Annual Report for Period: 08/2008 - 07/2009
Submitted on: 08/07/2009
Principal Investigator: Jones, Matthew B.
Award ID: 0743429
Organization: U of Cal Santa Barbara
Submitted By:
Jones, Matthew - Principal Investigator
Title:
Semantic Enhancements for Ecological Data Management

Project Participants

Senior Personnel

Name: Jones, Matthew
Worked for more than 160 Hours: Yes
Contribution to Project:
Jones is the project lead and coordinator for the project. He contributes to the design and conceptualization of software extensions to support semantic data management, and supervises development staff on the project.

Name: Schildhauer, Mark
Worked for more than 160 Hours: Yes
Contribution to Project:
Schildhauer is a co-PI on the project focusing on Knowledge Representation approaches, particularly with respect to observational data models. He is a leader in the development of OBOE, and is the principal liaison with the Scientfic Observations Network (SONet).

Name: Bowers, Shawn
Worked for more than 160 Hours: Yes
Contribution to Project:
Bowers is a co-PI on the project, helping to specify project goals especially with respect to semantic data representation and query capabilities.

Name: Madin, Joshua
Worked for more than 160 Hours: Yes
Contribution to Project:
Joshua Madin is a co-PI on the project and has helped to specify the overall semantic data management product suite. He also contributes to ontology development in the biological science area.

Name: O'Brien, Margaret
Worked for more than 160 Hours: Yes
Contribution to Project:
Margaret O'Brien is a co-PI on the project and is conceptualizing and leading the two scientific use case projects that will determine the capabilities needed by the proposed semantic data management tools.

Post-doc

Graduate Student

Undergraduate Student

Technician, Programmer

Name: Berkely, Chad
Worked for more than 160 Hours: Yes
Contribution to Project:

Berkley served as a software engineer on the project to develop the initial system prototype for semantic extensions to the Metacat system. Berkley was funded from the SEEK project but was working jointly on this project that extends SEEK concepts of semantic data query.

Other Participant

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

SONet: We have been collaborating with the Scientific Observations Network (SONet), which is an NSF-funded INTEROP project focused on advancing a core semantic model of scientific observations. The initial Semtools proposal outlined work on just such a semantic model of observations, but was dropped during budget cuts as we felt the work would be addressed by SONet. Nevertheless, the tools that we propose in Semtools depend on the core model from SONet being completed, and so Schildhauer, Bowers, Madin, and Jones are all collaborating to create a core model for SONet. The first step of this process is to produce effective comparisons of the existing semantic observations models, and so we have organized a workshop for Fall 2009 that will include an Interoperability Challenge that can be used to compare the various models in use in the science communities. We expect this workshop to produce a straw-man core scientific observations model that can be used as the basis of observational data modeling in the Semtools project.

Activities and Findings

Research and Education Activities: (See PDF version submitted by PI at the end of the report)

Activities 2008-2009

Data for ecological and environmental studies quantify, among other things, the distribution and abundance of organisms; the processes that influence biological populations, communities, and ecosystems; and the environmental and anthropogenic drivers of these processes. Scientists increasingly rely on accessing and analyzing these diverse data collected by cross-disciplinary communities of researchers to achieve synthetic, crosscutting insights into the environment that can address issues of fundamental importance to science and society.

Despite these needs, discovering these data is difficult. The precision and recall of data searches in data repositories is not satisfactory even at current collection sizes. Data archives like the Knowledge Network for Biocomplexity (KNB), the National Biological Information Infrastructure (NBII) Metadata Clearinghouse, and the Global Change Master Directory (GCMD) rely on semi-structured metadata with fields containing largely natural-language descriptions to provide search and browsing capabilities and to allow human use and interpretation of the data. These metadata enable simple keyword searches that return results generally related to the topics of interest, but they cannot be used to perform precise searches of the data archives. Ironically data sets with more extensive (natural language) metadata are included in search results simply due to the incidental mention of a term in an ancillary part of the metadata document. These extraneous results decrease the precision of the search, seriously reducing the efficiency in researchers? finding the data they need. In addition, because natural-language metadata does not generally rely on controlled vocabularies, researchers typically classify their data sets using ad-hoc descriptive terms, reducing recall. Given the number of synonyms and overlapping terms used in scientific disciplines, searches frequently miss relevant data because the search terms do not exactly match the terms used to classify the documents.

Activities during the first year of the project have been focused on the refinement of a core model for scientific observations (in collaboration with SONet), and on the development of a prototype semantic search system for Metacat. This prototype represents a proof-of-concept for semantic search approaches and will allow us to compare multiple search strategies. As shown in Figure 1, we have added support to Metacat for storing and managing OWL-DL ontologies and semantic annotations, and for reasoning and search services to support different semantic-search strategies.

To implement these extensions to Metacat, we used approaches that exploit the use of formal reasoning over an ontology designed to facilitate the semantic description of scientific observations (OBOE). In our current implementation, the Jena API is used to access ontologies and ontology terms within Metacat, and Pellet is used to provide reasoning services over these ontologies (e.g., to compute class subsumption

hierarchies and to ensure ontologies added to Metacat are consistent). We also extend Metcat?s XML management capabilities with support for managing semantic annotations. Ontologies and annotations added to Metacat are assigned unique identifiers (URIs), allowing both to be easily accessed through external applications (e.g., Prot?g?). Further, ontologies and annotations can be versioned using this URL scheme.

The Extensible Observation Ontology (OBOE) provides a high-level abstraction of scientific observations and measurements that facilitate the creation of domain-specific vocabularies for defining observation and measurement semantics. OBOE is represented using OWL-DL and enables data (or metadata) structures to be linked to domain-specific ontology concepts so that critical aspects of scientific observations can be documented?such as what kind of Entity was measured, which Characteristics of that entity were measured and by which Measurement Standards (e.g., kilograms/m^2), and what other observations provide Context for interpreting those measurements. In our approach, semantic annotations are then used to map these critical parts of a scientific observation to the data instances that are stored in a data repository (see Figure 2).

In addition to plain-text keyword search, we implemented three different search methodologies to investigate the utility of semantic methods for scientific data discovery: (i) simple term expansion against ontologies to broaden the search terms against the metadata corpus; (ii) term expansion against semantic annotations; and (iii) structured searches that pose queries against the components of an observation described via OBOE.

Findings:

Findings 2008-2009

Our semantic search system adds to Metacat the ability to store OWL-DL ontologies in addition to semantic annotations that link data set attributes to ontology terms. Our approach also extends Metacat to improve metadata search in multiple ways: (i) by expanding standard keyword searches with ontology term hierarchies; (ii) by allowing keyword searches to be applied to annotations in addition to traditional metadata; and (iii) by allowing more structured searches over annotations via ontology terms. We describe our implementation of these extensions, and compare and contrast these different types of search for a corpus of annotated documents. As data repositories continue to grow, these tools will be instrumental in helping scientists precisely locate and then interpret data for their research needs.

Figure 1 shows the primary components of our semantic-discovery framework. The bottom of Figure 1 consists of two simple, example data sets. Although different types of data are often used within ecological analysis (e.g., raster, GIS, etc.), data sets are predominantly tabular (relational) and denote sets of related observations and measurements that were either directly collected or were the result of aggregation or analysis. Although not obvious, the example data sets in Figure 2 contain largely similar information consisting of spatial locations divided into sub-locations (i.e., a plot or quadrat), fertilization treatment information, and weight measurements.

Metadata schemes such as the Ecological Metadata Language (EML) provide standard ways of describing implicit aspects of data sets. In Figure 2, we show a fragment of EML for describing the ?wt? and ?LL? attributes of the data sets. EML can be used to represent the basic structural aspects of data?the number of attributes, their names, and their allowable values? but the semantics of the data set?the types of entities observed, the characteristics of these entities that were measured, and how these entities were observed in relation to each other?is either indirectly described (e.g., within the methods section of the metadata document) or are altogether missing. Metadata alone would not reveal the closely related semantics of the highlighted attributes from our sample data sets.

Semantic annotations extend EML by providing a mechanism to describe data set attributes in terms of OBOE concepts. An annotation is a formal structure, which represents a mapping from data set values to ontology instances (i.e., individuals), and an XML-based syntax is used to represent annotation mappings. As shown in Figure 2, we can see that the two annotated attributes: (i) represent observations of leaf-litter entities; (ii) measure the weight of leaf-litter (although using different weight characteristics); and (iii) use compatible but different measurement units (kilograms and grams). Annotations can be used to find all data sets related to a particular concept, determine all of the concepts related to particular data set attributes, and compare data sets based on their corresponding OBOE structures (which can facilitate data integration). XML is used as an interchange format for representing annotations; in general, annotation providers will annotate data sets using higher-level metadata editors and interfaces provided through tools such as Morpho.

A more detailed example showing the various XML syntaxes used for representing EML attributes (bottom), semantic annotations (middle), and an OBOE ontology extension (top) are shown in Figure 3.

To improve overall precision and recall of Metacat searches we prototyped three new search strategies.

Keyword-Based Term Expansion. In this approach, we ?intercept? keyword queries issued to Metacat and expand them according to the term hierarchies of the stored ontologies. Specifically, if a given search keyword matches a class name (i.e., as specified by the rdf:label property of

the class), then the search is expanded to include the synonyms of the class as well as the names of subclasses. This form of search alleviates the problem with simple keyword searches of not returning data sets described with synonyms or more specialized terms of the user-entered keywords. In our implementation, when a user enters a keyword search, Metacat locates synonyms and corresponding subclasses for each keyword using an ontology manager. The query is then augmented by Metacat with the expanded terms according to the given search constraints (i.e., whether terms should exactly match document terms and whether all given keywords must be present in a document) and executed against the current Metacat keyword search service. Although this strategy improves recall for documents that may have been omitted with simple keyword searching, it can also cause additional false positives due to the addition of keywords. Thus, this approach generally increases recall, but not necessarily precision. Specifically, the set of metadata documents returned is always a superset of what is returned using the traditional Metacat keyword search.

Annotation-Enhanced Term Expansion. Semantic annotations allow individual data set attributes to be linked to one or more ontology classes. By applying keyword searches only to annotations, search results can potentially improve precision by returning fewer false positives. In annotation-enhanced search, each search term is first expanded using the ontology similar to the keyword-based term expansion. Here, when a search term matches an ontology class, we use the class and all subclasses to find matching annotations. An annotation is considered a match if it contains the corresponding classes according to the search constraints (see above). The metadata document linked to the matched annotation is returned by the search. For example, a metadata description of a data attribute described textually as ?counts of grasshoppers? would be annotated as a ?count per square meter? of ?Romalea guttata? by linking the attribute to the ontology classes that define these concepts. When the user searches for ?grasshopper?, the term is expanded to ?Romalea guttata? via the ontology?s class hierarchy, and the annotation linking the metadata attribute to the class ?Romalea guttata? becomes a match. Since the annotation is linked to a specific attribute reference within the metadata, data sets containing comments about ?grasshoppers? in other fields would not be matched. Moreover, recall is improved due to matches facilitated by descending the ontology?s class hierarchy.

Observation-Based Structured Query. Though the annotation-enhanced term expansion approach limits search to the relevant portions of metadata that describe data content (via attribute annotations), it does not take advantage of observation and measurement structures and relationships. In the observation-based query approach, users can search for data sets via their observed entities (organism, site, etc.) and the characteristics and standards used to measure them. In an observation-based search, queries are specified by explicitly filling in an observation ?template? where ontology classes are given for the observed entity, measurement characteristic, and measurement standard. To search for data sets containing tree lengths, we would fill in an observation query template, using the tree class as the observed entity type and the length class as the measurement characteristic type. Search is performed by finding matches between the observation types of annotations and the query template, where a match includes searching subclasses of the template classes. This type of search has both good recall ? hitting all relevant data through appropriate use of term expansion ? and good precision by exploiting the structure of OBOE annotations to find exactly the entity, characteristic, and context of interest to the user.

The search interfaces in Metacat were implemented rapidly to explore the implications of different search strategies. Our next stage of development is to design an effective user interface for composing semantic queries and then to use the semantic search engine to execute those queries.

Training and Development:

Outreach Activities:

Outreach activities for the project have principally been through talks at scientific conferences and workshops where we have discussed our approaches to semantically modeling scientific observations and the benefits of doing so. Presentations on Semtools and SONet related work included:

Jones, M. 2008. Directions in observational data organization: from schemas to ontologies. Biodiversity Information Standards (TDWG) Annual Conference. Freemantle, Australia. 19-25 October, 2008.

Schildhauer, M. 2008. Facilitating data interoperability within the environmental and ecological sciences through advanced semantic approaches. Biodiversity Information Standards (TDWG) Annual Conference. Freemantle, Australia. 19-25 October, 2008.

Schildhauer: Improving Data Discovery in Metadata Repositories through Semantic Search. CISIS/iSEEK. Fukuoka, Japan. March 18, 2009

Jones: Semantic Data Integration for Heterogeneous Scientific Data. Lifewatch WP5 Workshop on Semantic Data Integration. Amsterdam, Netherlands. May 18, 2009.

Journal Publications

Berkley C, Bowers S, Jones MB, Madin JS, Schildhauer M, "Improving Data Discovery in Metadata Repositories through Semantic Search", Proceedings of iSEEK'09. IEEE Computer Society., p. , vol. , (2009). Accepted,

Books or Other One-time Publications

Web/Internet Site

URL(s):

http://semtools.ecoinformatics.org

Description:

This is the main web site for the Semtools project. It is a collaborative wiki that was established to aid communication among project participants and help with organization and outreach for the project.

Other Specific Products

Contributions

Contributions within Discipline:

Through our work on Semtools, we have demonstrated improvements in the effectiveness of data discovery for large, heterogeneous data collections such as the Knowledge Network for Biocomplexity (KNB). These advances have been possible through the use of an semantic model of scientific observations (Extensible Observation Ontology) and an annotation language that is used to map relational data sources to the concepts in OBOE. The prototype system that we developed will form the basis for future work this year on a production system that will have broad applicability in the ecological and environmental sciences.

Contributions to Other Disciplines:

Contributions to Human Resource Development:

Contributions to Resources for Research and Education:

The project is helping to build the extensive Knowledge Network for Biocomplexity (KNB) repository, which provides thousands of data sets for use in research and educational contexts. Data from the KNB will become more accessible as the semantic search facilities that we have developed become incorporated into the production Metacat software used by the KNB. This will enable educators and researchers to more readily access KNB data and therefore facilitate science and education advances in many disciplines.

Contributions Beyond Science and Engineering:

Conference Proceedings

Special Requirements

Special reporting requirements:

During the first year of the award, we have spent only a small fraction of the awarded funds, and expect to have $465,557.27 remaining at the end of the first year. There are several reasons for this slow expenditure rate. First, due to budget cuts we removed the portion of the project dealing with the creation of a core model for scientific observations, instead opting to collaborate with the SONet project on that activity. This core model is required before tools can be developed under Semtools, so we wanted to delay work on the tools until the SONet activity was established. Second, we were able to leverage the end of the SEEK project to employ Chad Berkely as a software engineer to do the initial prototyping work on the semantic extensions to Metacat described in the Activities and Findings section. This allowed us to explore our initial prototypes without using Semtools funds during the first year, thereby increasing our overall ability to meet project goals. Third, the final budget from NSF did not allow us to fund a developer for three full years, and so we decided that it would benefit the project to delay the start

of the developer until our plans for tool extension were fully mature.

Our spending on this project will increase significantly in August, 2009, as we will be starting to employ Benjamin Leinfelder on the project to begin development in August. Leinfelder will continue full time on the project, and we will also begin to spend the PI summer salaries. We fully expect to spend the funds as originally allocated to complete project goals.

Change in Objectives or Scope: None

Animal, Human Subjects, Biohazards: None

Categories for which nothing is reported:

Organizational Partners

Activities and Findings: Any Training and Development

Any Book

Any Product

Contributions: To Any Other Disciplines

Contributions: To Any Human Resource Development

Contributions: To Any Beyond Science and Engineering

Any Conference

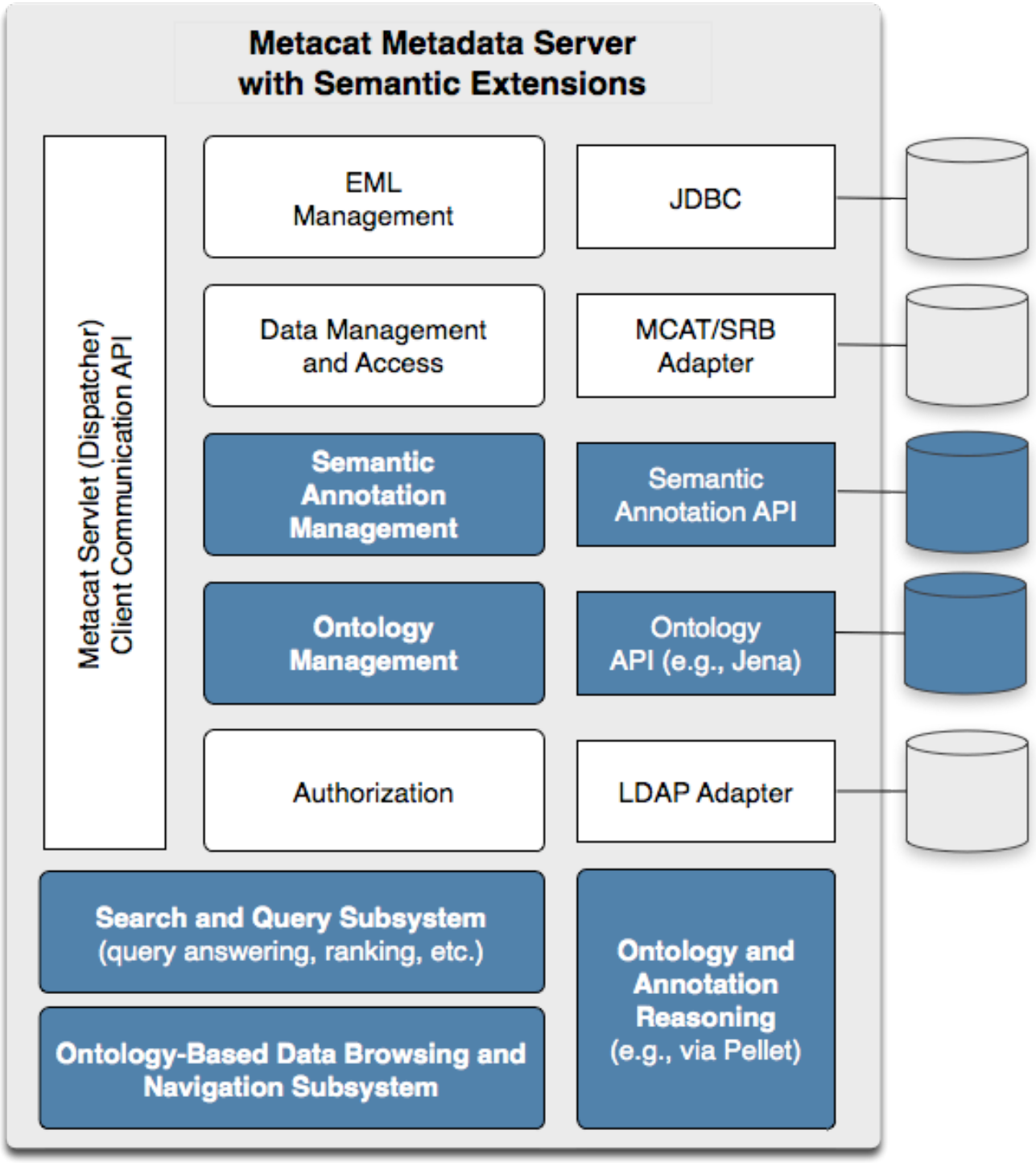

Figure 1: Semantic extensions (highlighted in blue) to the Metacat data and metadata repository support improved precision and recall in searches for scientific data sets.

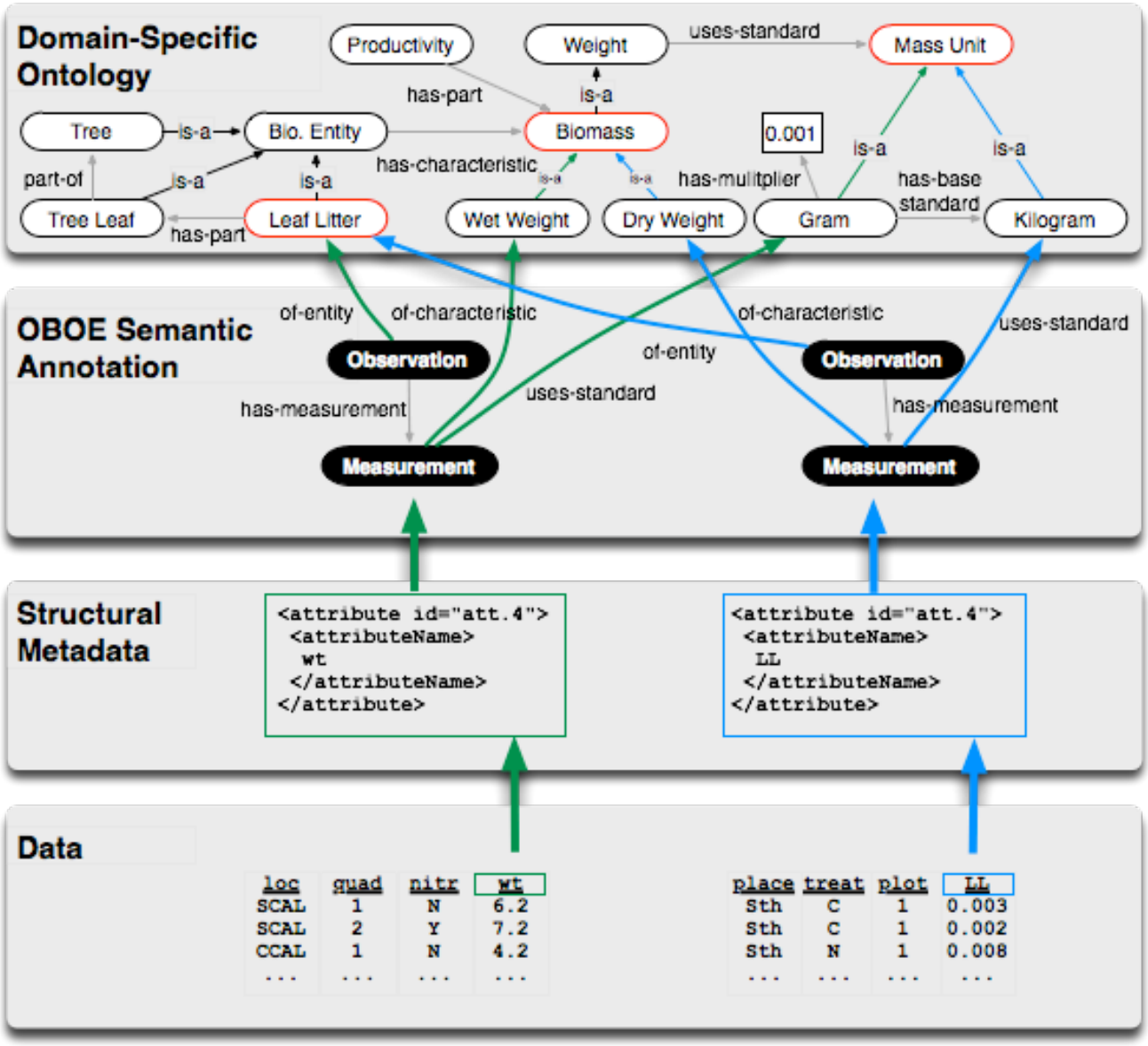

Figure 2. The components of our semantic-search framework including relational data, EML-based metadata, semantic annotations based on OBOE, and OBOE domain-ontology extensions.

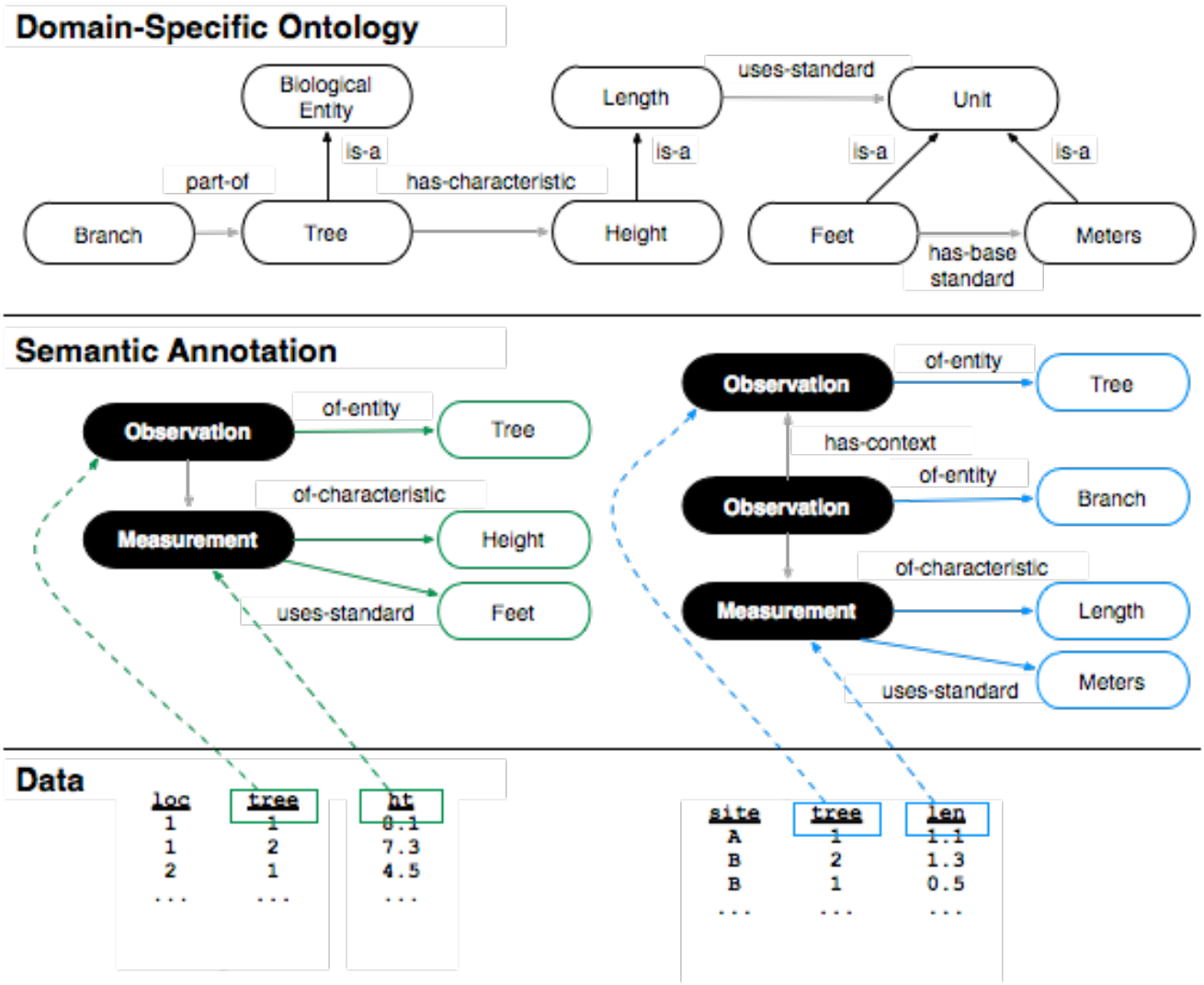

Figure 3. Example annotations demonstrating more precise search results for observation-based structured query.