IN31B-1437

Semantic technologies improving the recall and precision of the Mercury metadata search engine

Line C. Pouchard,* Robert B. Cook,* Jim Greene,* Natasha Noy,** Giri Palanisamy*

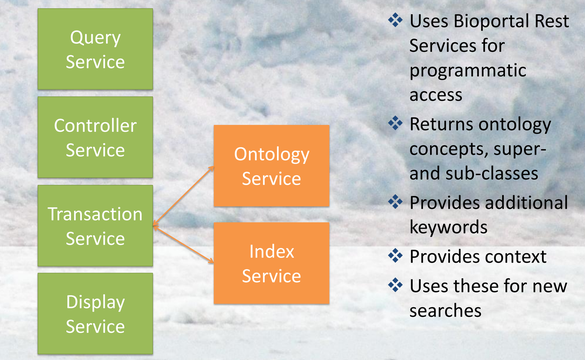
* Oak Ridge National Laboratory, ** Stanford University

Two Relevance Metrics from Information Retrieval

$$\text{Recall} = \frac{\text{Relevant Documents} \cap \text{Retrieved Documents}}{\text{Relevant Documents}}$$

$$\text{Precision} = \frac{\text{Relevant Documents} \cap \text{Retrieved Documents}}{\text{Retrieved Documents}}$$

- 100% recall means all relevant documents were retrieved, but maybe also many non-relevant ones.
- 100% precision means all retrieved documents were relevant, but where relevant documents not retrieved?



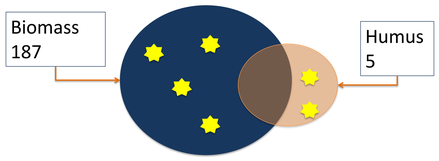
- Uses Bioportal Rest Services for programmatic access
- Returns ontology concepts, super- and sub-classes
- Provides additional keywords
- Provides context
- Uses these for new searches

Architecture Diagram

CURRENT DATA DISCOVERY

- Each dataset is represented by a metadata XML document
- Data discovery is based on the content of several XML elements
- Results are presented as facetted searches
- Mercury supports FGDC, Dublin and Darwin cores, Ecological Mark-up Language and ISO 19115
- The ORNL DAAC holds over 100,000 metadata records from several data providers
- Over 1000 datasets focus on biogeochemical dynamics, ecological data, and environmental processes, accounting for 2 TB
- Also holds MODIS land product subsets, accounting for 60 TB

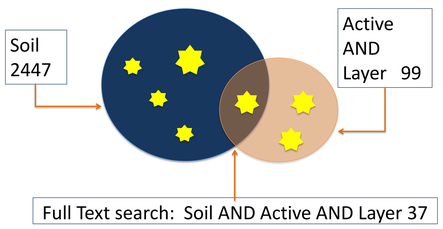
Why improve Recall?



Full Text search: Biomass OR Humus 192

Humus is a type of Biomass:
5 additional datasets are found
Humus is contained in their metadata but NOT Biomass

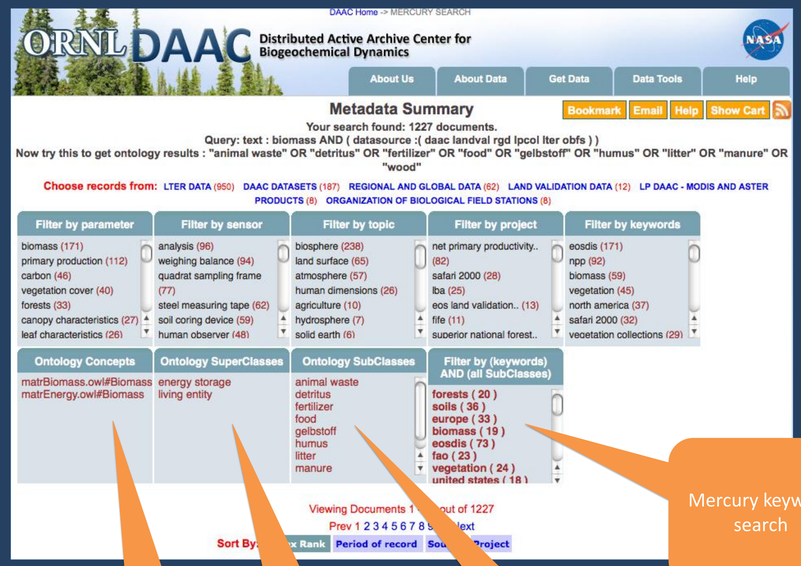
Why improve precision?



Full Text search: Soil AND Active AND Layer 37

Too many facets can be confusing
Active Layer DOES NOT appear as a facet
The user must enter a new query to find the 37 datasets

Our approach: re-use, interoperability, integration

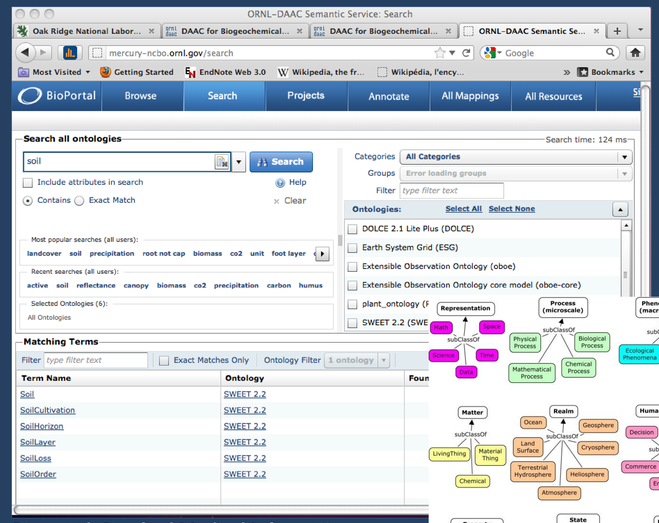


Mercury keyword search

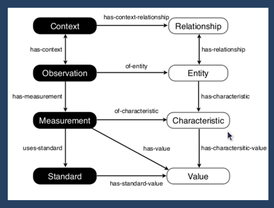
Concepts acquire context: biomass as Material or biomass as Energy

Super-classes have different properties

Additional search terms



Bioportal, Stanford Medical Informatics



OBOE: Extensible Observation Ontology, UCSB

SWEET: Semantic Web for Earth and Environmental Ontologies, JPL

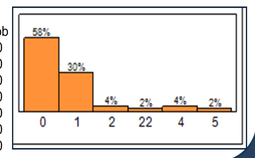
Plant Ontology PO Consortium

Discussion

- SWEET provides a good basis
- Needs to be further specified for this data center requirements
- Ontology concepts and terms match no keywords in the Mercury index
- The current display will not scale to more relationships.
- Many ontologies provide only a few relationships

Matching the top 50 keywords in Mercury to ontology terms

Frequencies n=50		
Level	Count	Prob
0	29	0.58000
1	15	0.30000
2	2	0.04000
22	1	0.02000
4	2	0.04000
5	1	0.02000
Total	50	1.00000



58% of top 50 keywords have no match: adding semantics to a keyword index can never replace annotations of the source datasets.

Rethink the approach of metadata capture to include ontology objects
Add ontologies and/or provide SWEET extensions
Dynamically driven display for improved clarity and comprehension

This work has been in part performed at Oak Ridge National Laboratory, Managed by UT Battelle, LLC under Contract No. De-AC05-00OR22725 for the U.S. Department of Energy