

Semantic Data Integration for Heterogeneous Scientific Data

Matthew B. Jones¹, Mark Schildhauer¹, Margaret O'Brien¹,
Chad Berkley¹

National Center for Ecological Analysis and Synthesis (NCEAS)
University of California, Santa Barbara¹

Shawn Bowers²

University of California, Davis²

Josh Madin³

MacQuarie University³



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 inaccessible
- Data sharing only via personal networks 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 Island, Georgia. The experiments were conducted from May 1996 to September 1997. Each mixture differed in plot elevation, soil water content, and soil salinity, so each was considered a separate habitat type. The experiments occurred in different geographic locations (i.e. Dean Creek on southern Sapelo Island, Marsh Landing on southwestern Sapelo Island, and Shell Island). The University of Georgia Marine Institute. In May 1996, 16 1mx1m plots were placed within each plant mixture and alternate plots were assigned to different fertilization treatments. Pelleted fertilizer (29% N, 3% P, 4% K) was broadcast into fertilization treatment plots by hand at the rate of 60g/m². The 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 constant weight, and 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)
- corrensantha (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
...





KNB Software Suite

```
<attribute id="att.5">  
<attributeName>avesr91</attributeName>  
<attributeLabel>Average Species Richness for 1991</attributeLabel>  
<attributeDefinition>The average species richness for the field in 1991</attributeDefinition>  
<storageType>float</storageType>  
<measurementScale>  
    <ratio>  
        <unit><standardUnit>none</standardUnit></unit>  
        <precision>0.1</precision>  
        <numericDomain id="att.5.1">  
            <numberType>real</numberType>  
            <bounds>  
                <minimum exclusive=>0</minimum>  
                <maximum>100</maximum>  
            </bounds>  
        </ratio>  
    </measurementScale>  
</attribute>
```

Metadata

<eml/>



The Knowledge Network for Biocomplexity

KNB is a network intended to facilitate ecological and environmental research. Using the KNB is an efficient way to discover, access, integrate, retrieve and analyze data from a highly distributed set of field stations, laboratories, research sites, and individual researchers.

View Interactive Map

Search Title, Abstract, Keywords, Personnel (Quicker)

Search all fields (detailed)

Taxonomy Measurements

Meta Cat

Data Storage



Data Package jscientist.7.3

File Edit Search Presentation Data Window Help

Joe Scientist - Population sample data for zooplankton in the Great Lakes, 2000 (fictitious sample data :)

Accession Number: jscientist.7.3

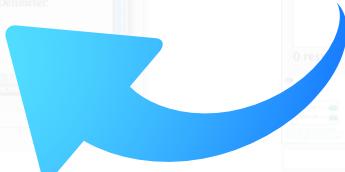
zooplankton in Dorn sampling, population survey, fictitious sample data, Great Lakes

Morpho

Text	Text	Text	Text	Text	Text	Text
Lake	site	sampledate	depths	depths	depths	depths
Lake Erie	N1	10JUN2000	1	1	1	1
Lake Erie	N1	10JUN2000	1	1	1	1
Lake Erie	N2	10JUN2000	1	1	1	1
Lake Erie	N2	10JUN2000	1	1	1	1
Lake Erie	N3	10JUN2000	1	1	1	1
Lake Erie	N3	10JUN2000	1	1	1	1
Lake Erie	N1	10JUN2000	1	1	1	1
Lake Erie	N1	10JUN2000	1	1	1	1
Lake Erie	N2	10JUN2000	1	1	1	1
Lake Erie	N2	10JUN2000	1	1	1	1
Lake Erie	N3	10JUN2000	1	1	1	1
Lake Erie	N3	10JUN2000	1	1	1	1
Lake Erie	N1	10JUN2000	1	1	1	1
Lake Erie	N1	10JUN2000	1	1	1	1
Lake Erie	N2	10JUN2000	1	1	1	1
Lake Erie	N2	10JUN2000	1	1	1	1
Lake Erie	N3	10JUN2000	1	1	1	1
Lake Erie	N3	10JUN2000	1	1	1	1
Lake Erie	N1	10JUL2000	1	1	1	1
Lake Erie	N1	10JUL2000	1	1	1	1
Lake Erie	N2	10JUL2000	1	1	1	1
Lake Erie	N2	10JUL2000	1	1	1	1
Lake Erie	N3	10JUL2000	1	1	1	1
Lake Erie	N3	10JUL2000	1	1	1	1

ExampleData.txt

Data Management



Data Analysis



CT Director

Time

XY Plotter

0.1

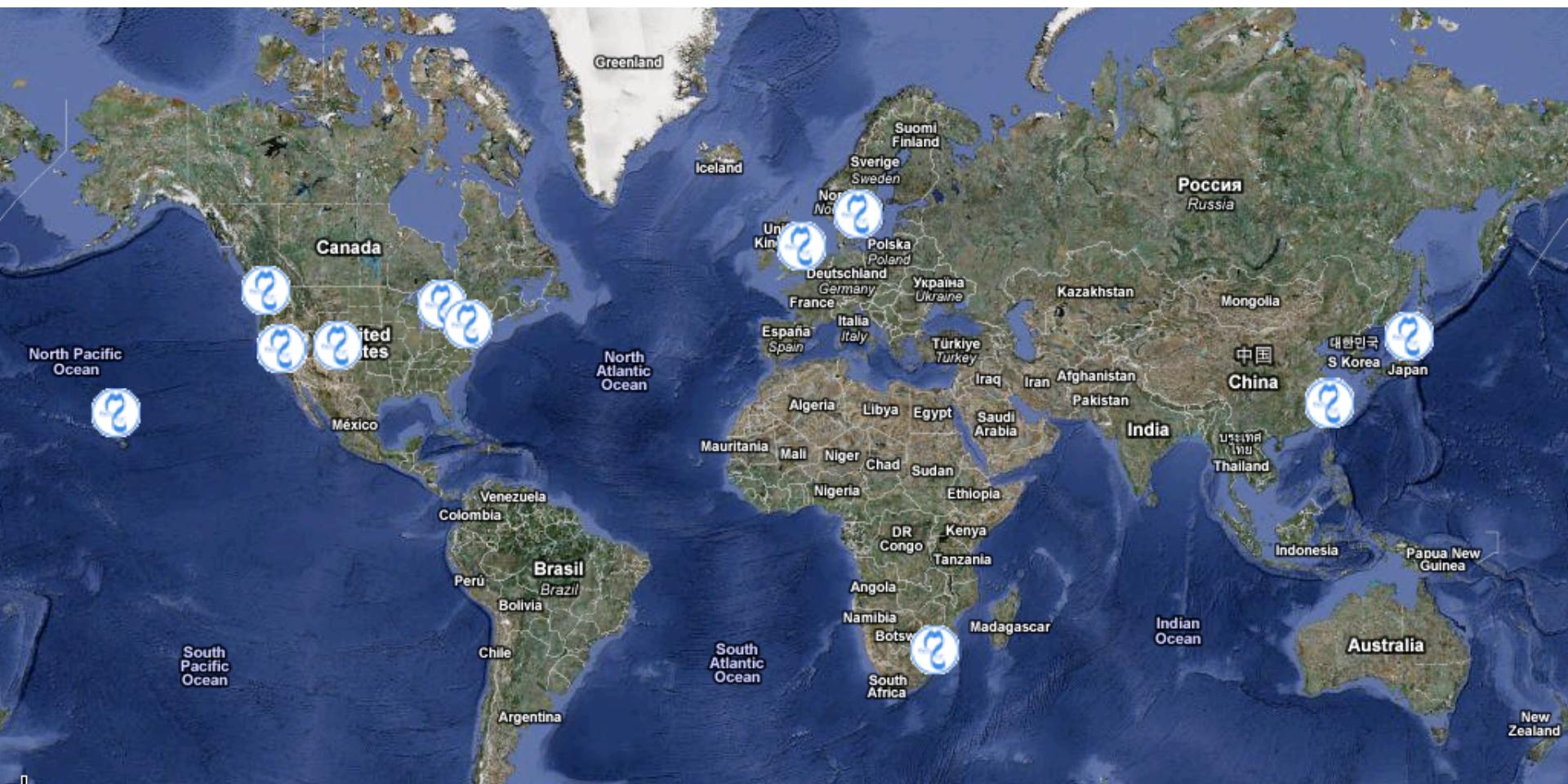
Integrate n

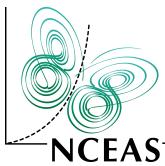
Integrator

Rich Williams, 2003, NCEAS

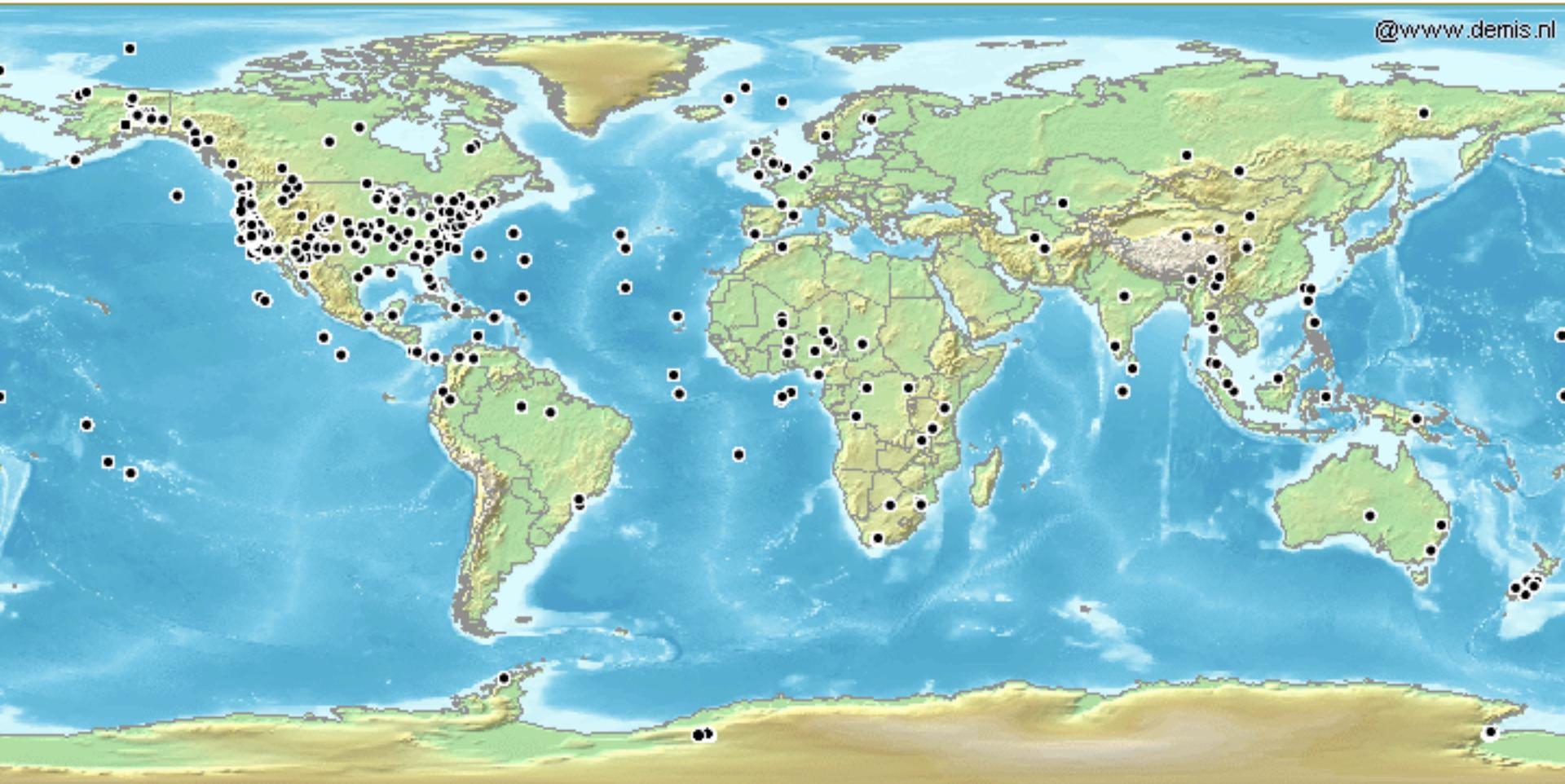
This model shows the solution to the classic Lotka-Volterra predator-prey dynamics model. It uses the Continuous Time domain to solve two coupled differential equations, one that models the predator population and one that models the prey population. The results are plotted as they are calculated showing both population change and a phase diagram of the dynamics.

Global Metacat deployments

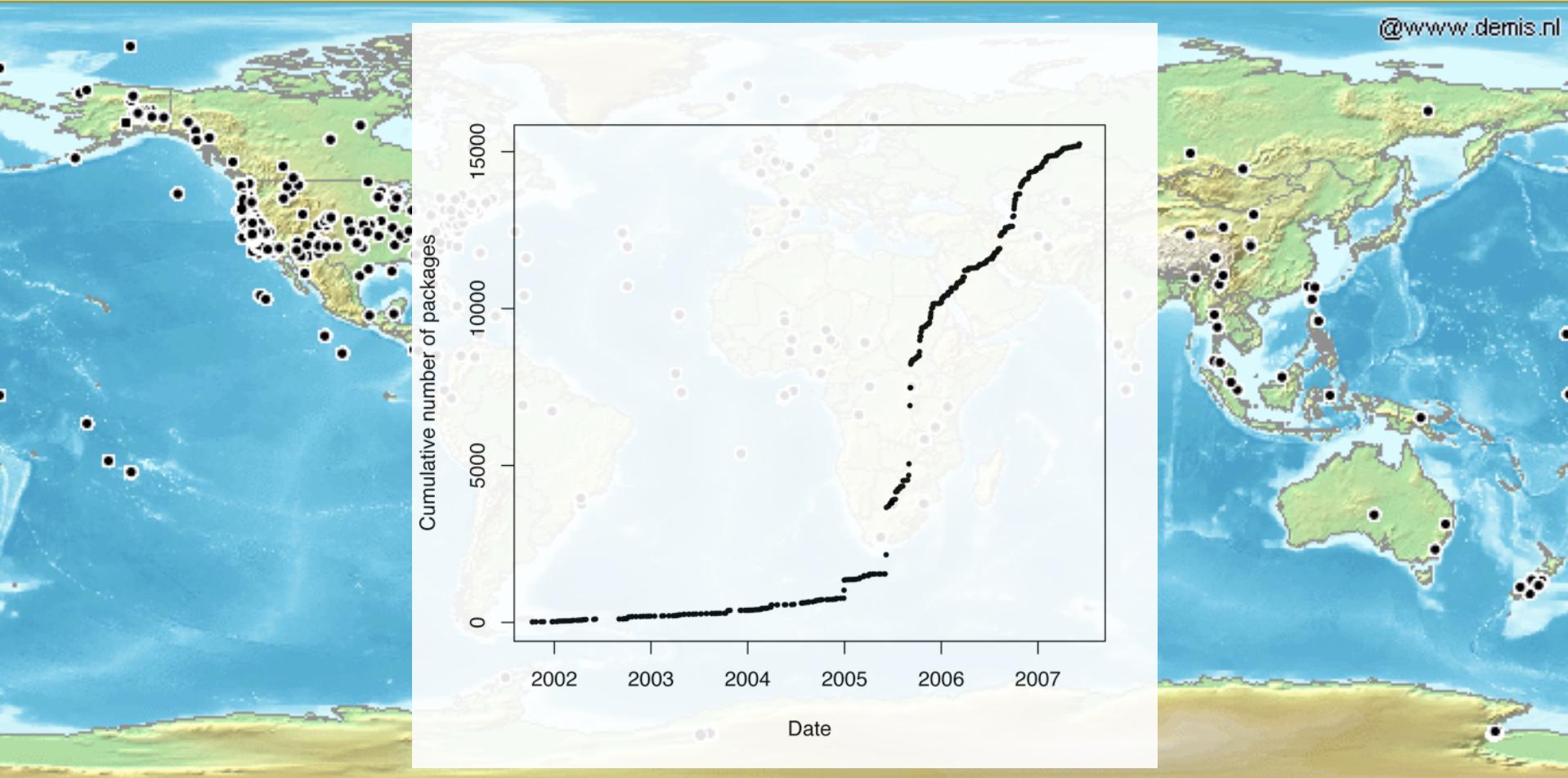




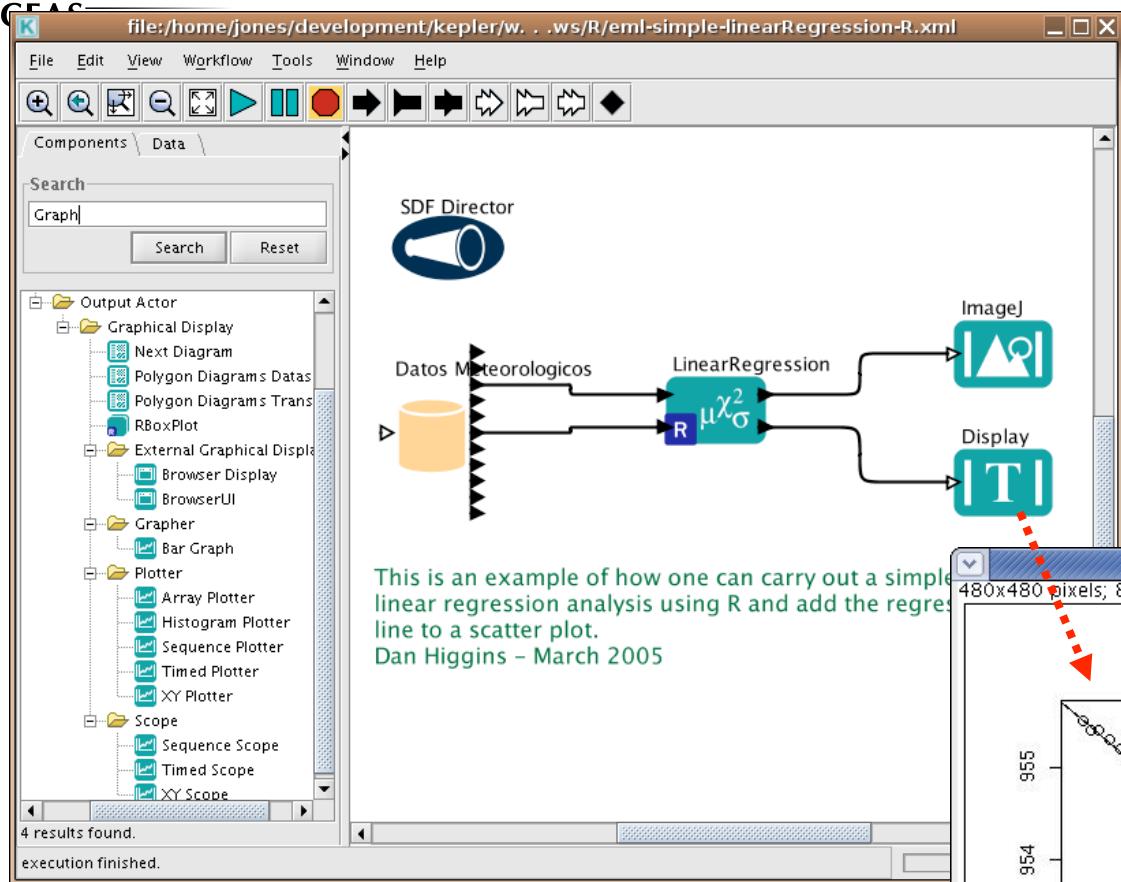
KNB Data Distribution



KNB Data Distribution

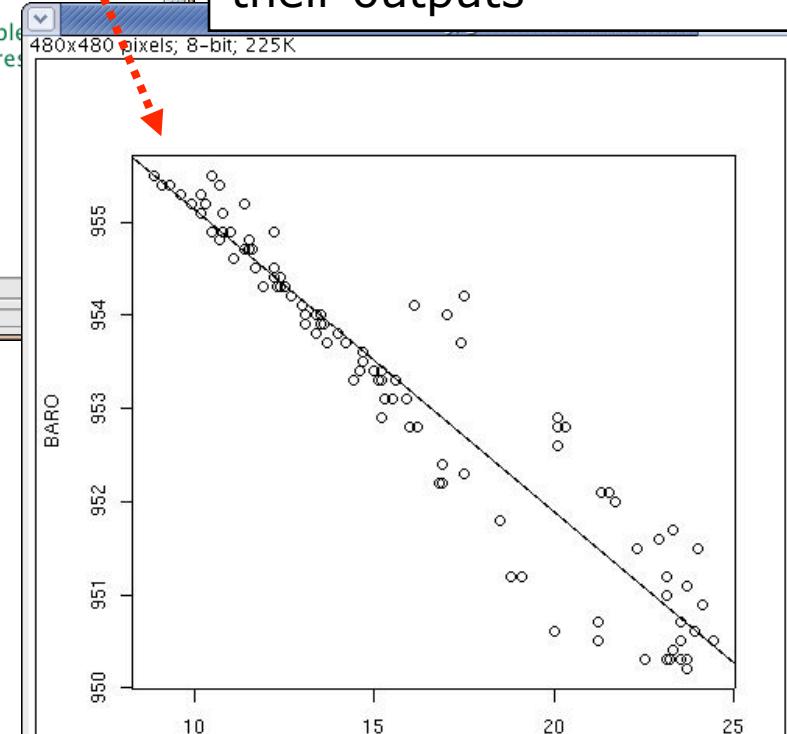


Kepler: scientific workflow system



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

Study A

METADATA (from EML)		Study A: White Mountains		
		Area col. units:		
		PIRU	=	<i>Picea rubens</i>
		BEPA	=	<i>Betula papyrifera</i>
DATA		date	site	species
		10/1/1993	N654	PIRU
		10/3/1994	N654	PIRU
		10/1/1993	N654	BEPA
DATA		area	count	
		2	26	
		2	29	
		1	3	

Study B

METADATA (from EML)		Study B: Green Mountains		
		Area sampled: 1 sq. meter		
		picrub	=	<i>Picea rubens</i>
		betpap	=	<i>Betula papyrifera</i>
DATA		date	site	picrub
		31 Oct 1993	1	13.5
		14 Nov 1994	1	8.4
DATA			betpap	
			1.6	
			1.8	

Integrated Data

study	date	site	species	density
A	10/1/1993	N654	<i>Picea Rubens</i>	13.0
A	10/3/1994	N654	<i>Picea Rubens</i>	14.5
A	10/1/1993	N654	<i>Betula papyrifera</i>	3.0
B	10/31/1993	1	<i>Picea Rubens</i>	13.5
B	10/31/1993	1	<i>Betula papyrifera</i>	1.6
B	11/14/1994	1	<i>Picea Rubens</i>	8.4
B	11/14/1994	1	<i>Betula papyrifera</i>	1.8

metadata 'promoted' to become data

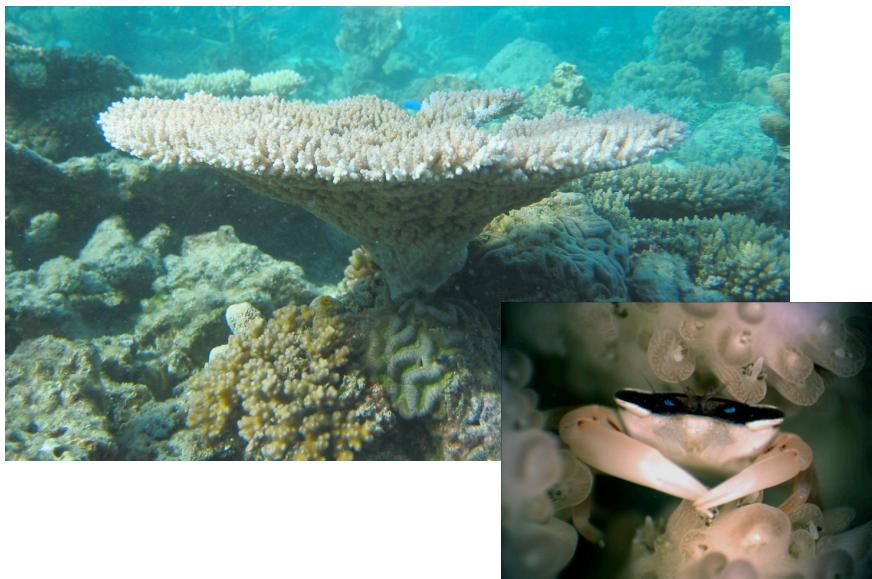
format normalized using metadata

species metadata from study B is now data (picrub/betpap column headings)

density calculated using metadata

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



<u>Site</u>	<u>Temp</u>	<u>Plot</u>	<u>species</u>	<u>Ht.</u>
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

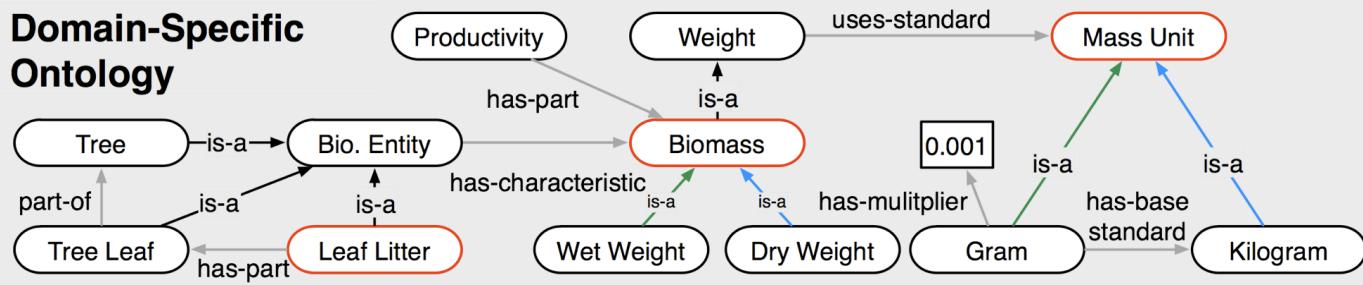
based on a common model of

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



Structural Metadata

```

<attribute id="att.4">
  <attributeName>
    wt
  </attributeName>
</attribute>
  
```

```

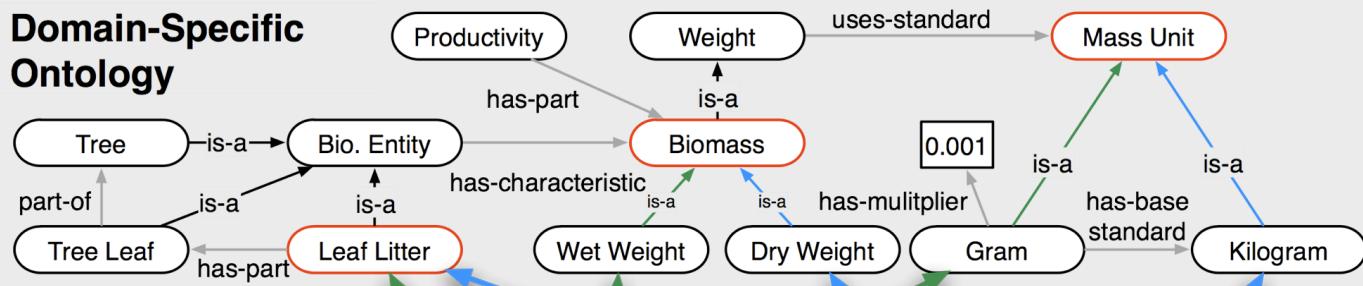
<attribute id="att.4">
  <attributeName>
    LL
  </attributeName>
</attribute>
  
```

Data

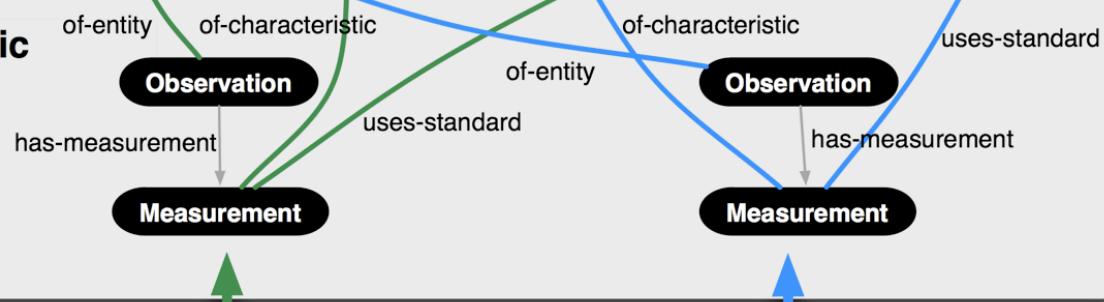
loc	quad	nitr	wt
SCAL	1	N	6.2
SCAL	2	Y	7.2
CCAL	1	N	4.2
...

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

Domain-Specific Ontology



OBOE Semantic Annotation



Structural Metadata

```

<attribute id="att.4">
  <attributeName>
    wt
  </attributeName>
</attribute>
  
```

```

<attribute id="att.4">
  <attributeName>
    LL
  </attributeName>
</attribute>
  
```

Data

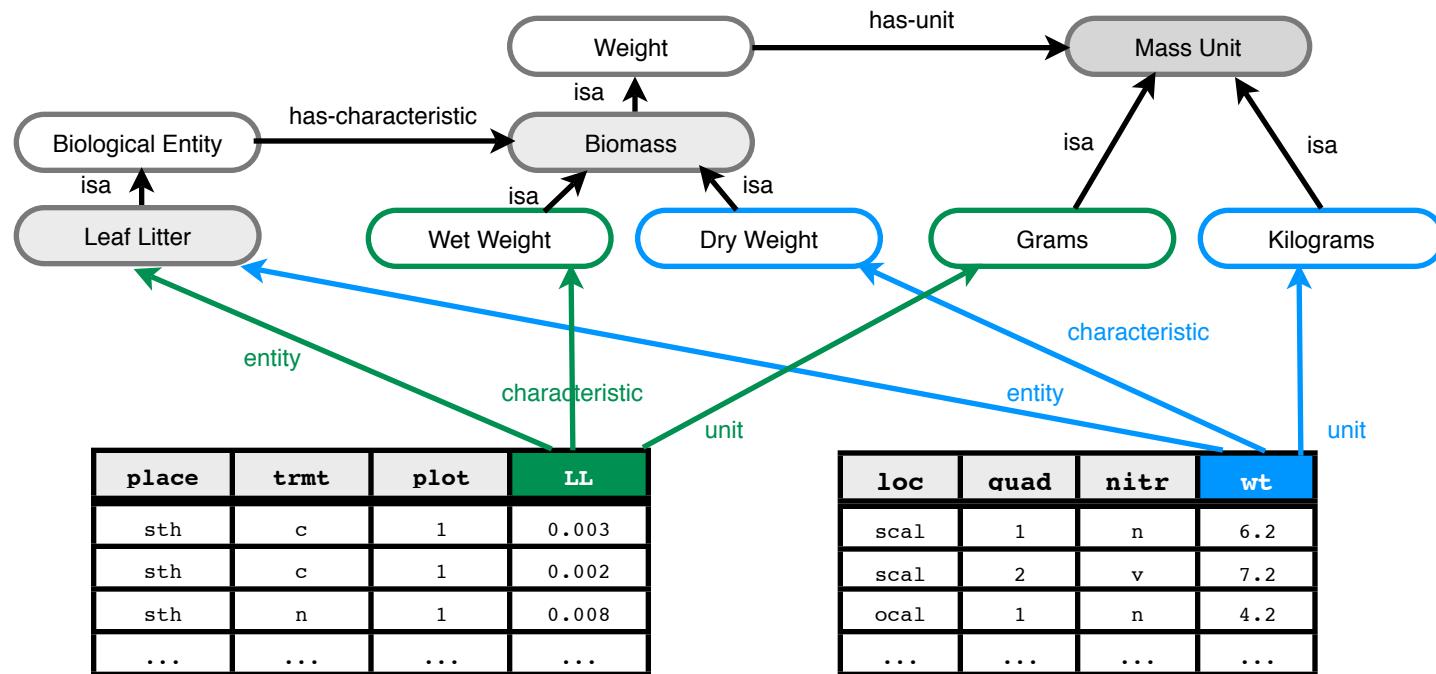
loc	quad	nitr	wt
SCAL	1	N	6.2
SCAL	2	Y	7.2
CCAL	1	N	4.2
...

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

Enhancing Structural Metadata



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

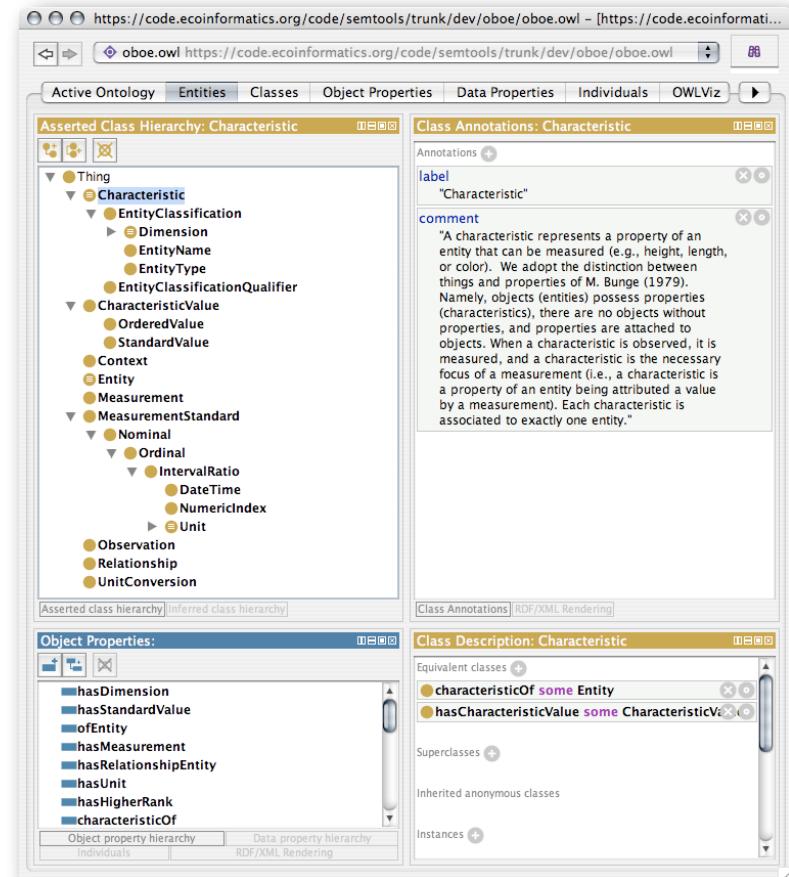
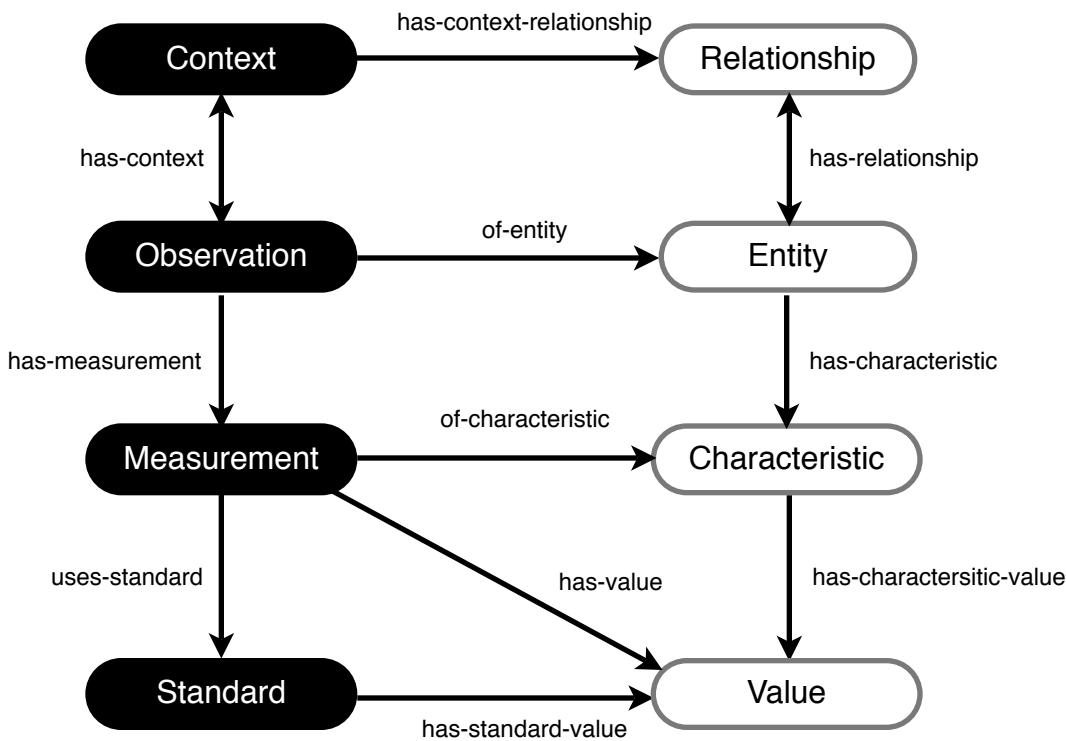


Can search for and compare columns via annotations

The OBOE Model



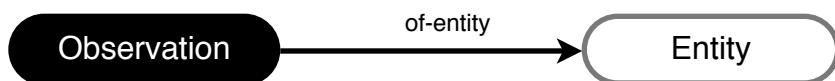
- Separates observation from what was observed



The OBOE Ontology



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



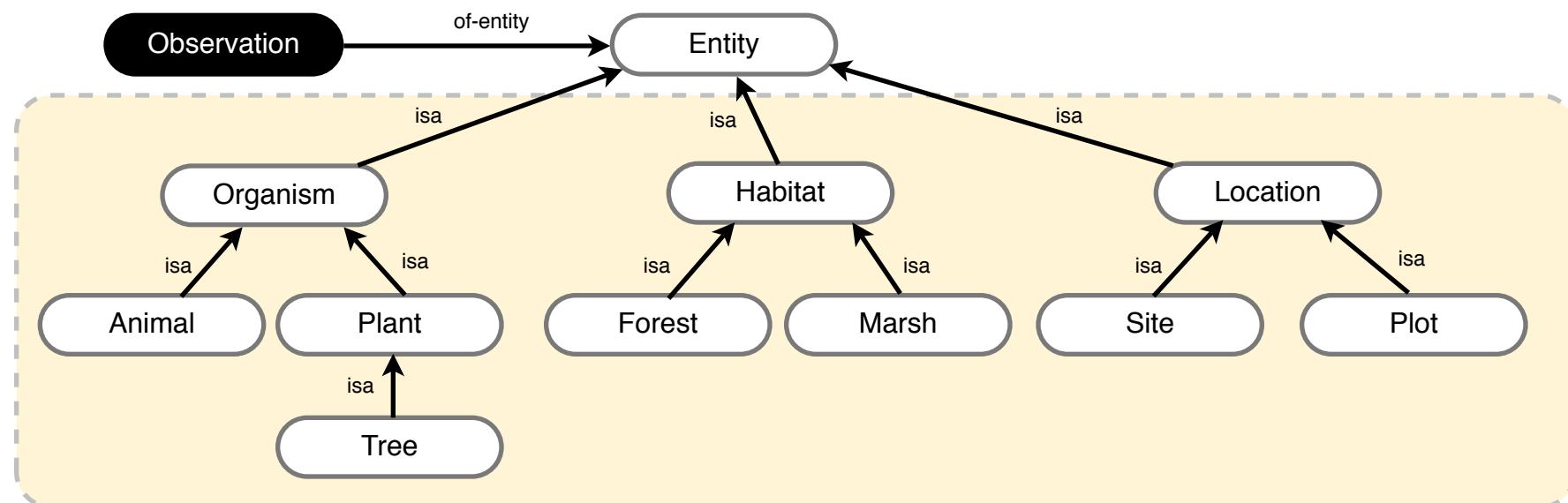
For example ...



The OBOE Ontology



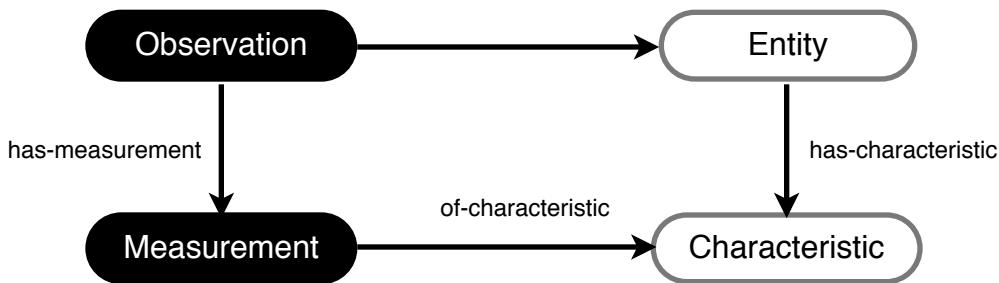
- Entities are OBOE extension points
 - extended by domain ontology terms



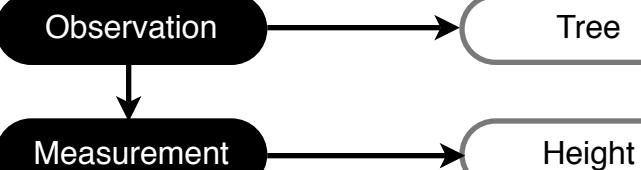
The OBOE Ontology



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



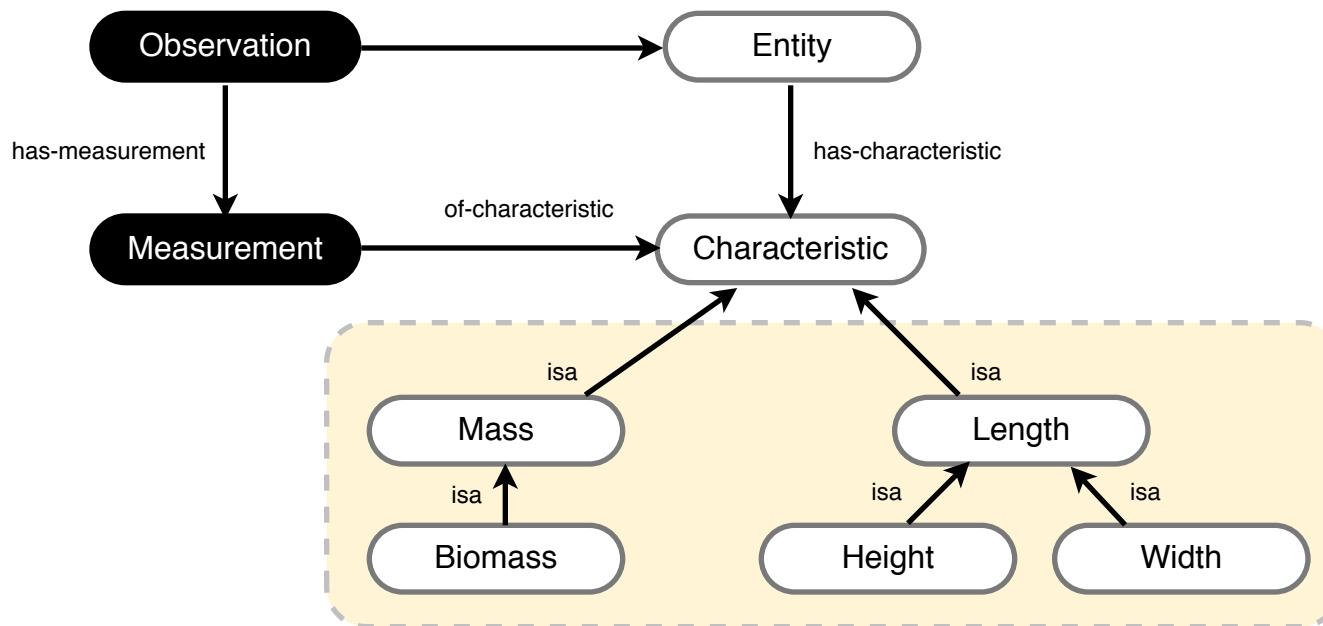
For example ...



The OBOE Ontology



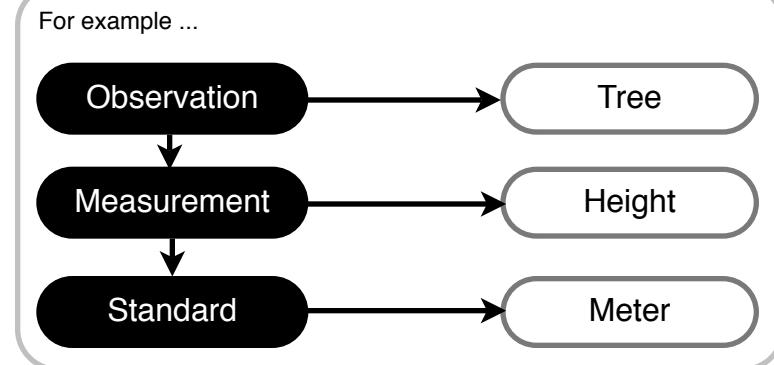
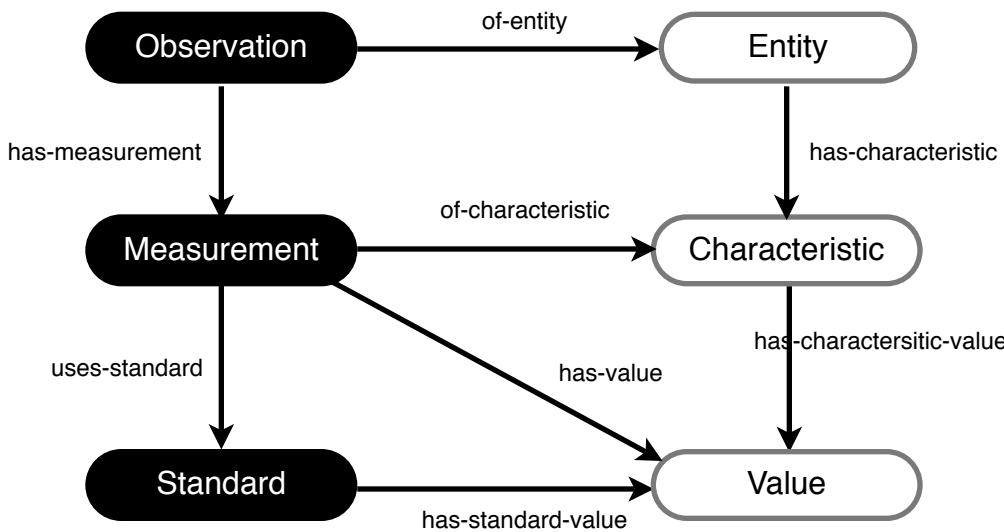
- Characteristics are another extension point



The OBOE Ontology



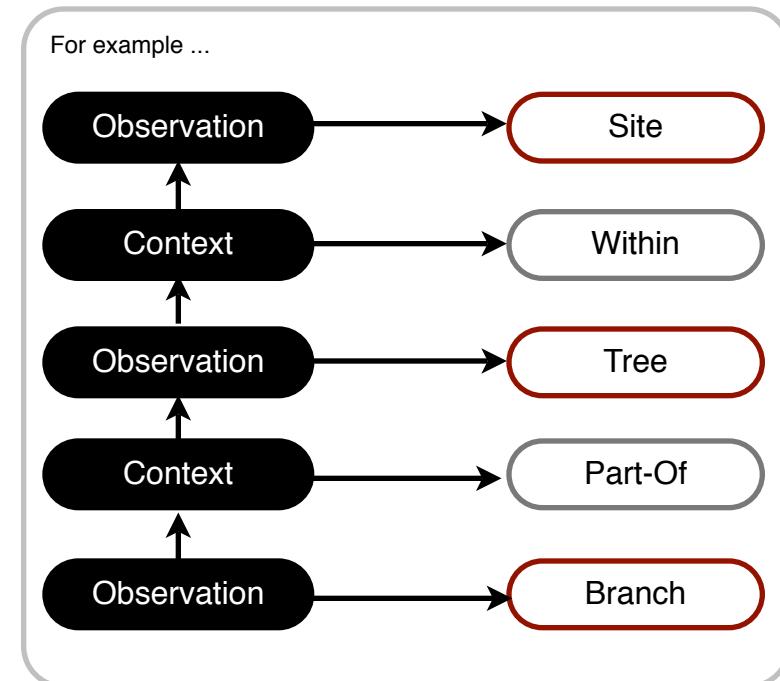
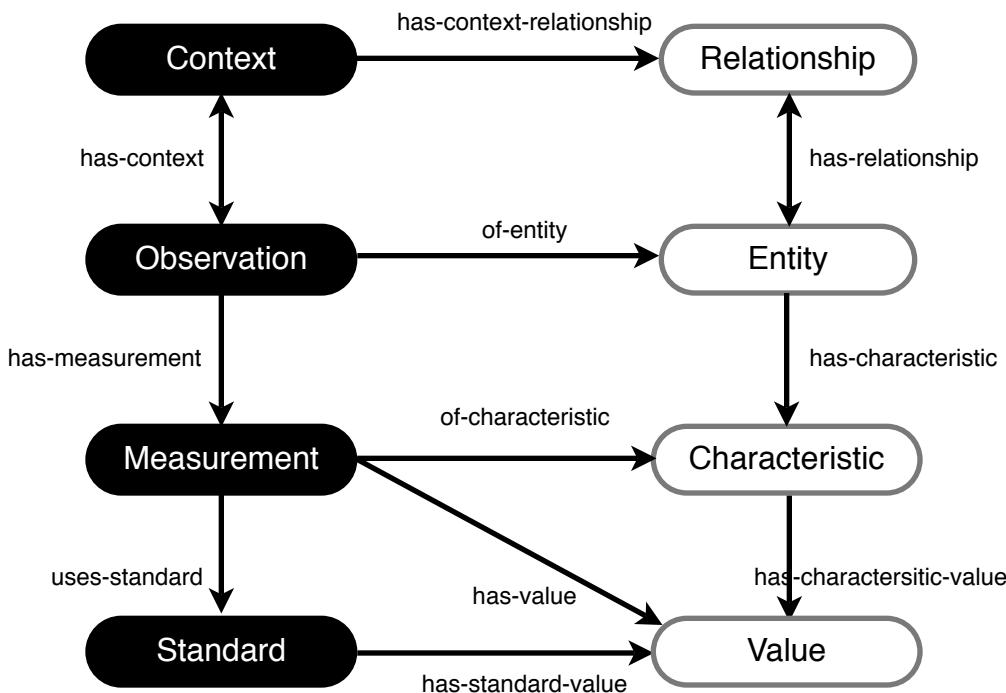
- Values assigned to characteristics according to **Measurement Standards** (e.g., units)
- Standards are another extension point
- Measurements also have precision



The OBOE Ontology



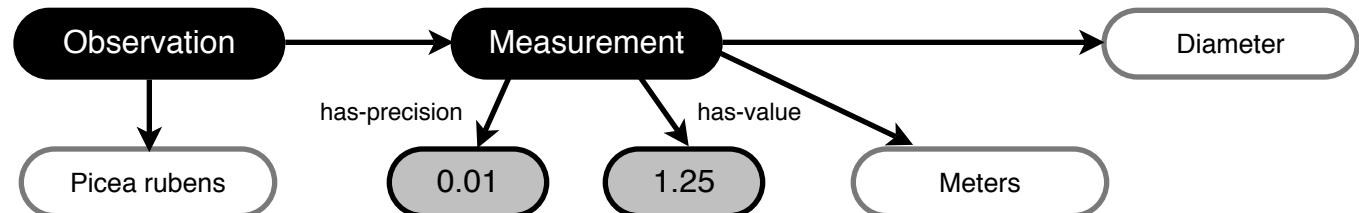
- Observations occur within a **Context** (e.g., spatial, temporal, ...)
- Context is denoted by other Observations
- Context is **transitive** (e.g., Branches also contextualized by a Site)



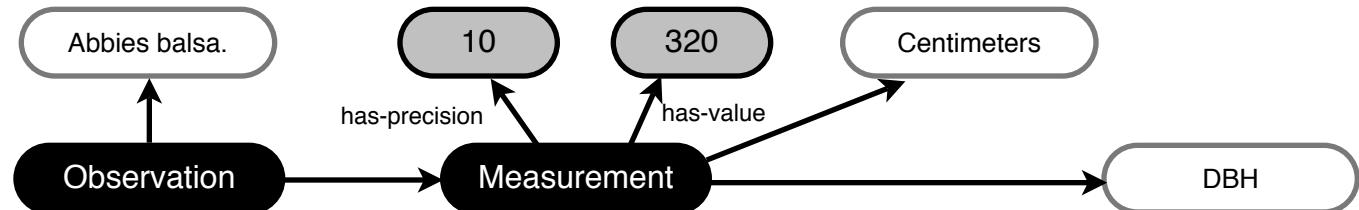
OBOE: Aligning Observations



- Observations can be aligned for data integration ...



Two similar observations of trees

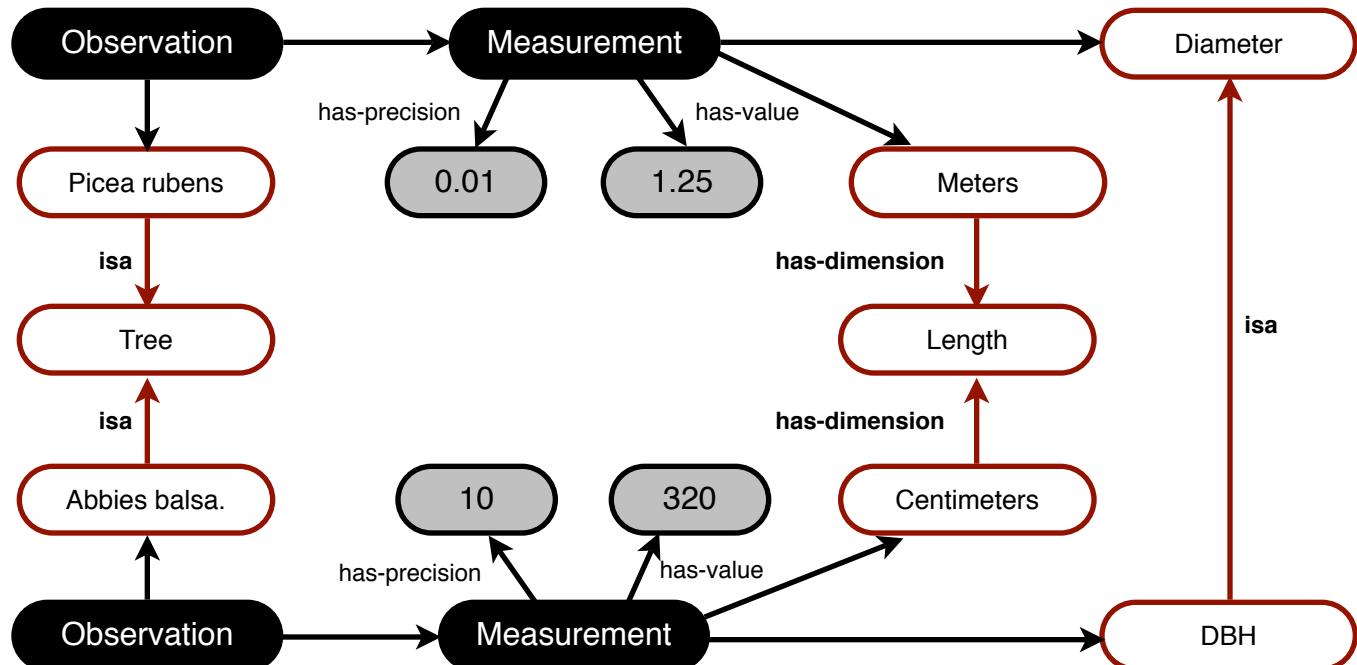


OBOE: Aligning Observations



- Observations can be aligned for data integration ...

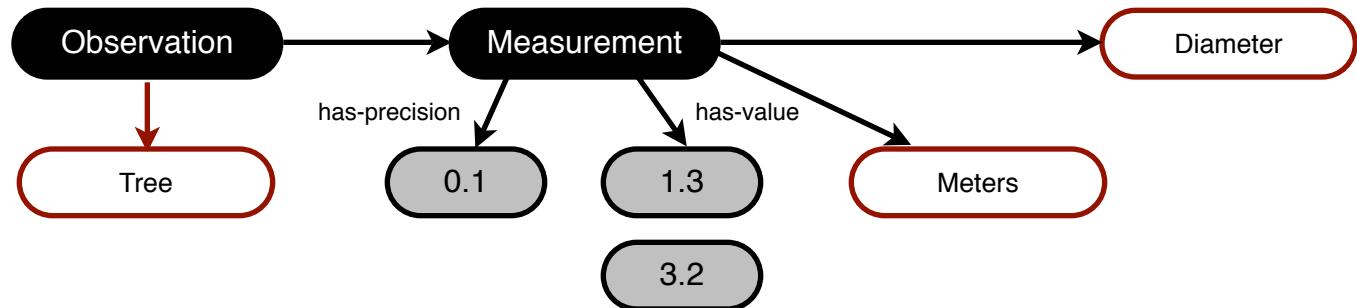
Align entities,
characteristics,
and standards



Data Integration with OBOE



- Observations can be aligned for data integration ...



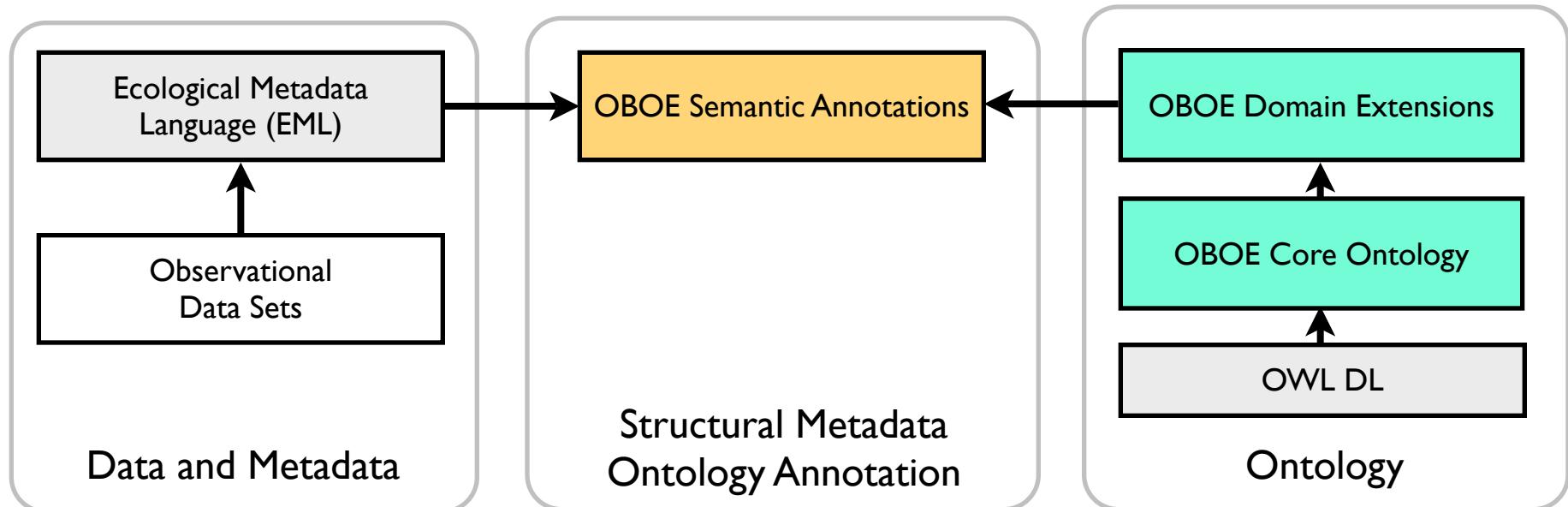
Apply conversions based on alignments, e.g.

- 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

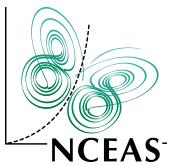
<EML>





Types of Implemented Searches

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



Structured Search

SEEK Semantic Mediation Tools

berkley Logged In [[logout](#)]

Search Upload Browse

Search Type: Structured search over annotations

Observation Entity

Ontology: All Ontologies

Class: Plant [Sort]

Measurement Characteristic

Ontology: All Ontologies

Class: DryWeight [Sort]

Measurement Standard

Ontology: All Ontologies

Class: GramsPerSquareMeter [Sort]

Search Terms: AND('http://linus.nceas.ucsb.edu/sms/metacat/oboegce.3#Plant'
'http://linus.nceas.ucsb.edu/sms/metacat/oboegce.3#DryWeight'
'http://linus.nceas.ucsb.edu/sms/metacat/oboegce.3#GramsPerSquareMeter')

[More Query Information](#)

Document ID	Title
eml.1.1	XML Plant allometry at GCE sampling sites 1–10 in October, 2002

[← previous](#) [next →](#)



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 Schildhauer¹, Shawn Bowers², Corina Gries³,
Deborah McGuinness⁴, Philip Dibner⁵, Josh Madin⁶,
Matt Jones¹, Luis Bermudez⁷, John Graybeal⁷

¹NCEAS UC Santa Barbara, ²UC Davis Genome Center

³CAP/LTER and Univ. of Arizona, ⁴McGuinness Associates,

⁵OGC Interoperability Institute, ⁶Macquarie University,

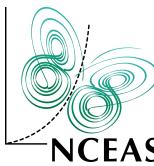
⁷Monterey Bay Aquarium Research Institute



Objectives of SONet

Broad Objectives

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

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.