

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

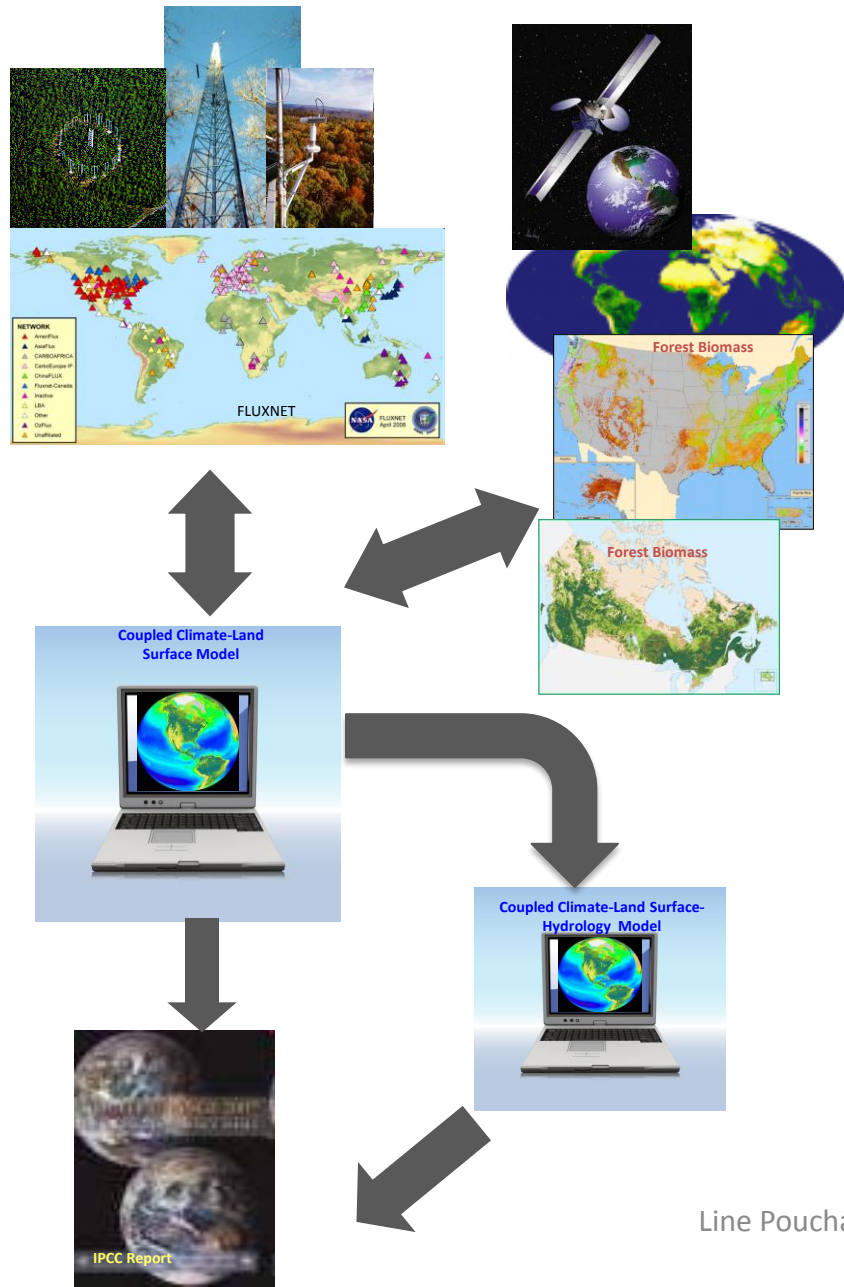


Linked Science

- 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



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

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

DAAC Home -> MERCURY SEARCH

ORNL DAAC Distributed Active Archive Center for Biogeochemical Dynamics

[About Us](#) [About Data](#) [Get Data](#) [Data Tools](#) [Help](#)

[Modify search](#) **Metadata Summary** [Bookmark](#) [Email](#) [Help](#) [Show Cart](#)

Your search found: 187 documents.
Query: (text : biomass) AND (datasource :(daac))
Similar search terms: (((tems : biomass)) AND ((datasource:cdiac)))

Filter by parameter	Filter by sensor	Filter by topic	Filter by project	Filter by keywords
biomass (134) primary production (97) canopy characteristics (19) vegetation cover (19) forest composition/vegetation structure (17)	weighing balance (94) quadrat sampling frame (76) steel measuring tape (61) soil coring device (59) analysis (45) human observer (44)	biosphere (169) atmosphere (32) land surface (26) human dimensions (11) agriculture (3) solid earth (3) radiance or imaerv (2)	net primary productivity.. (82) safari 2000 (28) lba (25) fife (11) superior national forest.. (11)	eosdis (126) npp (74) biomass (35) safari 2000 (32) biomass burning (17) fire (16) vegetation (14)

Viewing Documents 1 - 10 out of 187
Prev 1 2 3 4 5 6 7 8 9 10 Next

Sort By: [Index Rank](#) [Period of record](#) [Source](#) [Project](#)

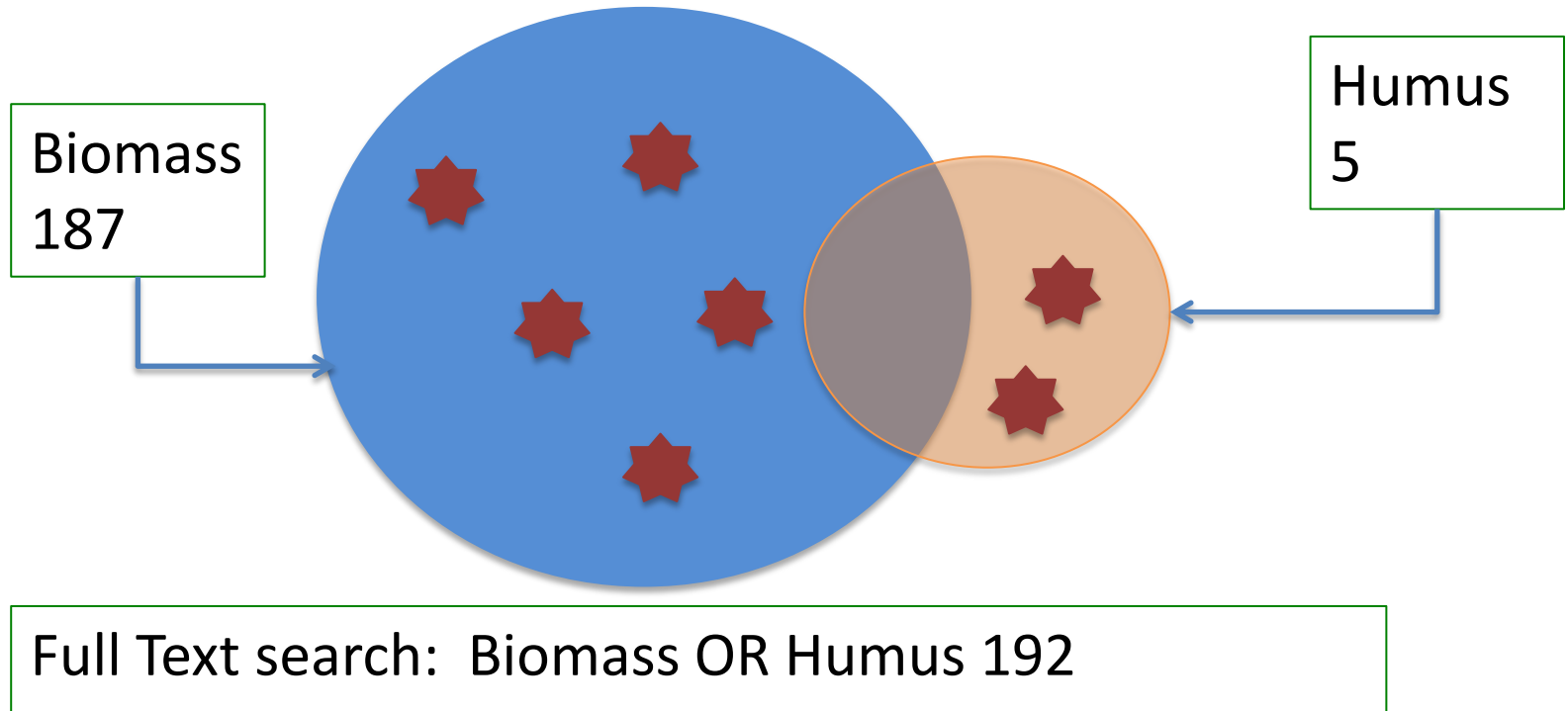
Recall and Precision: two relevance metrics from Information Retrieval

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

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

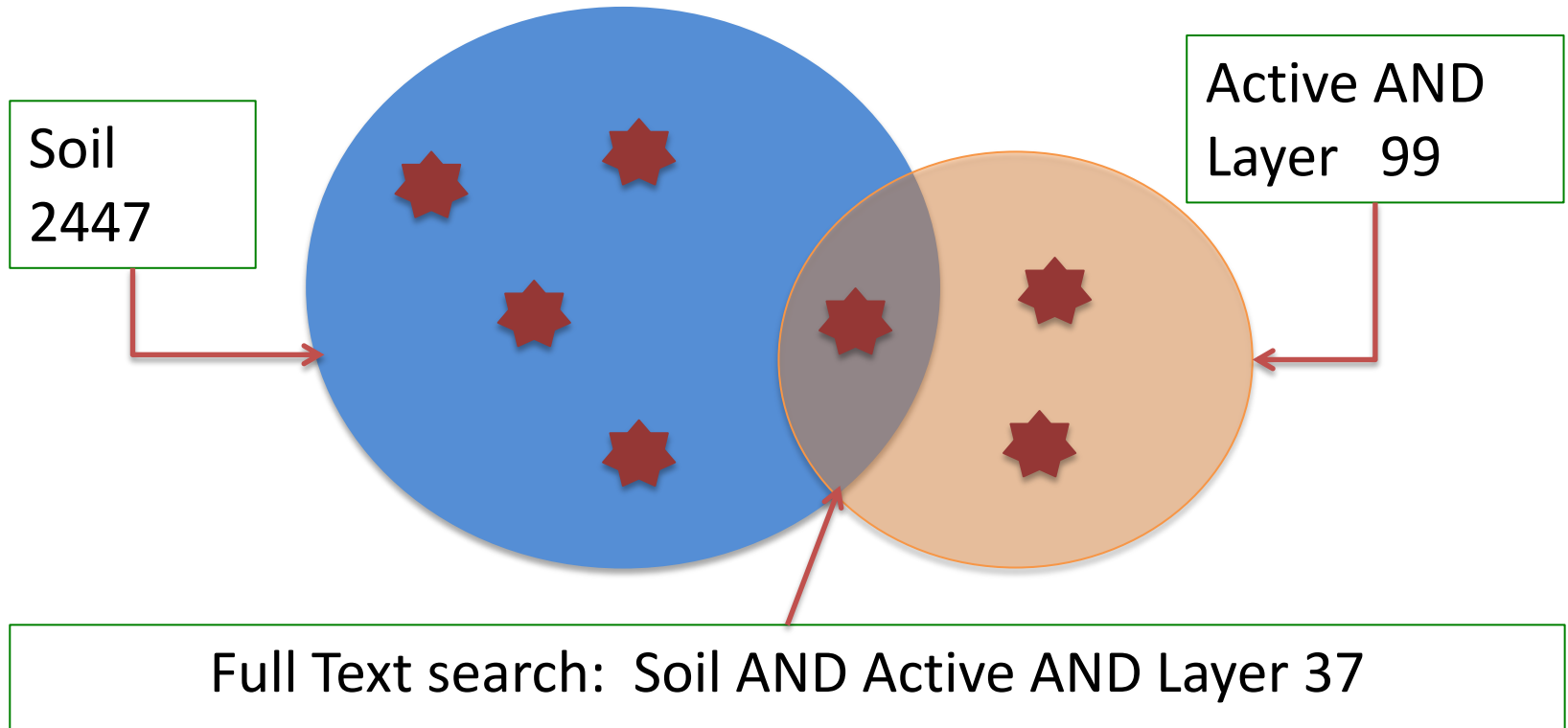
- 100% recall means all relevant documents were retrieved, but maybe also many non-relevant ones
- 100% precision means all retrieved documents were relevant, but where relevant documents not retrieved?

Why improve Recall for ORNL DAAC?



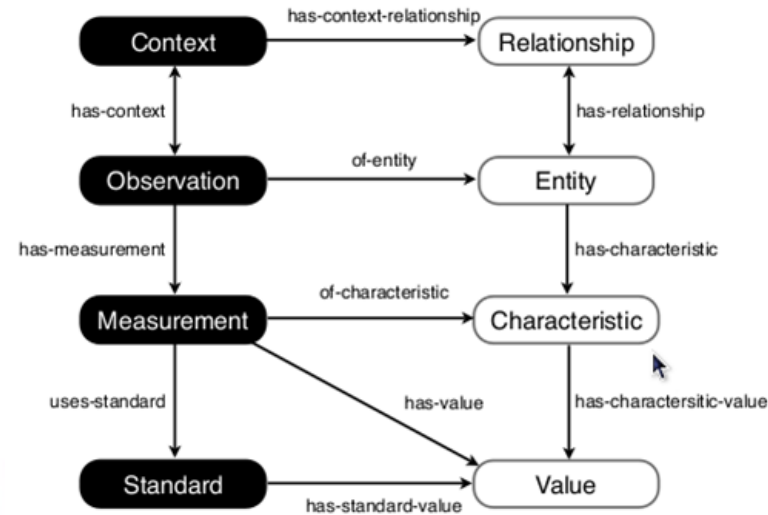
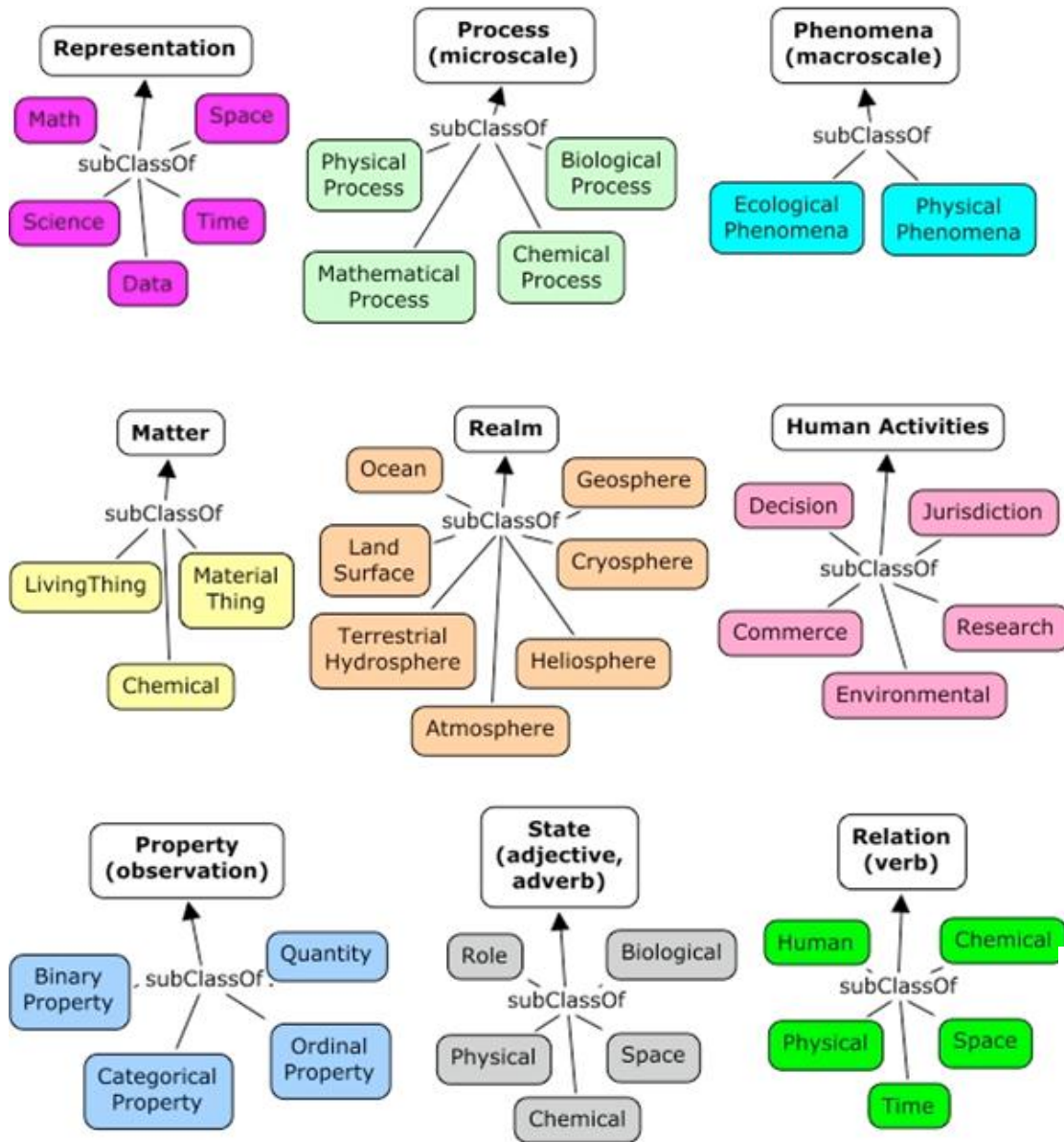
Humus is a type of Biomass:
5 additional datasets are found

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

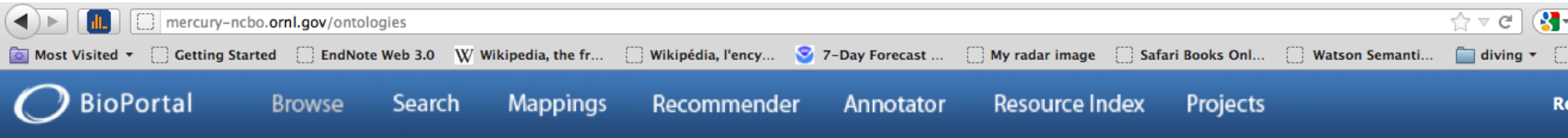
Using ontology entities



OBOE: Extensible
Observation Ontology,
Ben Leinfelder, UCSB

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

BioPortal provides access to ontologies



Browse

Access all ontologies that are available in NCBO BioPortal: You can filter this list by category to display ontologies relevant for a certain domain. You can also filter by group. [Subscribe to the NCBO BioPortal RSS feed](#) to receive alerts for submissions of new ontologies, new versions of ontologies, new notes, and new projects. You can subscribe to an individual ontology page. Add a new ontology to NCBO BioPortal using the [Submit New Ontology](#) link.

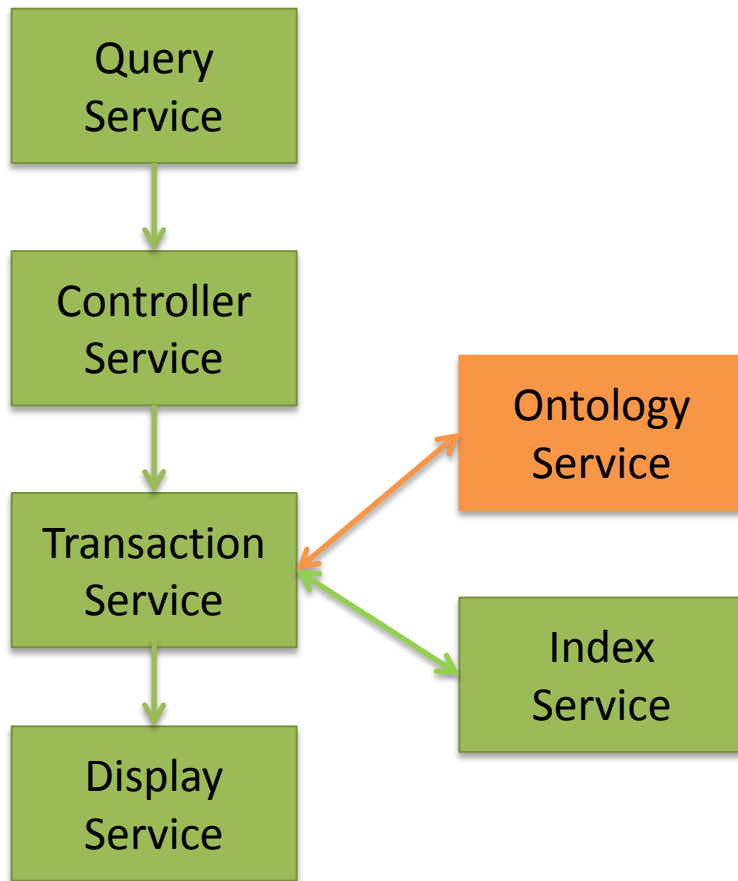
FILTER BY CATEGORY	<input type="text" value="All Categories"/>
FILTER BY GROUP ?	<input type="text" value="All Groups"/> ?
FILTER BY TEXT	<input type="text"/>

[Submit New Ontology](#)

ONTOLOGY NAME	VISIBILITY	TERMS	NOTES	REVIEWS	PROJECTS	UPLOADED
OBOE (OBOE)	Public	40	0	0	0	01/31/2012
OBOE-SBC (OBOE-SBC)	Public	630	0	0	0	01/31/2012
Plant Ontology (PO)	Public	1,448	0	0	0	01/31/2012
Semantic Web for Earth and Environment Terminology (SWEET)	Public	4,534	0	0	0	01/27/2012

Showing 1 to 4 of 4 entries

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

Ontology-based search results

[Bookmark](#) [Email](#) [Help](#)

Metadata Summary

Your search found: 1227 documents.

Query: text : biomass AND (datasource :(daac landval rgd lpcol lter obfs))

Now try this to get ontology results : "animal waste" OR "detritus" OR "fertilizer" OR "food" OR "gelbstoff" OR "humus" OR "litter" OR "manure"

Choose records from: [LTER DATA \(950\)](#) [DAAC DATASETS \(187\)](#) [REGIONAL AND GLOBAL DATA \(62\)](#) [LAND VALIDATION DATA \(12\)](#) [LP DAAC - MOI PRODUCTS \(8\)](#) [ORGANIZATION OF BIOLOGICAL FIELD STATIONS \(8\)](#)

Parameter	Filter by sensor	Filter by topic	Filter by project	Filter by keywords
analysis (112)	analysis (96)	biosphere (238)	net primary productivity.. (82)	forests (20)
weighing balance (94)	weighing balance (94)	land surface (65)	safari 2000 (28)	soils (36)
quadrat sampling frame (77)	quadrat sampling frame (77)	atmosphere (57)	lba (25)	europa (33)
steel measuring tape (62)	steel measuring tape (62)	human dimensions (26)	eos land validation.. (13)	biomass (19)
soil coring device (59)	soil coring device (59)	agriculture (10)	fife (11)	eosdis (73)
human observer (48)	human observer (48)	hydrosphere (7)	superior national forest (18)	fao (23)
		solid earth (6)		vegetation (24)
				united states (18)

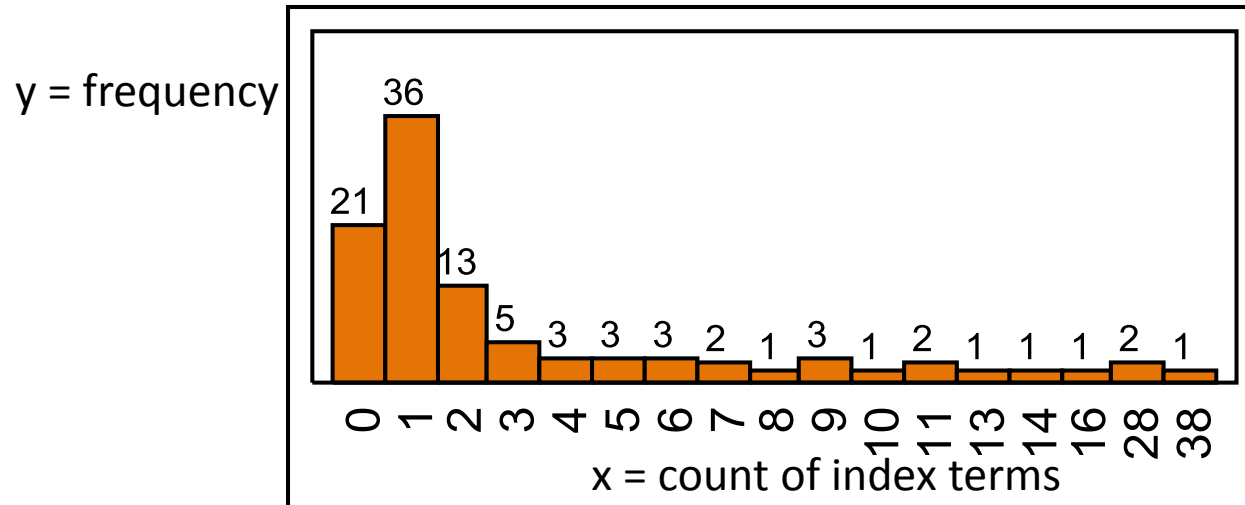
Ontology Concepts	Ontology SuperClasses	Ontology SubClasses	Filter by keywords (SubClasses)
matrBiomass.owl#Biomass matrEnergy.owl#Biomass	energy storage living entity	animal waste detritus fertilizer food gelbstoff humus litter manure	forests (20) soils (36) europa (33) biomass (19) eosdis (73) fao (23) vegetation (24) united states (18)

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

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

The submitted manuscript has been authored by the US Department of Energy, Office of Science of the Oak Ridge National Laboratory, managed for the U. S. DOE by UT-Battelle, LLC, under contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

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>