Structured Metadata at EDI

2018 March
Environmental Data Initiative (EDI)

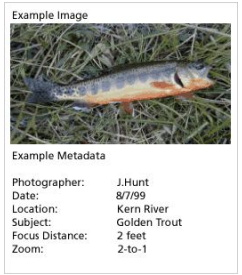
Why Use a Metadata Standard?

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:

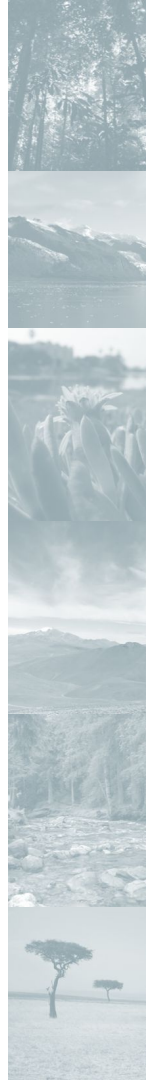
  Where *– geographical locations*

  When *– dates and times*

- Examples

  Darwin Core - organism occurrences (GBIF)

  ISO 19115/19139 - geospatial data



XML

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 it's 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

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <eml:eml xmlns:eml="eml://ecoinformatics.org/eml-2.1.1"
3   xmlns:stmml="http://www.xml-cml.org/schema/stmml-1.1"
4   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
5   packageId="knb-lter-sbc.19.22" scope="system" system="knb"
6   xsi:schemaLocation="eml://ecoinformatics.org/eml-2.1.1 http://nis.lternet.edu/schemas/EML/eml-2.1.1/eml.xsd">
```

A little bit of XML vocabulary:

General:	Components:
Prolog “Root” element Validate (an XML doc)	XML element - contains text, other elements XML attribute - text only Namespace Schema, schemaLocation

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

knb

LTER NETWORK

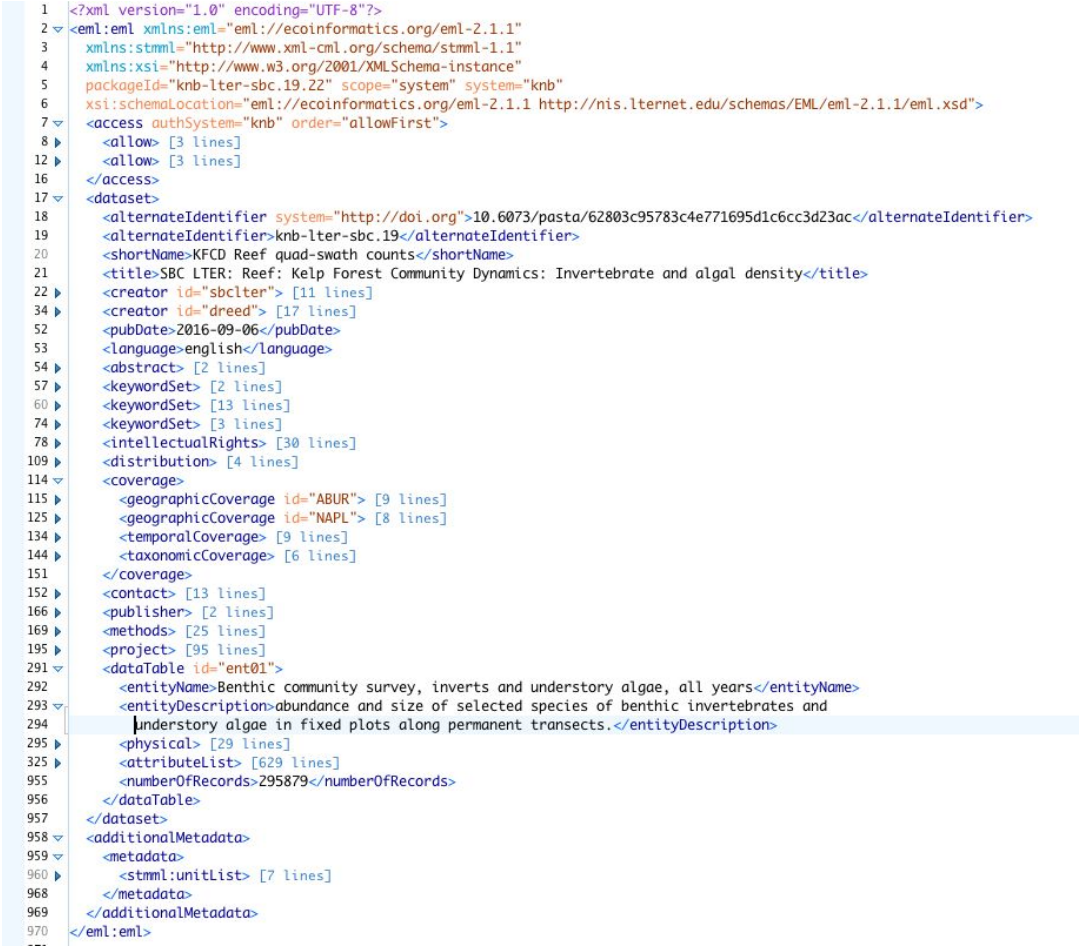
ILTER

NCEAS

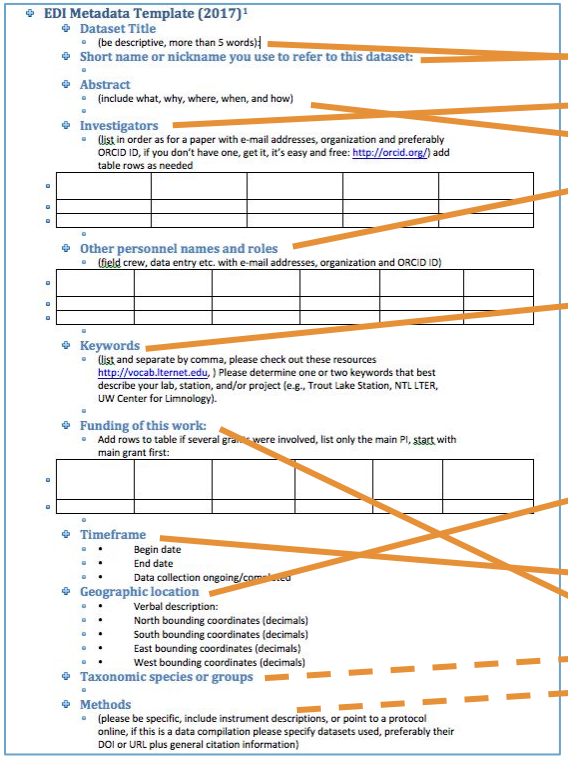
National Center for Ecological Analysis and Synthesis

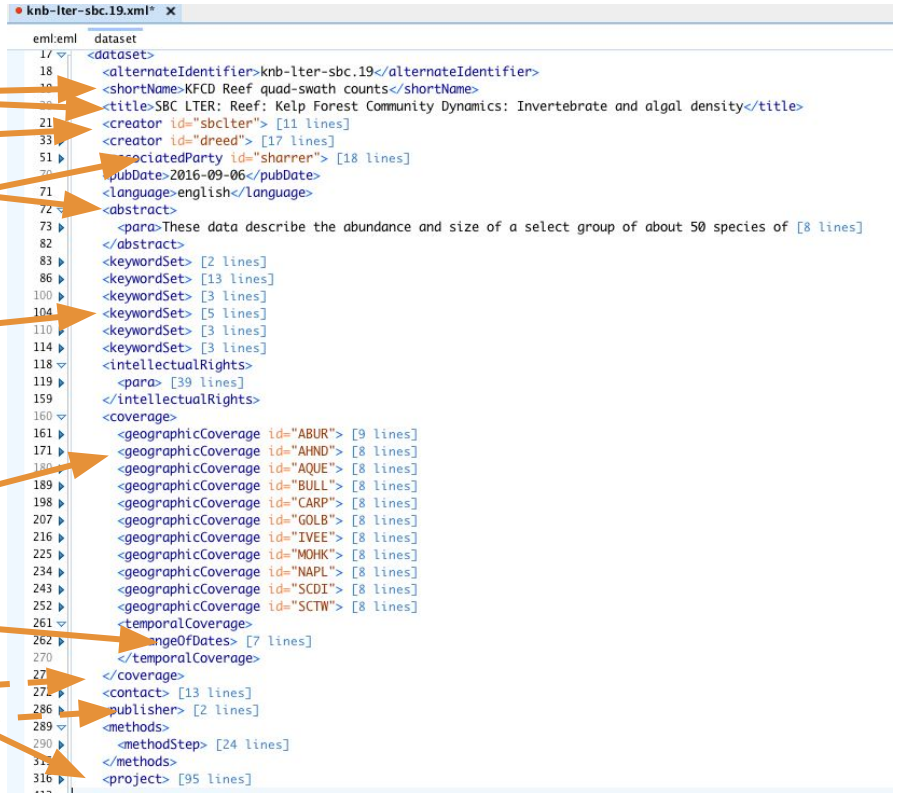
EML Anatomy

- 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



Metadata Template





REQUEST DATA:

Available Online:

Download data after acceptance of SBC 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 Name:	LTE_Quad_Swath_All_Years_20160315.csv	
Size:	55 megabyte	
Text Format:	Number of Header Lines:	1
	Record Delimiter:	\r\n
	Orientation:	column
	Simple Delimited:	Field Delimiter: ,
		Collapse Delimiters: no
		Quote Character: "

Table Column Descriptions

	Year	Month	Date	Site code	Transect name	Treatment	Quadrat Name	Transect side	Species code	Count	Observer code
Column Name	YEAR	MONTH	DATE	SITE	TRANSECT	TREATMENT	QUAD	SIDE	SP_CODE	COUNT	OBS_CODE
Definition	Calendar year	Month of data collection	Date of data collection	Code for reef site. See list of codes for meaning	40m transects defined by six permanent markers (stainless steel eyebolts or rebar stakes) at 0, 8, 16,	Experimental treatment, frequency of removal	Data collected in quadrats of 20mx1m called 20 or 40 (0-20, or 20-40m, also known as swaths), or 1mx1m quadrats at	Side of the transect the data was taken on. Either 'I' (inshore) or 'O' (offshore).	2-4 letter code used by SBC LTER. Numbers are used to differentiate size classes for Macrocystis.	Number of individuals in the area surveyed.	Numeric code indicating the SBCLTER data collector

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/

Image credit: https://thedeependbristol.wordpress.com

Data Table metadata

Some components can be automated

Attribute list: your knowledge of the data is essential

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. microSiemenPerCentimeter, microgramsPerLiter, absoptionPerMolePerCentimeter
- 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
- Data format: please tell us exactly how the date and time is formatted: e.g. mm/dd/yyyy hh:mm:ss 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

Notes and Comments

knb-lter-sbc.19.xml*

eml:eml dataset dataTable attributeList

```
<dataTable id="ent01">
  <entityName>Benthic community survey, inverts and understory algae, all years</entityName>
  <entityDescription>abundance and size of selected species of benthic invertebrates
    and understory algae in fixed plots along permanent transects.</entityDescription>
  <physical> [20 lines]
  <attributeList>
    <attribute id="ent1.att3"> [15 lines]
    <attribute id="ent1.att4"> [96 lines]
    <attribute id="ent1.att8">
      <attributeName>sp_code</attributeName>
      <attributeLabel>Species code</attributeLabel>
      <attributeDefinition>2-4 letter code generally composed of the first letter of the
        genus (G) name and the first three letters of the species (S) name (format=GSSS)</attributeDefinition>
      <storageType typeSystem="http://www.w3.org/2001/XMLSchema-datatypes">string</storageType>
      <measurementScale>
        <nominal>
          <nonNumericDomain>
            <textDomain>
              <definition>any text</definition>
            </textDomain>
          </nonNumericDomain>
        </nominal>
      </measurementScale>
      <missingValueCode>
        <code>-99999</code>
        <codeExplanation>no information available</codeExplanation>
      </missingValueCode>
    </attribute>
    <attribute id="ent1.att9">
      <attributeName>count</attributeName>
      <attributeLabel>Count</attributeLabel>
      <attributeDefinition>number of individuals in area</attributeDefinition>
      <storageType typeSystem="http://www.w3.org/2001/XMLSchema-datatypes">integer</storageType>
      <measurementScale>
        <ratio>
          <unit>
            <customUnit>number</customUnit>
          </unit>
          <precision>1</precision>
          <numericDomain>
            <numberType>real</numberType>
          </numericDomain>
        </ratio>
      </measurementScale>
      <missingValueCode>
        <code>-99999</code>
        <codeExplanation>no information available</codeExplanation>
      </missingValueCode>
    </attribute>
  </attributeList>
  <numberOfRecords>295879</numberOfRecords>
</dataTable>
```

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.

EML Attribute components

1 Attribute Name:
Usually the name you would give that column in a script

2 Attribute Label:
Longer, for display, Use whole words, capitals, etc.

3 Attribute Definition:
As complete and unambiguous as you need them to be, for the data to be understood

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

EML Attributes - Measurement Scale

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

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>`
			`...`