td>
<td>13.79</td>
<td>8.08</td>
<td>271</td>
<td>1013.5</td>
<td>15.3</td>
<td>14.3</td>
<td>999.0</td>
<td>99.0</td>
<td>99.00</td>
</tr>
<tr>
<td>2006</td>
<td>01</td>
<td>01</td>
<td>02</td>
<td>00</td>
<td>276</td>
<td>4.3</td>
<td>5.6</td>
<td>3.20</td>
<td>13.79</td>
<td>8.49</td>
<td>272</td>
<td>1013.7</td>
<td>14.9</td>
<td>14.4</td>
<td>999.0</td>
<td>99.0</td>
<td>99.00</td>
</tr>
</tbody>
</table>

**Macroecological data for fossil occurrences**

<table>
<thead>
<tr>
<th>coll_no</th>
<th>genus_name</th>
<th>occ_no</th>
<th>locomot</th>
<th>life_habit</th>
<th>diet</th>
<th>composition</th>
<th>10mybin</th>
<th>Carboniferous</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Australosutura</td>
<td>1</td>
<td>mobile</td>
<td>epifaunal</td>
<td>detritivore</td>
<td>chitin</td>
<td>10mybin</td>
<td>Carboniferous</td>
</tr>
<tr>
<td>1</td>
<td>Carbonocoryphe</td>
<td>2</td>
<td>mobile</td>
<td>epifaunal</td>
<td>detritivore</td>
<td>chitin</td>
<td>Carboniferous</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Thigriffides</td>
<td>3</td>
<td>mobile</td>
<td>epifaunal</td>
<td>detritivore</td>
<td>chitin</td>
<td>Carboniferous</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
“Dark” data in the long tail

Heidorn 2008
Data Heterogeneity

Low Heterogeneity

High Volume

High

Low

• Tight coupling
• Simple subsetting
• Explicit semantics

• Loose coupling
• Hard subsetting
• Limited semantics

VEGBANK

DRYAD

KNB
The Repository: Key features

- **Flexible** about data format, while encouraging the use and further development of community standards.
- Fits into the manuscript submission workflow of its partner journals, making data submission easy.
- Gives journals the option of making data privately available during peer review and of allowing submitters to set limited-term embargoes post-publication.
- Data are linked both to and from the corresponding publication and, where appropriate, to and from select specialized data repositories (e.g., GenBank).
- Assigns data Digital Object Identifiers (DOIs) to data so that researchers can gain professional credit through data citation.
- Promotes data visibility by allowing content to be indexed, searched and retrieved through interfaces designed for both humans and computers.
- Contents are free to download and have no legal barriers to reuse.
- Contents are curated to ensure the validity of the files and metadata.
- Submitters may update data files when corrections or additions are desired, without overwriting the original version linked from the article.
- **Long-term preservation** ... by migrating common file formats when older versions become obsolete, and partnering with DataONE to guarantee access to its contents indefinitely.

The Repository: Technology

- Dryad is built upon the open-source DSpace repository software. All customizations not available within the main DSpace distribution are available from the Dryad code repository under an open source (new BSD) license.
- Dryad supports multiple ways of receiving article or manuscript metadata from publishers. The simplest method involves reading email notifications, but we are also implementing a REST API for those desiring greater control over the data deposition process.
- Digital Object Identifiers provided by Crossref and EZID.

Learn more about:
- Membership
- Submission integration
- Pricing plans

Search for data

Enter keyword, DOI, etc.

Go

Advanced search
Knowledge Network for Biocomplexity Data Distribution

Total: 25,191 data sets

Data until: 07 Oct 2011
Software diversity
Solutions

• Preserve data

• Adopt standards

• Create networks

• Create interoperable software
Metadata and data heterogeneity

• Every community has
  – many data schemas
    • one for each project and person
  – many data formats
    • ASCII, NetCDF, HDF, GeoTiff, ...
  – many metadata schemas
    • Biological Data Profile, Darwin Core, Dublin Core, Ecological Metadata Language (EML), Open GIS schemas, ISO Schemas, ...

• Accepting this heterogeneity is critical
# Column metadata

<table>
<thead>
<tr>
<th>Attribute(s) Info:</th>
<th>sampleID</th>
<th>sampleDate</th>
<th>nominalMonth</th>
<th>monthNumber</th>
<th>year</th>
<th>trimester</th>
<th>birthday</th>
<th>ageInDays</th>
<th>agePerBobFoy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Column Label</strong></td>
<td>Code for Sample identification</td>
<td>Date sample was taken</td>
<td>Month of sampling cruise</td>
<td>Numeric month</td>
<td>Year sample was taken</td>
<td>Trimester sample was taken</td>
<td>Hypothetical birthdate of sampled fish</td>
<td>Age of sampled fish (days)</td>
<td>Adjusted age (per Bob Foy)</td>
</tr>
<tr>
<td><strong>Definition</strong></td>
<td>Alphanumeric sample identifier used to track samples; those with the format xHEy where x is a 2-digit year and y is a 1 to 3 digit number were samples initially processed by Dr. AJ Paul (U. of Alaska) who then provided Kline with the standard length and wet mass data along with the remaining sample</td>
<td>Calendar date when sample was collected</td>
<td>Cruise month name when sample was collected</td>
<td>Month number (January = 1, February =2, etc.) when sample was collected</td>
<td>Year when sample was collected</td>
<td>Trimester of calendar year when sample was collected, Trimesters A+B = A, and Trimester C = B in Kline (2007)</td>
<td>Date agreed upon by SEA investigators as a hypothetical birthday, used to compute age in days</td>
<td>Data used as the common independent variable to enable merging of data sets used in Norcross et al. (2001)</td>
<td>Herring age in years per instructions from Bob Foy (at the time a Ph. D. student of Dr. Brenda Norcross, U. of Alaska ), 0 or 1, otherwise A (putative adults)</td>
</tr>
<tr>
<td><strong>Type of Value</strong></td>
<td>nominal</td>
<td>dateTime</td>
<td>dateTime</td>
<td>dateTime</td>
<td>dateTime</td>
<td>dateTime</td>
<td>dateTime</td>
<td>ratio</td>
<td>interval</td>
</tr>
<tr>
<td><strong>Measurement Type</strong></td>
<td>Sample identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Measurement Domain</strong></td>
<td>Unit</td>
<td>Age</td>
<td>Unit</td>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Wizard to create metadata
Morpho highlights

• **Create metadata** in EML format
• **Manage data** in EML packages
• **Save, publish, and share** data

• **Search** for data
• **Multi-language**
  – English, Spanish, Chinese, French, Portuguese, Japanese
• **Export** data and metadata
• **Cross-platform**, and open source
How do we harness the long tail?

• Efficient data federation
  – Focus on individual contributors

• Late binding in informatics systems
  – Loose coupling
  – Schema-less storage

• Central search for discovery

• Interoperable software