Semantic Data Integration for Heterogeneous Scientific Data

Matthew B. Jones[1], Mark Schildhauer[1], Margaret O'Brien[1], Chad Berkley[1]

National Center for Ecological Analysis and Synthesis (NCEAS)

University of California, Santa Barbara[1]

Shawn Bowers[2]

University of California, Davis[2]

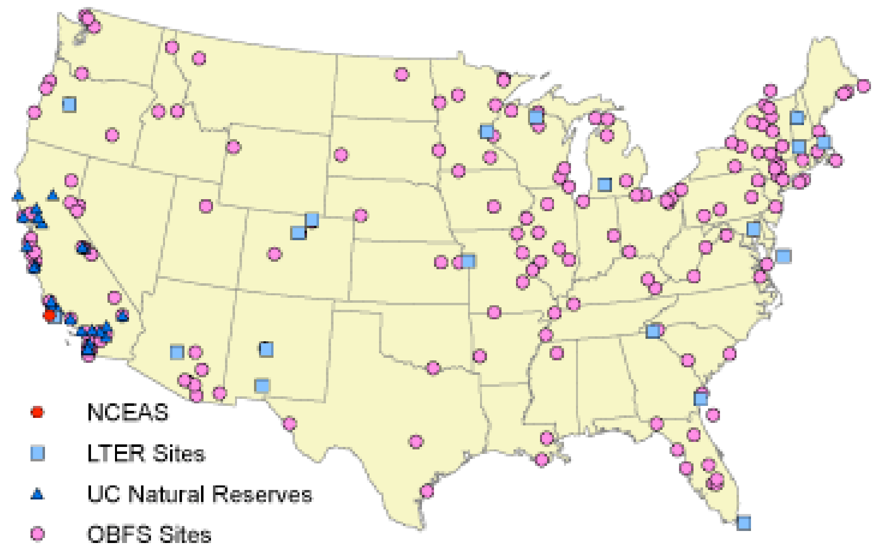
Josh Madin[3]

MacQuarie University[3]

NCEAS

Data Access Challenges

- Data are massively dispersed
  - Field stations (100’s)
  - Natural history museums (100’s)
  - Government agencies (10’s to 100’s)
  - **Individual scientists** (10,000’s)
- Data largely <u>inaccessible</u>
- Data sharing only via <u>personal networks</u> among scientists



- Data from many disciplines
  - Community ecology
  - Population ecology
  - Behavior, Genetics
  - Remote sensing
  - Environmental Science
  - Economics + Law
  - Human demographics

Descriptive Metadata

- Describe data set using natural-language text
  - information about the project, the location of data collection
  - information about data-collection methods and protocols

sth	c	1	0.003
sth	c	1	0.002
sth	n	1	0.008
...	...	...	...

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/People/Pennings/Pennings.html

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:	**Mr. Wade Sheldon**
Organization:	University of Georgia
Email Address:	sheldon@uga.edu
Role:	co-author

Abstract:

Parallel fertilization experiments were performed in five different types of perennial plant mixtures found in the salt marsh habitat around Sapelo
May 1996 to September 1997. Each mixture differed in plot elevation, soil water content, and soil salinity, so each was considered a separate ha
occurred in different geographic locations (i.e. Dean Creek on southern Sapelo Island, Marsh Landing on southwestern Sapelo Island, and Shell
the University of Georgia Marine Institute). In May 1996, 16 1mx1m plots were placed within each plant mixture and alternate plots were assig
fertilization treatments. Pelletized fertilizer (29% N, 3% P, 4% K) was broadcast into fertilization treatment plots by hand at the rate of 60g/m
central 0.5mx0.5m of each plot was harvested in September 1997 after two summers growth. Live plants were sorted to species, dried to a co
weighed to measure biomass. Standing dead shoots and litter were not weighed.

Keywords:

- Sapelo Island (place)
- Georgia (place)
- USA (place)

- GCE (theme)
- LTER (theme)
- Primary Production (theme)
- Batis maritima (theme)
- Borrichia frutescens (theme)

Structural Metadata

- Describes the structural aspects of a dataset
  - Number of columns
  - Name (informal "meaning") of columns
  - Allowable values (e.g., 'n' and 'c' are allowable for trmt)

place	trmt	plot	LL
sth	c	1	0.003
sth	c	1	0.002
sth	n	1	0.008
...	...	...	...

loc	quad	nitr	wt
scal	1	n	6.2
scal	2	y	7.2
ocal	1	n	4.2
...	...	...	...

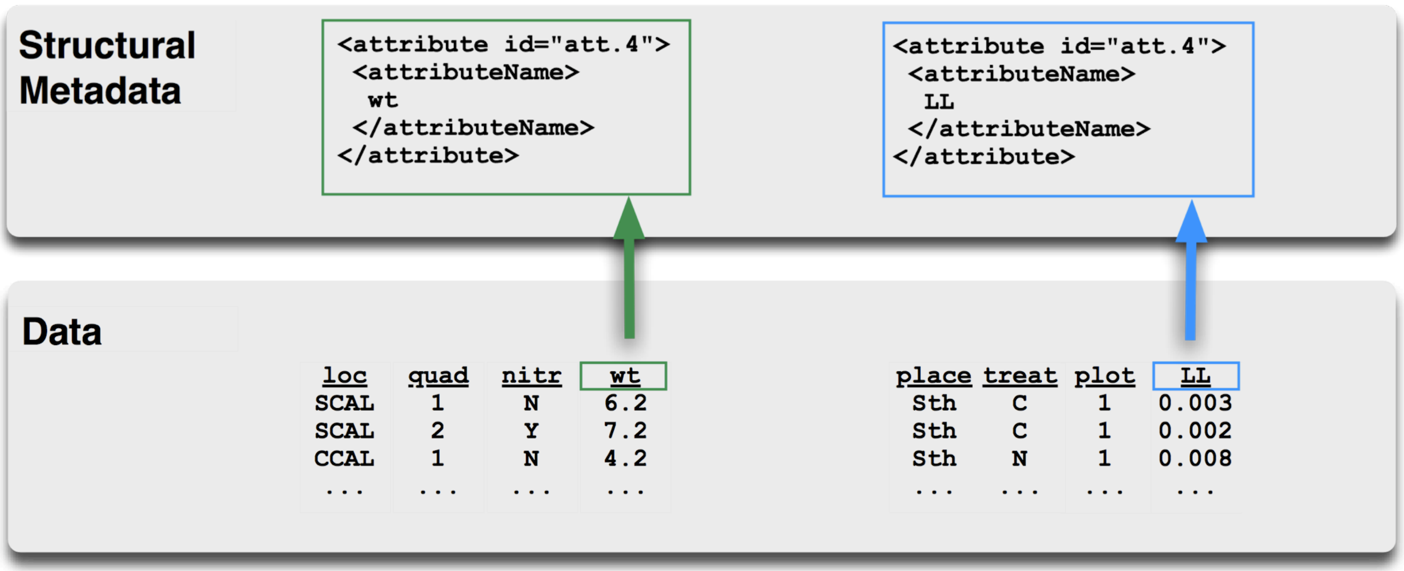
Hard to determine if columns are the same

Relationships between columns unclear

Data

loc	quad	nitr	wt
SCAL	1	N	6.2
SCAL	2	Y	7.2
CCAL	1	N	4.2
...	...	...	...

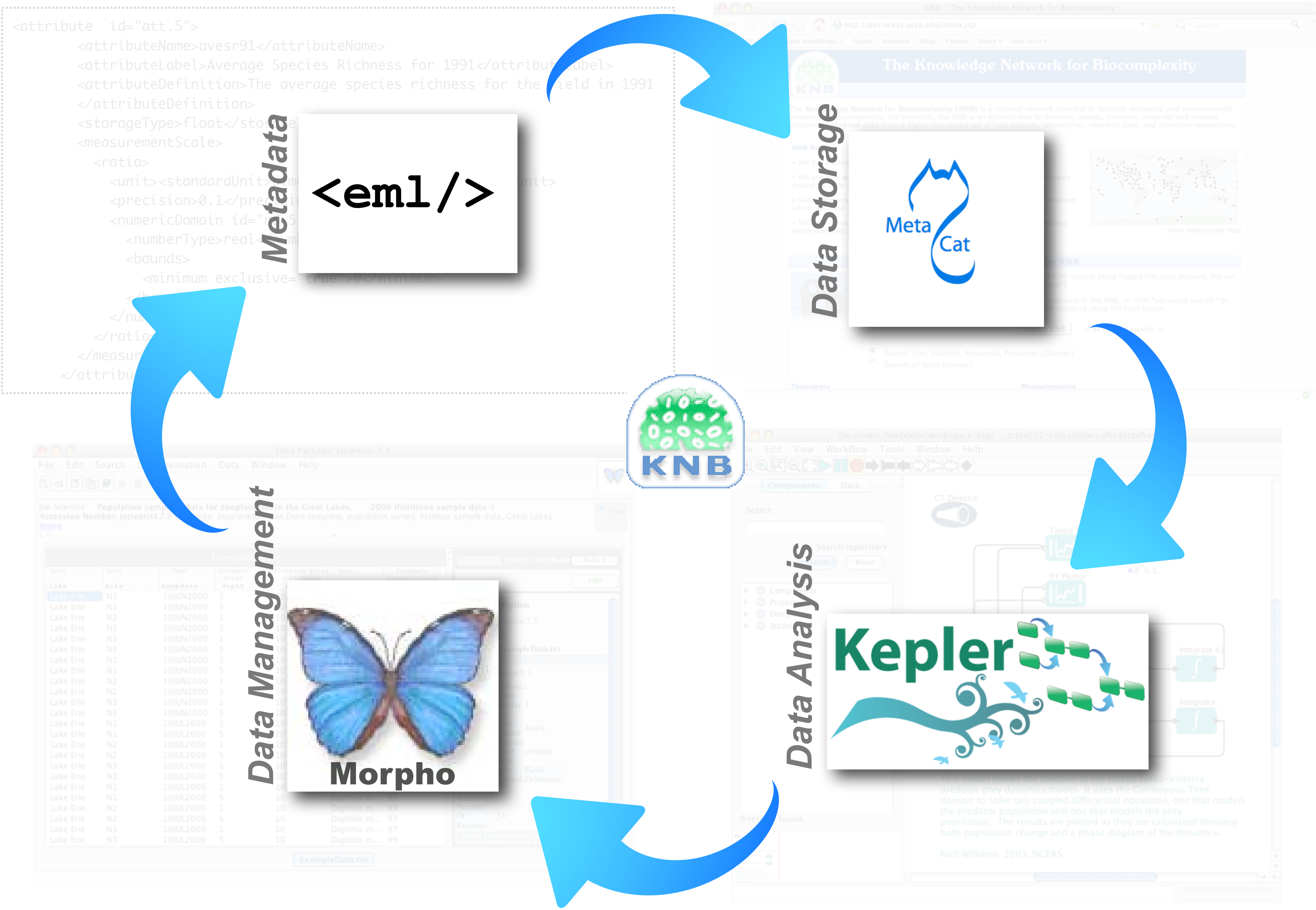
place	treat	plot	LL
Sth	C	1	0.003
Sth	C	1	0.002
Sth	N	1	0.008
...	...	...	...



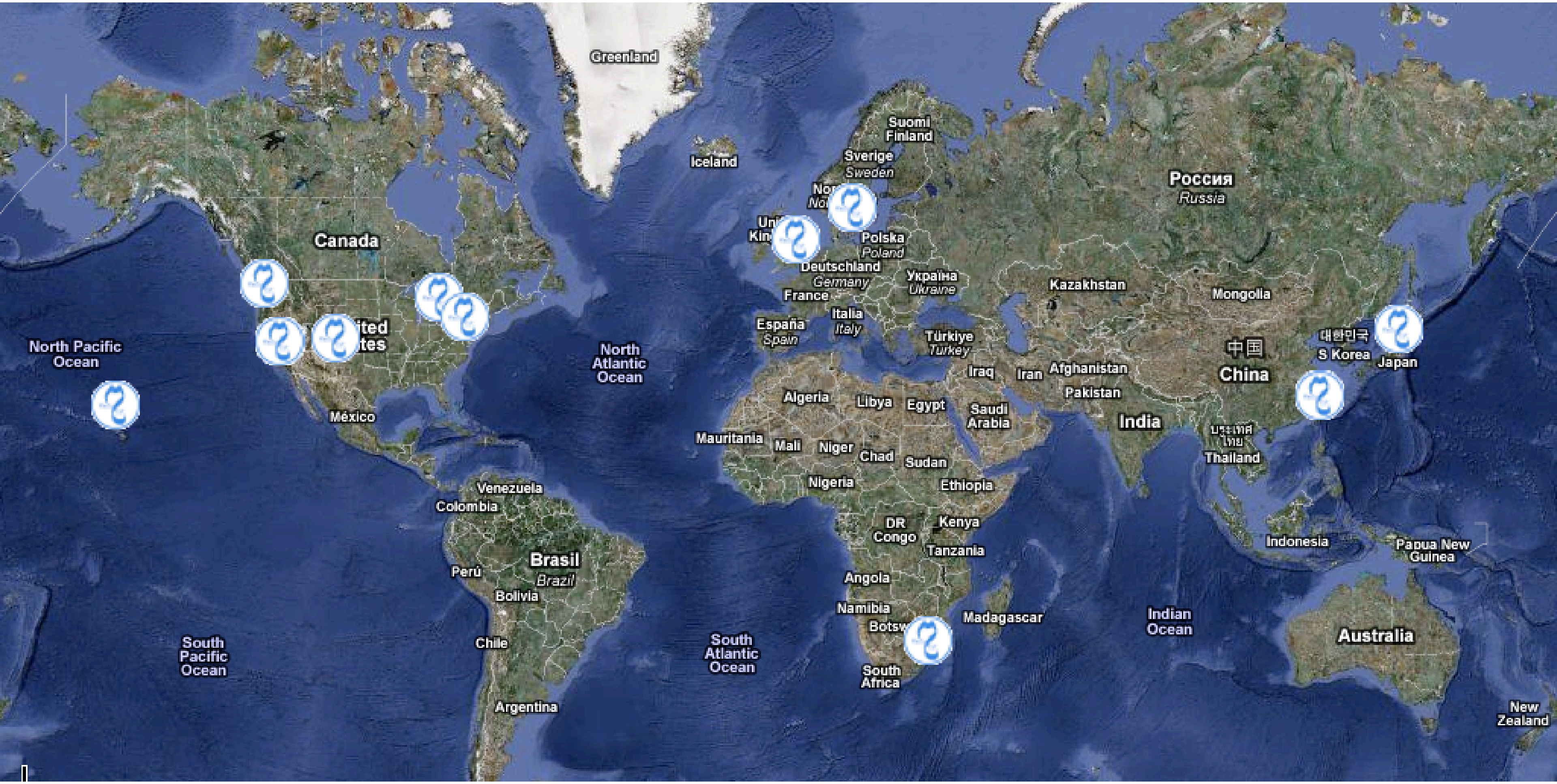
Structural Metadata
<attribute id="att.4">
<attributeName>
wt
</attributeName>
</attribute>
<attribute id="att.4">
<attributeName>
LL
</attributeName>
</attribute>
Data
loc quad nitr wt
SCAL 1 N 6.2
SCAL 2 Y 7.2
CCAL 1 N 4.2
... ...
place treat plot LL
Sth C 1 0.003
Sth C 1 0.002
Sth N 1 0.008
... ...

NCEAS

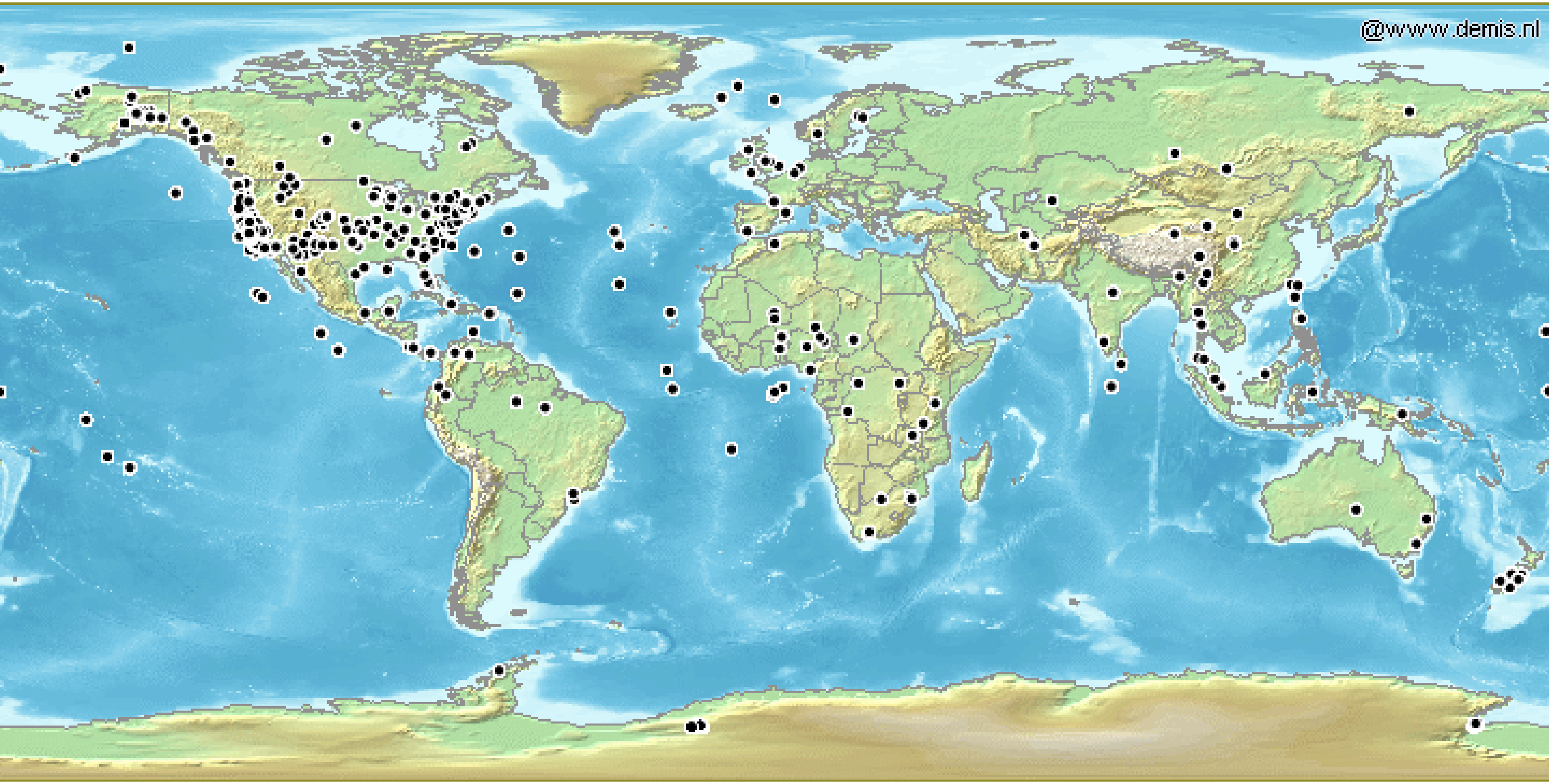
KNB Software Suite



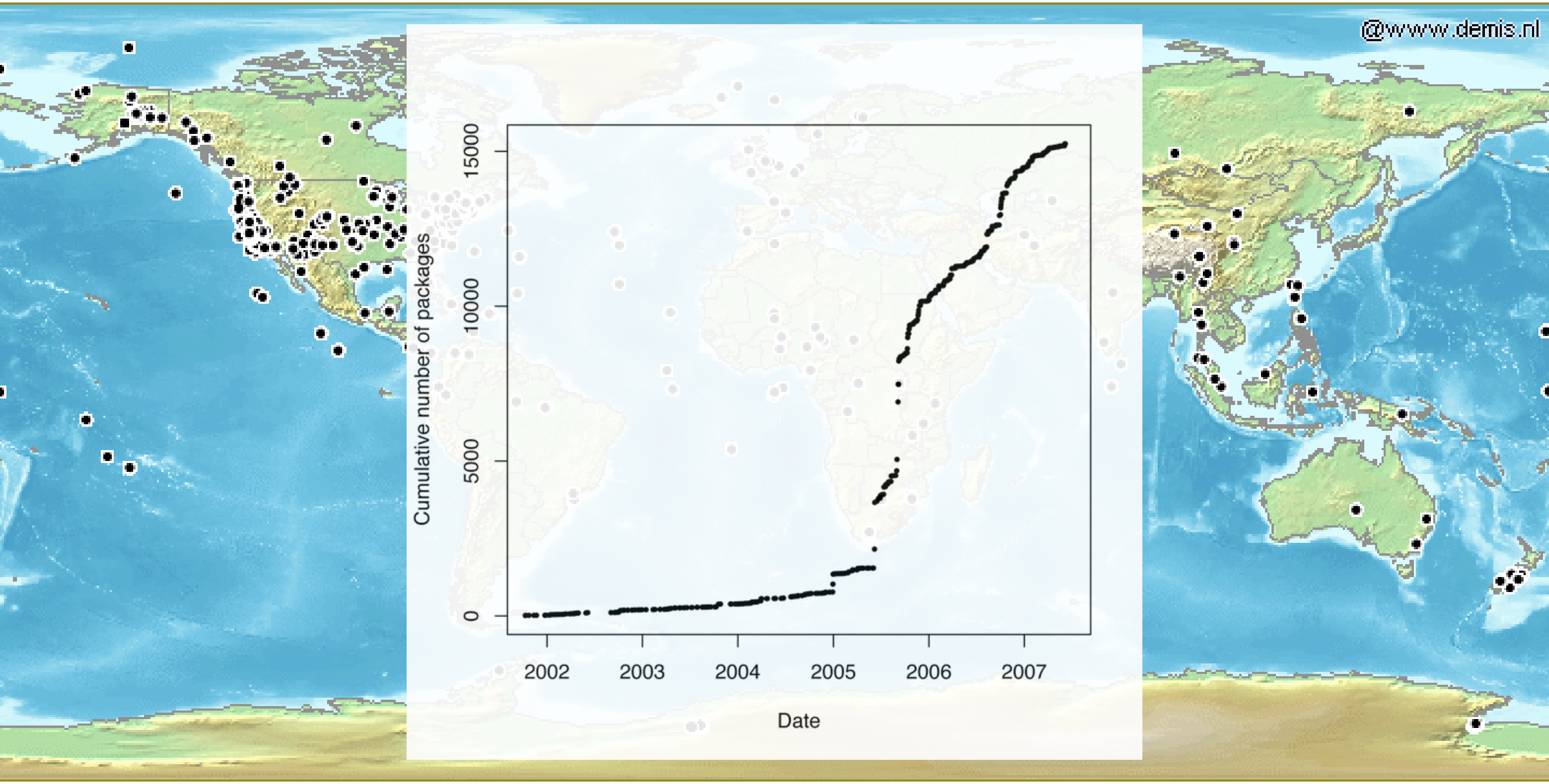
Global Metacat deployments



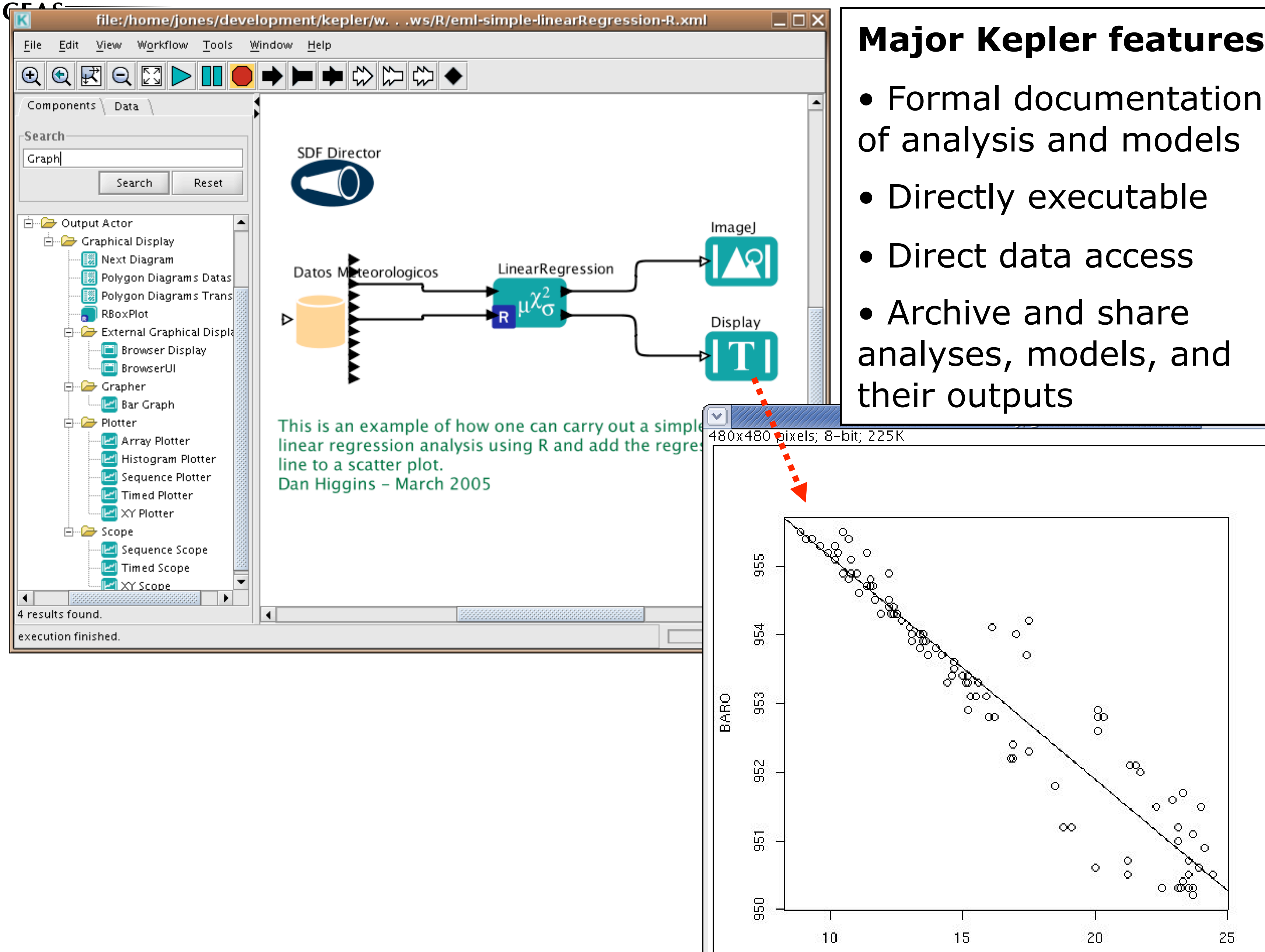
KNB Data Distribution



KNB Data Distribution



Kepler: scientific workflow system

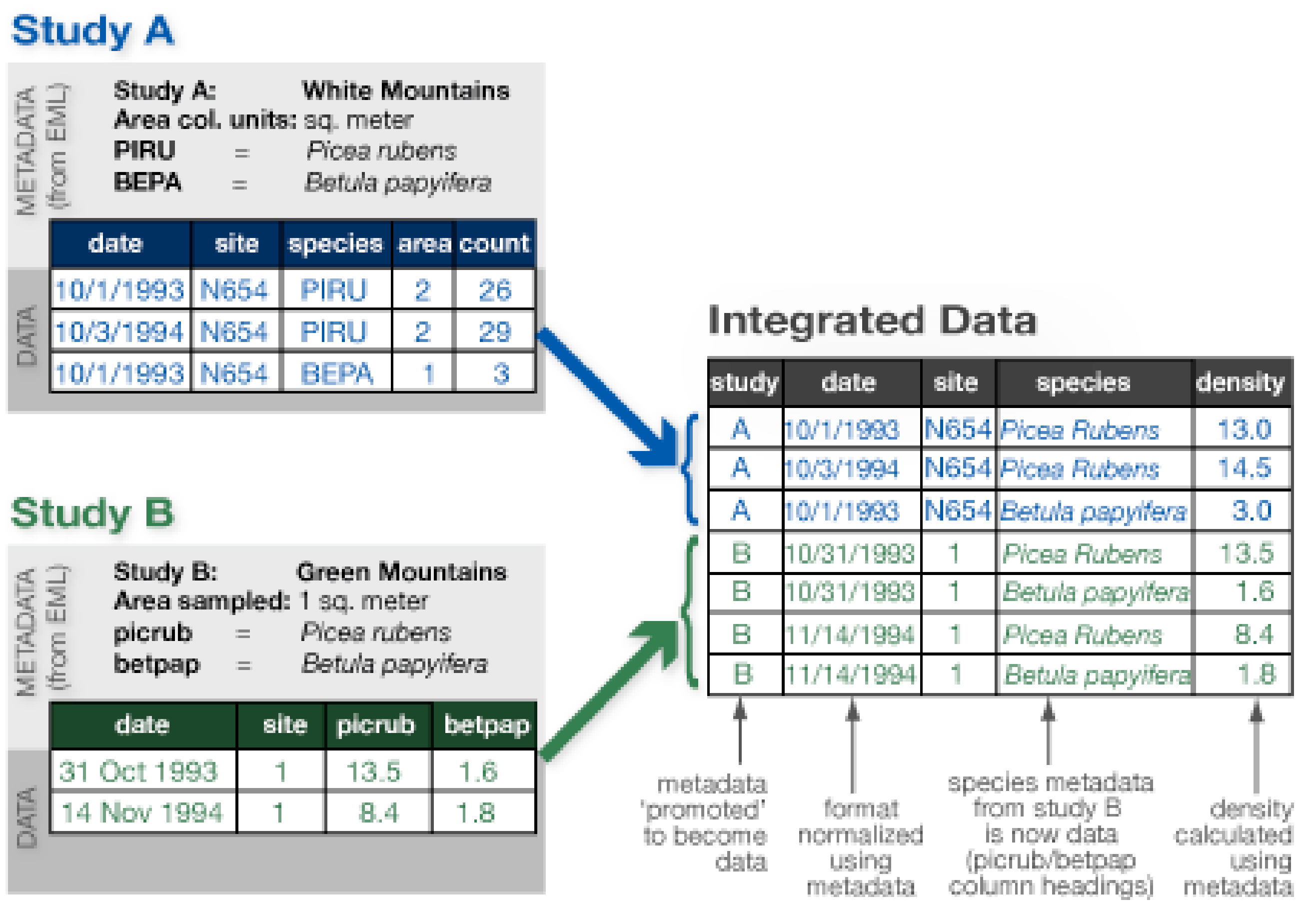


Major Kepler features:

- Formal documentation of analysis and models
- Directly executable
- Direct data access
- Archive and share analyses, models, and their outputs

Data Integration Challenges

- Data are heterogeneous
  - Differing formats, logical organization, and interpretation
  - Syntax
  - Schema
  - Semantics



Jones et al., 2006, AREES

Semantic annotation

- Tabular data lacks critical semantic information
  - no way for computer to determine that “Ht.” represents a “height” measurement
  - no way for computer to determine if Plot is nested within Site or vice-versa
  - no way for computer to determine if the Temp applies to Site or Plot or Species

Site	Temp	Plot	species	Ht.
1	21	A	AHYA	4.7
1	21	A	AGEM	3.4
1	21	B	AHYA	2.4
1	21	B	AGEM	6.2
2	15	A	AHYA	1.3
2	15	A	AGEM	4.5
2	15	A	APAL	2.0
2	15	B	AHYA	4.5
2	15	B	APAL	5.6
3	17	A	AGEM	9.2
...	...	...	...	...

Data set

slide from J. Madin

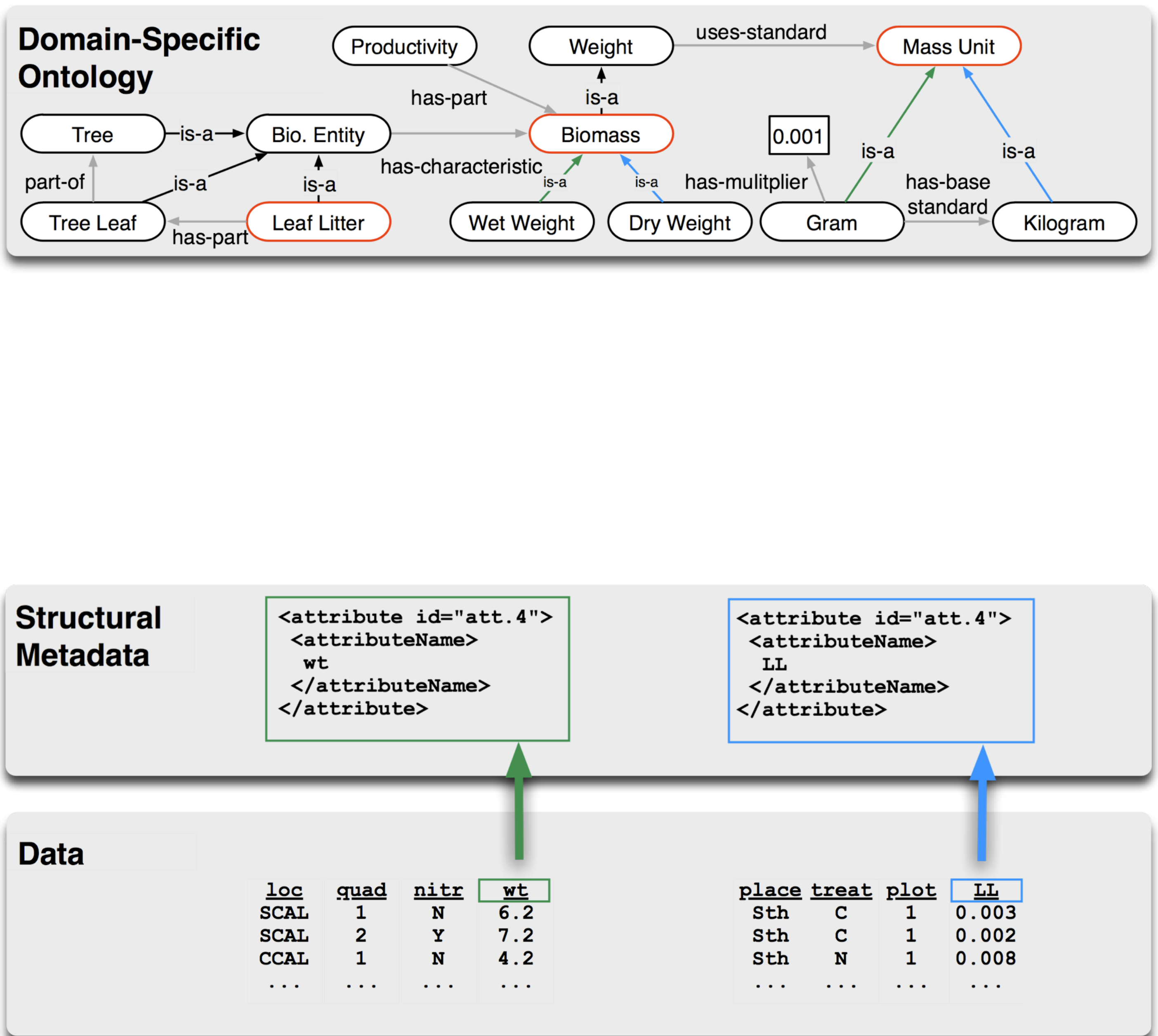
Semantic Data Integration

based on a common model of

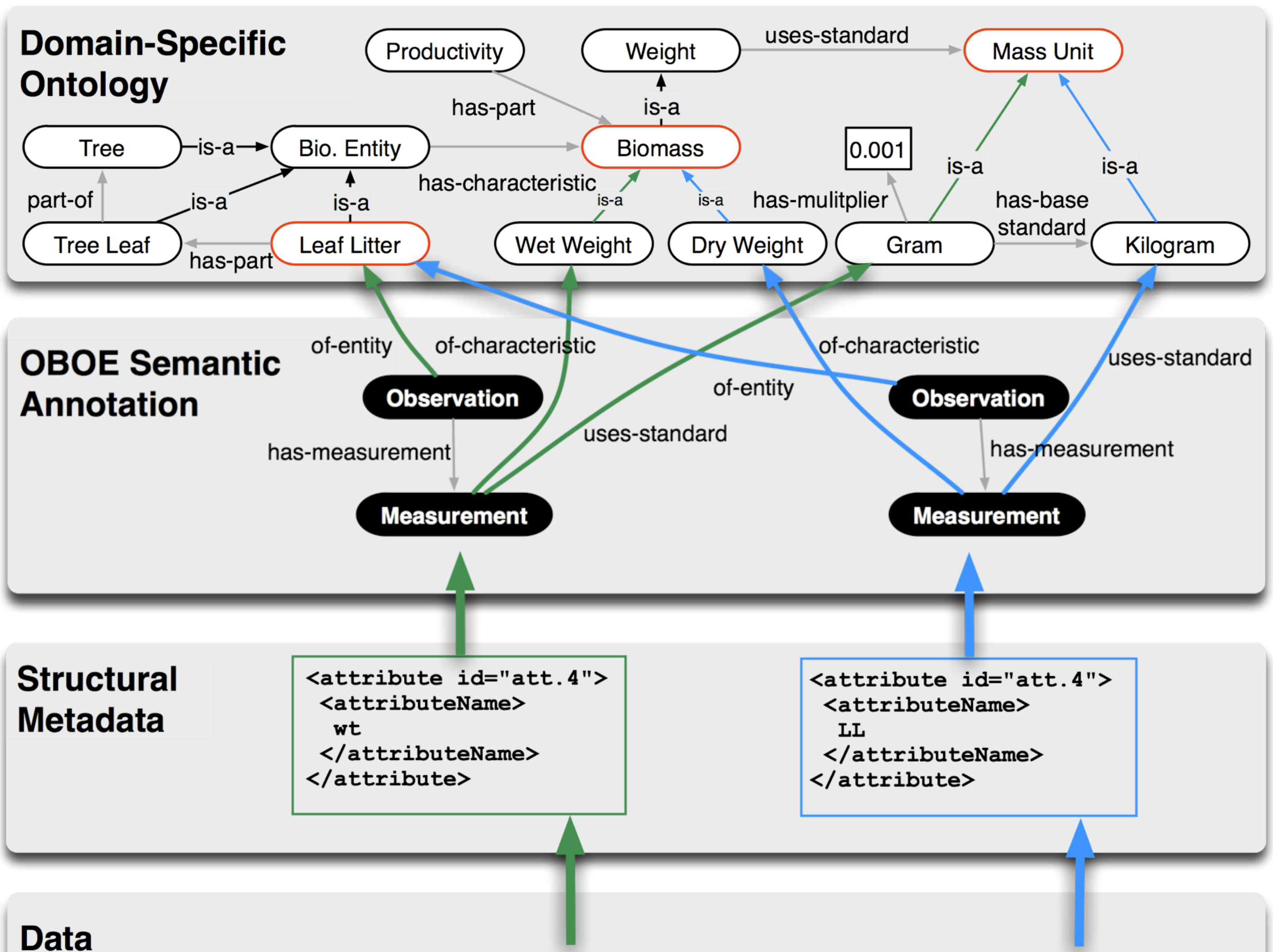
Scientific Observations

Scientific Observations

- A scientific **Observation** is the
  - **Measurement** of the **Value**
  - of a **Characteristic**
  - of some **Entity**
  - in a particular **Context**



Domain-Specific Ontology
Productivity
Weight
uses-standard
Mass Unit
has-part
is-a
Tree
is-a
Bio. Entity
Biomass
0.001
is-a
is-a
has-characteristic
part-of
is-a
is-a
is-a
is-a
has-mulitplier
has-base standard
Tree Leaf
has-part
Leaf Litter
Wet Weight
Dry Weight
Gram
Kilogram
Structural Metadata
<attribute id="att.4">
<attributeName>
wt
</attributeName>
</attribute>
<attribute id="att.4">
<attributeName>
LL
</attributeName>
</attribute>
Data
loc quad nitr wt
SCAL 1 N 6.2
SCAL 2 Y 7.2
CCAL 1 N 4.2
... ...
place treat plot LL
Sth C 1 0.003
Sth C 1 0.002
Sth N 1 0.008
... ...

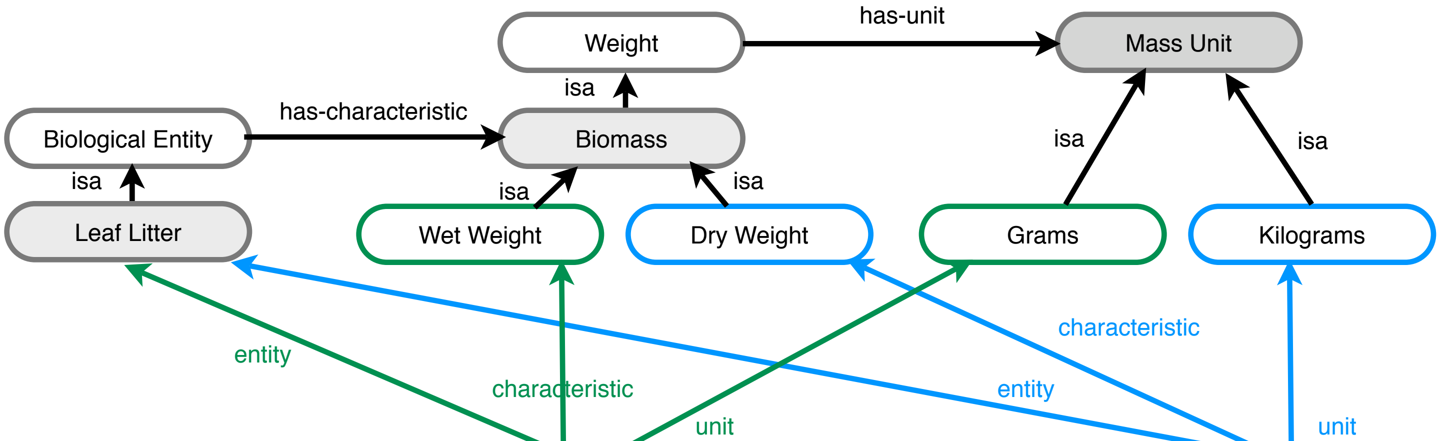


loc	quad	nitr	wt
SCAL	1	N	6.2
SCAL	2	Y	7.2
CCAL	1	N	4.2
...	...	...	...

place	treat	plot	LL
Sth	C	1	0.003
Sth	C	1	0.002
Sth	N	1	0.008
...	...	...	...

Enhancing Structural Metadata

- Ontologies can enhance structural metadata by providing
  - terms for "annotating" columns with concepts
  - concepts and properties for representing relationships



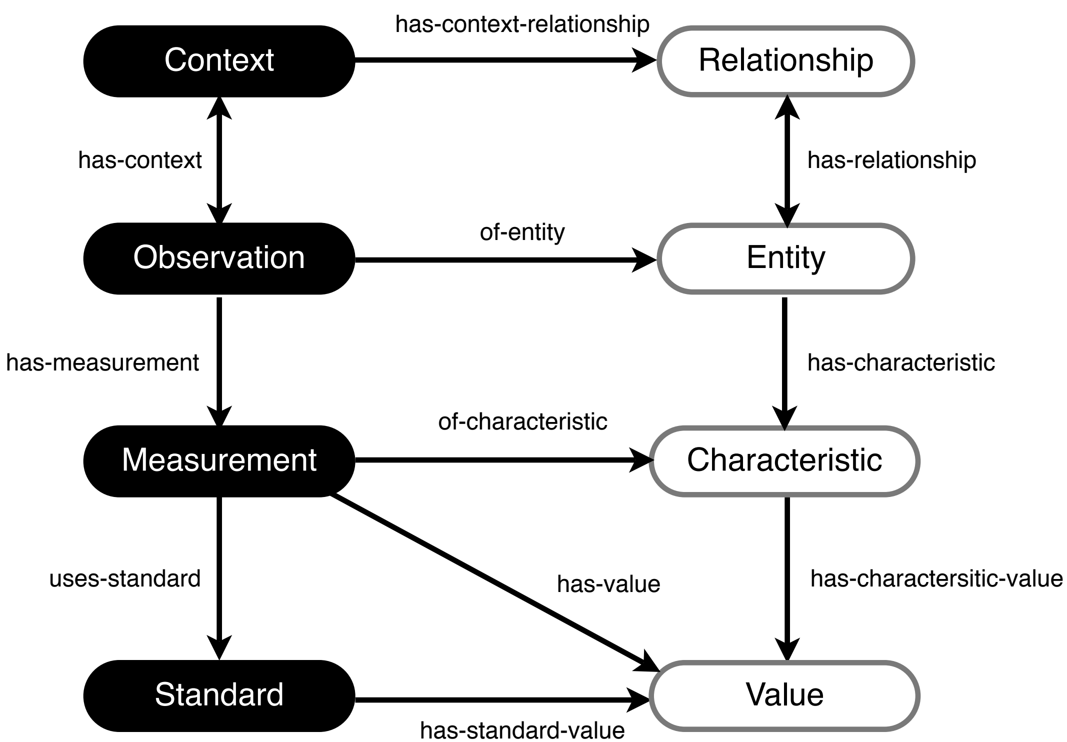
place	trmt	plot	LL
sth	c	1	0.003
sth	c	1	0.002
sth	n	1	0.008
...	...	...	...

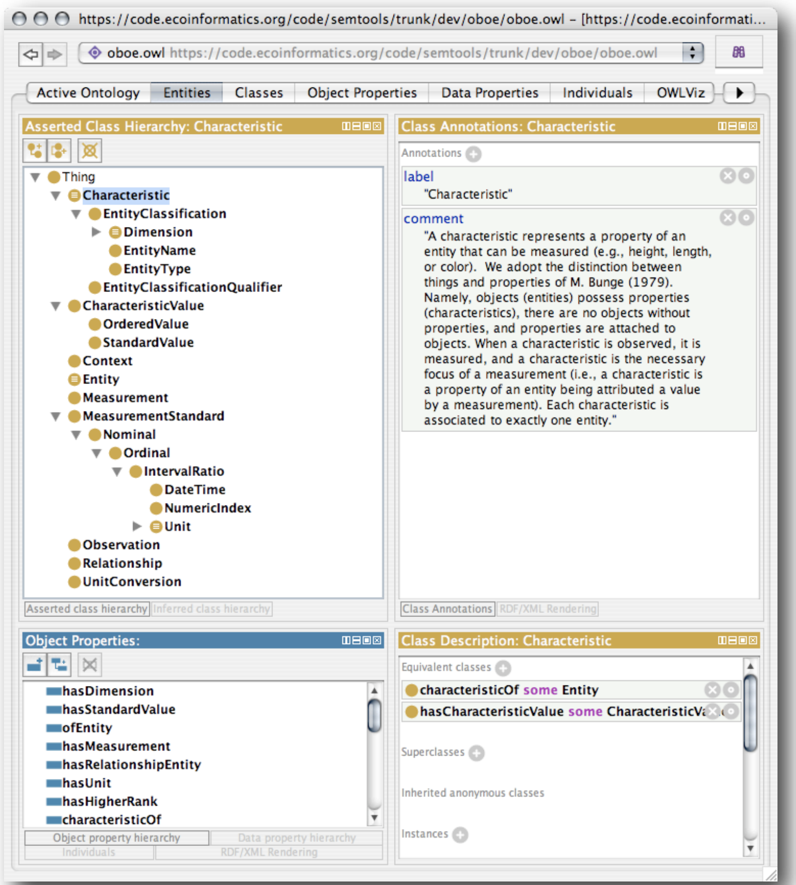
loc	quad	nitr	wt
scal	1	n	6.2
scal	2	v	7.2
ocal	1	n	4.2
...	...	...	...

Can search for and compare columns via annotations

The OBOE Model

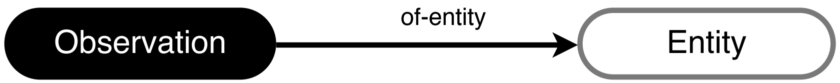
- Separates observation from what was observed

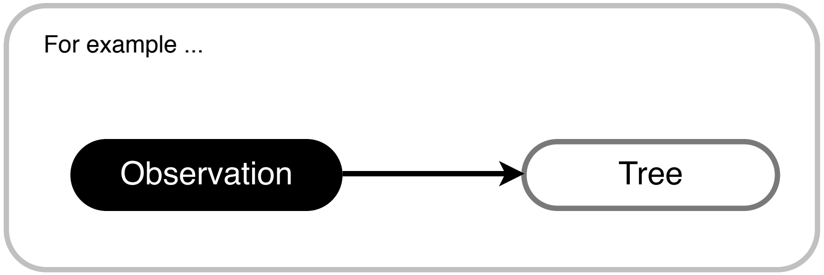




The OBOE Ontology

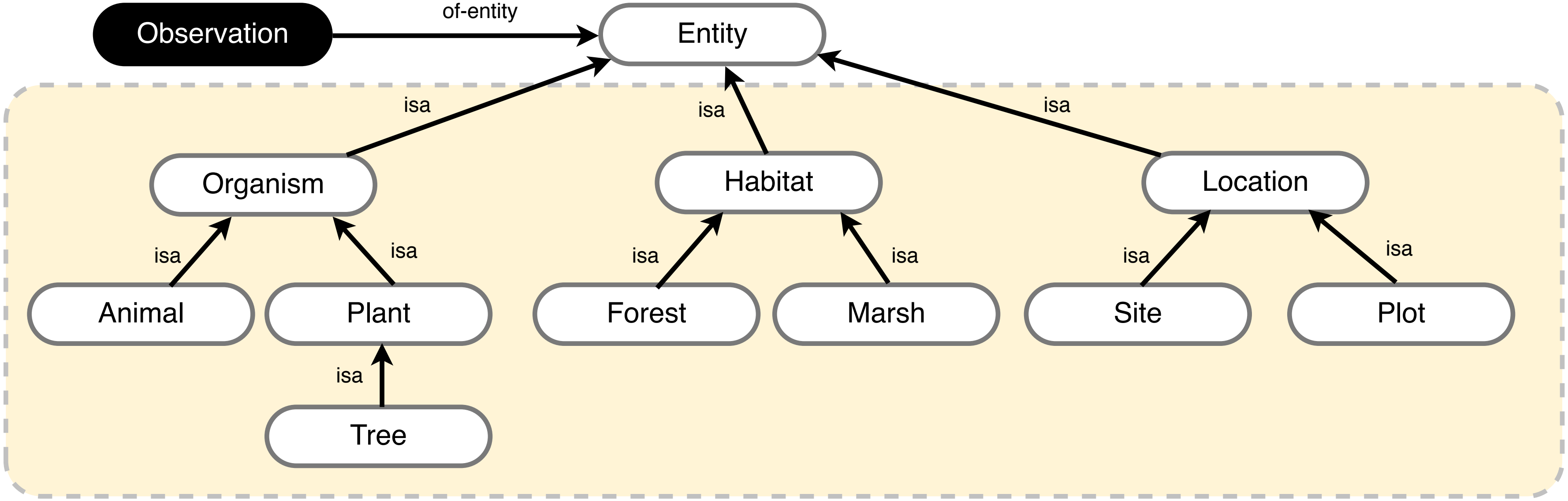
- An Observation is an assertion that something was observed
- Every observation is of some Entity





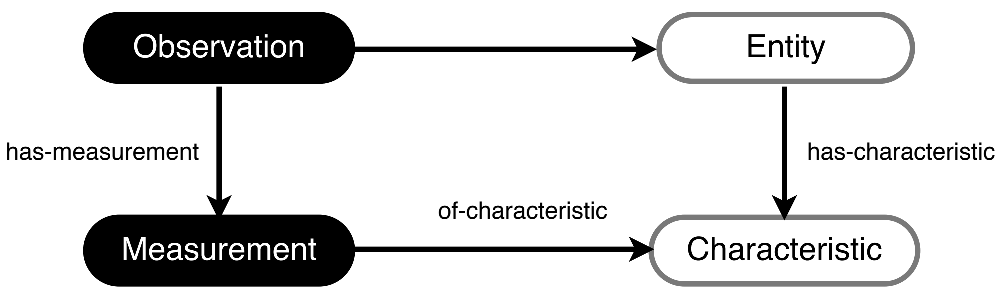
The OBOE Ontology

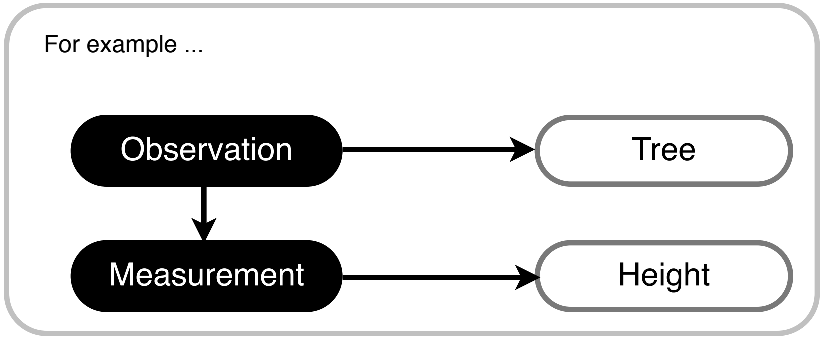
- Entities are OBOE extension points
  - extended by domain ontology terms



The OBOE Ontology

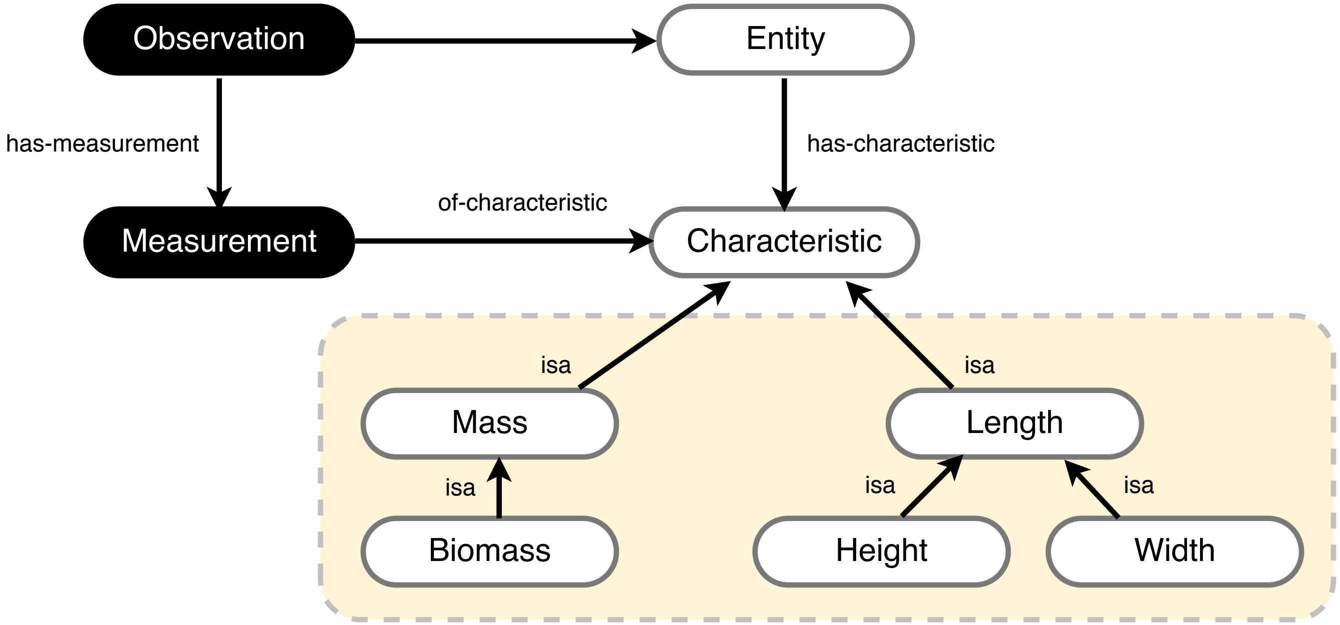
- Observations are composed of Measurements
- Measurements are of an entity Characteristic





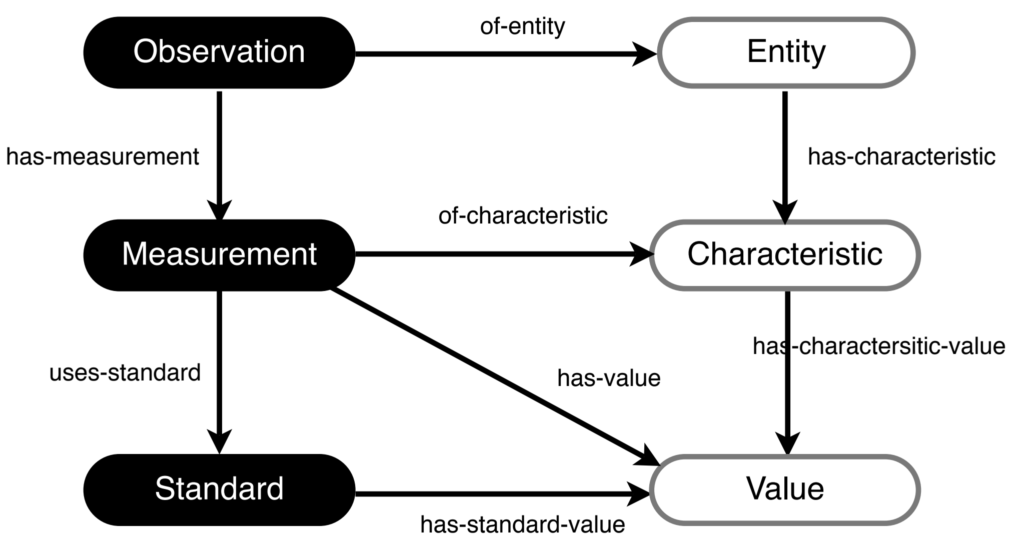
The OBOE Ontology

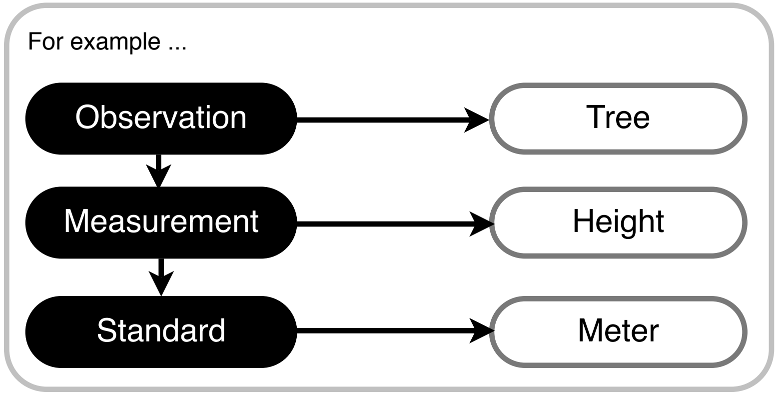
- Characteristics are another extension point



The OBOE Ontology

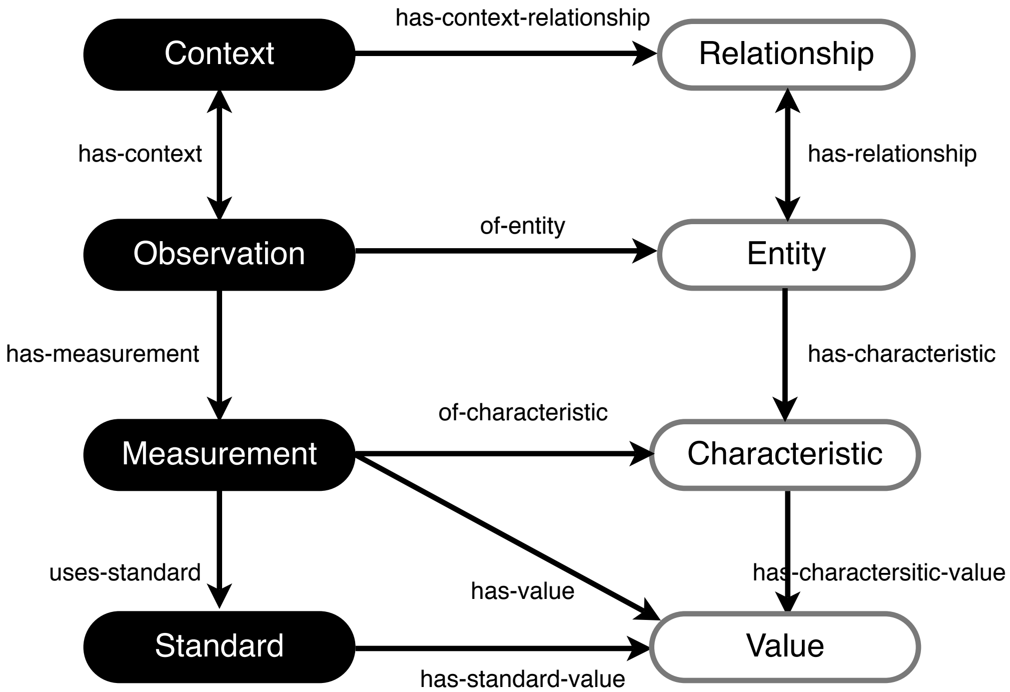
- Values assigned to characteristics according to Measurement Standards (e.g., units)
- Standards are another extension point
- Measurements also have precision

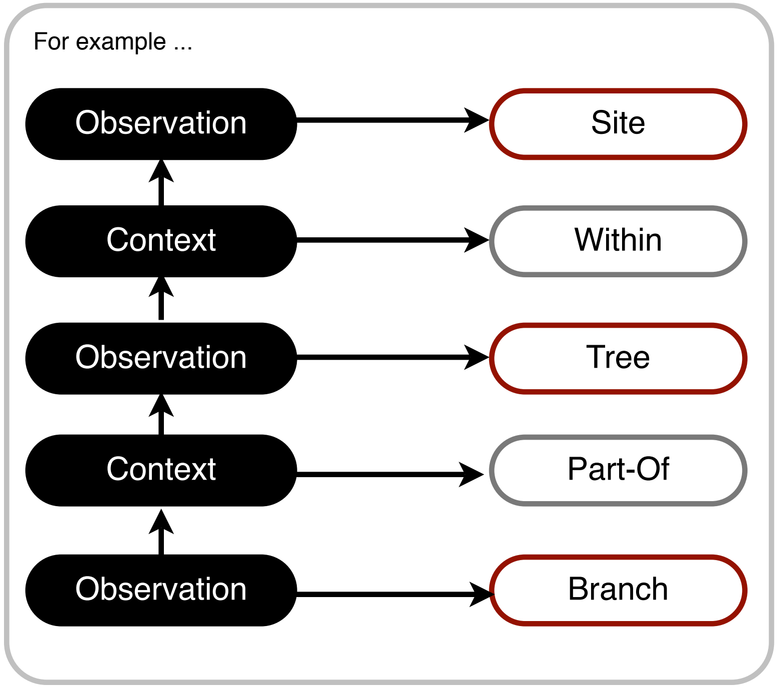




The OBOE Ontology

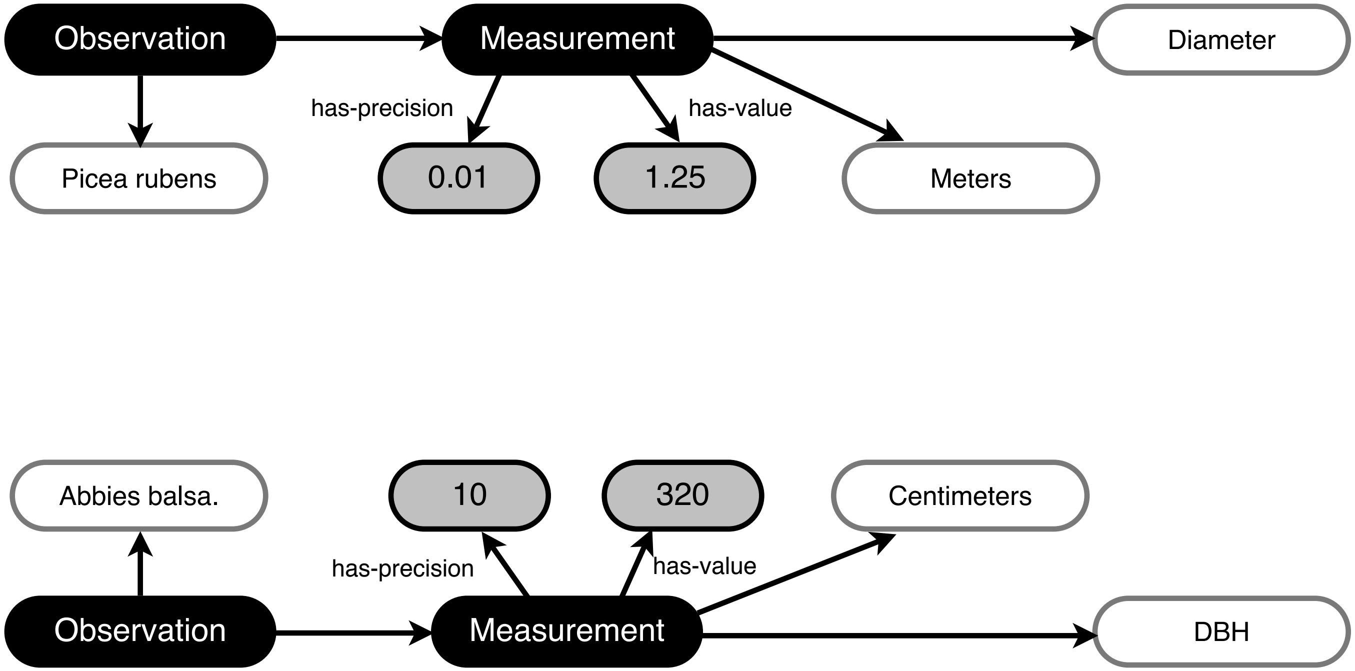
- Observations occur within a Context (e.g., spatial, temporal, ...)
- Context is denoted by other Observations
- Context is transitive (e.g., Branches also contextualized by a Site)





OBOE: Aligning Observations

- Observations can be aligned for data integration ...

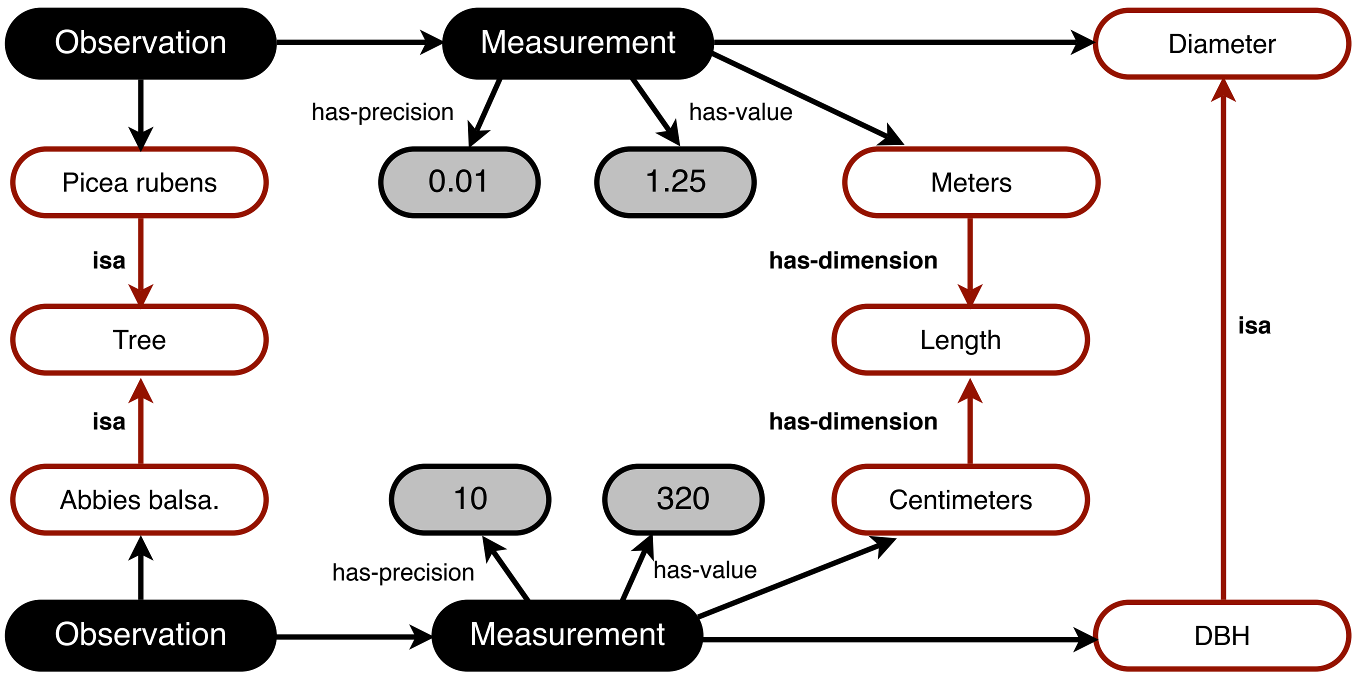


Two similar observations of trees

OBOE: Aligning Observations

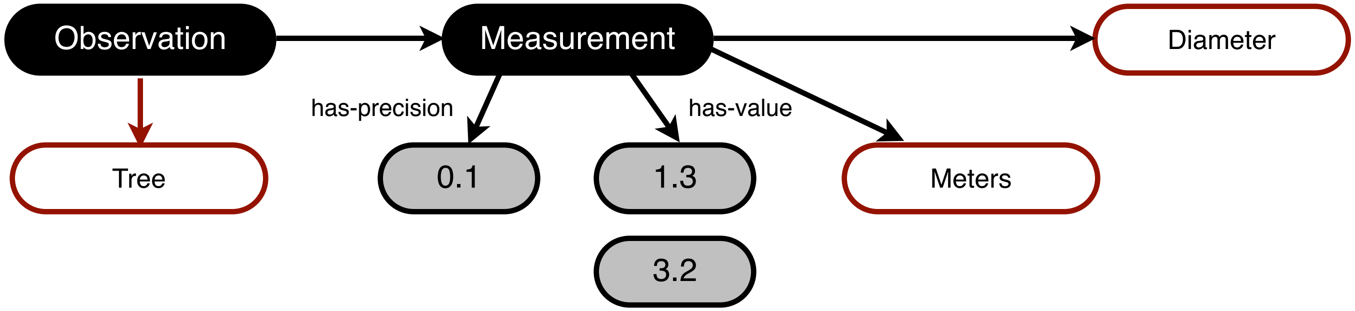
- Observations can be aligned for data integration ...

Align entities, characteristics, and standards



Data Integration with OBOE

- Observations can be aligned for data integration ...

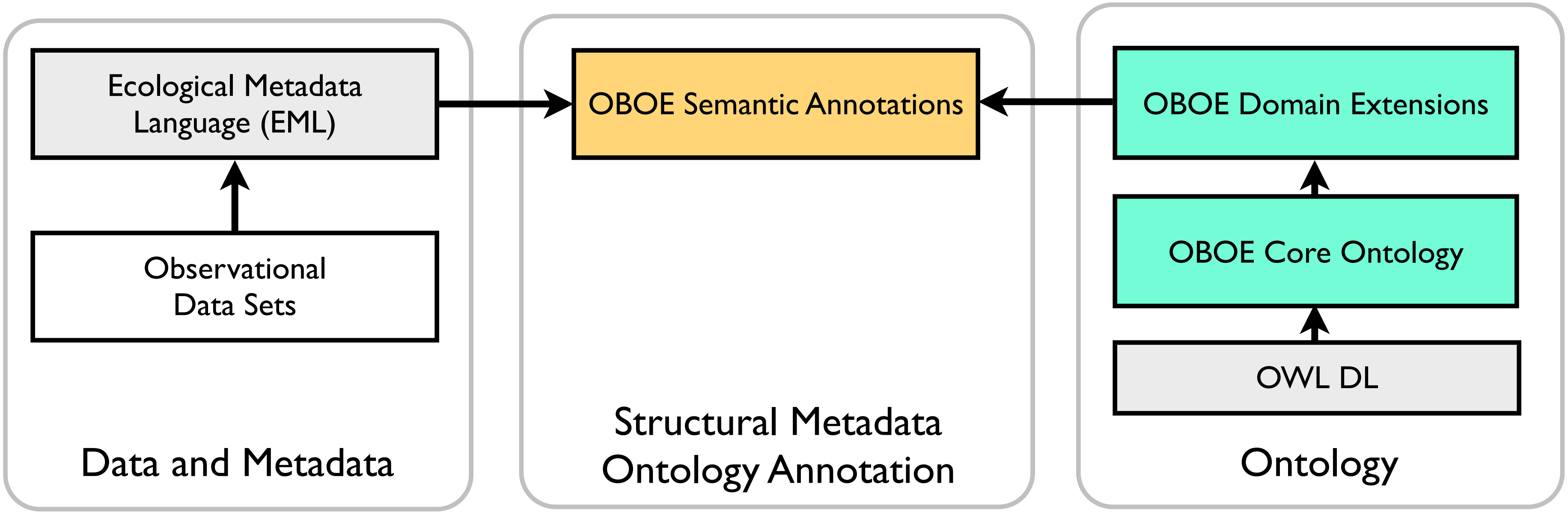


Apply *conversions* based on alignments, e.g.

- use common Entity and Characteristic concepts
- apply Unit conversions to values
- select lowest precision and apply

The OBOE Framework

- The Extensible Observation Ontology
  - represented in OWL-DL
  - generic concepts and properties for describing observations
  - explicit "extension points" for defining domain ontologies
  - support for annotating data sets via observation terms



Semantic Tools Prototypes

- Extend tools for Semantic Data Management

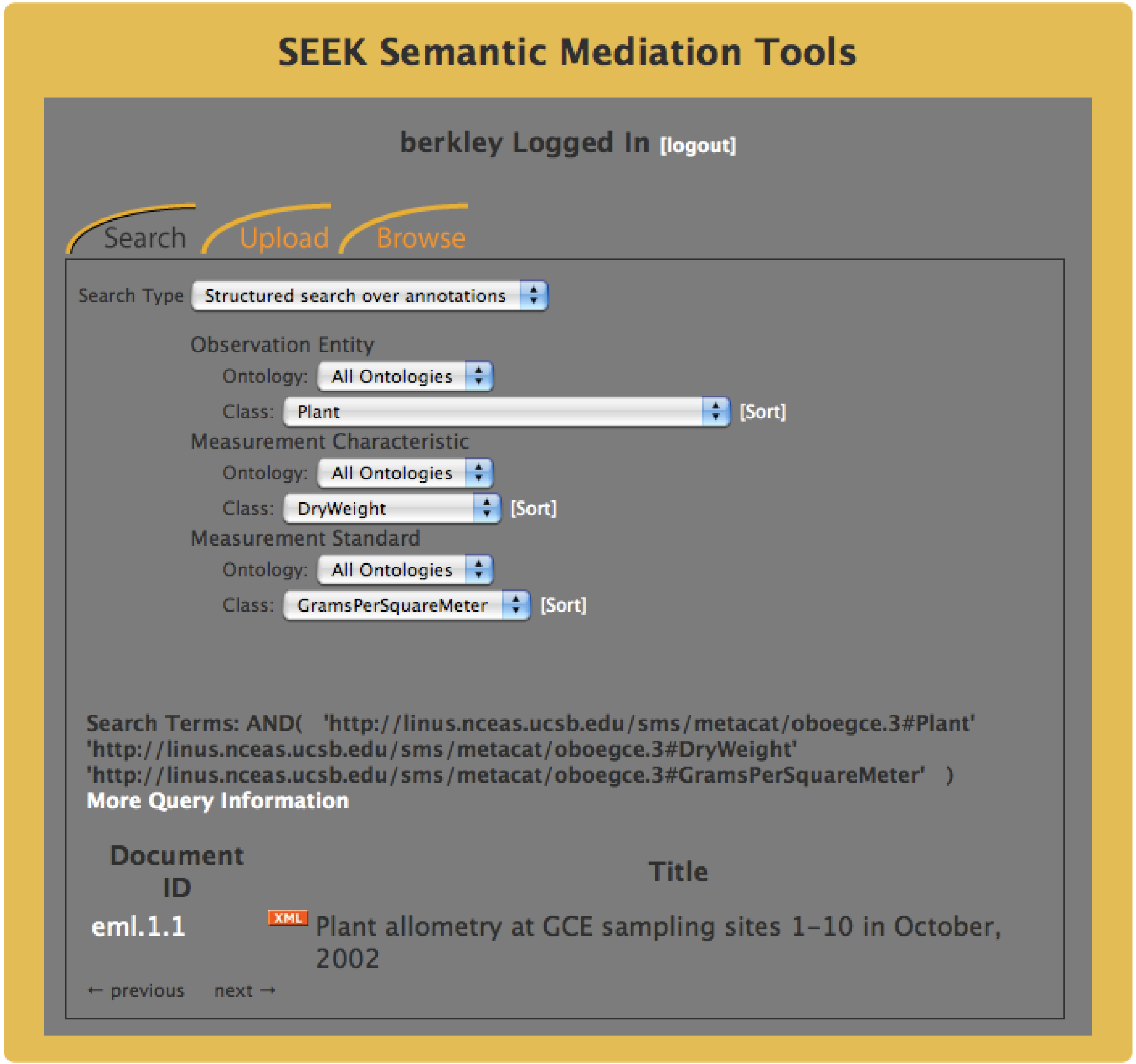
<EML>

Meta Cat

Types of Implemented Searches

- Simple Keyword (baseline)
- Keyword-based term expansion
- Annotation enhanced term expansion
- Observation based semantic query

Structured Search



Take-home points

- Generalized data integration is a phenomenally challenging problem for synthesis applications
- Metadata is a good start, but needs to be semantically enriched to truly enable data integration
- Annotation: provides system independence

SONet: A Community-Driven *Scientific Observations Network* to achieve Semantic Interoperability of

Project Organizers

Mark Schildhauer[1], Shawn Bowers[2], Corina Gries[3], Deborah McGuinness[4], Philip Dibner[5], Josh Madin[6], Matt Jones[1], Luis Bermudez[7], John Graybeal[7]

[1]NCEAS UC Santa Barbara, [2]UC Davis Genome Center
[3]CAP/LTER and Univ. of Arizona, [4]McGuinness Associates,
[5]OGC Interoperability Institute, [6]Macquarie University,
[7]Monterey Bay Aquarium Research Institute

Objectives of SONet

Broad Objectives

- Address *semantic interoperability* issues in environmental and ecological data [sharing, discovery, integration]
- Build a *network of practioners*

- **Immediate Goals to Develop:**
- An extensible and open *observations data model* to unify existing domain-specific approaches
- A semantic (ontology) framework for *scientific terminology*, and corresponding domain extensions
- *Demonstration prototypes* using these to address current interoperability issues

- Please join SONet to make it a success!

Questions?

- Madin, Bowers, Schildhauer, and Jones. 2008. **Advancing ecological research with ontologies**. Trends in Ecology and Evolution 23(3): 159-168.

- http://www.nceas.ucsb.edu/ecoinformatics/
- http://sonet.ecoinformatics.org
- http://knb.ecoinformatics.org/
- http://kepler-project.org/

NCEAS

Acknowledgments

- This material is based upon work supported by:
- The National Science Foundation under Grant Numbers 9980154, 9904777, 0131178, 9905838, 0129792, and 0225676.
- The National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant Number 0072909), the University of California, and the UC Santa Barbara campus.
- The Andrew W. Mellon Foundation.

- Thanks to Shawn Bowers, Mark Schildhauer, and Josh Madin for many slides.