Semantics for LTER Carbon Cycling Data

Margaret O'Brien[1], Mark Schildhauer[2], Matt Jones, Deborah McGuinness, Ben Leinfelder, Sophie Hou, Chase LeCroy, Lauren Walker

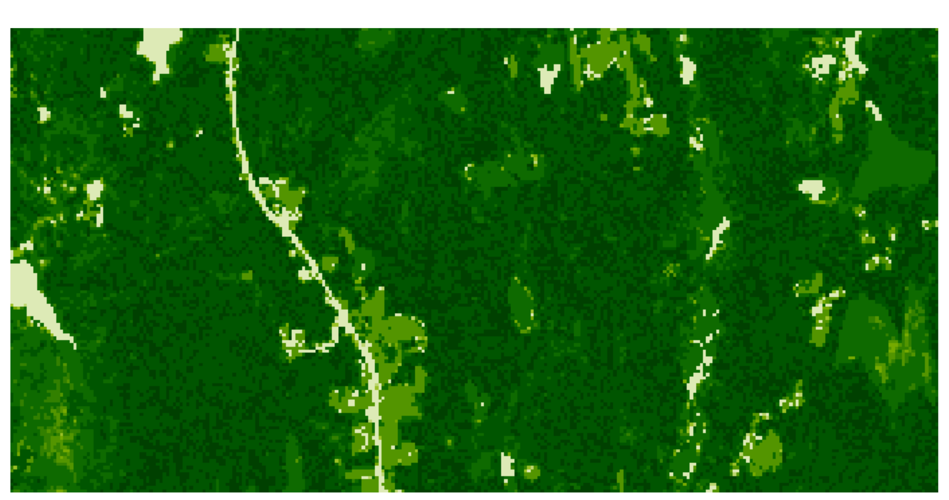
Overview

LTER has a wealth of carbon cycling data from many biomes, but its diversity– in terms of methodology, and terminology used to describe the data-- hampers effective discovery for synthesis or reuse. W3C Semantic Web technologies offer languages (OWL/ RDF) to formally describe content, and mechanisms for attaching "concepts" to datasets, that enhance discovery. The DataONE project, of which the LTER Network is a Member Node, is exploring the use of semantics to enhance discovery of data, using LTER carbon cycling-related datasets as a lead Use Case.

Obstacles to understanding complex data

Carbon fixation and flux measurements are obtained locally, and described or labeled in datasets in many different ways: NPP, biomass, "primary production", PPROD, kg/ m^2/yr, "Dry Weight", etc.

Above, a chamber for measuring in situ NPP in a benthic algal community at the Santa Barbara Coastal LTER.



Left, satellite image depicting NPP values from the BigFoot site, Harvard Forest (image from ORNL DAAC).

Compare metadata at http://portal.lternet.edu
knb-lter-hfr.103.27
knb-lter-sbc.37.4

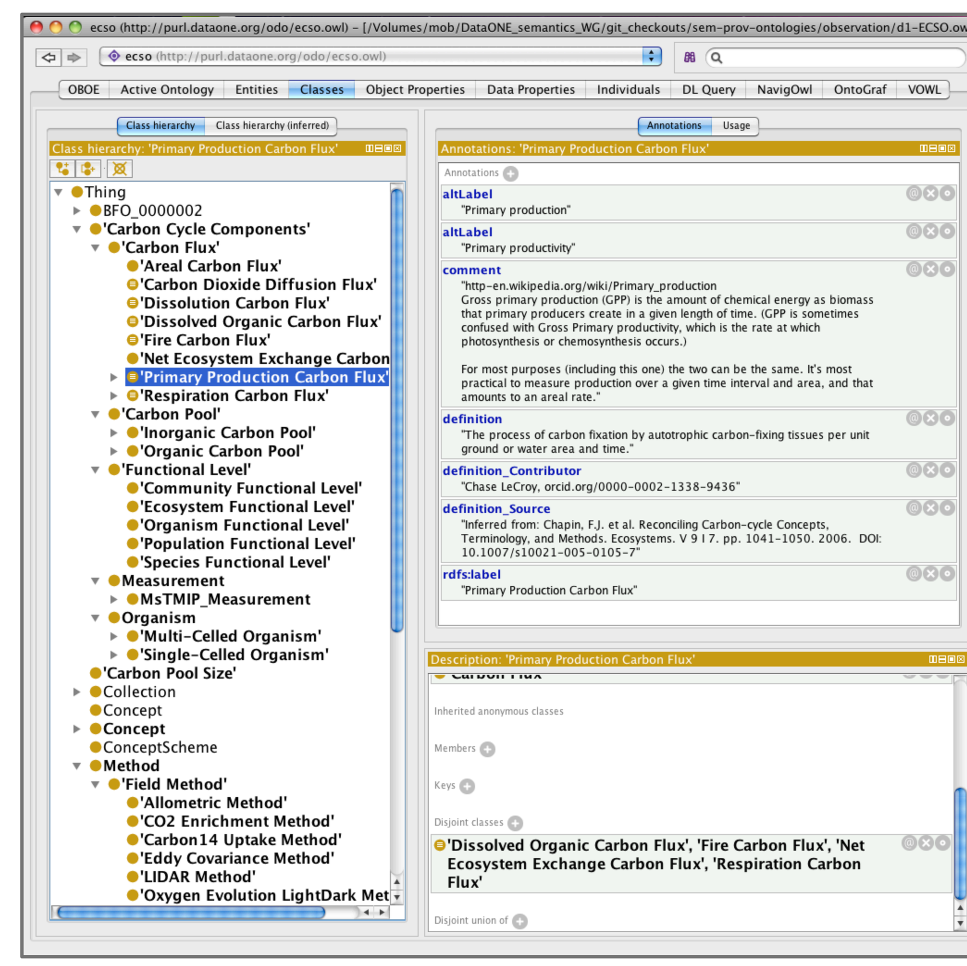
Both datasets have rich metadata & values in "mass per area per time" Does that tell you enough?

LTER's SKOS-based controlled vocabulary supports "narrower terms" and "synonyms"–helping refine search results in the Network catalog

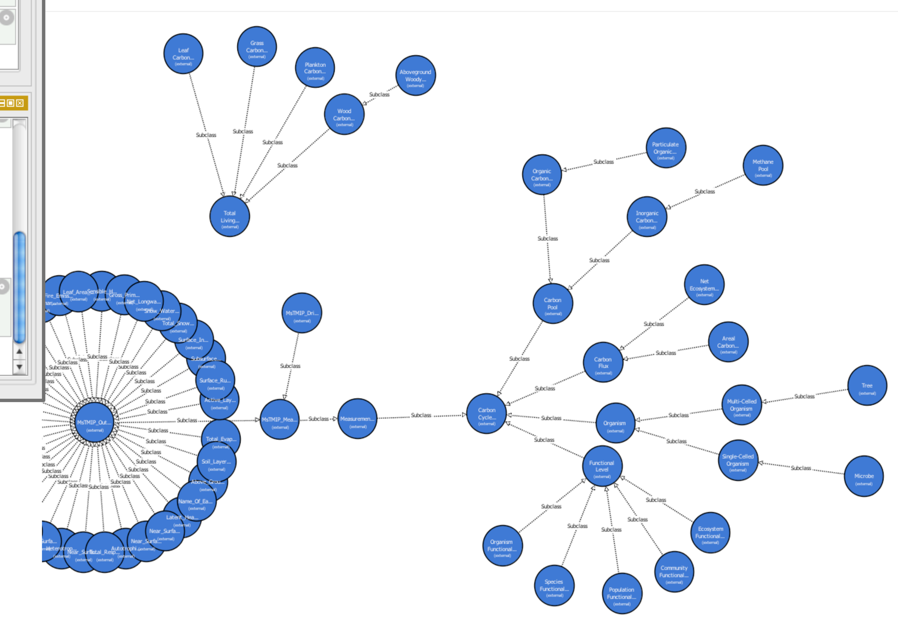
However carbon flux-related terms in the LTER vocabulary number fewer than 20, and searches are based on matches to any metadata, not for specific "measurements"

Issues surrounding descriptions of carbon cycling data are not unique to LTER. A more complete, robust solution will require collaborations of scientists with informatics specialists and knowledge modelers.

The Ecosystems Ontology



An Ontology is a set of structured definitions for concepts, with logic to define the relationships between them, which can be interpreted by computers. We are using the W3C standard "OWL"



ECSO's initial focus is on concepts for carbon flux measurements. It is designed to be expanded to other scientific domains.

Development process

Existing ontologies were examined for terms related to environmental processes and measurements, and for structural elements already in common use. The Ecosystems Ontology (ECSO) imports parts of these ontologies:

- **EnvO** - Environment Ontology: environmental features, materials, systems and conditions
- **ChEBI** – Chemical Entities of Biological Interest: molecular and small chemical compounds
- **UO** - Units Ontology: measurement dimensions and units
- **SKOS** - Simple Knowledge Organization System: synonyms and alternate labels

These and other ontologies are available from the OBO Foundry, which preserves and promotes science-based ontologies (http://www.obofoundry.org). ECSO plans to participate in these efforts.

Annotation and Testing

Goal: Determine the annotation strategy that improves discoverability of datasets through measured recall and precision.

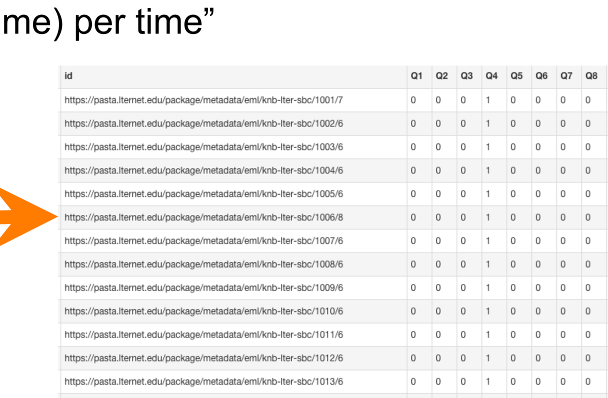
$$recall = \frac{relevant\_retrieved}{total\_relevant}$$

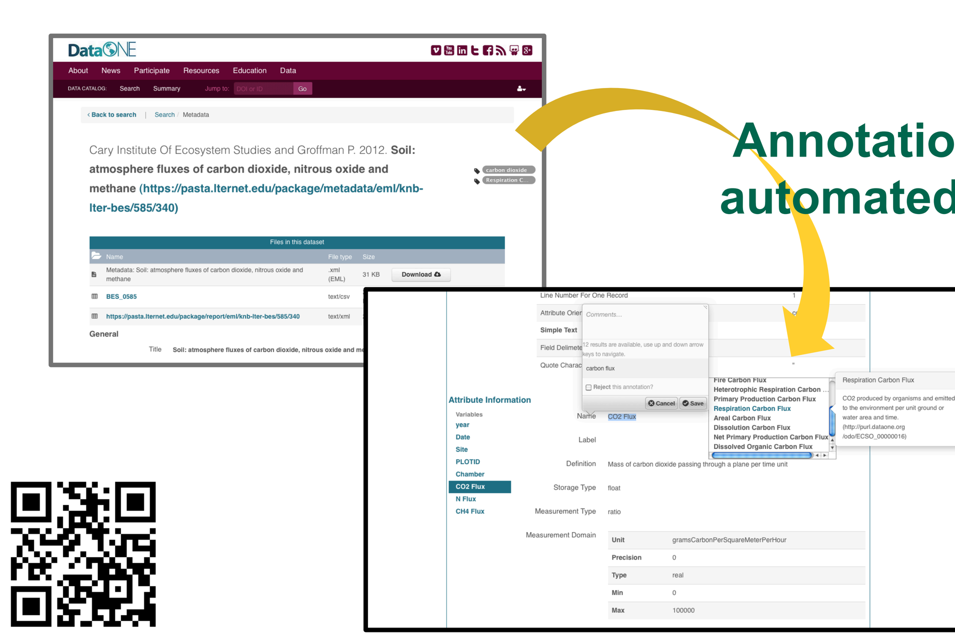
$$precision = \frac{relevant\_retrieved}{total\_retrieved}$$

Method:

1. Construct sample search queries
2. Determine relevant datasets for each query ("ground truth")
3. Compute recall and precision from search results

Find datasets with…
"heterotrophic soil respiration as carbon dioxide flux in dimensions of (amount or mass) per (area or volume) per time"





Annotations may be automated or manual

Benefits

Discovering relevant data from LTER's diversity of measurements, biomes, and methodologies, is a serious challenge. This work will:

- improve LTER capability to retrieve one of its more complex data types at finer semantic grain
- allow for comparing the effectiveness of multiple "semantic" approaches
- create ontology modeling patterns general enough to accommodate future work on other scientific measurements and domains

Our approach builds on pre-existing community practices for structuring ontologies and annotations

[1]margaret.obrien@ucsb.edu, [2]schild@nceas.ucsb.edu

This research was supported by the US National Science Foundation under Grant Numbers 0753144, 0830944, and 1430508