LTER EML Best Practices

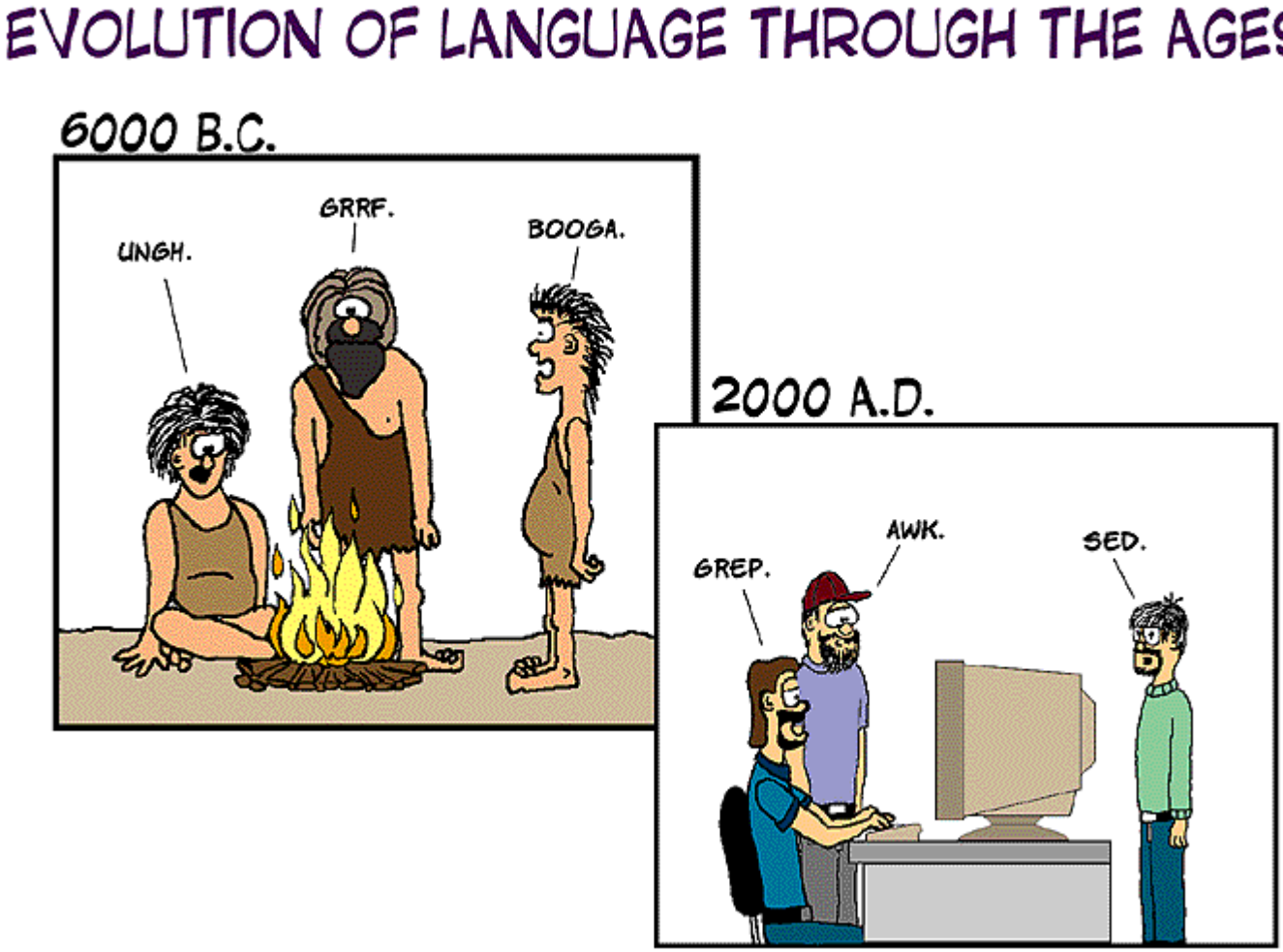
Second KNB Data Management Workshop
2-4 February 2005

Mark Servilla
LTER Network Office
University of New Mexico, Albuquerque

Agenda

- Introduction
- Goals & Motivation – Why "*EML Best Practices*"?
- LTER Metadata Tiers and recommended EML elements
- Additional Recommendations

Introduction



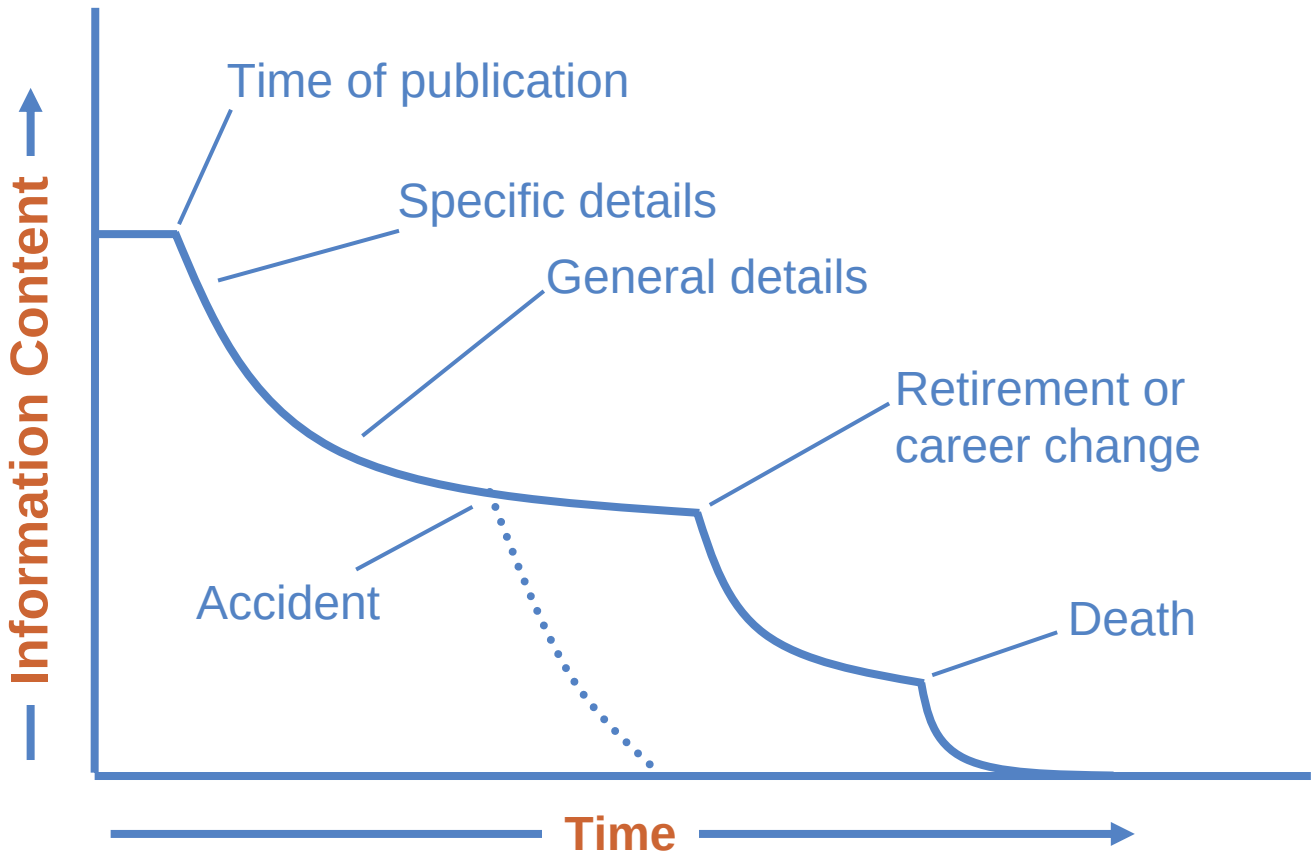
Goals & Motivation

Why do we need an "EML Best Practices" document?

- Guidelines to achieve the following goals
  - Maximize interoperability of LTER EML documents to facilitate data synthesis
  - Minimize heterogeneity of LTER EML documents to simplify development and re-use of software tools and style sheets
  - Identify useful subsets of the EML to support specific functionality tiers targeted by the LTER NIS Advisory Committee (NISAC)
  - Provide guidance to sites in their initial implementation of EML, and a roadmap for improving their implementation to achieve higher functionality

Information Entropy over Time

en·tro·py : a process of degradation or running down or a trend to disorder – Merriam-Webster



after Michener et al., 1997

LTER Tiered Trajectory for Metadata

Tiered Trajectory

Metadata completeness →

Metadata structure ↓

	Discovery	Access	Usability
Tier 1	• Unstructured, online site catalog with minimal metadata • Data discovery through manual searches	• Establish data access policy • Data and metadata access requires human intervention	• Unstructured, machine-readable metadata and data • Data use requires human intervention
Tier 2	• Online, enhanced metadata with consistent internal structure • Data discovery through machine search	• Automated access to data • Access to site data and metadata does not require human intervention	• Structured, comprehensive metadata and data • Data use does not require human intervention
Tier 3	• Discovery-enabling metadata structured in EML • Data discovery integrated across network	• Access-enabling metadata structured in EML • Data access is integrated across network	• Complete validated EML • Data analysis is integrated across the network
Future Outcome	Semantic-based discovery through machine-based searches	Data access through a knowledge-based query process	Semi-automated knowledge extraction

Best Practices – Metadata completeness

- Identification
- Discovery
- Evaluation
- Access
- Integration
- Semantic Use

Completeness Level	Description and Major Elements Added
1: Identification	Minimum content for adequate data set discovery in a general cataloging system or repository (functionally equivalent to LTER DTOC): • title • creator • contact • publisher • pubDate • keywords • abstract (recommended) • dataset/distribution (i.e. url for general dataset information)
2: Discovery	Level 1 content, plus coverage information to support targeted searches • geographicCoverage • taxonomicCoverage • temporalCoverage
3: Evaluation	Level 2 content, plus data set details to enable end-user evaluation of the methodology and data entities: • project • methods • entity • attributes (strongly recommended, as possible) • intellectualRights
4: Access	Level 3 content plus data access details to support computer-assisted data retrieval: • access • physical
5: Integration	Level 4 content plus complete attribute and QA/QC details to support computer-assisted data integration and re-sampling • attribute (required) • measurementScale • units • constraint • qualityControl
6: Semantic Use	Level 5 content plus semantic information (currently under development by SEEK, and may require extension to the EML schema)

Level 1 - Identification

- ***Description*** – Minimum content for adequate data set discovery
- ***Major Elements Added*** **:**
  - Title
  - Creator
  - Contact
  - Publisher
  - Publication Date
  - Keywords
  - Abstract
  - Dataset/distribution (i.e. URL for dataset information)

Level 1 Code Example

```
<?xml version="1.0" encoding="UTF-8"?>
<eml:eml xmlns:eml="eml://ecoinformatics.org/eml-2.0.1"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="eml://ecoinformatics.org/eml-2.0.1
    http://someserver.fls.edu/eml-2.0.1/eml.xsd"
    packageId="knb-lter-fls.1.1" system="FLS” scope="system">

     <dataset id="FLS-1" system="FLS">
         <alternateIdentifier>FLS-1</alternateIdentifier>
         <shortName>Arthropods</shortName>
         <title>
             Long-term Ground Arthropod Monitoring Dataset at
             Silver City, NM USA from 1998 to 2004
         </title>
                  . . .
     </dataset>
```

Level 1 Code Example cont.

```
<creator id="pers-1" system="FLS">
    <individualName>
        <givenName>John</givenName>
        <surName>Ecologist</surName>
    </individualName>
    <organizationName>FLS LTER</organizationName>
    <address id="addr-1" system="FLS">
        <deliveryPoint>Department of Ecology</deliveryPoint>
        <deliveryPoint>University of New Mexico</deliveryPoint>
        <deliveryPoint>PO Box 1234</deliveryPoint>
        <city>Albuquerque</city>
        <administrativeArea>NM</administrativeArea>
        <postalCode>87131-1234</postalCode>
    </address>
    <phone phonetype="voice">(505) 999-9999</phone>
    <electronicMailAddress>jeco@unm.edu</electronicMailAddress>
    <onlineUrl>http://www.unm.edu/~jeco</onlineUrl>
</creator>
```

Level 2 - Discovery

- ***Description*** – Level 1 content, plus coverage information to support targeted searches
- ***Major Elements Added*** **:**
  - Geographic Coverage
  - Taxonomic Coverage
  - Temporal Coverage

Level 2 Code Example

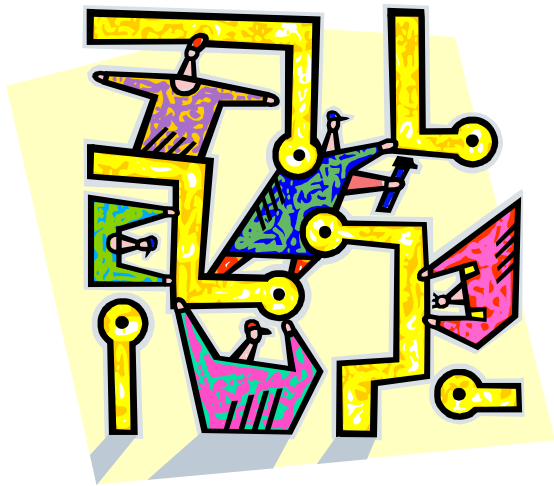
```
<coverage>
    <geographicCoverage>
        <geographicDescription>
            Silver City, NM USA
        </geographicDescription>
        <boundingCoordinates>
            <westBoundingCoordinate>-112.373634</westBoundingCoordinate>
            <eastBoundingCoordinate>-111.612936</eastBoundingCoordinate>
            <northBoundingCoordinate>+33.708829</northBoundingCoordinate>
            <southBoundingCoordinate>+33.298975</southBoundingCoordinate>
            <boundingAltitudes>
                <altitudeMinimum>304</altitudeMinimum>
                <altitudeMaximum>627</altitudeMaximum>
                <altitudeUnits>meter</altitudeUnits>
            </boundingAltitudes>
        </boundingCoordinates>
    </geographicCoverage>
...
```

Level 2 Code Example cont.

```
...
    <temporalCoverage>
        <rangeOfDates>
            <beginDate>
                <calendarDate>1998-11-12</calendarDate>
            </beginDate>
            <endDate>
                <calendarDate>2003-12-31</calendarDate>
            </endDate>
        </rangeOfDates>
    </temporalCoverage>
    <taxonomicCoverage>
        <generalTaxonomicCoverage>
            Orthopteran insects (grasshoppers) were id using
            the 2004 BigKey to Orthoptera
        </generalTaxonomicCoverage>
        <taxonomicClassification>
            <taxonRankName>Kingdom</taxonRankName>
            <taxonRankValue>Animalia</taxonRankValue>
            <taxonomicClassification>
                <taxonRankName>Phylum</taxonRankName>
                <taxonRankValue>Arthropoda</taxonRankValue>
            </taxonomicClassification>
        </taxonomicClassification>
    </taxonomicCoverage>
</coverage>
```

Level 3 - Evaluation

- ***Description*** – Level 2 content, plus data set details to enable end-user evaluation of the methodology and data entities
- ***Major Elements Added*** **:**
  - Intellectual Rights
  - Project
  - Methods
  - Data Table/Entity Group
  - Data Table/Attributes (constrained by current version of EML)



Level 3 Code Example

```
<intellectualRights>
    <section>
        <para>
            The dataset is released to the public and
            may be used for academic or commercial purposes
            subject to the following restrictions:
        </para>
        <para>
            <itemizedlist>
                <listitem>
                    <para>
                        LTER will make every effort possible
                        to control and document the quality of
                        the data it publishes. Data are made
                        available "as is"...
                    </para>
                </listitem>
                ...
            </itemizedlist>
        </para>
    </section>
</intellectualRights>
```

Level 3 Code Example cont.

```
...
<project>
    <title>Fictitious LTER Site (FLS) permanent monitoring program</title>
    <personnel id="pers-30" system="FLS">
        <individualName>
            <salutation>Dr.</salutation>
            <givenName>Eva</givenName>
            <surName>Scientist</surName>
        </individualName>
        <address>
            <reference>addr-1</reference>
        </address>
        <role>principalInvestigator</role>
    </personnel>
    <abstract>
       <para>
           The FLS basic monitoring program consists of monitoring of
           arthropod populations, plant net primary productivity, and bird
           populations. Monitoring takes place at 3 sites, 4 times a year.
           Climate parameters are continuously measured at all stations.
       </para>
    </abstract>
</project>
```

Level 3 Code Example cont.

```
<methods>
    <methodStep>
        <description>
            <para>
                FSL Protocol for Surveying Ground Arthropods has been...
            </para>
        </description>
        <protocol>
            <title>
                FLS Protocol for Surveying Ground Arthropods
            </title>
            <creator>
                <references>pers-1</references>
            </creator>
            <pubDate>2000-02-23</pubDate>
            <abstract>
                <para>
                    This protocol is being used by FLS arthropod...
                </para>
            </abstract>
            <keywordSet>
                <keyword keywordType="theme">Ecology</keyword>
              ...
            </keywordSet>
            <distribution>
                <online>
                    <url>http://fls.univ.edu/protocols/arthro.html</url>
                </online>
            </distribution>
        </protocol>
    </methodStep>
    ...
```

Level 3 Code Example cont.

```
    <methodStep>
        <instrumentation>
            SBE MicroCAT 37-SM (S/N 1790); manufacturer: Sea-Bird
            Electronics (model: 37-SM MicroCAT); parameter: Conductivity
            (accuracy: 0.0003 S/m, readability: 0.00001 S/m, range:
            0 to 7 S/m); last calibration: Feb 28, 2001
        </instrumentation>
        <instrumentation>
            SBE MicroCAT 37-SM (S/N 1790); manufacturer: Sea-Bird
            Electronics (model: 37-SM MicroCAT); parameter: Pressure (water)
            (accuracy: 0.2m, readability: 0.0004m, range: 0 to 20m); last
            calibration: Feb 28, 2001
        </instrumentation>
        <instrumentation>
            SBE MicroCAT 37-SM (S/N 1790); manufacturer: Sea-Bird
            Electronics (model: 37-SM MicroCAT); parameter: Temperature
            (water)(accuracy: 0.002C, readability: 0.0001C, range: -5
            to 35C); last calibration: Feb 28, 2001
        </instrumentation>
    </methodStep>
        ...
</methods>
```

Level 3 Code Example cont.

```
...
<dataTable>
    <entityName>arthro_hab</entityName>
    <entityDescription>
         Habitat description for the sampling locations
    </entityDescription>
    <attributeList>
        <attribute>
            <attributeName>temp</attributeName>
            <attributeDefinition>Water Temperature</attributeDefinition>
            <storageType>float</storageType>
            <measurementScale>
                <interval>
                    <unit>
                        <standardUnit>celsius</standardUnit>
                    </unit>
                    <precision>0.001</precision>
                    <numericDomain>
                        <numberType>real</numberType>
                    </numericDomain>
                </interval>
            </measurementScale>
            <missingValueCode>
                <code>NaN</code>
                <codeExplanation>
                    value not recorded or invalid
                </codeExplanation>
            </missingValueCode>
        </attribute>
        ...
```

Level 3 Code Example cont.

```
    <attribute>
        <attributeName>cond</attributeName>
        <attributeLabel>Conductivity</attributeLabel>
        <attributeDefinition>
            measured with SeaBird Electronics CTD-911
        </attributeDefinition>
        <storageType>float</storageType>
        <measurementScale>
            <ratio>
                <unit>
                    <customUnit>siemensPerMeter</customUnit>
                </unit>
                <precision>0.0001</precision>
                <numericDomain>
                    <numberType>real</numberType>
                    <bounds>
                        <minimum exclusive="false">0</minimum>
                        <maximum exclusive="false">40</maximum>
                    </bounds>
                </numericDomain>
            </ratio>
        </measurementScale>
    </attribute>
</attributeList>
...
```

Level 3 Code Example cont.

```
...
<addtionalMetadata>
    <unitList>
        <unit id="siemensPerMeter" name="siemensPerMeter"
         unitType="conductance" parentSI="siemen" multiplerToSI="1">
            <description>
                electrical conductance of a solution (conductivity)
            </description>
        </unit>
    </unitList>
</additionalMetadata>
...
```

Level 4 - Access

- ***Description*** – Level 3 content plus data access details to support automated data retrieval
- ***Major Elements Added*** **:**
  - Access
  - Physical

Level 4 Code Example

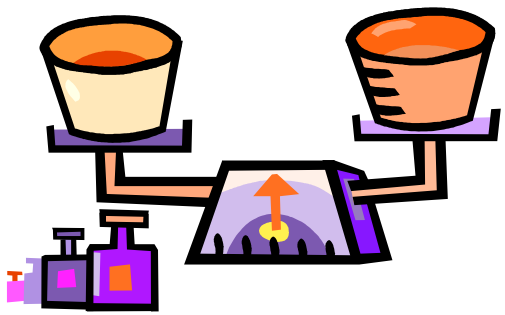
```
<access authSystem="FLS">
    <allow>
        <principal>PUBLIC</principal>
        <permission>read</permission>
    </allow>
    <allow>
        <principal>uid=fls,o=LTER,dc=ecoinformatics,dc=org </principal>
        <permission>all</permission>
    </allow>
</access>
```

Level 4 Code Example

```
<dataTable>
    ...
    <physical>
        <objectName>flslter.299.1</objectName>
        <size unit="bytes">59847</size>
        <dataFormat>
            <textFormat>
                <numHeaderLines>1</numHeaderLines>
                <attributeOrientation>column</attributeOrientation>
                <simpleDelimited>
                    <fieldDelimiter>,</fieldDelimiter>
                </simpleDelimited>
            </textFormat>
        </dataFormat>
        <distribution>
            <online>
                <url>http://fls.unm.edu/flslter.296.1</url>
            </online>
        </distribution>
    </physical>
    ...
```

Level 5 - Integration

- ***Description*** – Level 4 content plus complete attribute and quality control details to support computer-assisted data integration and re-sampling; Integration-level metadata should support computer-mediated access and processing of data, and therefore requires that all aspects of the data package be fully described.
- ***Major Elements Added*** **:**
  - Attribute List (full descriptions)
  - Measurement Scale
  - Units
  - Constraint
  - Quality Control



Level 5 Code Example

```
...
<constraint id="pkarthro_taxa">
    <primaryKey>
        <constraintName>pkarthro_taxa</constraintName>
        <key>
            <attributeReference>dbo.arthro_taxa.taxon</attributeReference>
        </key>
    </primaryKey>
</constraint>
<constraint id="arthro_taxa.taxonNotNull">
    <notNullConstraint>
        <constraintName>arthro_taxa.taxonNotNull</constraintName>
        <key>
            <attributeReference>dbo.arthro_taxa.taxon</attributeReference>
        </key>
    </notNullConstraint>
</constraint>
...
```

Level 5 Code Example

```
</measurementScale>
    <method>
        <qualityControl>
            <description>
                <para>
                    Passage of clouds during a profile reduces the incident
                    radiation, and leads to erroneous estimates of Kd.
                    Variation of incident irradiance was described in two
                    ways (before binning): 1) the coefficient of variation
                    (cv) over the 10m depth interval, and 2) difference...
                </para>
            </description>
        </qualityControl>
    </method>
    ...
```

Level 6 - Semantic

- ***Description*** – Level 5 content plus semantic information (currently under development by SEEK, and may require extension to the EML schema)

Additional Recommendations

- packageID and Metacat document naming convention

  Metacat and by extension the Metacat harvester rely on numerical data set ids and revision numbers for document management and synchronization - packageId attributes for EML contributed to the KNB Metacat should be formed as follows:

  knb-lter-[site].[dataset number].[revision], e.g. knb-lter-sev.187.4
  Scope **UniqueID** **Revision#**

- LDAP access control in Metacat

  Metacat access control format conforms to the LDAP Distinguished Name concept:

  <principal>uid=FLS,o=lter,dc=ecoinformatics,dc=org</principal>

- Organizational citation

  The “Organization” field on the Metacat query results page is populated using the first eml:eml/dataset/creator/organizationName element in the document, so it is recommended that for LTER-contributed data sets the LTER site be included as the first creator:

  <organizationName>Sevilleta LTER</organizationName>

Credits

- James Brunt (LNO)
- Corinna Gries (CAP)
- Jeanine McGann (LNO)
- Margaret O’Brien (SBC)
- Ken Ramsey (JRN)
- Wade Sheldon (GCE)

Acknowledgements

This material is based upon work supported by:

The National Science Foundation under Grant Numbers 9980154, 9904777, 0131178, 9905838, 0129792, and 0225676.

The National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant Number 0072909), the University of California, and the UC Santa Barbara campus.

The Andrew W. Mellon Foundation.

PBI Collaborators: NCEAS, University of New Mexico (Long Term Ecological Research Network Office), San Diego Supercomputer Center, University of Kansas (Center for Biodiversity Research)

Kepler contributors: SEEK, Ptolemy II, SDM/SciDAC, GEON