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

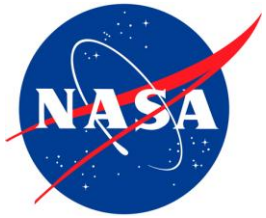
Line C. Pouchard,* Robert B. Cook,* Jim Green,*
Natasha Noy,** Giri Palanisamy*

Oak Ridge National Laboratory*

Stanford University**

Presented to the Woods Hole Coastal and Marine Science Center

February 7, 2012



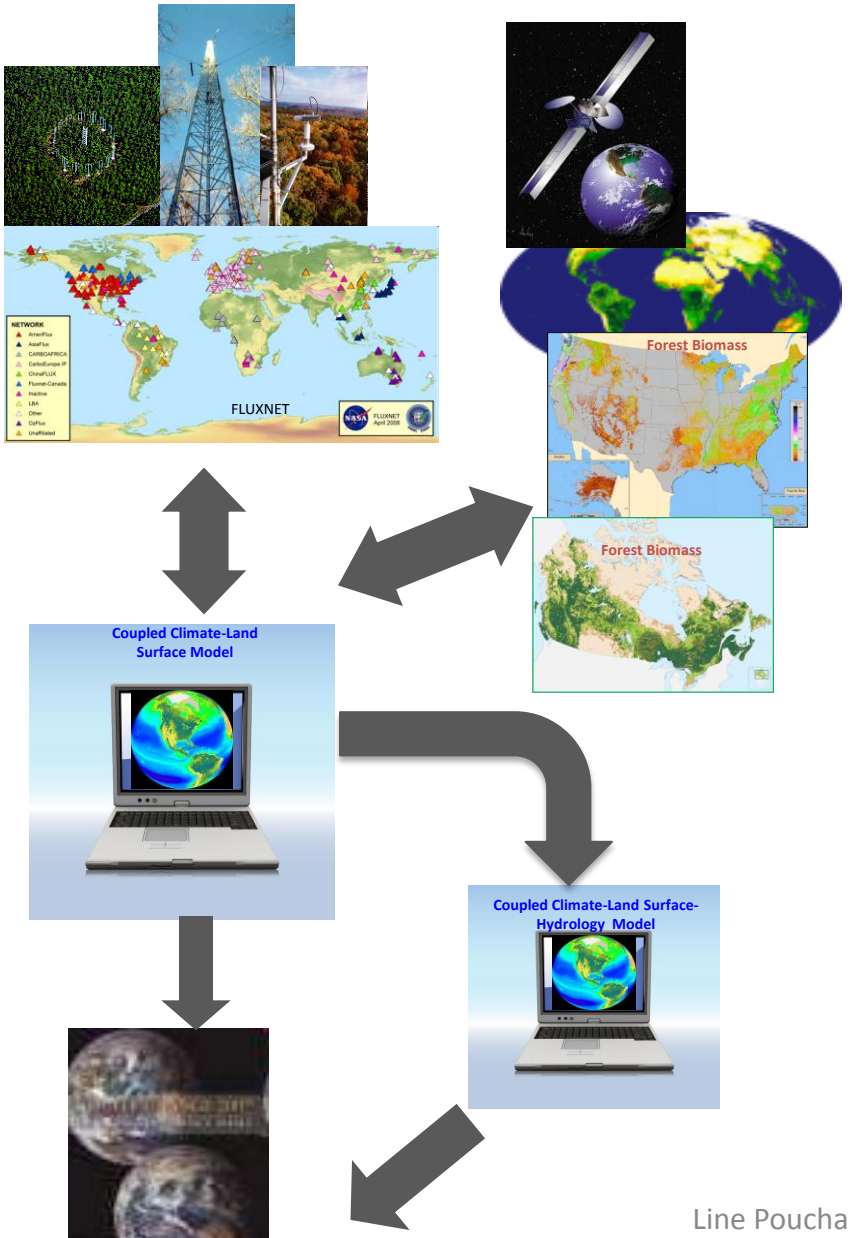
Line Pouchard, pouchardlc@ornl.gov

Linked Science

- Scientific collaborations are growing and becoming more interdisciplinary
  - key trends in the role of data loosely referred to as “linked Science”
  - end-to-end perspective
  - means to communicate are key
- First Linked Science Workshop, International Semantic Web Conference, October 2011
- Second Linked Science Workshop, ISWC, October 2012, Boston

ISWC 2011

What is Linked Science?

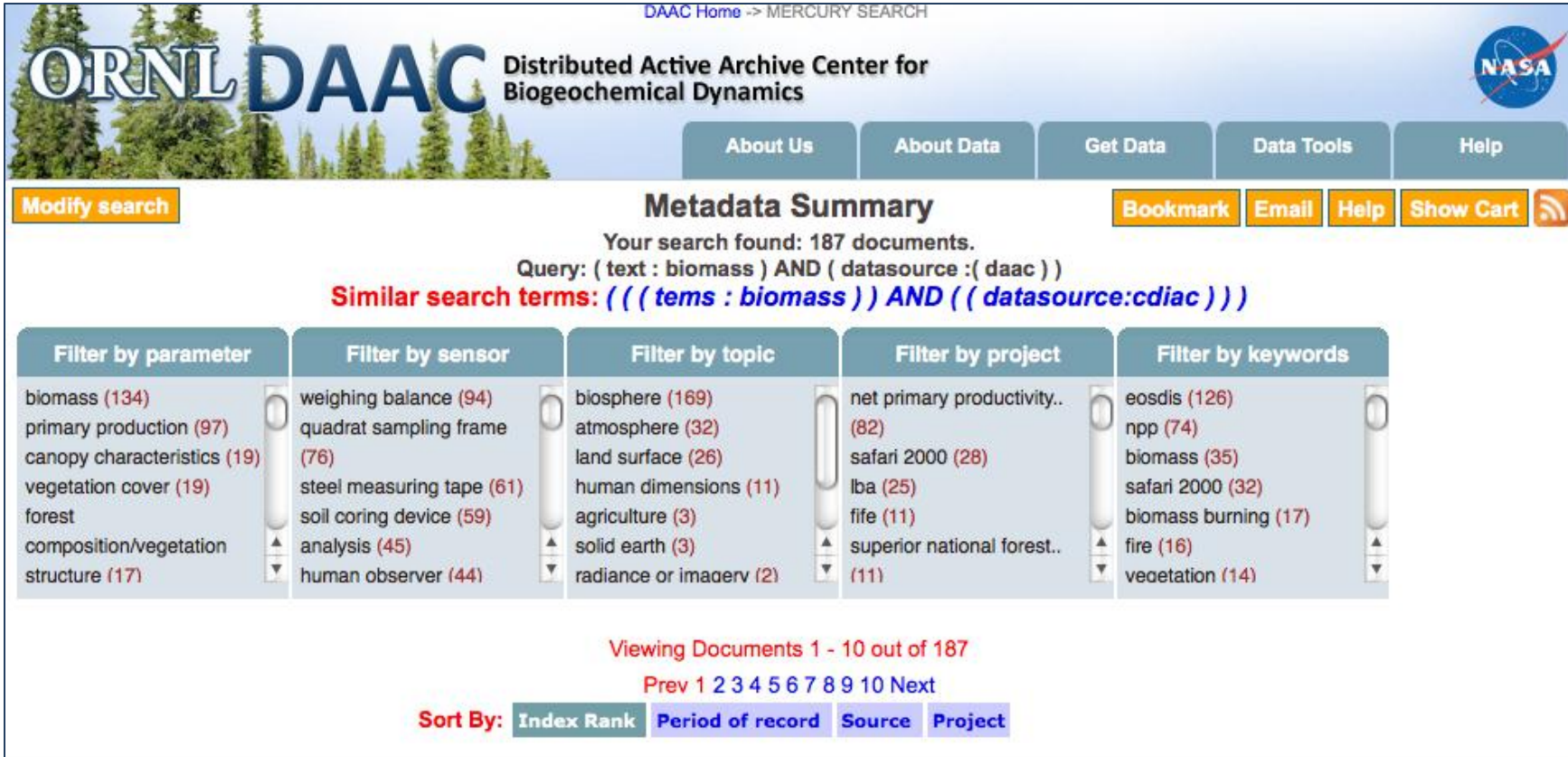


- Discovery and access from heterogeneous sources
  - Simulations, models, experiments, remote sensing, GIS, molecular and omics databases, publications
- Metadata and semantics integration
- Workflows, scenario development, data and process re-use, provenance
- Engaging communities of scientists, educators, librarians, developers, volunteers
- Relies upon cyber-infrastructure promoting open source
- Complex systems of systems, networks of projects, repositories, archives, publishers

Line Pouchard, pouchardlc@ornl.gov

Current data discovery system

- Each dataset is represented by a metadata XML document
- Data discovery is based on the content of several XML elements
- Mercury supports Federal Geographic Data Committee keywords, Dublin and Darwin cores, Ecological Mark-up Language and ISO 19115 (location)
- Mercury holds over 100,000 metadata records from several providers
- Over 1000 datasets focus on biogeochemical dynamics, terrestrial ecology and environmental processes, accounting for 2 TB in ORNL DAAC



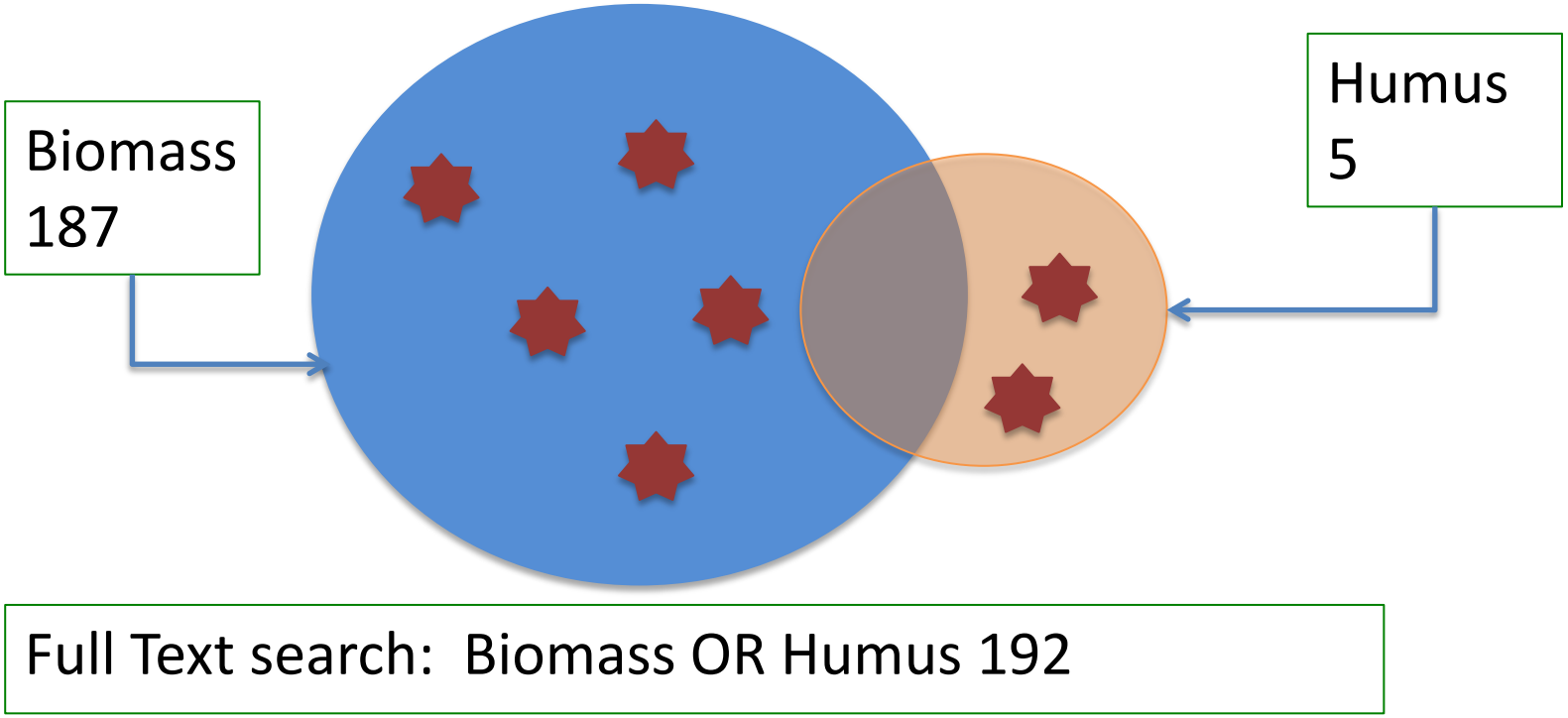
Recall and Precision: two relevance metrics from Information Retrieval

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

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

- 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?

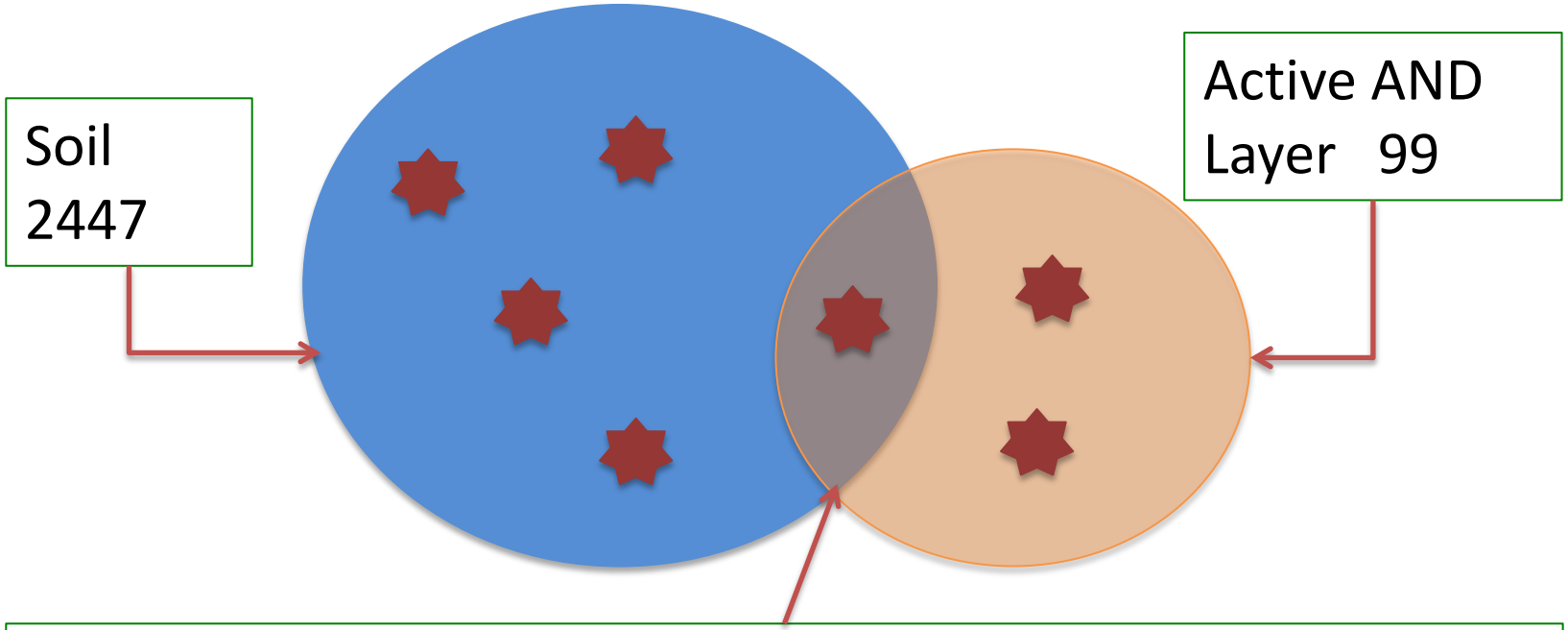
Why improve Recall for ORNL DAAC?



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

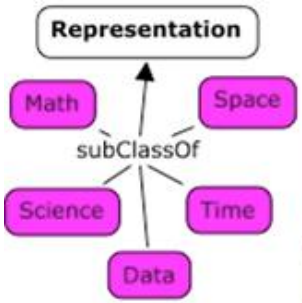
Line Pouchard, pouchardlc@ornl.gov

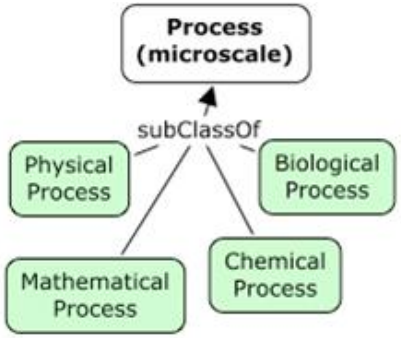
Why improve precision?

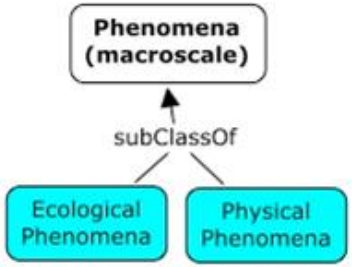


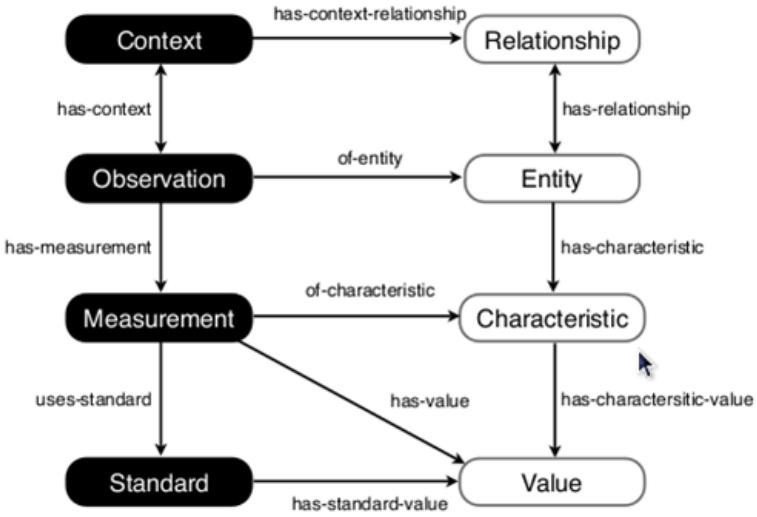
There are dozens of facets to choose from
Active Layer DOES NOT appear as a facet
The user must enter a new query to find the 37 datasets

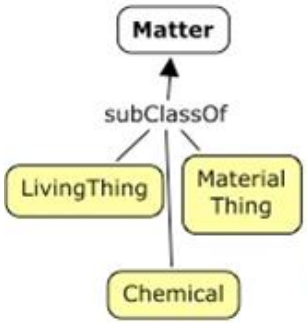
Using ontology entities

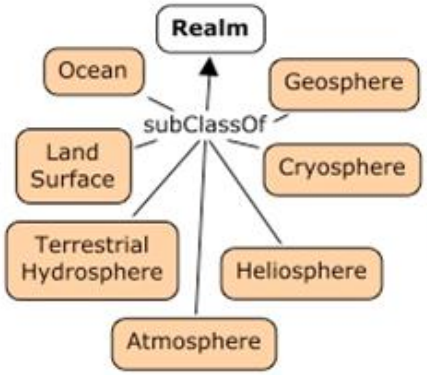


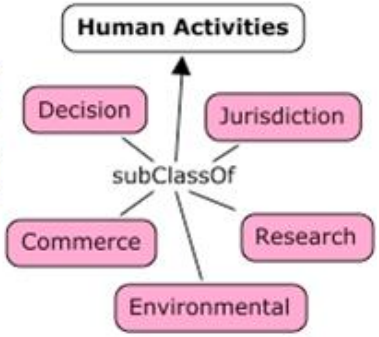




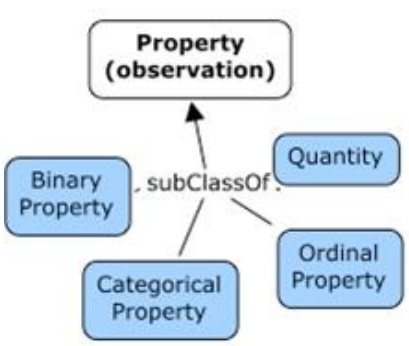


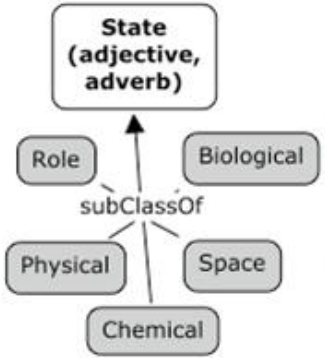


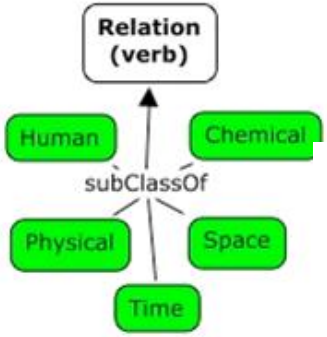




OBOE: Extensible Observation Ontology, Ben Leinfelder, UCSB

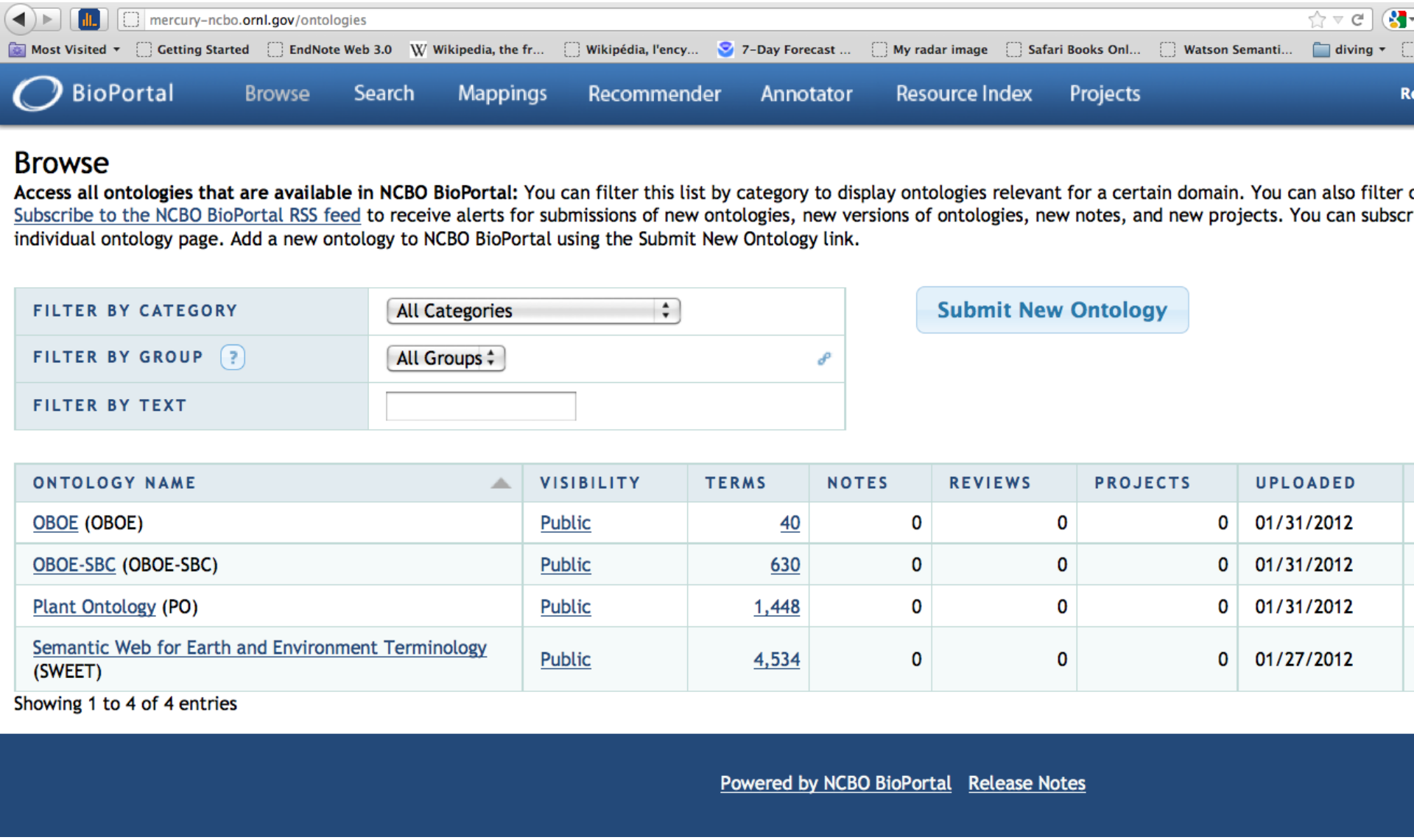




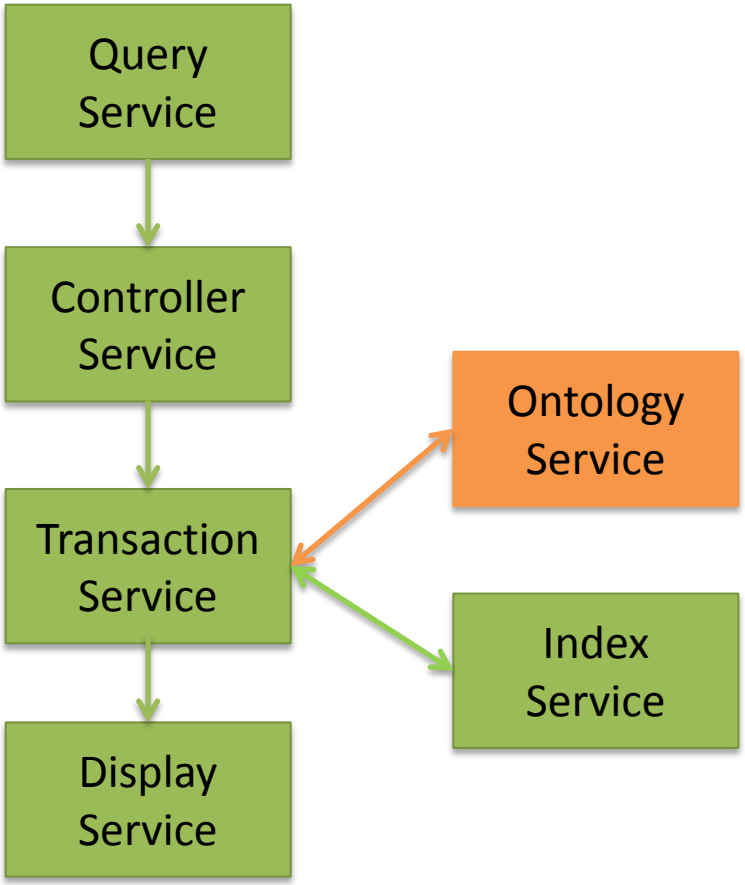


SWEET: Semantic Web for Earth and Environmental Terminology, Rob Raskin, JPL

BioPortal provides access to ontologies



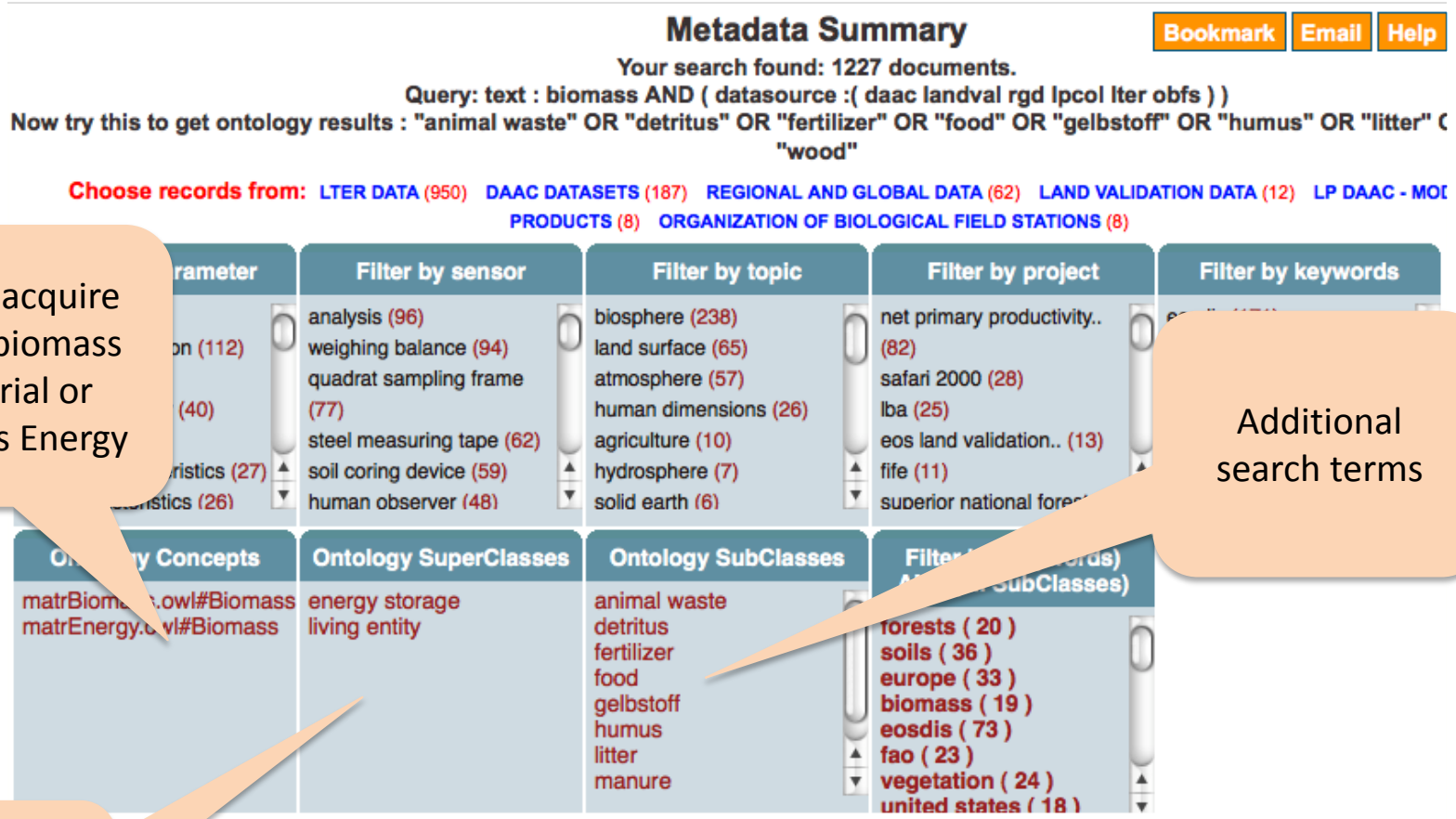
Coupling Mercury and BioPortal



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

Line Pouchard, pouchardlc@ornl.gov

Ontology-based search results

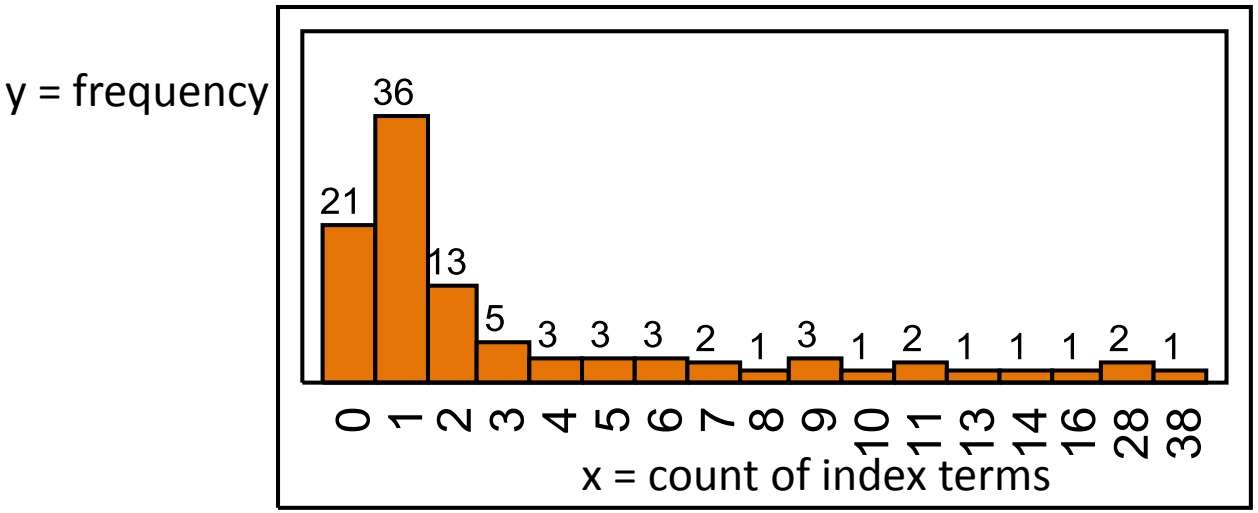


Concepts acquire context: biomass as Material or biomass as Energy

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, pouchardlc@ornl.gov

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

Future Work

- Rethink the approach of metadata capture to include ontology concepts and relations
- Extend SWEET and/or add ontologies for common use cases in the ORNL DAAC
- Dynamically driven display for improved clarity and comprehension

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