Improving Metadata Searches by Enabling Semantic Queries

Chad Berkley[1], Shawn Bowers[2], Matthew Jones[1], Mark Schildhauer[1]
[1] National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara
[2] University of California, Davis

Rationale

- Increasing numbers of digitized ecological datasets make it critically important to improve techniques for more precisely locating and delivering relevant information from scientific searches.
- Ontology-enabled search can help to increase the number of relevant search results while decreasing the number of irrelevant results, allowing users to browse a smaller and more focused resultset.

Future Goals

- Enhance structured search to be able to search for multiple observations
- Develop more comprehensive ecological ontologies
- Create a larger corpus of annotated EML documents
- Further extend the ontology and annotation support in Metacat
- Do precision/recall studies to determine efficacy of different searches
- Extend current search approaches and add new ones
- Implement the ability to browse datasets via observation/measurement annotations and ontologies
- Improve the feedback describing why a result was a match

Types of Searches Implemented

Basic Keyword Search - Search all metadata text fields for user-entered terms.

Ontological Term Expansion - User-entered keywords are expanded using subsumption hierarchies defined in an ontology, then each expanded term is run against the metadata document base. Terms can be combined with various boolean functions.

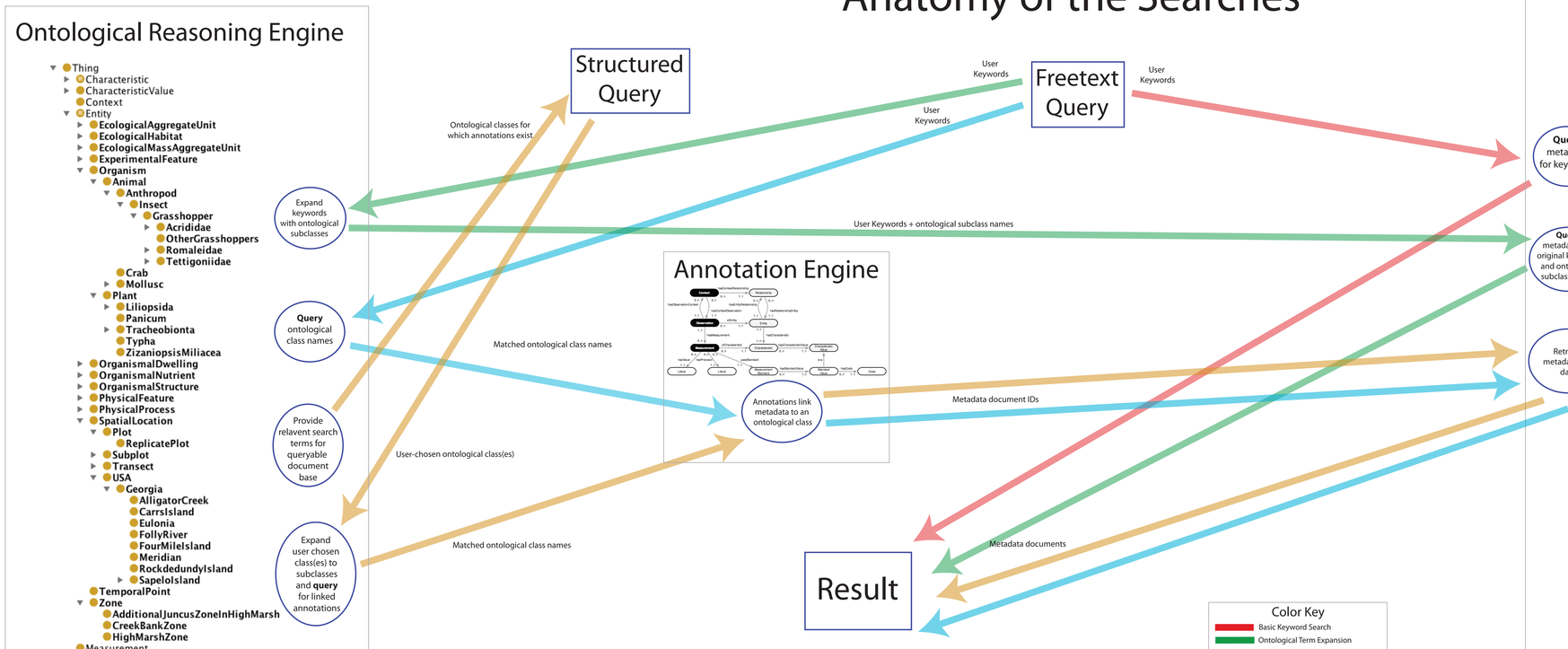
Annotation Based Searching - Metadata fields are annotated against the structure of an ontology. The annotation is searched for user-entered keywords instead of the metadata document itself. When a result is found, any metadata document linked via the annotation is returned.

Structured Annotation Search with Term Expansion - The user selects fields to search based on the types of observations and measurements within the datasets (given by the annotations). In the search, the user specifies what was observed and what characteristics and standards were measured of it.

Anatomy of the Searches

Ontological Reasoning Engine

- Thing
  - Characteristic
  - CharacteristicValue
  - Context
  - Entity
    - EcologicalAggregateUnit
    - EcologicalHabitat
    - EcologicalMassAggregateUnit
    - ExperimentalFeature
    - Organism
      - Animal
        - Anthropod
          - Insect
            - Grasshopper
              - Acrididae
              - OtherGrasshoppers
              - Romaleidae
              - Tettigoniidae
            - Crab
        - Mollusc
      - Plant
        - Liliopsida
        - Panicum
        - Tracheobionta
        - Typha
        - ZizaniopsisMiliacea
    - OrganismalDwelling
    - OrganismalNutrient
    - OrganismalStructure
    - PhysicalFeature
    - PhysicalProcess
    - SpatialLocation
      - Plot
        - ReplicatePlot
      - Subplot
      - Transect
      - USA
        - Georgia
          - AlligatorCreek
          - CarrsIsland
          - Eulonia
          - FollyRiver
          - FourMileIsland
          - Meridian
          - RockdedundyIsland
          - SapeloIsland
    - TemporalPoint
    - Zone
      - AdditionalJuncusZoneInHighMarsh
      - CreekBankZone
      - HighMarshZone
  - Measurement
  - MeasurementStandard
  - Observation
  - Relationship
  - UnitConversion



Metacat Metadata Repository

Biocomplexity Data Search

Home

Data Set Citation

Pennings SC. 2003. **Fall 2000 plant monitoring survey -- biomass calculated from shoot height and flowering status of plants in permanent plots at GCE sampling sites 1-10**.
KNB Data Registry: urn:lsid:knb.ecoinformatics.org:knb-lter-gce:92:26 (http://knb.ecoinformatics.org/knb/metacat/knb-lter-gce.92.26/knb).

Data Tables, Images, and Other Entities:

Metadata download:	Ecological Metadata Language (EML) File
Data Table:	PLT-GCEM-0303a_3_0.TXT (View Metadata | Download File)
Alternate Identifier:	PLT-GCEM-0303a.3.0

Online Distribution Info:

Download File:	http://gce-lter.marsci.uga.edu/lter/asp/db/dataset_details.asp?accession=PLT-GCEM-0303a

Data Set Owner(s):

Organization:	Georgia Coastal Ecosystems LTER Project
Address:	Dept. of Marine Sciences, University of Georgia, Athens, Georgia 30602-3636 USA
Email Address:	gcelter@uga.edu
Web Address:	http://gce-lter.marsci.uga.edu/lter/
Individual:	Dr. Steven Pennings
Organization:	University of Houston
Address:	Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204-5513 USA
Email Address:	spennings@uh.edu
Web Address:	http://www.bchs.uh.edu/~steve/

Metadata Provider(s):

Organization:	Georgia Coastal Ecosystems LTER Project
Address:	Dept. of Marine Sciences, University of Georgia, Athens, Georgia 30602-3636 USA
Email Address:	gcelter@uga.edu
Web Address:	http://gce-lter.marsci.uga.edu/lter/

Associated Party:

Individual:	Wade Sheldon
Organization:	University of Georgia

Data linked via EML metadata

Data

This work was supported by the National Science Foundation under awards 0533368 and 0225676 (SEEK). Support was also given by the National Center for Ecological Analysis and Synthesis (a Center funded by NSF grant #DEB-0553768, the University of California, Santa Barbara, and the State of California), and the University of California, Davis.