Annotation with EML 2.2

Margaret O'Brien 0000-0002-1693-8322
Steven Chong 0000-0003-1264-1166
Mark Schildhauer 0000-0003-0632-7576
Matt Jones 0000-0003-0077-4738

Outline

Review annotation in EML 2.2

RDF triples

Ontologies and the vocabulary spectrum

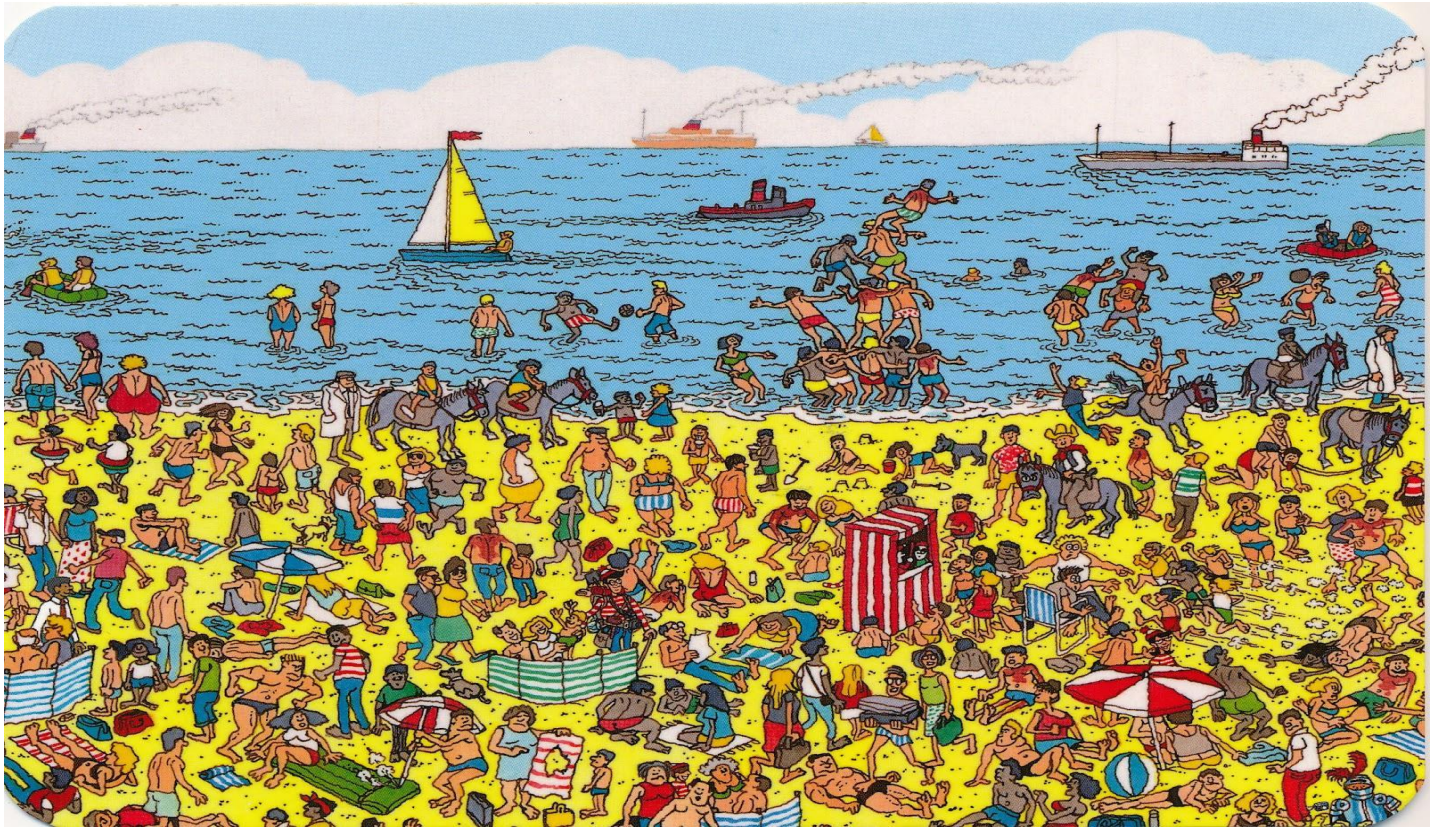
Annotations workflows, experiences so far

Upcoming EDI/ADC workshop

Why Annotate?

Amount of data in repositories is *increasing*

Finding and understanding data are *difficult*



Why Annotate?

Synonyms: CO2 flux == Carbon dioxide flux

Homonyms: Litter <> Litter

Hierarchy: NPP *subClass of* Carbon flux

Why Annotate?

Annotations hold links to external vocabularies

Used to construct an “RDF Triple”

Machine readable

EML Annotation element

```
<annotation>
     <propertyURI label="property label here">property URI here</propertyURI>
     <valueURI label="value label here">value URI here</valueURI>
</annotation>
```

Shared Definitions

Accessed through resolvable http URI's

this attribute ("umolesCO2") is_a " Carbon Dioxide Flux"

this attribute ("CO2_Flux_Rate") is_a " Carbon Dioxide Flux"

```
<annotation>
  <propertyURI label="type">http://www.w3.org/1999/02/22-rdf-syntax-ns#rdf:type</propertyURI>
  <valueURI label="Carbon Dioxide Flux">http://purl.dataone.org/odo/ECSO_00000536</valueURI>
</annotation>
```

Carbon Dioxide Flux: http://purl.dataone.org/odo/ECSO_00000536

Annotations in datasets

- **top-level**:
  - an `<annotation>` element is a child of `<dataset>`
- **entity-level**:
  - an `<annotation>` element is a child of <dataTable>
- **attribute**:
  - an `<annotation>` element is a child of an entity's `<attribute>` element

- **eml/annotations**:
  - a container for a group of `<annotation>` elements using references
- **eml/additionalMetadata**:
  - `<annotation>` elements that reference a main-body element by its `id` attribute

Dataset Annotation

```
<keywordSet>
  <keyword>Forest</keyword>
  <keywordThesaurus>LTER Controlled Vocabulary V 1.0</keywordThesaurus>
</keywordSet>
```

EML ^ 2.1

EML 2.2 V

```
  </taxonomicCoverage>
</coverage>

<!-- dataset level, last element in the Resource Group, after coverage -->
<annotation>
  <propertyURI label="is about">http://purl.obolibrary.org/obo/IAO_0000136</propertyURI>
  <valueURI label="Forest Biome">http://purl.obolibrary.org/obo/ENVO_01000174</valueURI>
</annotation>

<contact>
```

Attribute

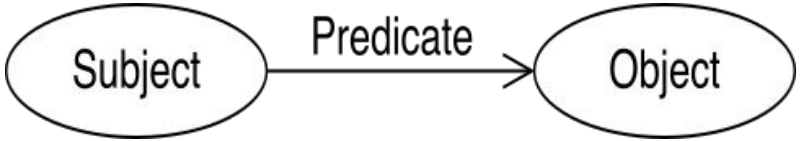
```
<attribute>
  <attributeName>NEE</attributeName>
  <attributeDefinition>net ecosystem exchange</attributeDefinition>
  <measurementScale>
    ...
  </measurementScale>
  <missingValueCode>
    ...
  </missingValueCode>
</attribute>
```

EML ^ 2.1

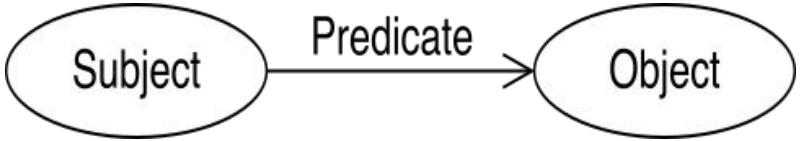
EML 2.2 V

```
<attribute id="dsid_01.att_01">
  <attributeName>NEE</attributeName>
  <attributeDefinition>net ecosystem exchange</attributeDefinition>
  <measurementScale>
    ...
  </measurementScale>
  <missingValueCode>
    ...
  </missingValueCode>
  <annotation>
    <propertyURI label="contains measurements of type"
      >http://ecoinformatics.org/oboe/oboe.1.2/oboe-core.owl#containsMeasurementsOfType</propertyURI>
    <valueURI label="Net Ecosystem Exchange Carbon Flux"
      >http://purl.dataone.org/odo/ECSO_00000014</valueURI>
  </annotation>
</attribute>
```

RDF Triple

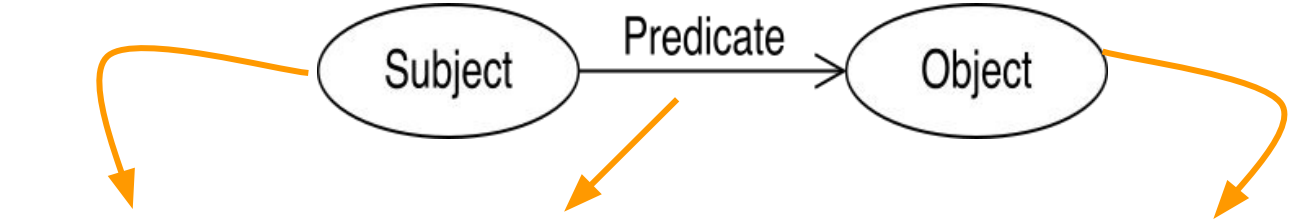


RDF Triple



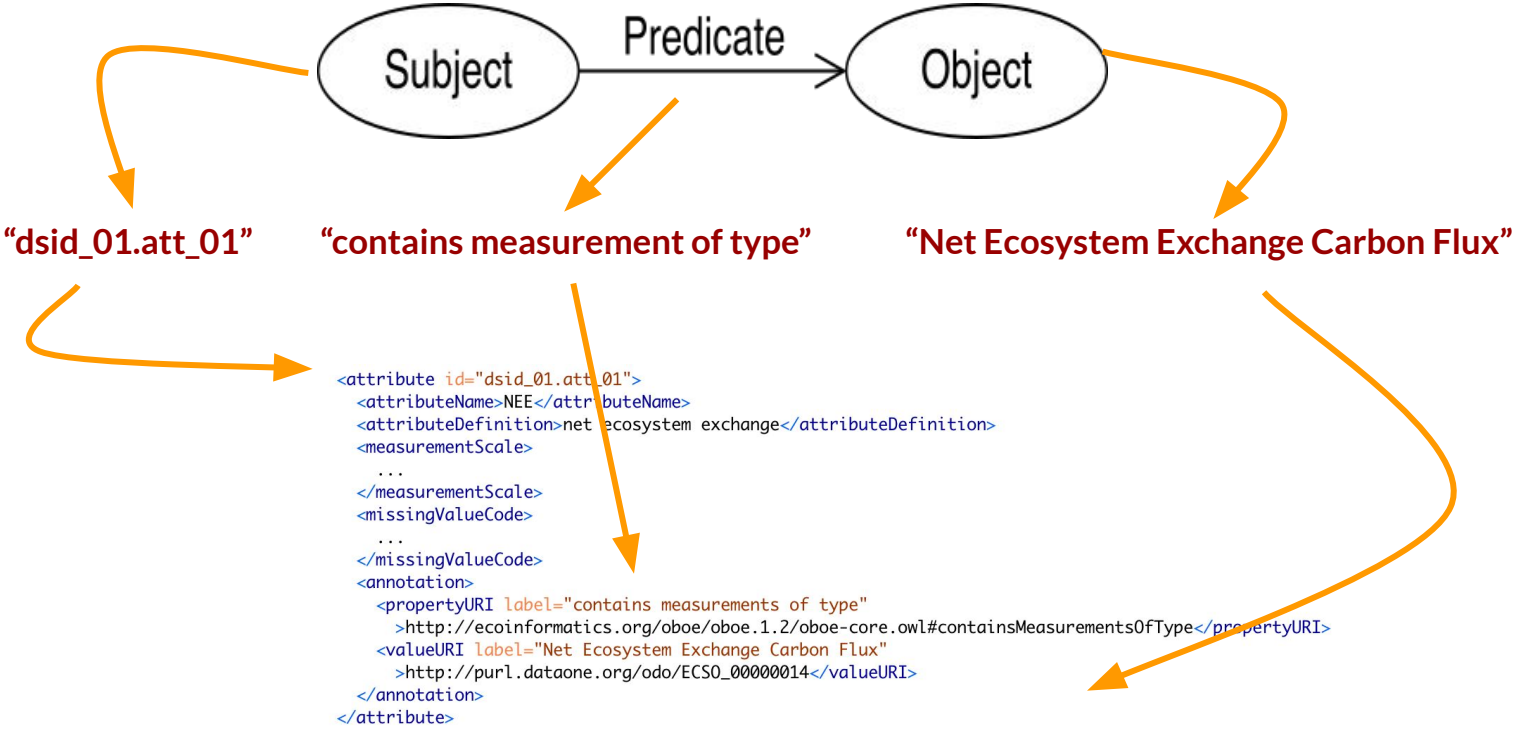
“dsid_01.att_01” “contains measurement of type” “Net Ecosystem Exchange Carbon Flux”

RDF Triple

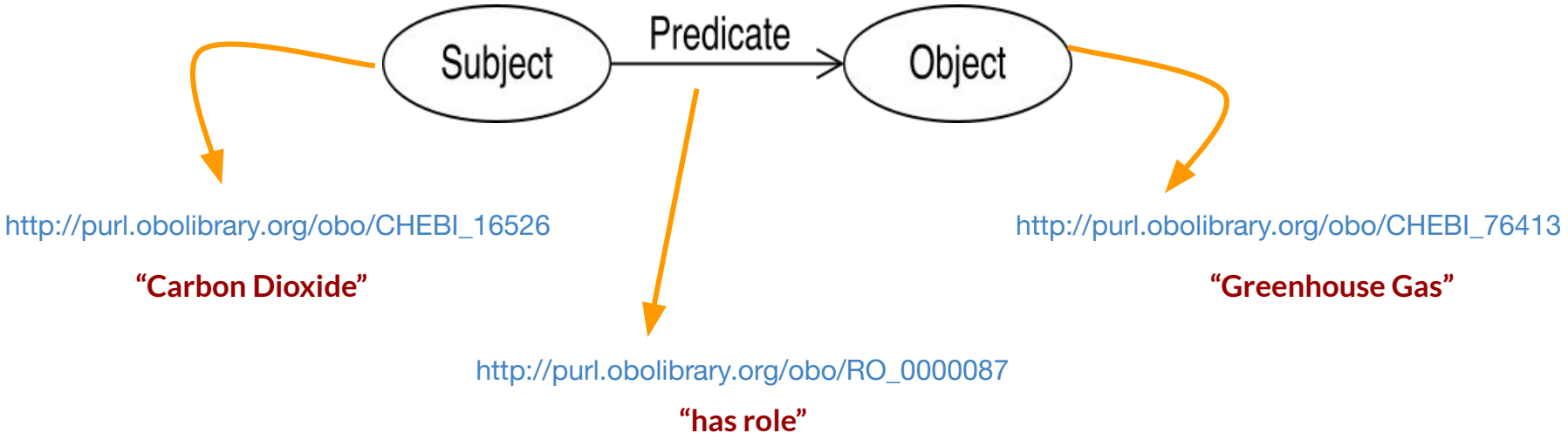


“dsid_01.att_01” “contains measurement of type” “Net Ecosystem Exchange Carbon Flux”

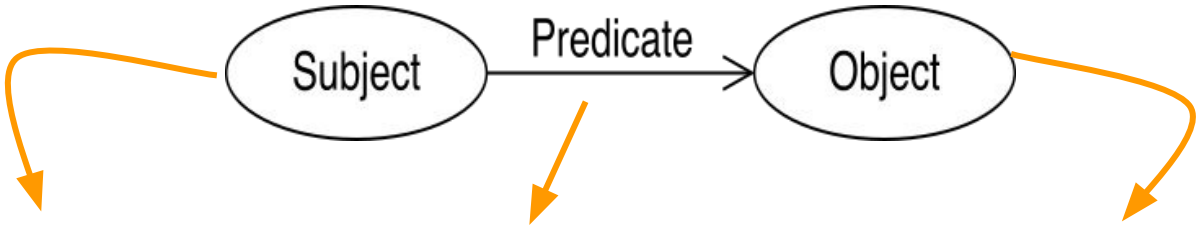
Annotation - attribute



What the computer uses



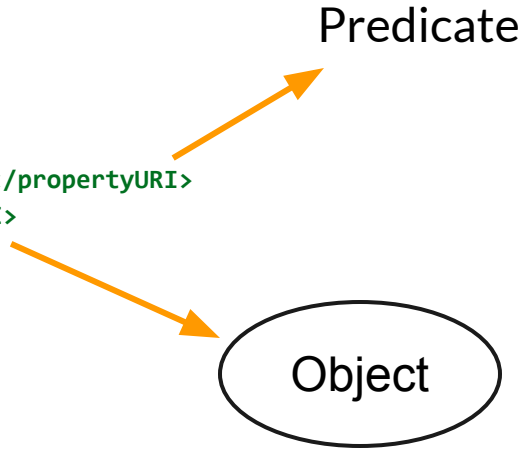
Annotation - dataset

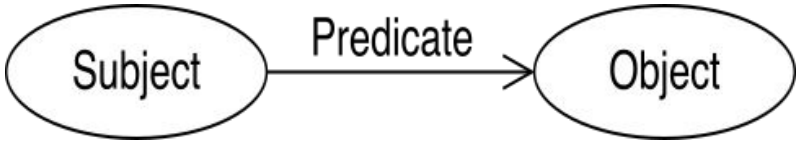



```
  </taxonomicCoverage>
</coverage>

<!-- dataset level, last element in the Resource Group, after coverage -->
<annotation>
  <propertyURI label="is about">http://purl.obolibrary.org/obo/IAO_0000136</propertyURI>
  <valueURI label="Forest Biome">http://purl.obolibrary.org/obo/ENVO_01000174</valueURI>
</annotation>

<contact>
```

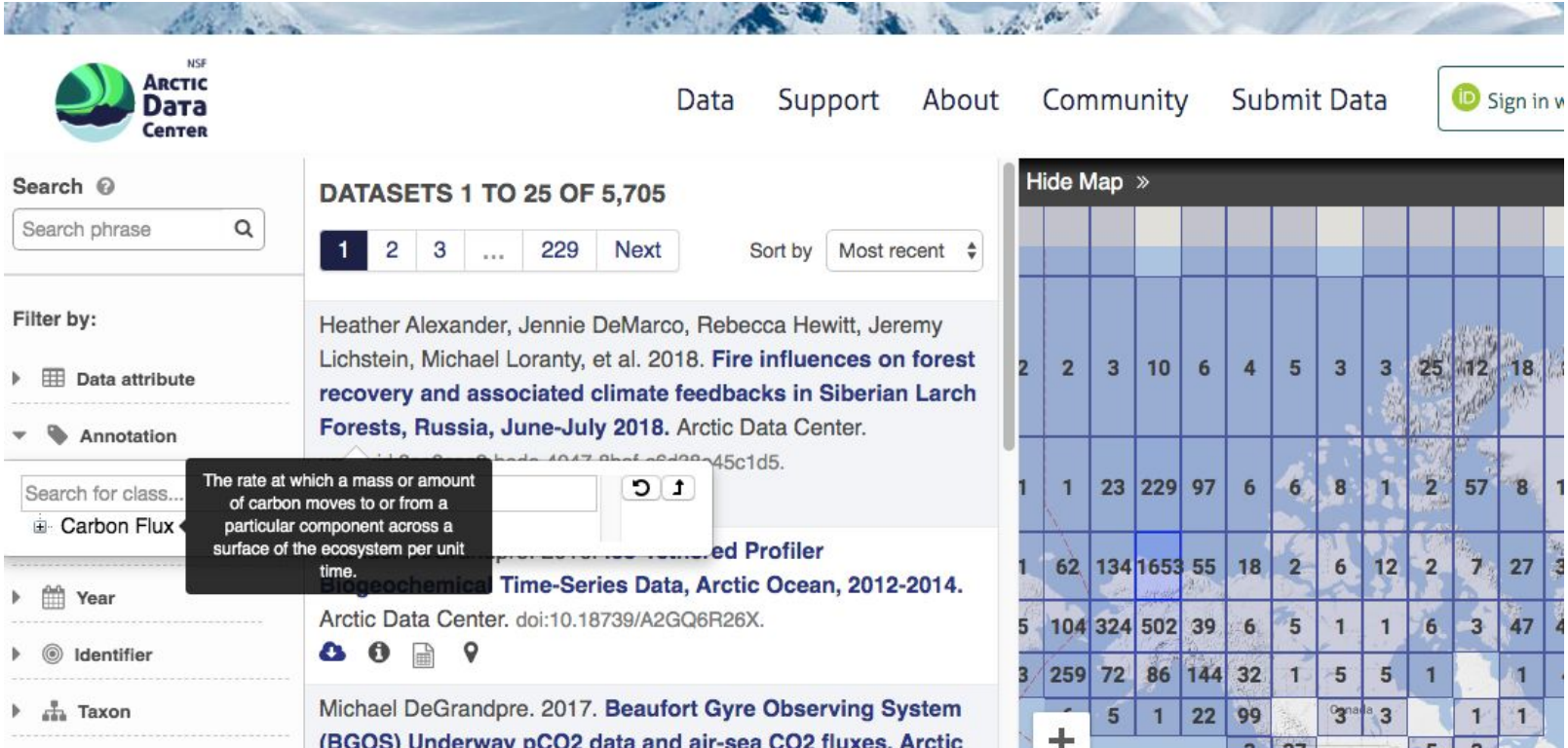
Summary

```
<annotation>
     <propertyURI label="property label here">property URI here</propertyURI>
     <valueURI label="value label here">value URI here</valueURI>
</annotation>
```





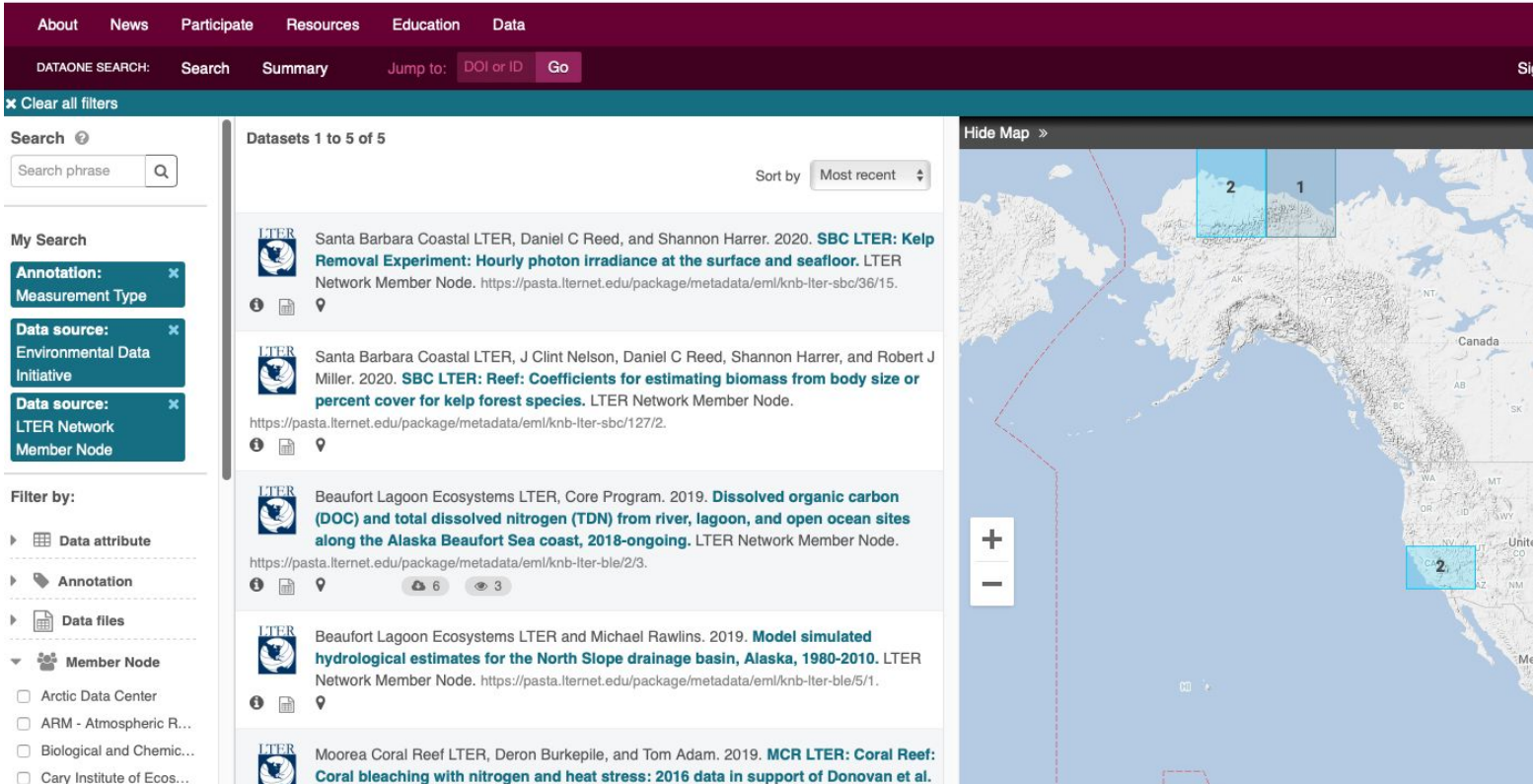
Support at Arctic Data Center



Support at DataONE

More at: [January 2020 DataONE Webinar (Schildhauer/O'Brien)]

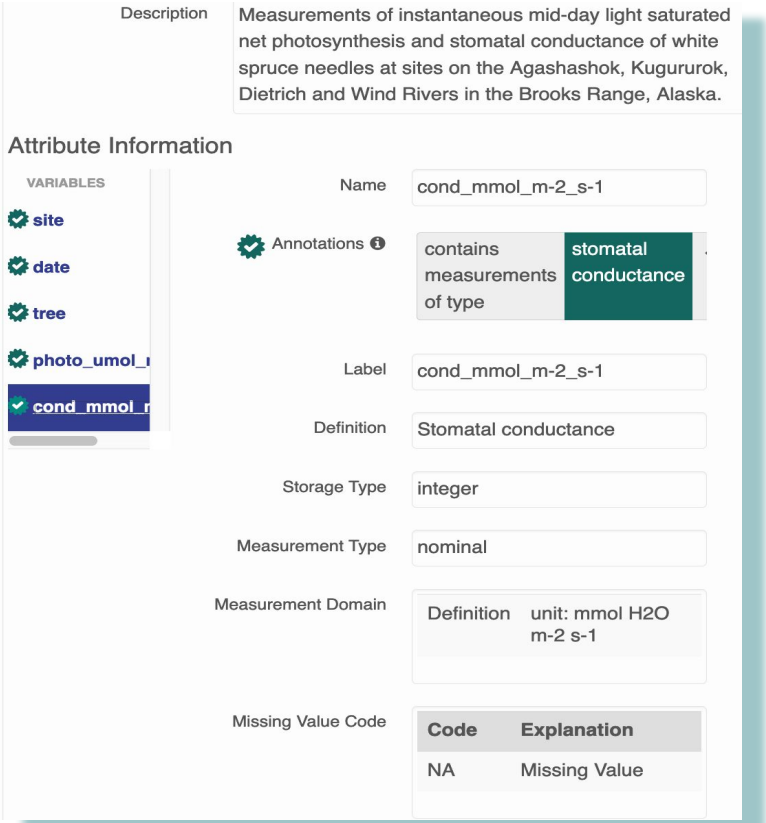
Filter by annotation



Reusable Measurements

Patrick Sullivan. 2016. Light Saturated Needle Photosynthesis and Stomatal Conductance. Arctic Data Center. doi:10.18739/A28911R0D.

http://purl.dataone.org/odo/ECSO_00003084

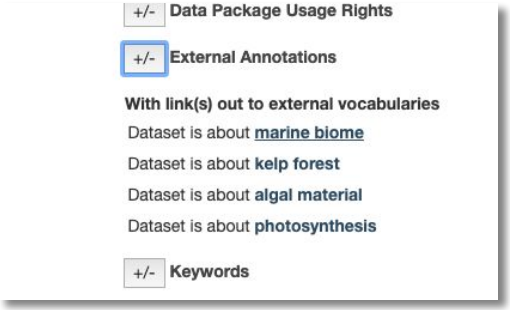


Annotation Support at EDI

Phase 1:

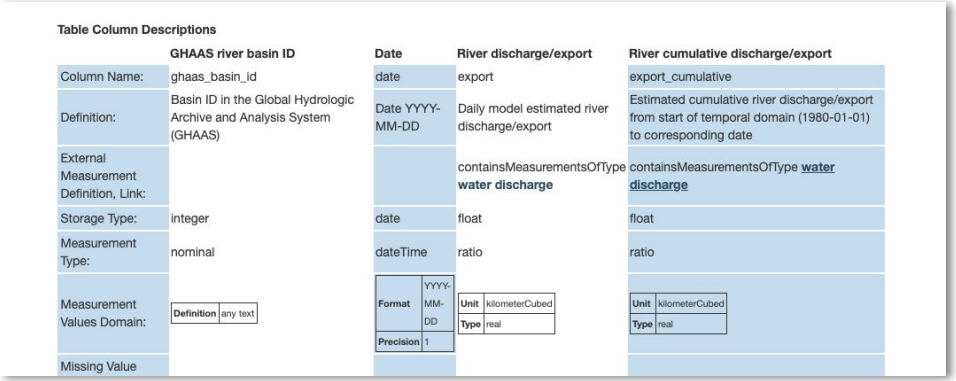
HTML - display and link at 2 levels

- Dataset
- Attribute



Phase 2:

TBD



Dataset-level: knb-lter-sbc.57.9

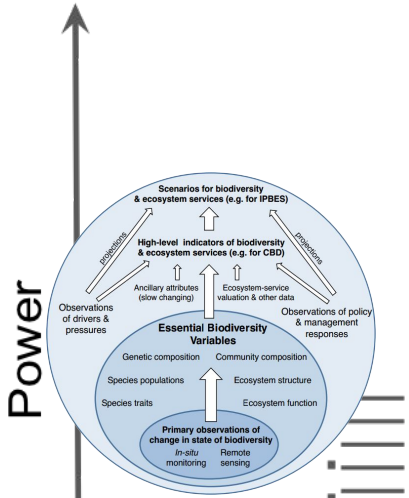
Attribute-level: knb-lter-ble.5.2

URIs in a Darwin Core Archive

Measurement URIs can populate the measurementID field, here using a defined measurement for volumetric abundance in a **Darwin Core Archive for GBIF**

eventID	date	Taxon	taxonID	measurementOrFact	measurementID	measurementValue	measurementUnit
80116_1_5m	16-Jan-2008	Pseudo-nitzschia	ITIS:584561	cells_liter	http://purl.dataone.org/odo/ECSO_00001201	5960	numberPerLiter
80116_1_5m	16-Jan-2008	Leptocylindrus	ITIS:2394	cells_liter	http://purl.dataone.org/odo/ECSO_00001201	3040	numberPerLiter
80116_1_5m	16-Jan-2008	Thalassiosira	ITIS:2484	cells_liter	http://purl.dataone.org/odo/ECSO_00001201	2680	numberPerLiter
80116_1_5m	16-Jan-2008	Chaetoceros	ITIS:2758	cells_liter	http://purl.dataone.org/odo/ECSO_00001201	4740	numberPerLiter

Vocabulary Spectrum



Power

Glossary

List

DwC Terms

The Global Ocean Observing System

CF Standard Names

Taxonomy

Thesaurus

LTER NETWORK

PCO

CYVERSE

Envo

Ontology

NCEAS

ECSO

BODC British Oceanographic Data Centre

NATURAL ENVIRONMENT RESEARCH COUNCIL

Complexity

SKOS and Ontology Compared

Semi-formal KOS	Ontology
Semi-formal, for thesauri and classification schemes may also be arranged in structures (e.g. hierarchies); relationships are built-in structures have no formal semantics	Represents a logical view of a domain Asserts axioms or facts; designers must create these Automated reasoning Import other ontologies

ENVO - Environment Ontology

Hierarchies for
- Biomes
- Environmental Features
- Environmental Materials

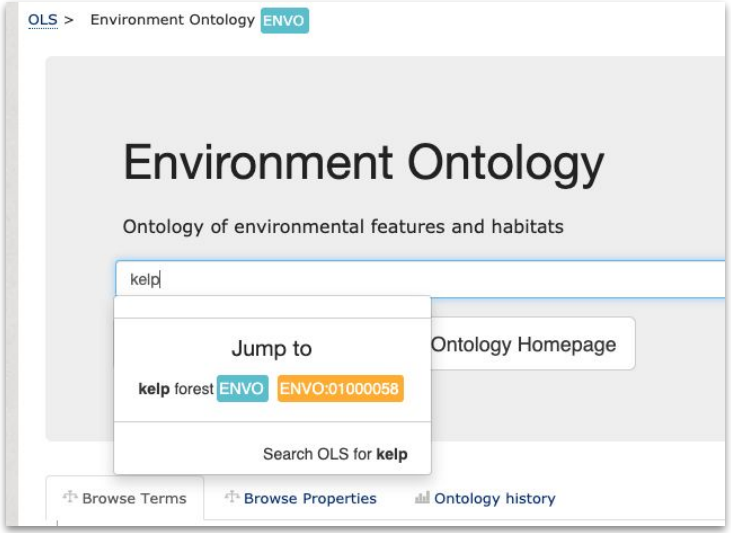
International

Cross-references LTER vocabulary terms

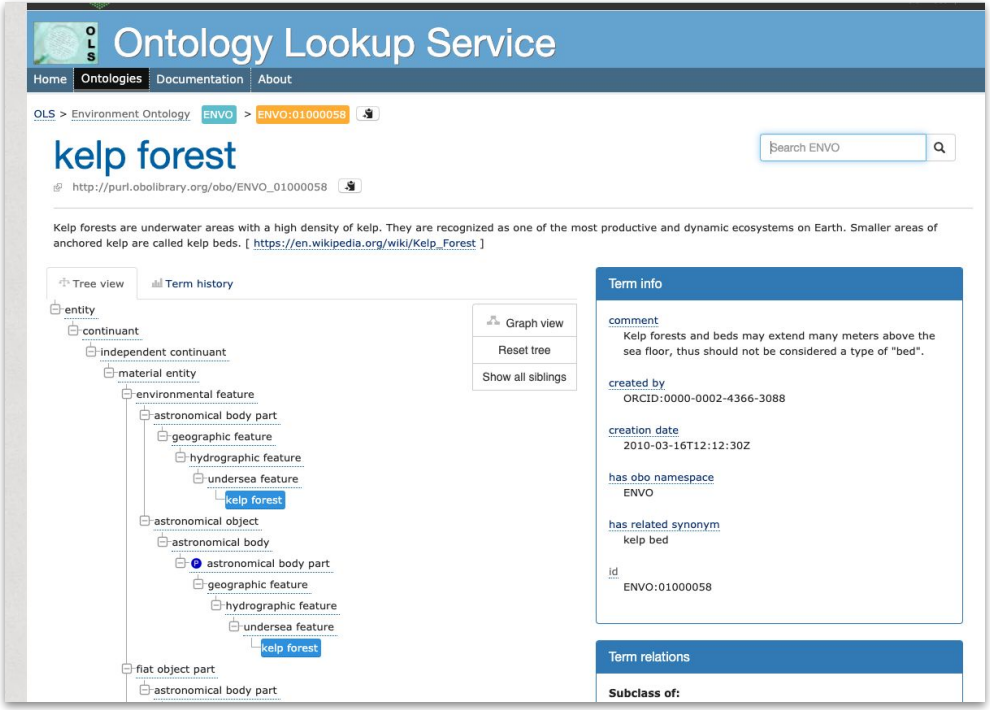
http://environmentontology.org/

obofoundry.org/

ENVO - Environment Ontology

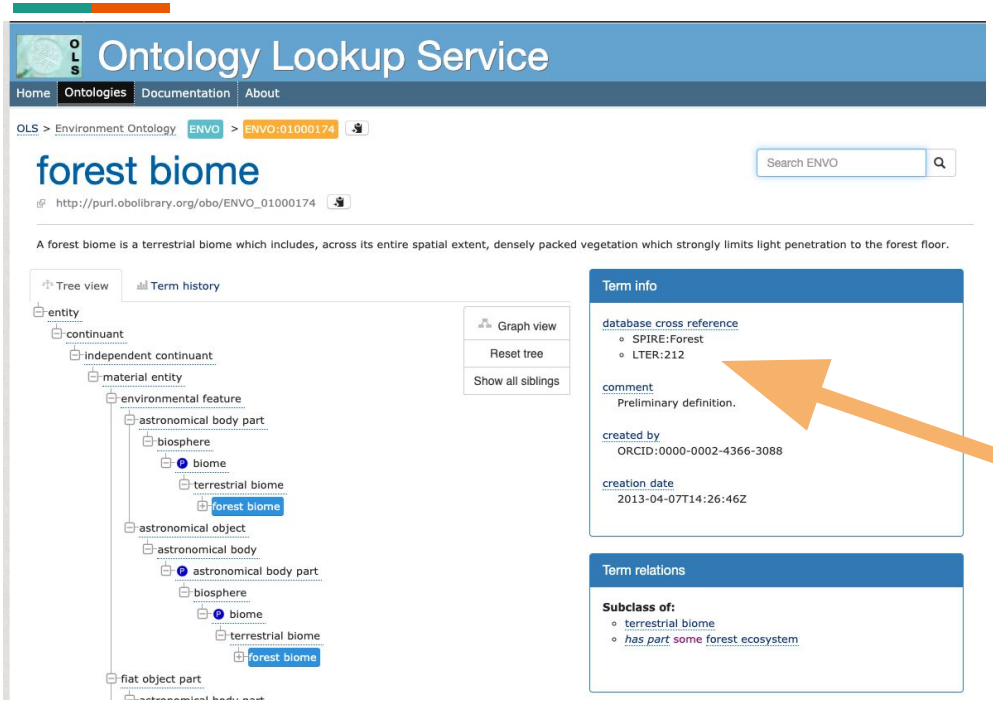


http://www.environmentontology.org/Browse-EnvO



Concept "Kelp Forest" (ENVO:01000058) in Ontology Lookup Service

ENVO - Environment Ontology



Cross-reference to term “Forest” in the LTER Controlled Vocabulary

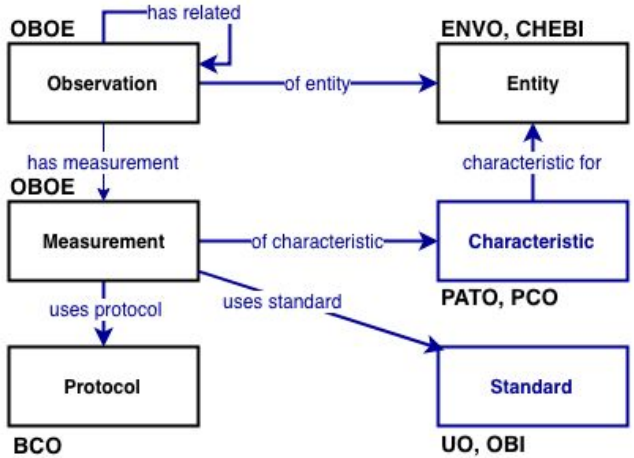
Concept "Forest Biome" (ENVO:01000174) in Ontology Lookup Service

ECSO - EcoSystem Ontology

Defines “Measurement Types”

Imports existing ontologies into an observation framework, “OBOÉ ”

Initial Focus on Carbon Cycling



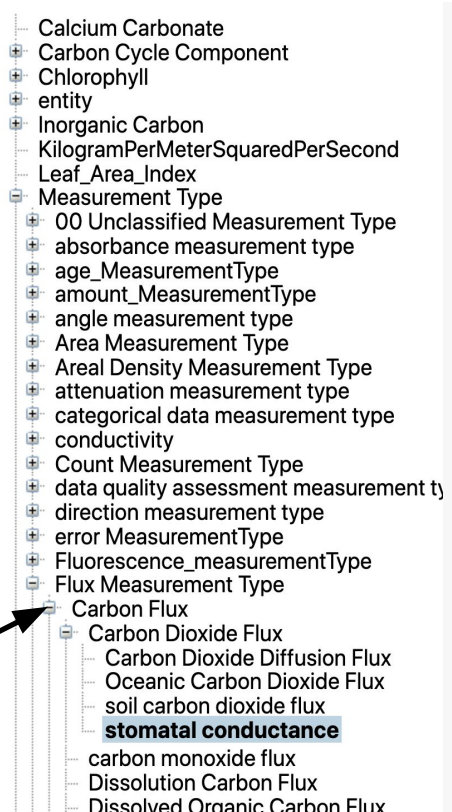
http://bioportal.bioontology.org/ontologies/ECSO

Interoperable Measurements

http://purl.dataone.org/odo/ECSO_00003084

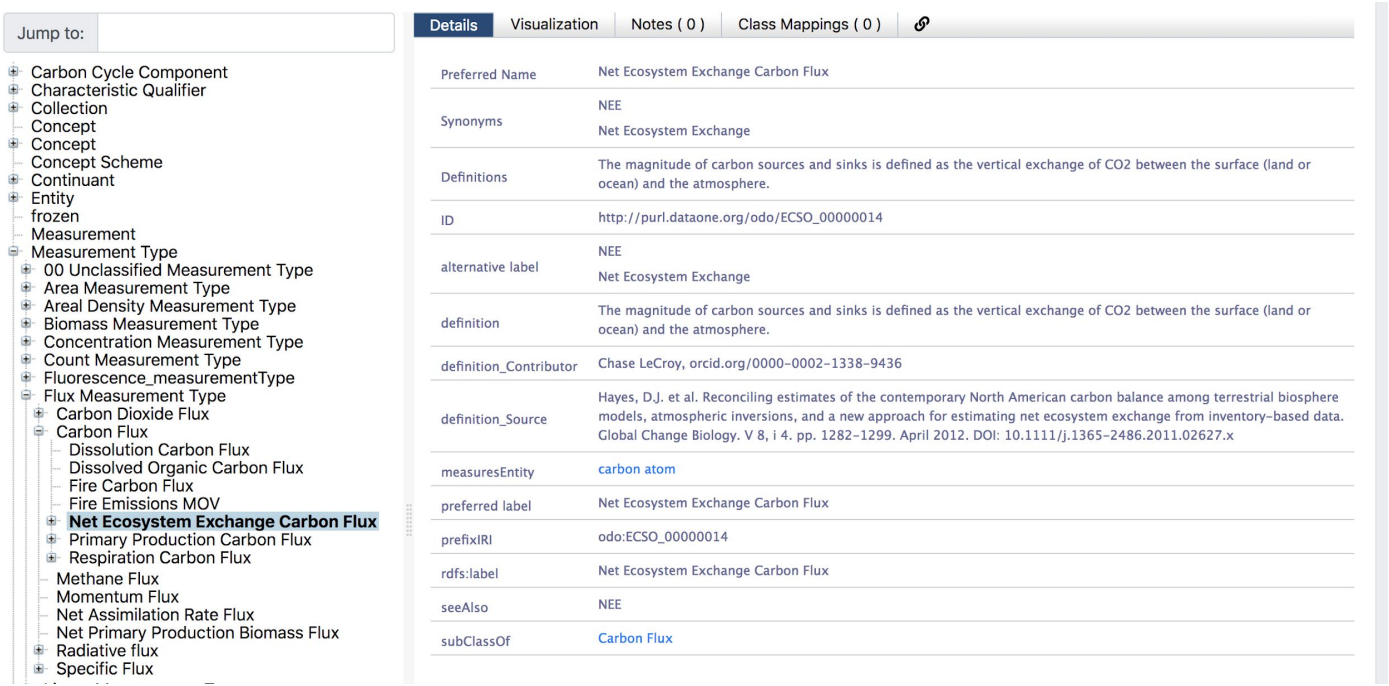
Above HTTP URI dereferences to a term in the ECSO ontology

“stomatal conductance” measurements are a type of “carbon flux” in the ECSO ontology



Preferred Name	stomatal conductance
Definitions	The rate of carbon dioxide entering or water vapor exiting through the stomata of a leaf.
ID	http://purl.dataone.org/odo/ECSO_00003084
creator	http://orcid.org/0000-0003-1264-1166
date	2019-07-19T00:06:38Z
definition	The rate of carbon dioxide entering or water vapor exiting through the stomata of a leaf.
label	stomatal conductance
prefixIRI	odo:ECSO_00003084
prefLabel	stomatal conductance
subClassOf	Carbon Dioxide Flux

http://purl.dataone.org/odo/ECSO_00000014



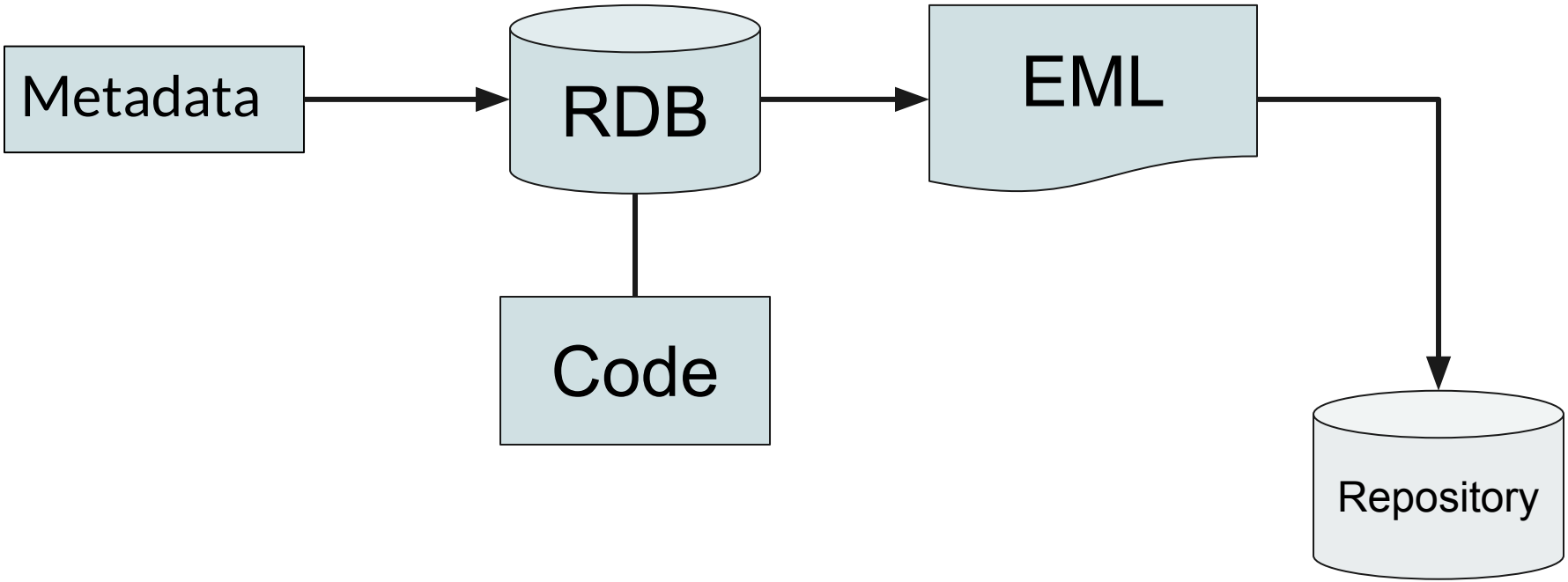
Attribute, with annotation

```
<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/oboe/oboe.1.2/oboe-core.owl#containsMeasurementsOfType</propertyURI>
    <valueURI label="Net Ecosystem Exchange Carbon Flux">http://purl.dataone.org/odo/ECSO_00000014</valueURI>
  </annotation>
</attribute>
```

Annotation in a curation workflow

LTER	Metadata has multiple uses, stored centrally (e.g., RDB with export scripts)
Ad hoc EML assembly	Metadata collected for datasets only
Enhance existing EML	EML exists, let's annotate it

LTER Workflow



LTER Core Metabase

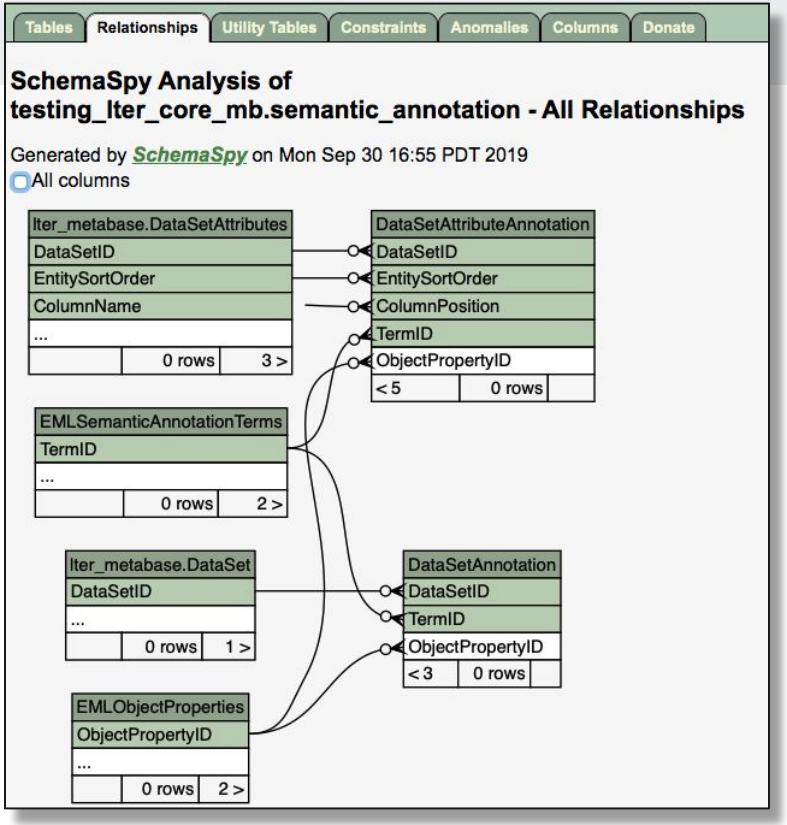
You will need

RDB Tables for annotation

- Dataset-level
- Attribute-level

Code to support

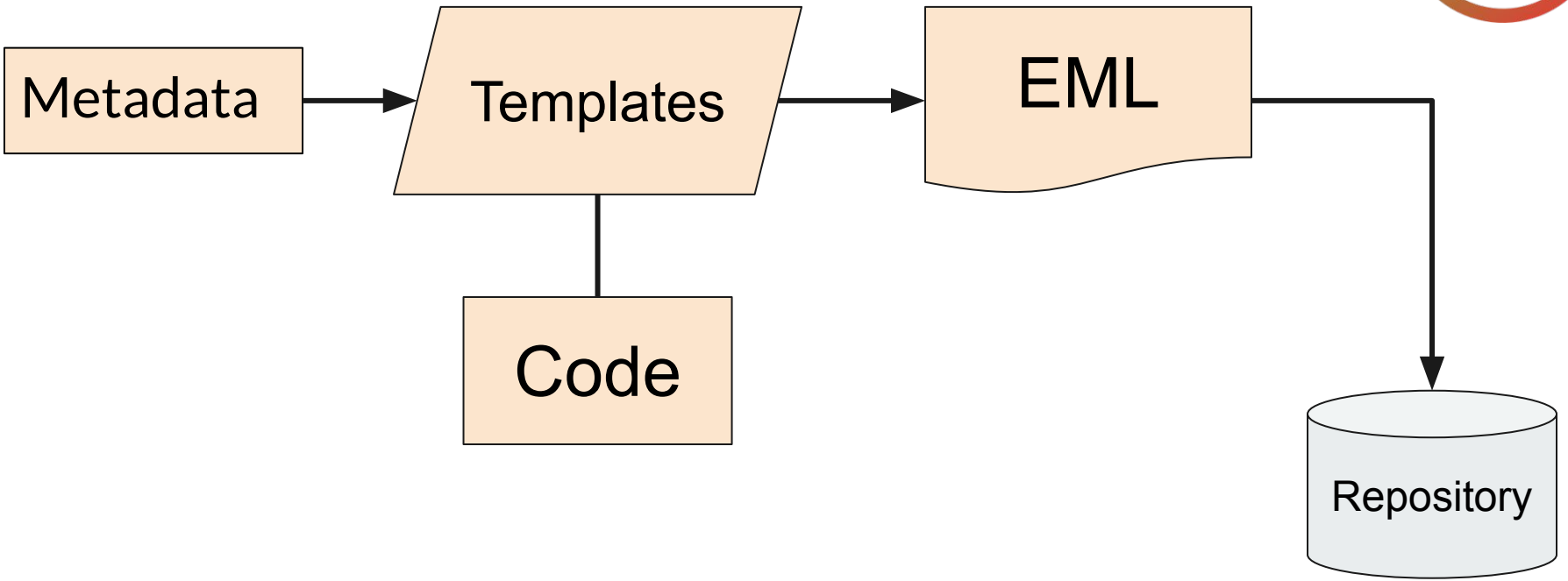
- EML 2.2
- Annotations



https://github.com/lter/LTER-core-metabase
https://github.com/BLE-LTER/MetaEgress

LTER NETWORK

Ad Hoc Workflow



EML Assembly Line

Option near end of an EML-creation workflow

Or

Read in existing EML, add annotations

Annotation Template

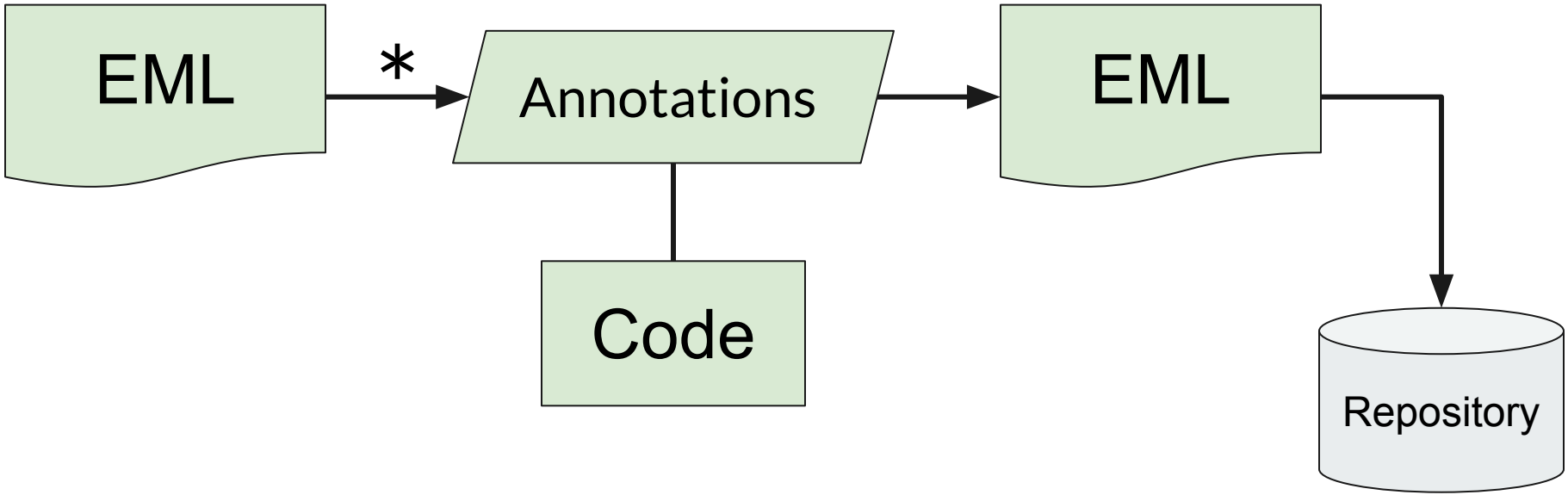
- Single file, per other EAL templates

Output to

- Dataset level
- Entity level
- Attribute level
- Repeated at dataset/annotation

ADC Annotation Project, 2018-2019

NSF Arctic Data Center



ADC Annotation Project, 2018-2019

NSF Arctic Data Center

*

Focus on attribute annotation

Identify datasets with "carbon cycling" measurements
(e.g. query carbon dioxide flux, CO2 flux, carbon flux, NEE, Net Ecosystem Exchange, CH4 flux, methane flux, gas flux ...)

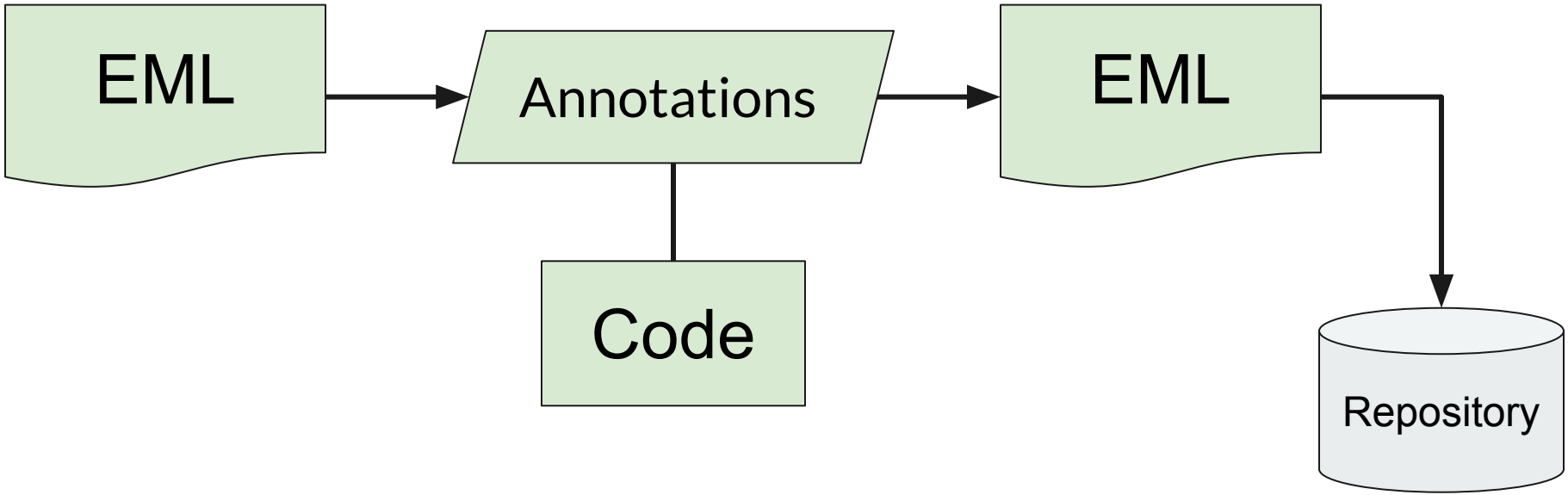
Extract attribute-level metadata from EML to a spreadsheet
(e.g., name, description, unit)

Build candidate terms for ECSO ontology

Add URIs for candidate annotations to the spreadsheet

ADC Annotation Project, 2018-2019

NSF Arctic Data Center



Lessons Learned

Consider a group of attributes together

Dataset-level and Attribute-level attotation have high return (discovery, understanding)

The ontology will be missing some terms

Annotation workshop

Late May 2020

EDI + ADC

Workshop activities

1. Examine, compare ontologies
2. Pick candidate concepts
3. Edit, refine with peer and expert input
4. Export EML in your preferred manner

Goal: annotations are added to *your* curation workflow

References

EML 2.2 Primer on semantic annotation in datasets
https://eml.ecoinformatics.org/semantic-annotation-primer.html

Background on URIs
https://www.w3.org/TR/cooluris/
https://www.w3.org/2001/tag/awwsw/issue57/20110625/
https://www.gbif.org/document/80575/a-beginners-guide-to-persistent-identifiers

ECSO presentation, ESIP 2017 (describes Dataone project):
https://docs.google.com/presentation/d/1Rh60ACyeaSew9ci2D3KSD4OjSCCL3biDHjPwK62j2uo/edit#slide=id.g23fe2b6583_0_41

Terms for the dataset level (URIs for LTER Controlled Vocabulary Terms):
http://www.environmentontology.org/Browse-EnvO

Thank you