Annotation in EML 2.2

EML Dev Committee
2018
EML 2.2 New Features

Backward compatible

EML you create now (2.1, 2.0) is fine as is or re-label as EML 2.2
EML 2.2 New Features

● Choice of TextType or markdown
● New elements for data papers
● Annotations at 3 levels
● Taxonomic classification has IDs
● Project tree includes more data background
  ○ e.g., data papers
● Units
  ○ Expanded list
  ○ Name strings standardized and aligned with LTER recommendations
  ○ conversion with udunits2 synonyms
EML 2.2 New Features

Collectively:

- Datasets will be easier to search
- Allow targeted queries, e.g., funding codes
- EML datasets can approach completeness of data papers
EML 2.2 - EDI

EDI will support EML 2.2
Display stylesheets
Congruence checking to be worked out
EML 2.2 New Features

Advanced features will take additional attention

LTER should plan to

- Identify the most important datasets, deserving specialized attention
- Set priorities - science-driven
Some intro text in abstract, then break into subsections.

## Level 2 heading

We use a level 2 heading because Level 1 would be at the same level as the main sections of the paper.

## Another level 2 heading

With some information.

Plus, it can include all of the other features of [Github Flavored Markdown (GFM)](https://github.github.com/gfm/).

Note that this version of GFM is a superset of CommonMark, and is intended to eventually be an official extension of CommonMark.

Follow: https://github.com/NCEAS/eml/issues/275
Taxonomic Classification

Follow: https://github.com/NCEAS/eml/issues/141
Project Tree - Additions

EML 2.1 -><br/>

```
<funding>
  <section>
    <para>wholly or in part by NSF Awards OCE-1233283, 1233288, 1233839</para>
  </section>
</funding>
```

Follow: https://github.com/NCEAS/eml/issues/266

EML 2.2 -><br/>

```
<project>
  ...
  <funding>
    <para>Funding is from a grant from the National Science Foundation.</para>
  </funding>
  <award>
    <funderName>National Science Foundation</funderName>
    <funderIdentifier>https://doi.org/10.13039/00000001</funderIdentifier>
    <awardNumber>1546024</awardNumber>
    <title>SciLine Arctic: A Knowledge Archive for Discovery and Reproducible Science in the Arctic</title>
  </award>
</project>
```
Fields to Support Data Papers

EML 2.1 includes...

- /eml/dataset/title
- /eml/dataset/creator
- /eml/dataset/keywordSet
- /eml/dataset/coverage/geographicCoverage
- /eml/dataset/coverage/temporalCoverage
- /eml/dataset/abstract
- /eml/dataset/methods/sampling/studyExtent/description
- /eml/dataset/methods/sampling/samplingDescription
- /eml/dataset/methods/methodStep
- /eml/dataset/methods/qualityControl
- /eml/dataset/creator/@userId
- /eml/dataset/creator/electronicMailAddress
- /eml/dataset/creator/address

EML 2.2 adds...

- Citation /eml/dataset/referencePublication
- Authors //citation/creator (consistent with /eml/dataset/creator)
- Year //citation/pubDate (consistent with /eml/dataset/pubDate)
- Title //citation/title
- Journal //citation/article/journal
- Journal number //citation/article/volume
- Paper number (from Ecosphere) //citation/article/issue
- DOI /eml/@packageId
- Introduction /eml/dataset/introduction
- Data synopsis /eml/dataset/purpose, expanded for tables, figs
- Getting started /eml/dataset/gettingStarted
- Acknowledgements /eml/dataset/acknowledgements
- Literature cited /eml/dataset/literatureCited
- Literature citing the data /eml/dataset/usageCitation

Follow:
https://github.com/NCEAS/eml/issues/269,
Follow: https://github.com/NCEAS/eml/issues/289
Annotations

```
<attribute>
  <attributeName>NEE</attributeName>
  <attributeDefinition>net ecosystem exchange</attributeDefinition>
  <measurementScale>
    ...
  </measurementScale>
  <missingValueCode>
    ...
  </missingValueCode>
</attribute>
```

Follow: https://github.com/NCEAS/eml/issues/25
**Net Ecosystem Exchange Carbon Flux**

- **Preferred Name**: Net Ecosystem Exchange Carbon Flux
- **Synonyms**: NEE, Net Ecosystem Exchange
- **Definitions**: The magnitude of carbon sources and sinks is defined as the vertical exchange of CO2 between the surface (land or ocean) and the atmosphere.
- **ID**: http://purl.dataone.org/odo/ECSO_00000014
- **alternative label**: NEE, Net Ecosystem Exchange
- **definition**: The magnitude of carbon sources and sinks is defined as the vertical exchange of CO2 between the surface (land or ocean) and the atmosphere.
- **definition_Contributor**: Chase LeCroy, orcid.org/0000-0002-1338-9436
- **measuresEntity**: carbon atom
- **preferred label**: Net Ecosystem Exchange Carbon Flux
- **prefixRI**: odo:ECSO_00000014
- **rdfs:label**: Net Ecosystem Exchange Carbon Flux
- **seeAlso**: NEE
- **subClassOf**: Carbon Flux
Next Steps

LTER & EDI
EML Best Practices updated
  Data Papers
  Annotation
Next Steps

Advanced features will take additional attention

LTER should plan to

- Identify the most important datasets, deserving specialized attention
- Set priorities - science-driven
RDF Triple

Subject → Predicate → Object
RDF Triple

“dsid_01.att_01” “contains measurement of type” “Net Ecosystem Exchange Carbon Flux”
RDF Triple

Subject: “dsid_01.att_01"
Predicate: “contains measurement of type"
Object: “Net Ecosystem Exchange Carbon Flux”
RDF Triple

Subject: "dsid_01.att_01"
Predicate: "contains measurement of type"
Object: "Net Ecosystem Exchange Carbon Flux"
Annotation - attribute

```xml
<attribute id="dsid_01.att_01">
  <attributeName>NEE</attributeName>
  <attributeDefinition>net ecosystem exchange</attributeDefinition>
  <measurementScale>
    <interval>
      <unit>
        <customUnit>micromolsPerSquareMeterPerSecond</customUnit>
      </unit>
      <numericDomain>
        <numberType>real</numberType>
      </numericDomain>
    </interval>
  </measurementScale>
  <missingValueCode>
    <code>-9999</code>
    <codeExplanation>NA</codeExplanation>
  </missingValueCode>
  <annotation>
    <propertyURI label="contains measurements of type">http://ecoinformatics.org/obo/obo.1.2/obo-core.owl#containsMeasurementsOfType</propertyURI>
    <valueURI label="Net Ecosystem Exchange Carbon Flux">http://purl.dataone.org/odo/ECS0_00000014</valueURI>
  </annotation>
</attribute>
```
Annotation - dataset

"knb-lter-bnz.373.17"  "is about"  "Forest Biome"

<!-- dataset level, last element in the Resource Group, after coverage -->
<annotation>
  <propertyURI label="is about">http://purl.obolibrary.org/obo/IAO_00000136</propertyURI>
  <valueURI label="Forest Biome">http://purl.obolibrary.org/obo/ENVO_01000174</valueURI>
</annotation>

<contact>
Annotation - entity

```
<dataTable id="dsid_01.dt_01">
  <entityName>BEC_2005_12JuneAug_v.txt</entityName>
  <physical_scope>document</physical_scope>
  <objectName>BEC_2005_12JuneAug_v.txt</objectName>
  <size unit="bytes">30565</size>
  <authentication method="SHA1">7e2014eb643583682c8fadcc24e31b9b503d9d2ef</authentication>
  <dataFormat>
    <externallyDefinedFormat>
      <formatName>text/plain</formatName>
    </externallyDefinedFormat>
  </dataFormat>
  <distribution scope="document">
    <online>
      <url functions="download">
        https://cn.dataone.org.cn/v2/resolve/urn:uuid:bd77143e-7328-413b-a44f-f44ac615f2c7</url>
      </url>
    </online>
    <physical>
  <distribution>
</dataTable>
```

"dsid_01.dt_01" "has file type" "Tabular data (.csv, .txt, etc.)"
Annotation - via id

Subject "Donatella Zona" Predicate "is a" Object "Person"

<additionalMetadata>
  <describes>dsid_01.creator_01</describes>
  <metadata>
    <annotation>
      <propertyURI label="is a">http://www.w3.org/1999/02/22-rdf-syntax-ns#type</propertyURI>
      <valueURI label="Person">https://schema.org/Person</valueURI>
    </annotation>
  </metadata>
</additionalMetadata>
Why Annotate?

- Datasets will be easier to search
- Users typically look for data based on
  - Creator
  - Coverage (spatial, temporal)
  - Measurements
## Taxonomic Authorities

<table>
<thead>
<tr>
<th>Example</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITIS</td>
<td></td>
</tr>
<tr>
<td>Catalog of Life</td>
<td>&gt; 100 expert taxonomic DBs</td>
</tr>
<tr>
<td>WoRMS</td>
<td>Temperate marine</td>
</tr>
<tr>
<td>GBIF Backbone Taxonomy</td>
<td>Aggregates several databases</td>
</tr>
</tbody>
</table>
Discussion
References

https://www.w3.org/TR/cooluris/
https://www.w3.org/2001/tag/awwsw/issue57/20110625/

ESIP 2017 session:
https://docs.google.com/presentation/d/1Rh60ACyeaSew9ci2D3KSD4OjSCCL3biDHjPwK62j2uo/edit#slide=id.g23fe2b6583_0_41
Appendix

Slides following have other info that may be useful during discussion

- highlight some aspects missing from external vocabularies.
- show how EDI’s data harmonization also needs external vocabs
Harmonization

Summary Table
- High-level evaluation
- Content may be elevated to metadata for discovery

Variable Mappings
- Link to external vocabularies
Key-Value Pairs

Flexible models tend to use key-value pairs, but ...

Values: lack typing
Keys: lack a vocabulary

Variable Mappings - hold links to external vocabularies
## Example 1 - Variables which need further description

Many aspects of "biological measurements" are not well described. Data cannot be fully understood until nuances are described.

<table>
<thead>
<tr>
<th>Variable name</th>
<th>N</th>
<th>Unit</th>
<th>Unknown aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>abundance</td>
<td>1</td>
<td>NA</td>
<td>Areal? Volumetric?</td>
</tr>
<tr>
<td>count, number_of_plants, number_of_arthropods, number_size_class_*</td>
<td>14</td>
<td>NA, number</td>
<td>opportunities for QC here</td>
</tr>
<tr>
<td>CPUE</td>
<td>1</td>
<td>NA</td>
<td>Ratio of two measurements (catch, effort); QC steps</td>
</tr>
<tr>
<td>LOGCPUE</td>
<td>1</td>
<td>NA</td>
<td>Units of original measurements</td>
</tr>
<tr>
<td>relative abundance</td>
<td>1</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>cover_amount</td>
<td>1</td>
<td>NA</td>
<td></td>
</tr>
</tbody>
</table>
Example 2 - Variables which need further description

<table>
<thead>
<tr>
<th>Location</th>
<th>Taxon</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>moose.cage</td>
<td>behavior</td>
<td>Surge</td>
</tr>
<tr>
<td>park_acreage</td>
<td>biogeographic.affinity</td>
<td>surveys_notes</td>
</tr>
<tr>
<td>park_code</td>
<td>Clade</td>
<td>surveys_observation_notes</td>
</tr>
<tr>
<td>park_district</td>
<td>Coarse_Trophic</td>
<td>Swell</td>
</tr>
<tr>
<td>point_code</td>
<td>colony.size</td>
<td>Temperature</td>
</tr>
<tr>
<td>point_location</td>
<td>common_name</td>
<td>time_end</td>
</tr>
<tr>
<td>restored</td>
<td>feeding.preference</td>
<td>time_start</td>
</tr>
<tr>
<td>treatment</td>
<td>Fine_Trophic</td>
<td>trap.num</td>
</tr>
<tr>
<td>urbanized</td>
<td>Fish_length</td>
<td>trap.type</td>
</tr>
<tr>
<td>water</td>
<td>hl</td>
<td>Visibility</td>
</tr>
<tr>
<td></td>
<td>Kingdom</td>
<td>wind</td>
</tr>
<tr>
<td></td>
<td>Lineage</td>
<td>Wind_Velocity</td>
</tr>
<tr>
<td></td>
<td>nest.substrate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Order</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Phylum</td>
<td></td>
</tr>
<tr>
<td></td>
<td>primary.habitat</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rel</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rll</td>
<td></td>
</tr>
<tr>
<td></td>
<td>secondary.habitat</td>
<td></td>
</tr>
<tr>
<td></td>
<td>seed.disperser</td>
<td></td>
</tr>
<tr>
<td></td>
<td>slavemaker.sp</td>
<td></td>
</tr>
<tr>
<td></td>
<td>source</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total_Length</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tribe</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accession_Number</td>
<td></td>
</tr>
<tr>
<td></td>
<td>air_temp_F</td>
<td></td>
</tr>
<tr>
<td></td>
<td>area</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cloud</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cloud_Cover</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Date</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diver</td>
<td></td>
</tr>
<tr>
<td></td>
<td>End</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gear Type</td>
<td></td>
</tr>
<tr>
<td></td>
<td>height</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of replicate samples</td>
<td></td>
</tr>
<tr>
<td></td>
<td>observer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pH</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sample_subtype</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sea_State</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Secchi Depth</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Start</td>
<td></td>
</tr>
<tr>
<td></td>
<td>subproject</td>
<td></td>
</tr>
</tbody>
</table>