Structured Metadata at EDI

2018 March
Environmental Data Initiative (EDI)
Why Use a Metadata Standard?

A Standard provides a structure to describe data with:

• Common terms to allow consistency between records
• Common structure to quickly locate components

In search and retrieval, standards provide:

• Reliable, predictable format for computer interpretation
• A uniform summary description of the dataset
Metadata standards are optimized

• Content

Most metadata standards include:

Who – information on who to contact
What – a description of the available item

Some include:

Where – geographical locations
When – dates and times

• Examples

Darwin Core - organism occurrences (GBIF)
ISO 19115/19139 - geospatial data
XML: a set of hierarchical custom elements for a particular community’s use

Relatively verbose, and requires larger storage than some other formats

Does not have strong data typing or access control (on its own)

Language requires some training, manual editing can be tedious

Common exchange format for web services

Easy programmatic access

Platform-independent, both human- and machine-readable

Editing tools are improving
XML Basics

XML Schema

- describes the structure of an XML document
- "XML Schema language" - also referred to as "XML Schema Definition" or XSD

So XML is (kind of) a language, and EML is a dialect
A community will write up its specification for a standard in XML Schema
# XML Basics

<table>
<thead>
<tr>
<th>General:</th>
<th>Components:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prolog</td>
<td>XML element - contains text, other elements</td>
</tr>
<tr>
<td>“Root” element</td>
<td>XML attribute - text only</td>
</tr>
<tr>
<td>Validate (an XML doc)</td>
<td>Namespace</td>
</tr>
<tr>
<td></td>
<td>Schema, schemaLocation</td>
</tr>
</tbody>
</table>

A little bit of XML vocabulary:

```xml
<?xml version="1.0" encoding="UTF-8"?>
<eml:eml xmlns:eml="eml://ecoinformatics.org/eml-2.1.1"
  xmlns:stmml="http://www.xml-cml.org/schema/stmml-1.1"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  packageId="knb-lter-sbc.19.22" scope="system" system="knb"
```
Ecological Metadata Language - EML

- Based on Dublin Core, FGDC, STMML
- Rich, customized structures
- Widely used for ecological and environmental data
- Machine readable
  - Read into Matlab, SAS, R, SPSS
  - Parser and DB loader
- EML records number ~90,000
EML Anatomy

- Access rules
- Dataset metadata
  - Identifier
  - *Title
  - *Creator(s)
  - Other metadata contributors
  - Associated parties
  - Publication date
  - Language
  - Abstract
  - Keywords
  - Intellectual rights statement
  - Geographic, temporal and taxonomic coverage
  - *Contact
  - Publisher
  - Methods
  - Project
  - Data table (with its own sub-hierarchy)
- Other metadata
**REQUEST DATA:**

Available Online:
Download data after acceptance of SSC LTER Data Use Agreement:

**DOWNLOAD DATA: LTE Benthic inverts and understory algae, all years**

Name: LTE Benthic inverts and understory algae, all years

Description: Abundance and size of selected species of benthic invertebrates and understory algae in fixed plots along permanent transects in kelp removal experiment

Number of Records: 271982

Number of Columns: 29

**Table Structure:**

Object: LTE_Quad_Swath_All_Years_20160315.csv
Name: 55 megabyte
Size: Number of Header Lines: 1
Record Delimiter: \n
Orientation: column
Simple Delimited: no Delimiters: Field Delimiter:
Quote Character: "

**Table Column Descriptions**

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>Calendar year</td>
</tr>
<tr>
<td>Month</td>
<td>Month of data collection</td>
</tr>
<tr>
<td>Date</td>
<td>Date of data collection</td>
</tr>
<tr>
<td>Site code</td>
<td>Code for reef site. See list of codes for meaning</td>
</tr>
<tr>
<td>Transect name</td>
<td>40 m transects defined by six permanent markers (stainless steel eyebolts or rebar stakes) at 0, 8, 16</td>
</tr>
<tr>
<td>Treatment</td>
<td>Experimental, frequency of removal</td>
</tr>
<tr>
<td>Quadrat Name</td>
<td>Data collected in quadrats of 20mx1m called 20 or 40 (0-20, or 20-40m, also known as swaths), or 1mx1m quadrats at 0-40m</td>
</tr>
<tr>
<td>Transect side</td>
<td>Side of the transect the data was taken on. Either 'T' (inshore) or 'O' (offshore).</td>
</tr>
<tr>
<td>Species code</td>
<td>2-4 letter code used by SBC LTER. Numbers are used to differentiate size classes for Macrocycts.</td>
</tr>
<tr>
<td>Count</td>
<td>Number of individuals in the area surveyed.</td>
</tr>
<tr>
<td>Observer code</td>
<td>Numeric code indicating the SBC LTER data collector</td>
</tr>
</tbody>
</table>
Resources

Dataset in examples, “SBC LTER Time-Series of Kelp Forest Invertebrate and Algal Density”
https://portal.edirepository.org/nis/mapbrowse?scope=knb-lter-sbc&identifier=19

Best Practices for EML datasets (originated by LTER IMC)

EML Project
Download the current release, read basic documentation: https://www.nceas.ucsb.edu/ecoinfo/tools
Maintained in GitHub: https://github.com/NCEAS/eml

Find resources at:
https://environmentaldatainitiative.org > Resources > 5 Phases of data publishing > Phase 3. create EML metadata

https://www.liquid-technologies.com/xml-tutorial https://www.w3schools.com/xml/
Data Table metadata

**Data Table**
- Column name: exactly as it appears in the dataset. Please avoid special characters, dashes and spaces.
- Description: please be specific, it can be lengthy
- Unit: please avoid special characters and describe units in this pattern: e.g. microgramPerCentimeter, microgramsPerLiter, AbsorptionPerMolePerCentimeter
- Code explanation: if you use codes in your column, please explain in this way: e.g. LR=Little Rock lake, A=Sample suspect, J=Nonstandard routine followed
- Date format: please tell us exactly how the date and time is formatted: e.g. mm/dd/yyyy plus the time zone and whether or not daylight savings was observed.
- If a code for ‘no data’ is used, please specify: e.g. -99999
- Please add rows as needed

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes and Comments**

Some components can be automated
Attribute list: your knowledge of the data is essential
Attributes & Units

- **Attribute**: a “property” of an object (data table)
  - In databases, a table column is called an “attribute”
  - Often referred to as “variables”, “parameters”, “columns” or “field names”

- **Unit**: a particular physical quantity
  - Defined and adopted by convention
  - Comparable

To describe a data table you need a moderate understanding of:

A. **how to define the table’s attributes**,  
B. **when and how to define a unit**, and  
C. **the relationship between the two**.
<table>
<thead>
<tr>
<th><strong>1 Attribute Name:</strong></th>
<th>Usually the name you would give that column in a script</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2 Attribute Label:</strong></td>
<td>Longer, for display. Use whole words, capitals, etc.</td>
</tr>
<tr>
<td><strong>3 Attribute Definition:</strong></td>
<td>As complete and unambiguous as you need them to be, for the data to be understood</td>
</tr>
</tbody>
</table>
| **4 Measurement Scale:** | - **Nominal** - attribute can be considered a category  
- **Ordinal** - categories that have a logical or ordered relationship to one another  
- **Interval** - the magnitude between the steps is known; equidistant points  
- **Ratio** - have a meaningful zero, which allows ratios between values to have meaning  
- **Datetime** - Gregorian dates and times |
| **5 Unit:** | Interval and Ratio measurements only  
Choose from: [http://unit.lternet.edu](http://unit.lternet.edu) |
# EML Attributes - Measurement Scale

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Values are members of a category</td>
<td>string</td>
<td>Place and taxon names, coded values (e.g., 1=male, 2=female), text comments</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Nominal categories that have a logical or ordered relationship to one another</td>
<td>string</td>
<td>Academic grades, quality rankings (e.g., 1=high, 2=medium, 3=low)</td>
</tr>
<tr>
<td>Interval</td>
<td>Ordinal, but the magnitude between the steps is known; equidistant points</td>
<td>numeric</td>
<td>Celsius scale, pH</td>
</tr>
<tr>
<td>Ratio</td>
<td>Interval, with a meaningful zero, so ratios between values to have meaning</td>
<td>numeric</td>
<td>Temperature in Kelvin, lengths, concentrations, organism densities</td>
</tr>
<tr>
<td>Datetime</td>
<td>Gregorian dates and times</td>
<td>datetime</td>
<td>Points in time, e.g., with formats like YYYY-MM-DD, hh:mm:ss.s</td>
</tr>
</tbody>
</table>
## EML Attributes - code lists

| Nominal | Values are members of a category | Place and taxon names, coded values (eg, 1=male, 2=female) | <codeDefinition>
<code>ABUR</code>
<definition>Arroyo Burro Reef</definition>
</codeDefinition>
<codeDefinition>
<code>NAPL</code>
<definition>Naples Reef</definition>
</codeDefinition> |
|---------|---------------------------------|----------------------------------------------------------|---------------------------------------------------------------|
| Ordinal | Nominal categories that have a logical or ordered relationship to one another | Academic grades, quality rankings (eg, 1=high, 2=medium, 3=low) | <codeDefinition>
<code>A</code>
<definition>scored higher than 90%</definition>
</codeDefinition>
<codeDefinition>
<code>B</code>
<definition>score 80 - 89%</definition>
</codeDefinition>
<codeDefinition>
<code>C</code>
<definition>score 70 - 79%</definition>
</codeDefinition> |