Semantic technologies improving the recall and precision of the Mercury search interface

Line C. Pouchard,* Natasha Noy,** Giri Palanisamy***

Purdue University Libraries*

Google**

Oak Ridge National Laboratory***

Presented to the Geospatial Semantic Workshop

June 2, 2014

Mercury is the search engine for the ORNL DAAC

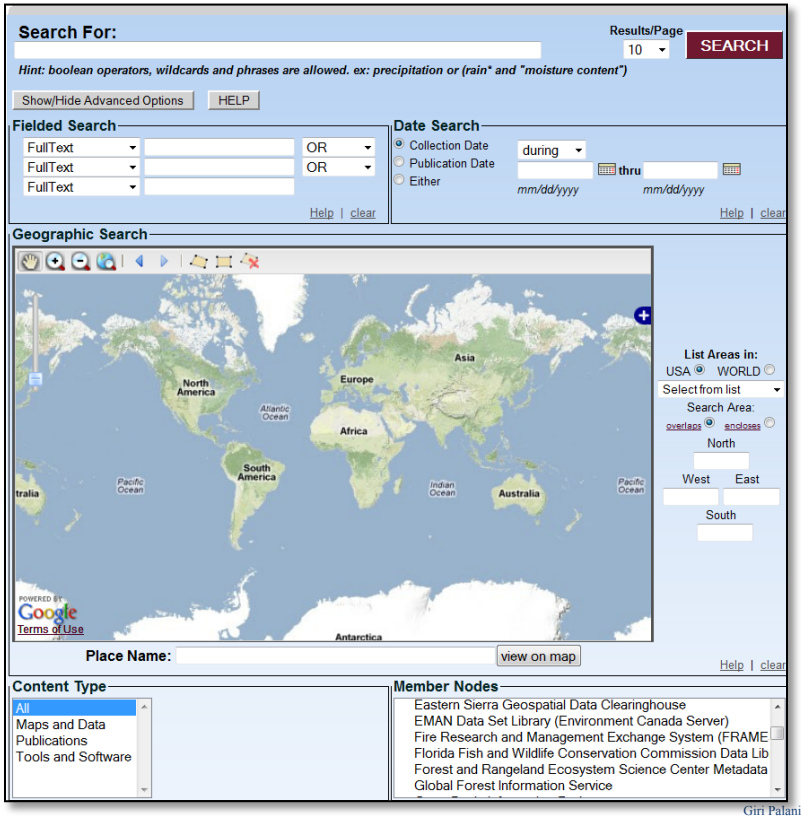
- Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center (DAAC) for Biogeochemical Dynamics
- Ground-based and remote-sensing measurements related to biogeochemical and ecosystem processes.
- NASA-funded field campaigns, selected relevant measurements from NASA Earth Observing System (EOS) satellites, model output, as well as other biogeochemical dynamics data useful to the global change research community.
- Spatial and tabular data on terrestrial ecology and biogeochemical dynamics.
- 1018 datasets having ~ 1 TB in volume
- 1.76 million data files distributed to ~ 23700 distinct users representing 11.65 TB in 2012.

Distributed data search using Mercury

- Distributed metadata management, data discovery and access system.
- Based on a combination of open source and ORNL developed software
- Provide a single portal to information contained in disparate data management systems
- Supports various metadata standards including ISO 19115, FGDC, EML, GCMD DIF, DC
- Allow PIs and database managers to distribute their data while maintaining complete control and ownership

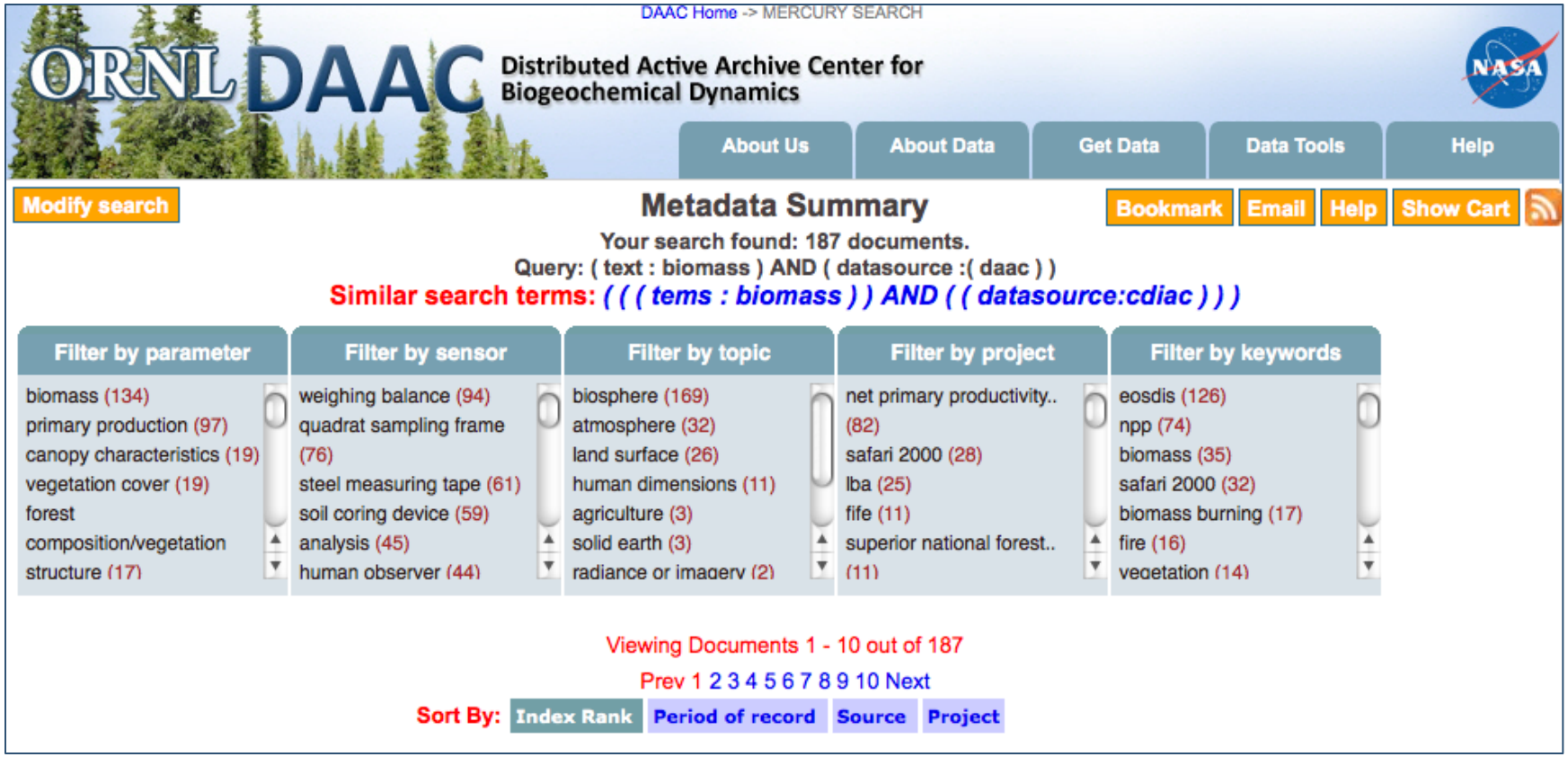
Powered by MERCURY

Oak Ridge National Laboratory

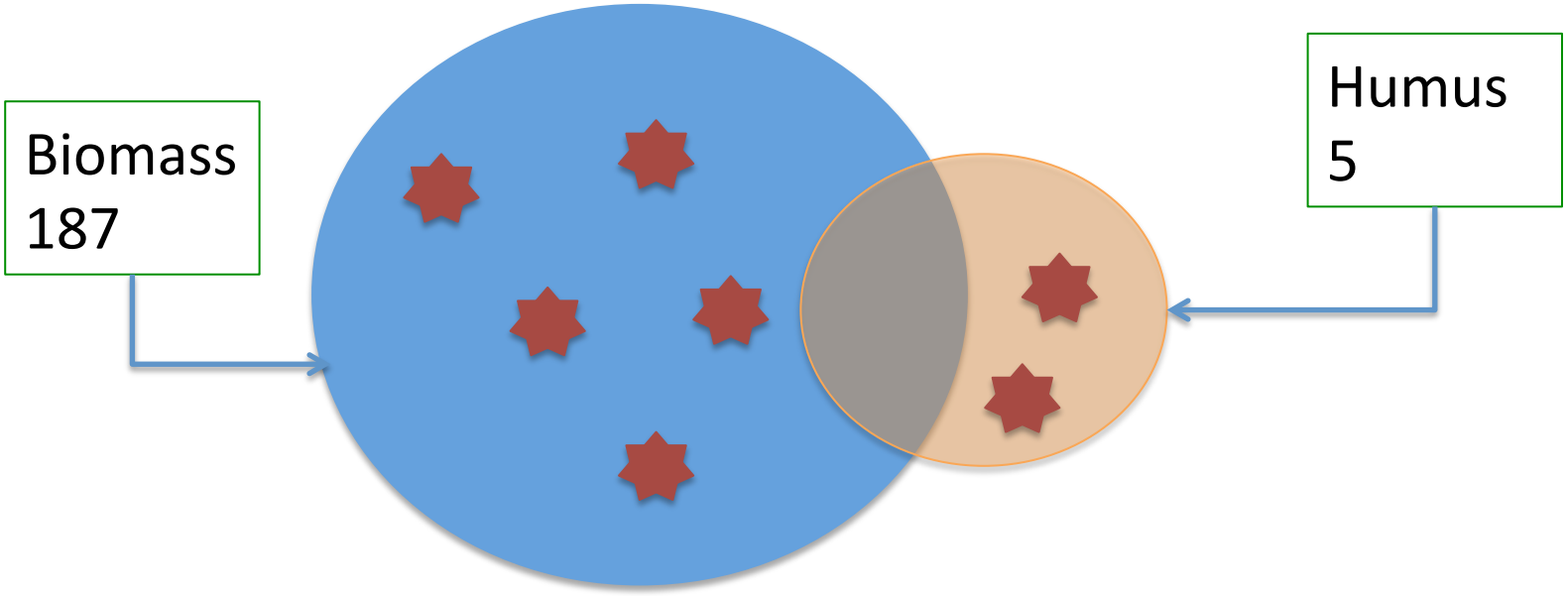


Facetted search results

- Each dataset is represented by a metadata XML document
- Data discovery is based on the content of several XML elements
- Mercury holds over 100,000 metadata records from several providers



Why improve Recall for ORNL DAAC?

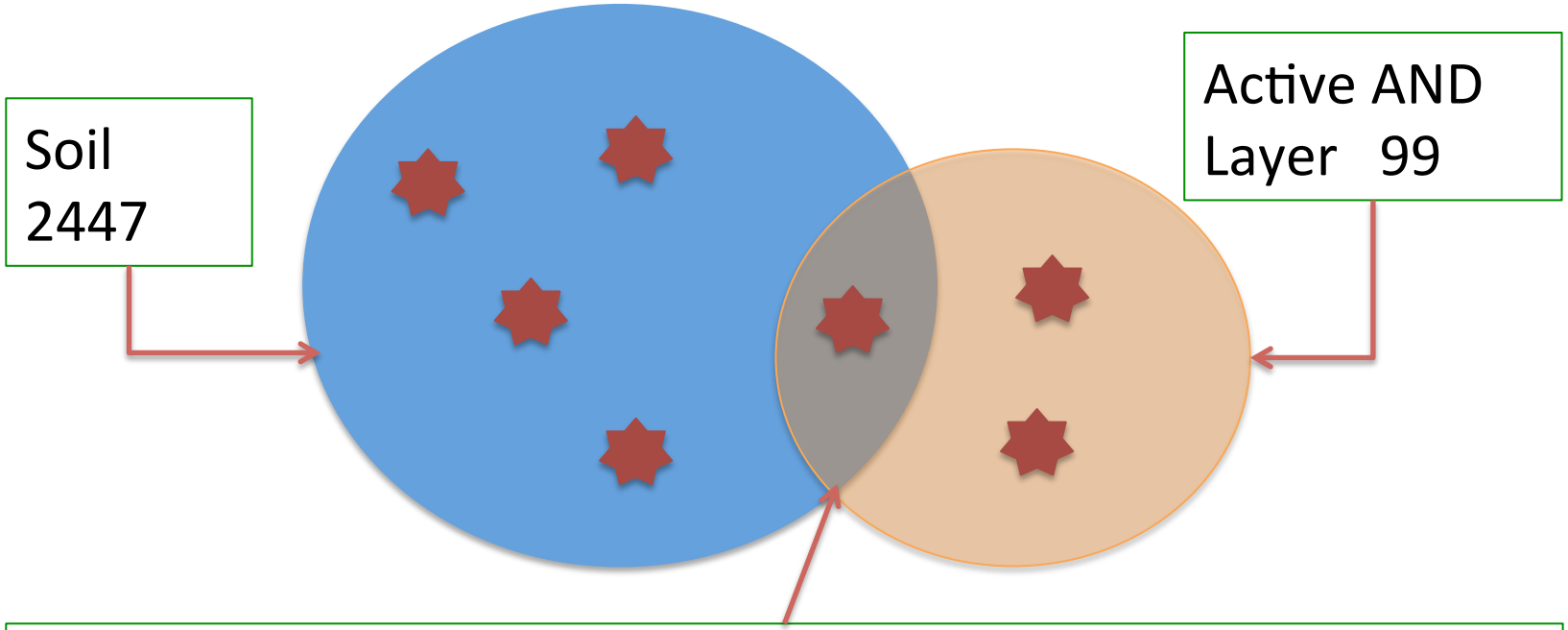


Full Text search: Biomass OR Humus 192

Humus is a type of Biomass:
5 additional datasets are found

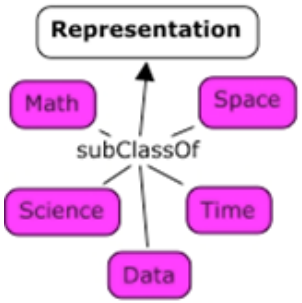
Line Pouchard, pouchard@purdue.edu

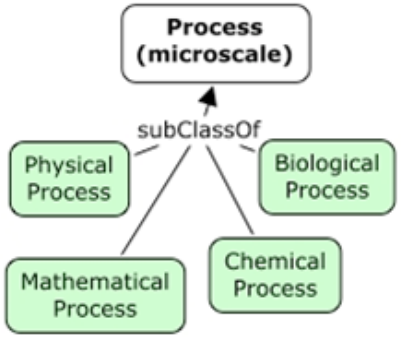
Why improve precision?

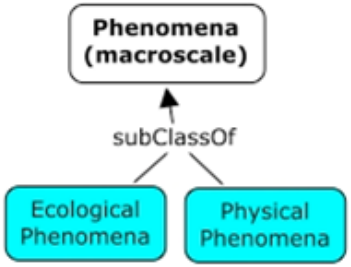


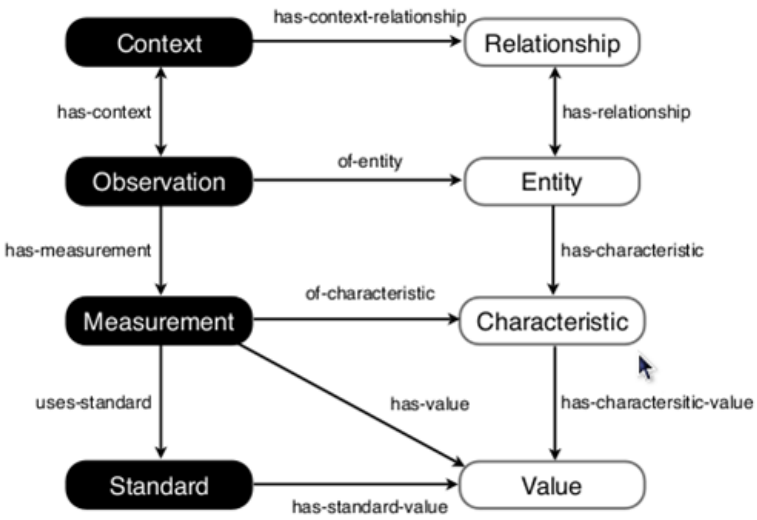
There are dozens of facets to choose from
Active Layer DOES NOT appear as a facet
The user must enter a new query to find the 37 datasets

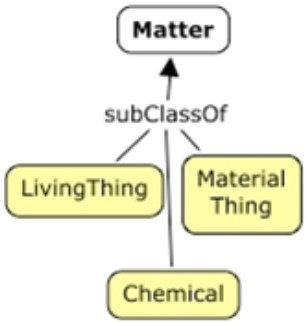
Using ontology entities

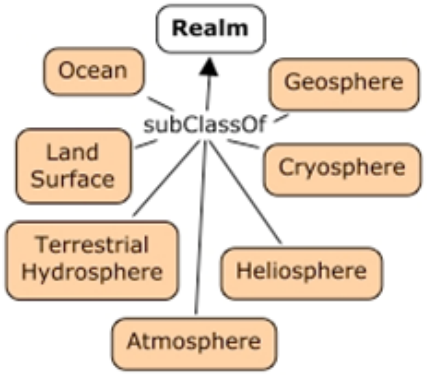


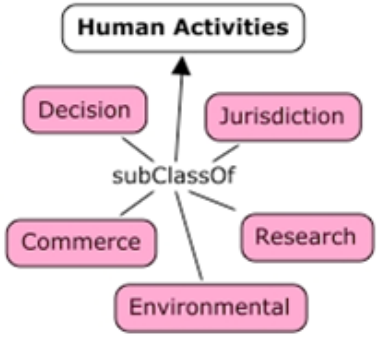




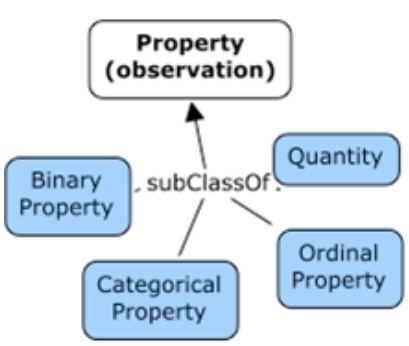


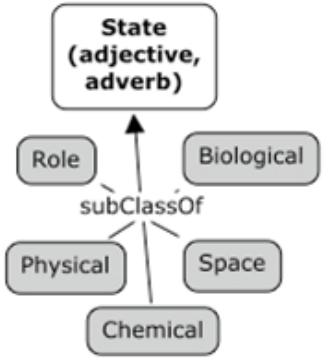


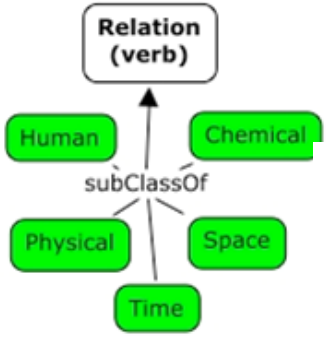




OBOE: Extensible Observation Ontology, Ben Leinfelder, UCSB

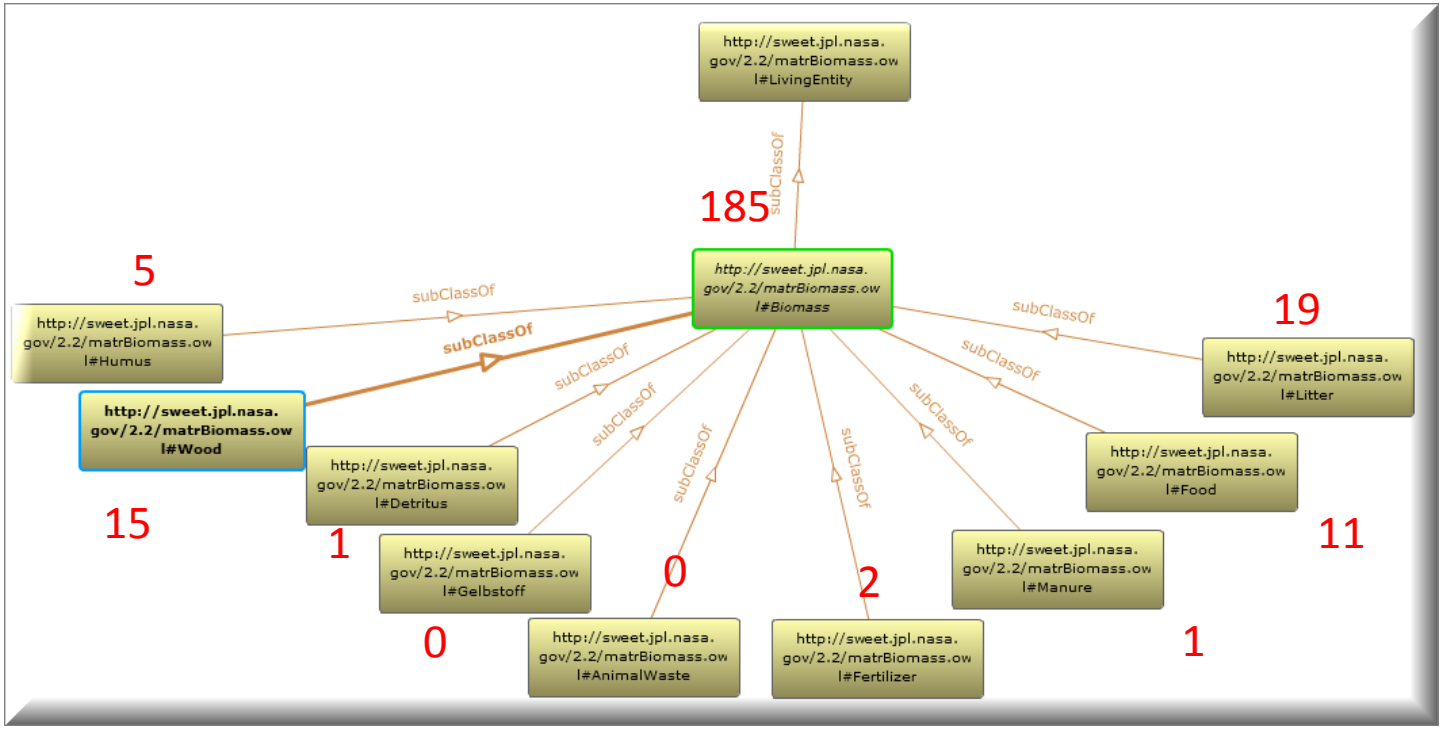






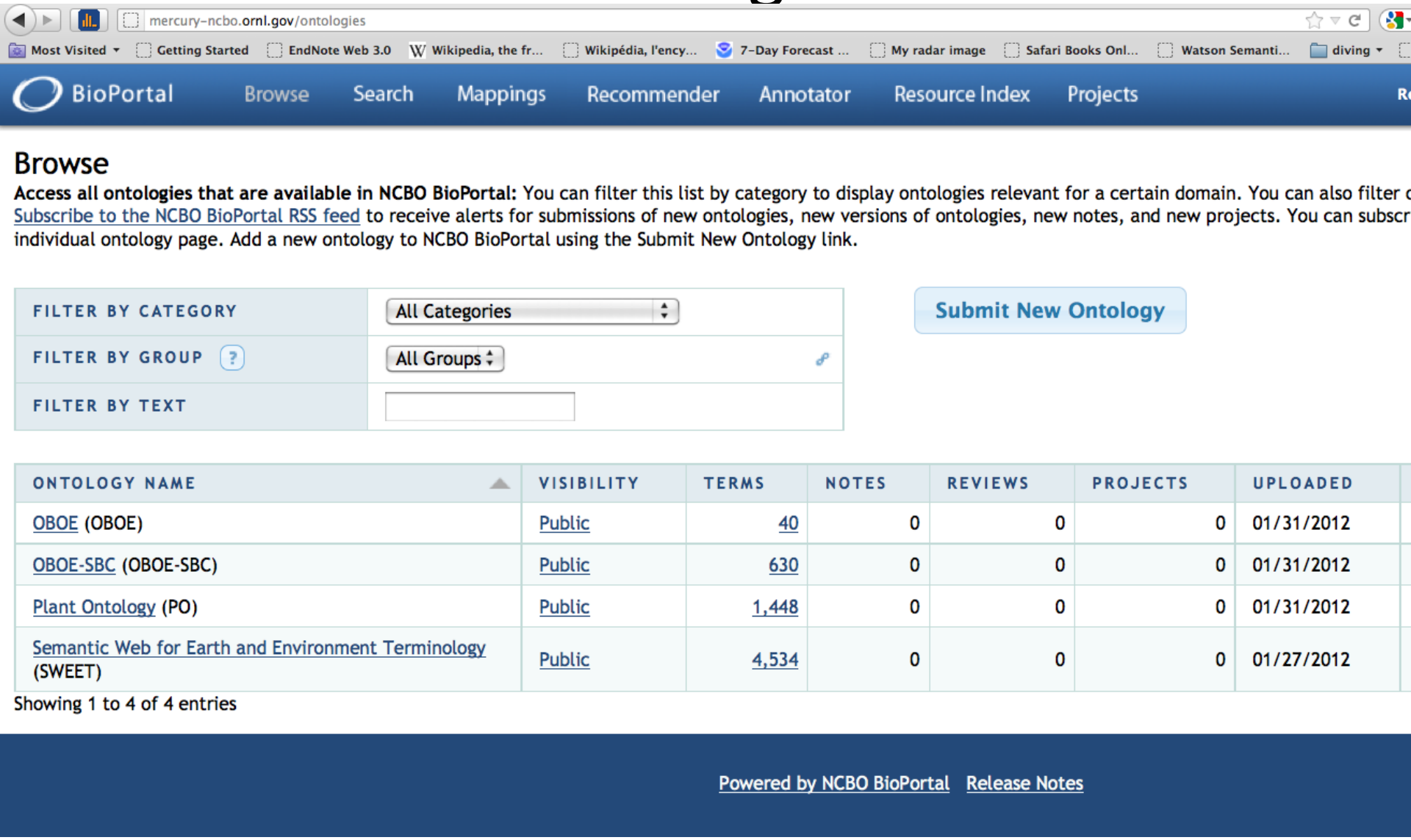
SWEET: Semantic Web for Earth and Environmental Terminology, Rob Raskin et al, NASA JPL

Search for “biomass” in SWEET ontologies

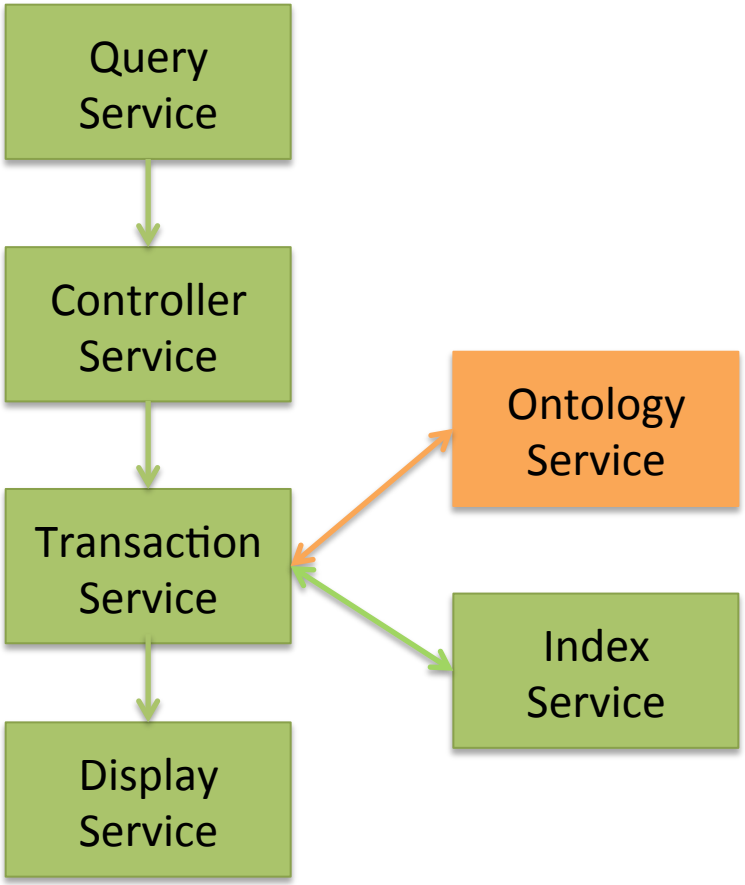


- Subclasses: Fertilizer, Living Entity, Litter, Food, Wood, Humus, Detritus, Manure, Animal Waste
- Numbers represent count of metadata documents per concepts
- Total: 239 records

BioPortal provides access to ontologies

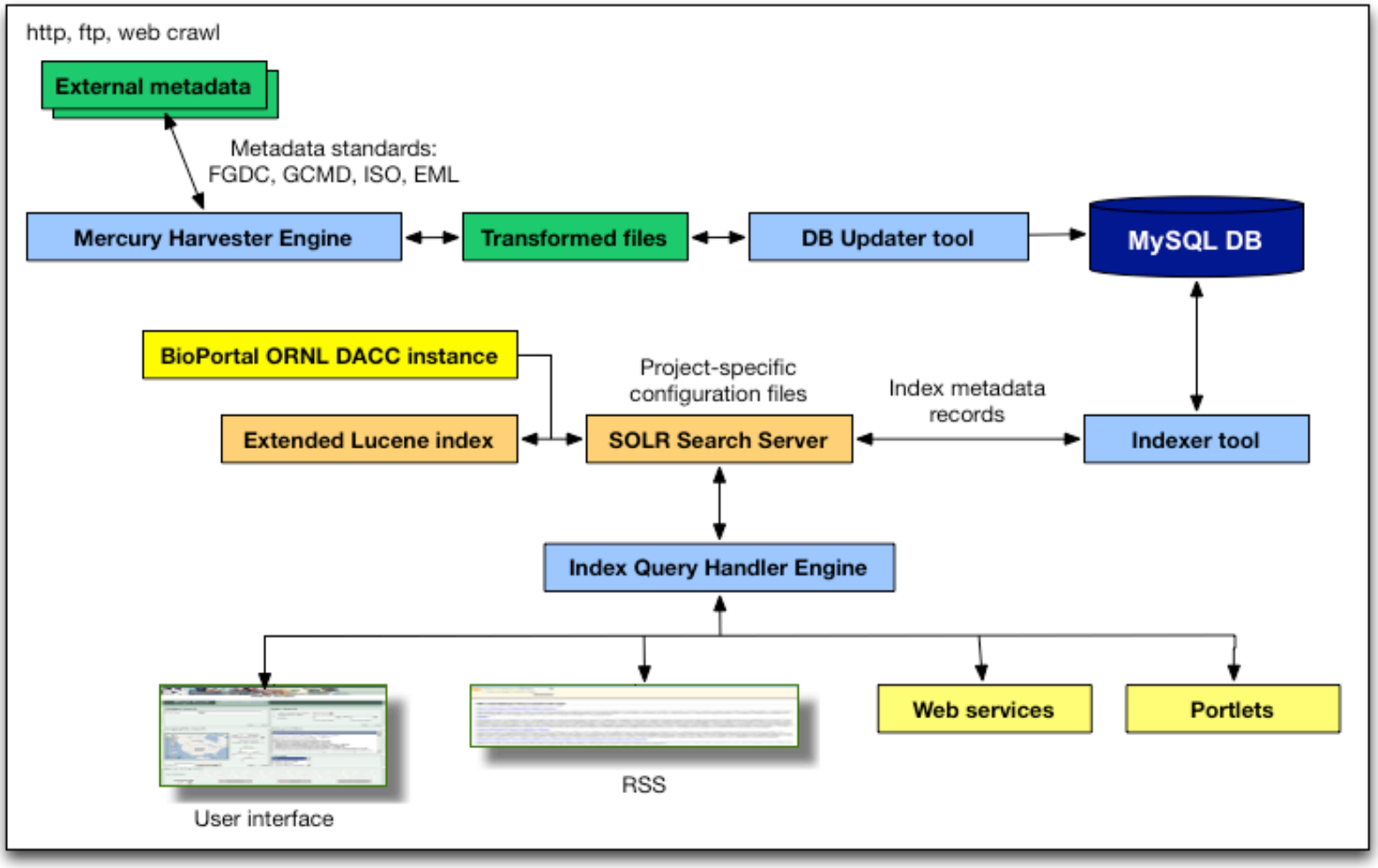


Coupling Mercury and BioPortal

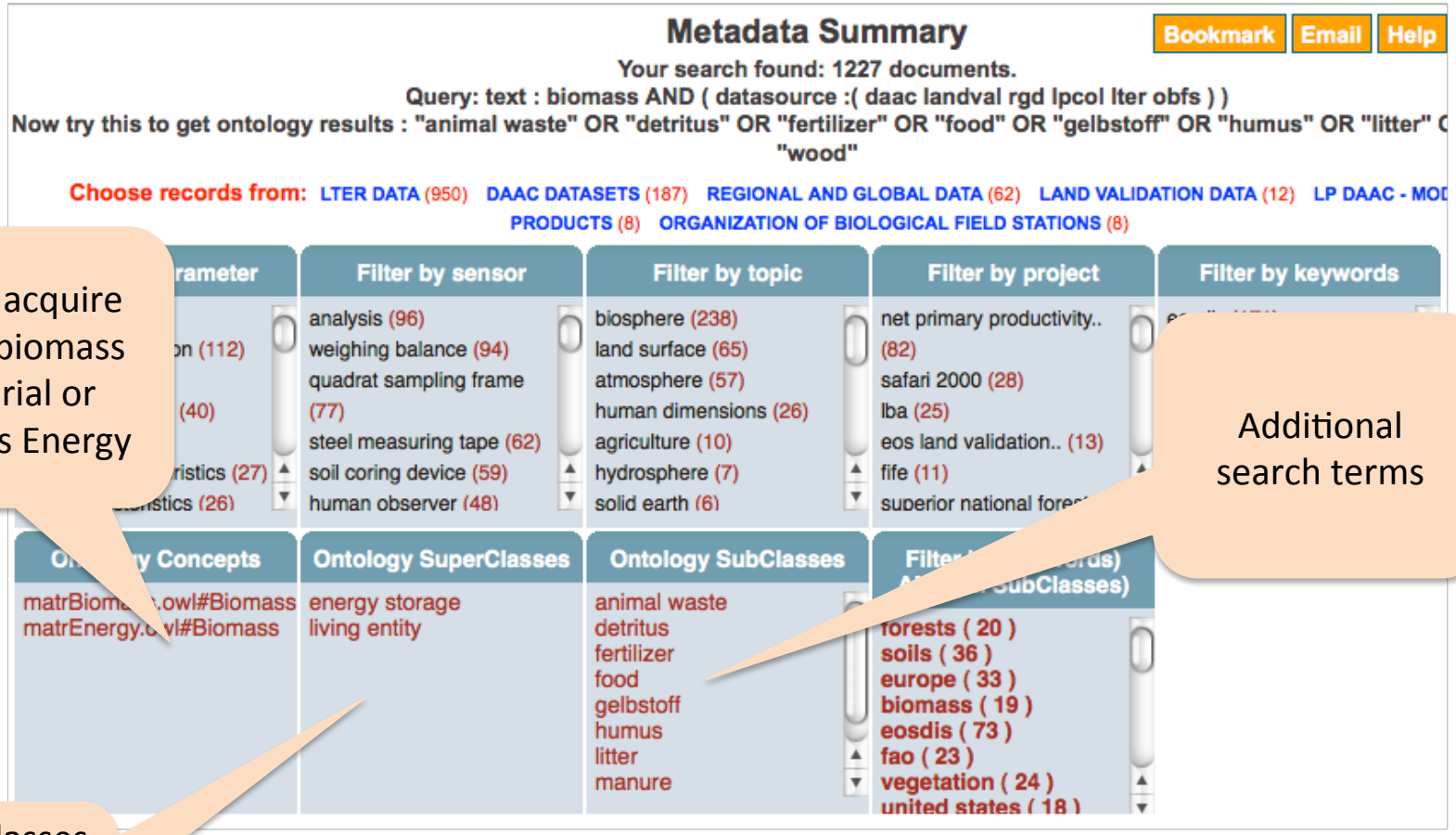


- Uses BioPortal Rest Services for programmatic access
- Returns ontology concepts, super- and sub-classes
- Provides additional keywords
- Provides context
- Uses these for new searches

Internal architecture



Ontology-based search results

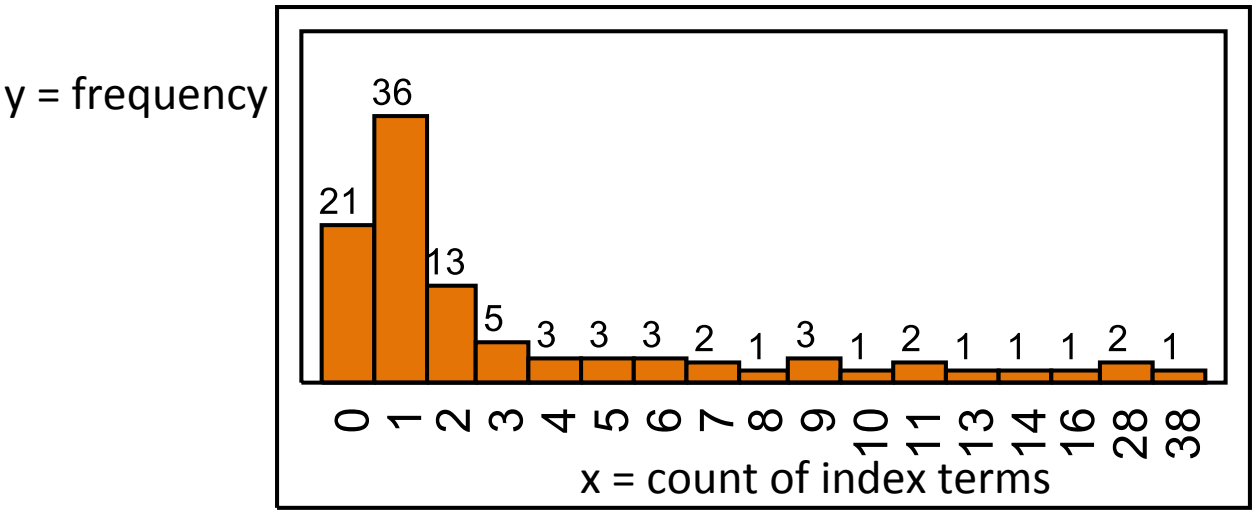


Concepts acquire context: biomass as Material or biomass as Energy

Additional search terms

Super-classes may have different properties

Matching the top 100 Mercury parameters to ontology terms



- Frequency count: 79% of the Top 100 keywords have at least one match in the chosen ontologies
- N = 99, 2 values missing (plant, leaf)
- water : 38
- air, carbon = 28

Line Pouchard, pouchard@purdue.edu

Limitations

User-friendly display

- Current display may be confusing. What are the options?
  - send the user to a new page
  - implement a new display dynamically driven by ontology relationships

Ontology content

- SWEET provides a good basis, but needs to be further specified for the needs of this Data Center
- Many ontologies provide only few relationships

Implementation

- Adding ontology entities to a keyword index helps with recall but cannot substitute for semantic annotations of the metadata documents

Thank you

- ORNL DAAC and Mercury
  - http://mercury.ornl.gov
- ORNL DAAC ontology service
  - http://mercury.ornl.gov/OntologyDemo
- ORNL DAAC instance of BioPortal
  - http://mercury-ncbo.ornl.gov
- Stanford Center for Biomedical Informatics Research BioPortal
  - http://bioportal.bioontology.org
- Stanford Center for Biomedical Informatics Research Protégé ontology editor
  - http://protege.stanford.edu