

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

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

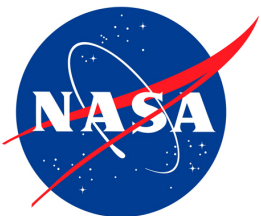
Purdue University Libraries*

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



Search For: Results/Page: 10 **SEARCH**

Hint: boolean operators, wildcards and phrases are allowed. ex: precipitation or (rain and *moisture content*)*

Show/Hide Advanced Options **HELP**

Fielded Search

FullText OR
 FullText OR
 FullText OR

Date Search

Collection Date during thru
 Publication Date thru
 Either mm/dd/yyyy mm/dd/yyyy

Geographic Search

List Areas in:
 USA WORLD
 Select from list
 Search Area:
 overlaps encloses
 North
 West East
 South

Place Name: view on map

Content Type

All
 Maps and Data
 Publications
 Tools and Software

Member Nodes


Eastern Sierra Geospatial Data Clearinghouse
 EMAN Data Set Library (Environment Canada Server)
 Fire Research and Management Exchange System (FRAME)
 Florida Fish and Wildlife Conservation Commission Data Lib
 Forest and Rangeland Ecosystem Science Center Metadata
 Global Forest Information Service

Ocean CO₂ CDIA C
 Office of Science
 WENDI
 DAAC
 USGS
 DataONE
 LBA
 CPTEC
 Brazil
 DADDI
 npon
 iRobin


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

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

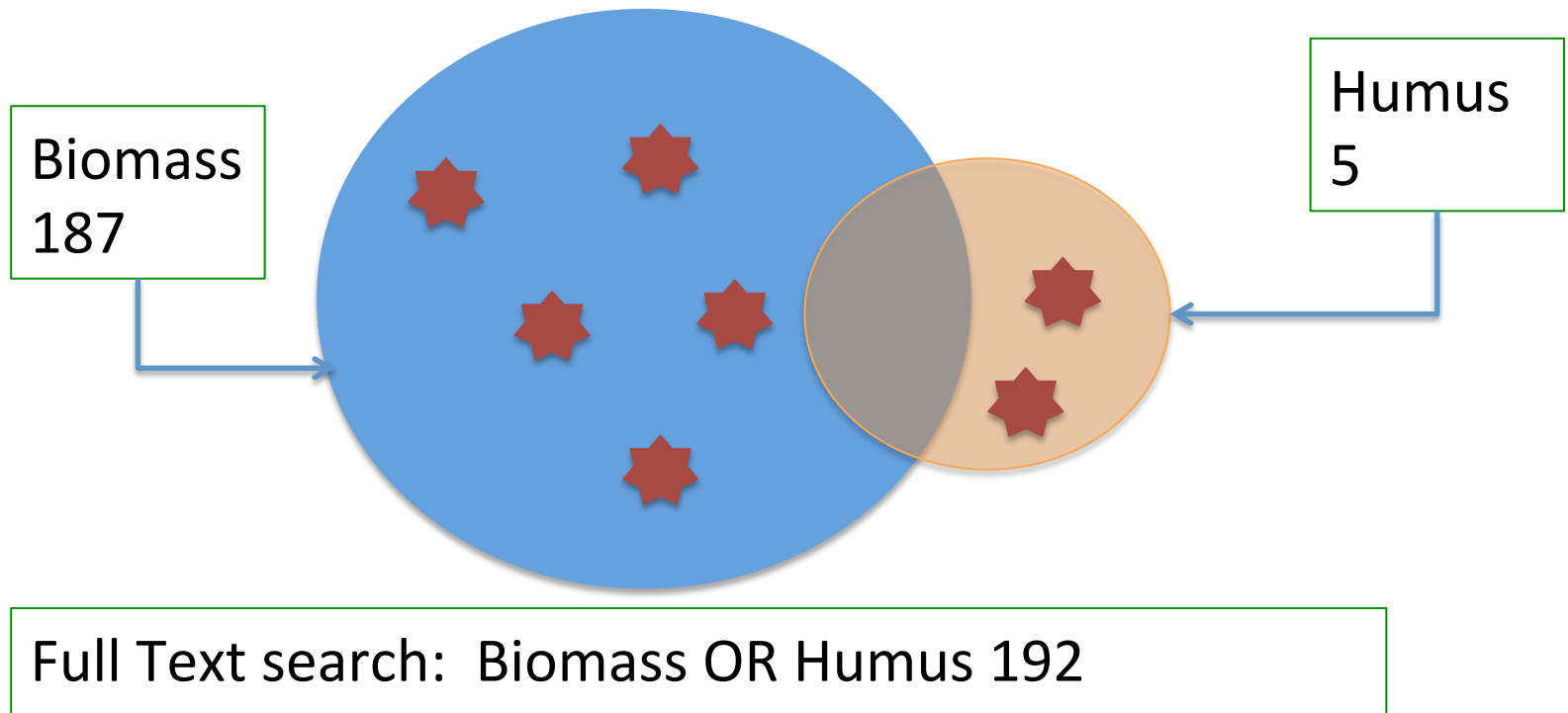
Your search found: 187 documents.
Query: (text : biomass) AND (datasource :(daac))
Similar search terms: (((tems : biomass)) AND ((datasource:cdiac)))

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

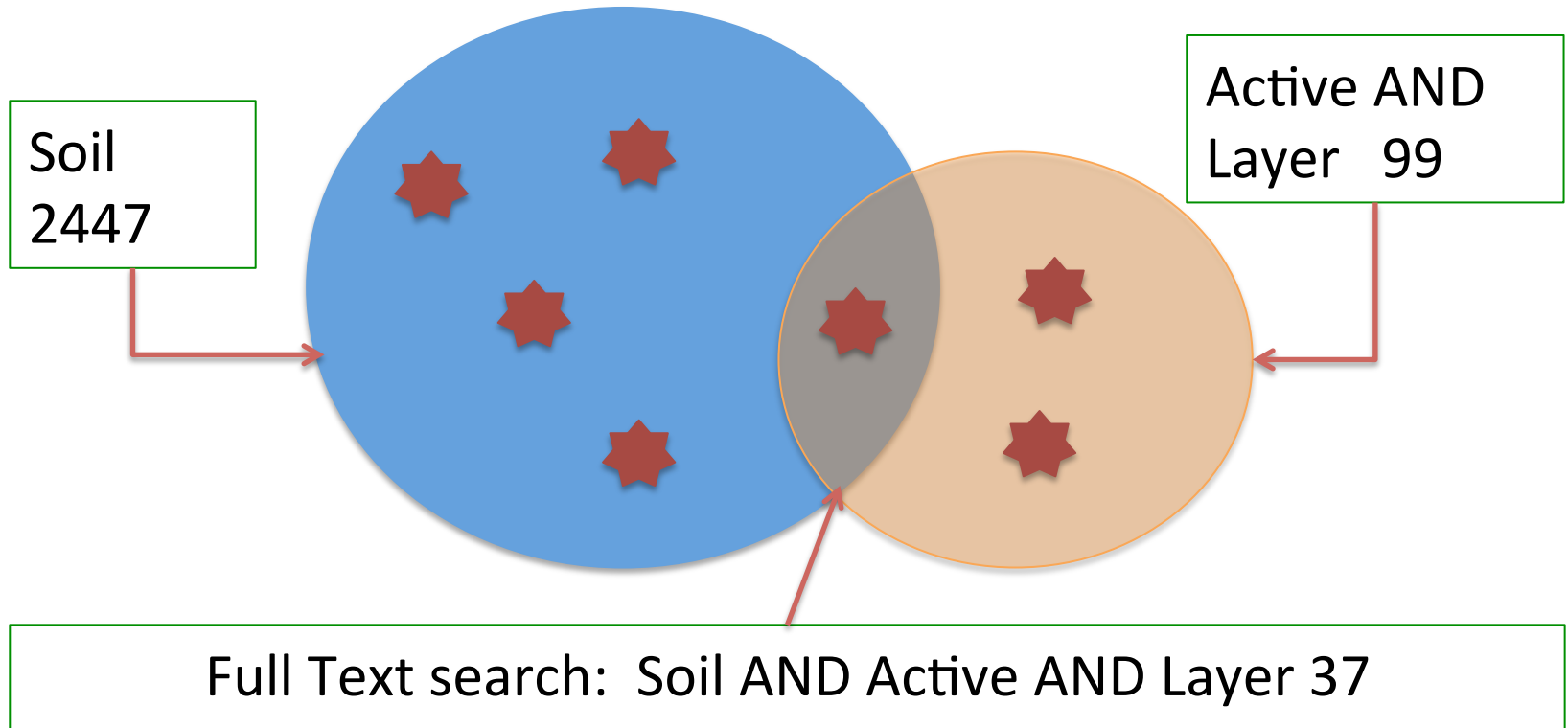
Sort By: Index Rank Period of record Source Project

Why improve Recall for ORNL DAAC?



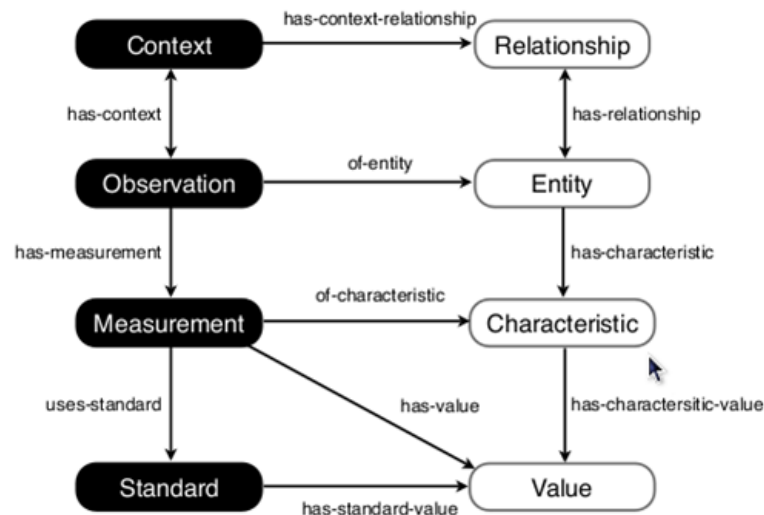
Humus is a type of Biomass:
5 additional datasets are found

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

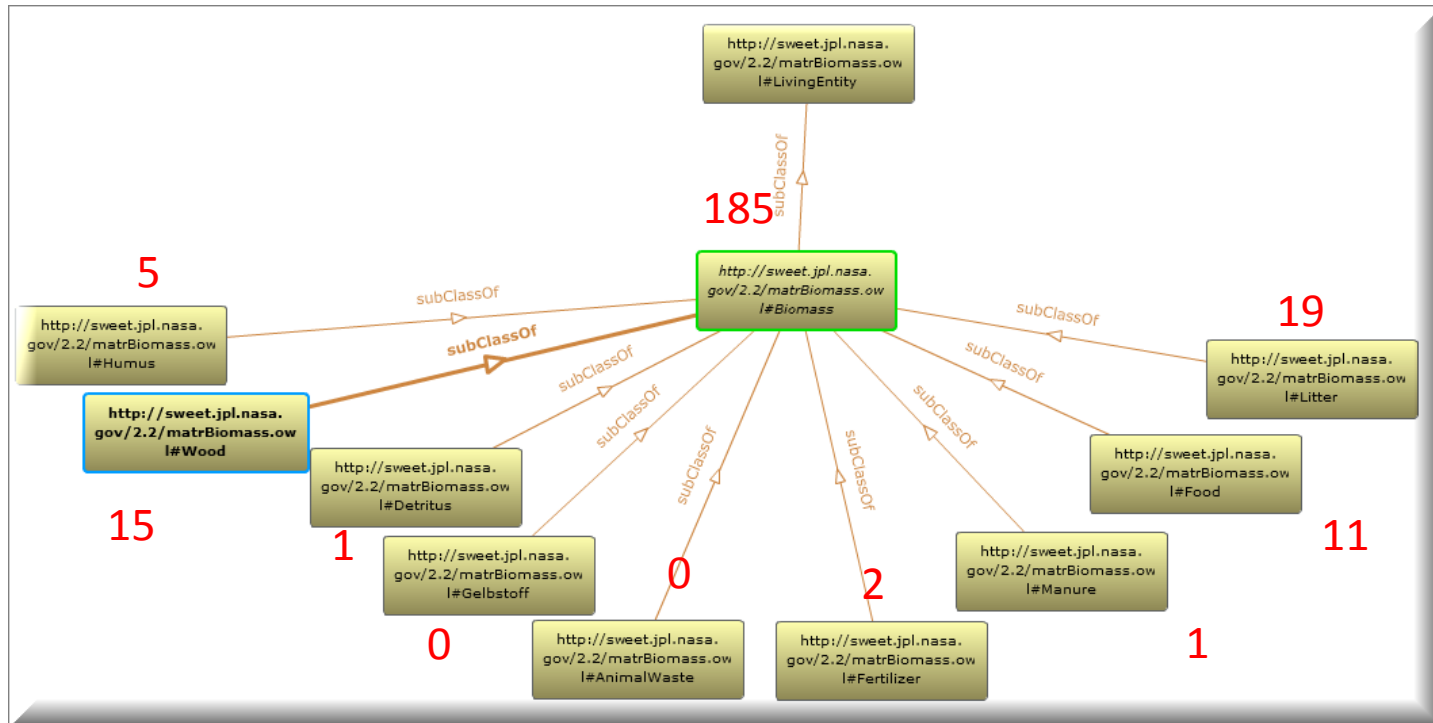
Using ontology entities



OBOE: Extensible
Observation Ontology,
Ben Leinfelder, UCSB

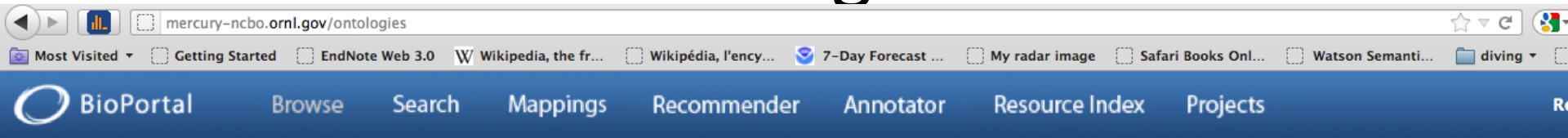
SWEET: Semantic Web for Earth and
Environmental Terminology, Rob
Raskin et al, NASA JPL

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



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.

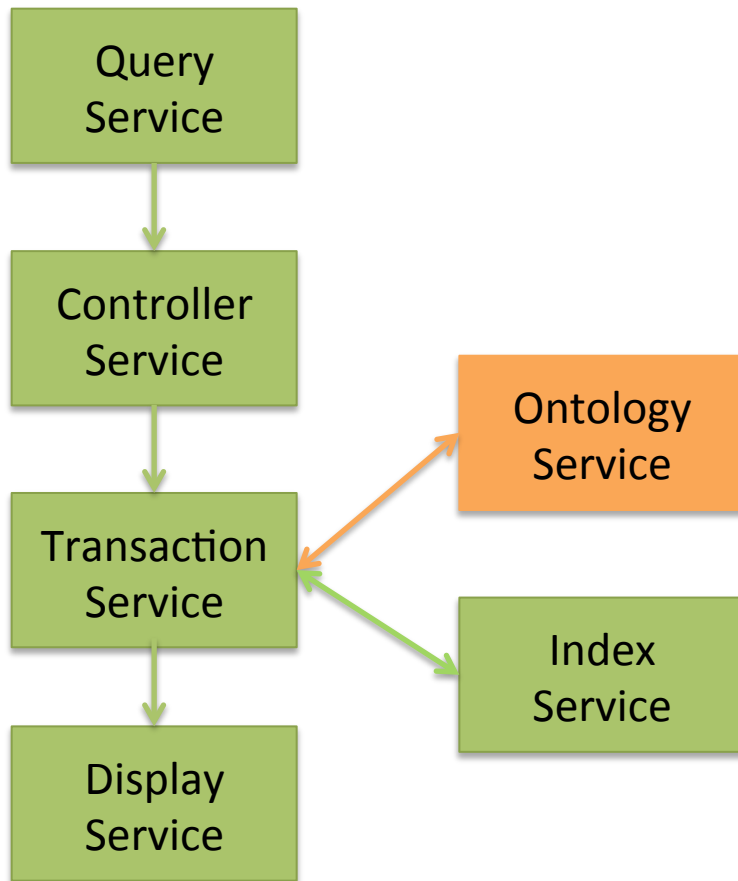
FILTER BY CATEGORY	<input type="text" value="All Categories"/>
FILTER BY GROUP ?	<input type="text" value="All Groups"/> ?
FILTER BY TEXT	<input type="text"/>

[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

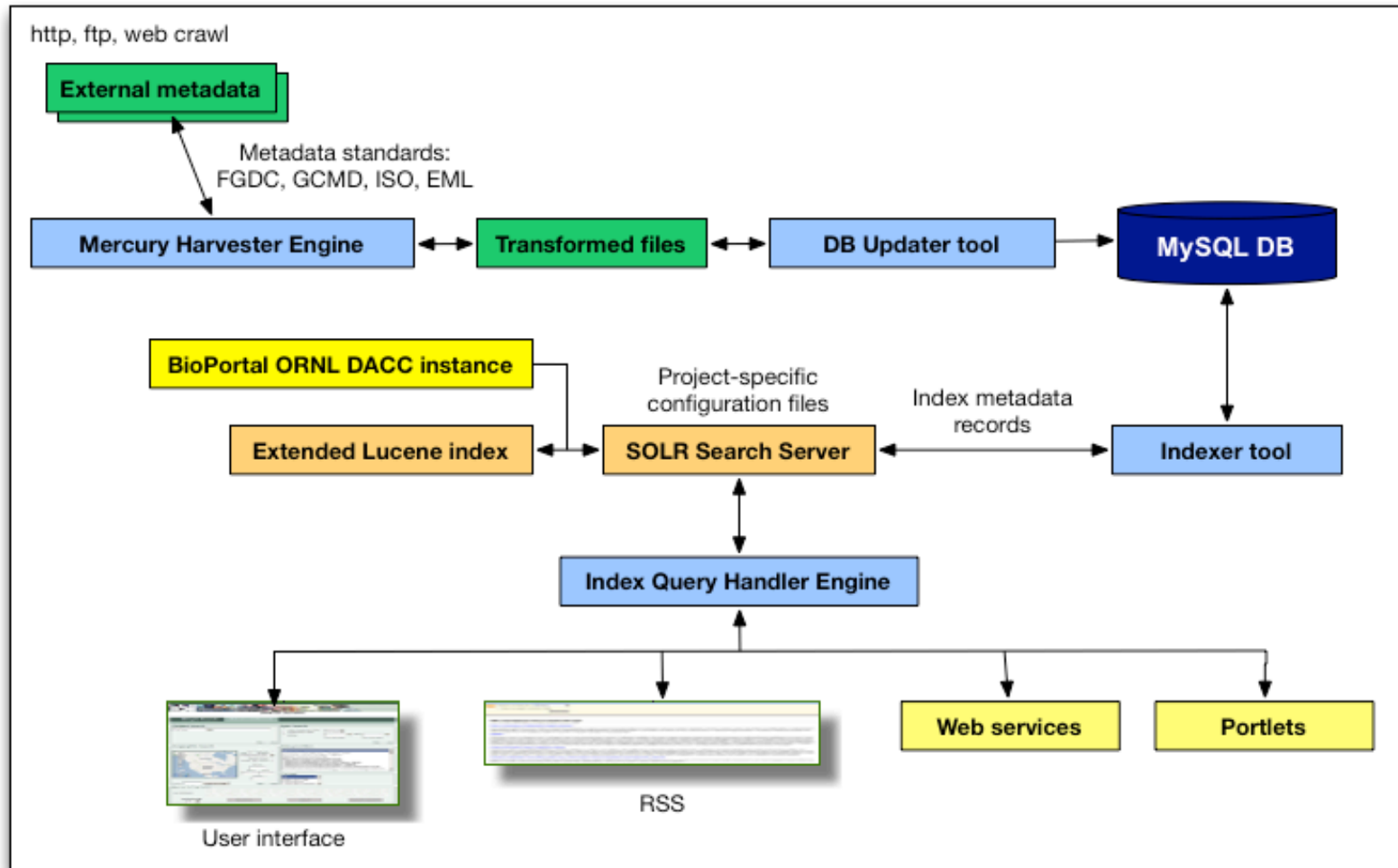
Showing 1 to 4 of 4 entries

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

Internal architecture



Ontology-based search results

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

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 "wood"

Choose records from: [LTER DATA \(950\)](#) [DAAC DATASETS \(187\)](#) [REGIONAL AND GLOBAL DATA \(62\)](#) [LAND VALIDATION DATA \(12\)](#) [LP DAAC - MO...](#)
[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)	...
weighing balance (94)	weighing balance (94)	land surface (65)	safari 2000 (28)	...
quadrat sampling frame (77)	quadrat sampling frame (77)	atmosphere (57)	lba (25)	...
steel measuring tape (62)	steel measuring tape (62)	human dimensions (26)	eos land validation.. (13)	...
soil coring device (59)	soil coring device (59)	agriculture (10)	fife (11)	...
human observer (48)	human observer (48)	hydrosphere (7)	superior national forest...	...
solid earth (6)		solid earth (6)		

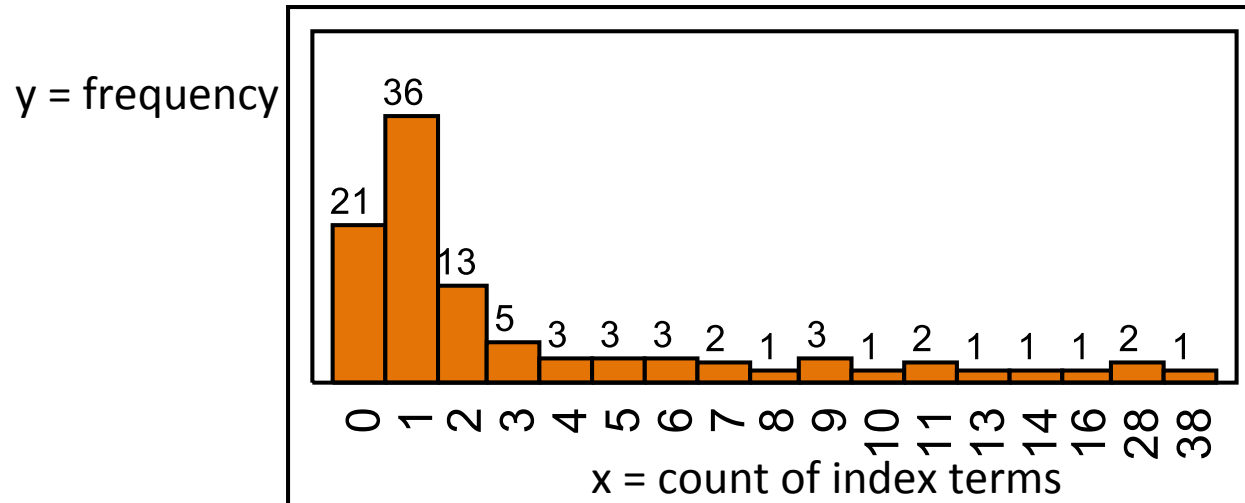
Ontology Concepts	Ontology SuperClasses	Ontology SubClasses	Filter by keywords (SubClasses)
matrBiomass.owl#Biomass matrEnergy.owl#Biomass	energy storage living entity	animal waste detritus fertilizer food gelbstoff humus litter manure	forests (20) soils (36) europe (33) biomass (19) eosdis (73) fao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

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>