Using semantic technologies to help manage scientific data

Christopher Jones, Mark Schildhauer, Matthew Jones,
Margaret O'Brien, Ben Leinfelder
National Center for Ecological Analysis and Synthesis (NCEAS)
University of California, Santa Barbara

Shawn Bowers
University of California, Davis

Josh Madin
MacQuarie University

Mara Zimmerman
Washington Department of Fish and Wildlife

February 21, 2012

Today's talk

- Scope of work & goals
- Problems in integrating data
- Federated databases & semantic approaches
- Pilot juvenile salmon ontology
- Demonstrated use
- Future possibilities

Scope of Work

- Early discussions on data heterogeneity
  - many data formats, many different types of measurements: huge heterogeneity
    - enable synthetic, integrative salmon research
- Envisioned longer-term initiative
- Scaled to pilot level juvenile salmon ontology
- NCEAS continued with other semantic efforts

Goals

- Tackle larger issues related to scientific data integration
- Federating 1000's of diverse databases
- Provide software tools for individual scientists

What's getting in the way?

- Data syntax differences
- Data structure differences
- Data semantic differences

	A	B	C	D	I	J	K	L	M	N	O	P
1												
2			**Tucannon Smolt**									
3			Pulled									
4			Debris: L=1, M=2, H=3		**Wild Chinook**				**Hatchery Chinook**			
5	start	start	end	end	Wild				Hatchery			
6	date	time	Date	time	Spr.	Mort.	Fall	Mort	Blue VIE	Mort.	Purple VIE	Mort.
7	10/8/07	1015	10/9/07	830	0	0	0	0	0	0	0	0
8	10/9/07	831	10/10/07	830	1	0	0	0	0	0	0	0
9	10/10/07	831	10/11/07	915	0	0	0	0	0	0	0	0
10	10/11/07	916	10/12/07	930	0	0	0	0	0	0	0	0
11	10/12/07	931	10/15/07	1000	2	0	0	0	0	0	0	0
12	10/15/07	1001	10/16/07	1200	0	0	0	0	0	0	0	0
13	10/16/07	1201	10/17/07	930	1	0	0	0	0	0	0	0
14	10/17/07	931	10/18/07	1015	6	0	0	0	0	0	0	0
15	10/18/07	1016	10/19/07	1130	5	0	0	0	0	0	0	0
16	10/19/07	1131	10/20/07	1245	3	1	0	0	0	0	0	0
17	10/20/07	1246	10/21/07	1230	2	0	0	0	0	0	0	0
18	10/21/07	1231	10/22/07	1100	9	0	0	0	0	0	0	0
19	10/22/07	1101	10/23/07	845	1	0	0	0	0	0	0	0
20	10/23/07	846	10/24/07	845	0	0	0	0	0	0	0	0
21	10/24/07	846	10/25/07	915	0	0	0	0	0	0	0	0
22	10/25/07	916	10/26/07	1000	2	0	0	0	0	0	0	0
23	10/26/07	1001	10/28/07	1030	12	1	0	0	0	0	0	0
24	10/28/07	1031	10/29/07	1130	20	0	0	0	0	0	0	0
25	10/29/07	1131	10/30/07	900	5	0	0	0	0	0	0	0
26	10/30/07	901	10/31/07	830	6	0	0	0	0	0	0	0
27	10/31/07	831	11/1/07	1300	14	0	0	0	0	0	0	0
28	11/1/07	1301	11/2/07	900	19	0	0	0	0	0	0	0
29	11/2/07	901	11/4/07	1200	13	9	0	0	0	0	0	0

	A	B	D	E	F	G	L	M	N	O	P
1	**Site**	**Date**	**Start**	**Stop**	**Status**	**Trap**	**YCW**	**YCH-CWTO**	**YCH-ADCWT**	**YCH-HFN**	**SCH-CWTO**
2	Methow	5/15/04	2130	530	R	5+	1	1	0	0	
3	Methow	5/16/04			P		0	0	0	0	
4	Methow	5/17/04	2000	500	R	8+	4	4	6	0	
5	Methow	5/17/04	2000	500	R	5+	0	0	1	0	
6	Methow	5/18/04	2100	500	R	8+	2	5	0	0	
7	Methow	5/18/04	2100	500	R	5+	2	0	0	0	
8	Methow	5/19/04	2045	1045	R	8	2	1	0	0	
9	Methow	5/20/04	2045	530	R	8	1	0	0	0	
10	Methow	5/21/04	2130	500	R	8	0	0	0	0	
11	Methow	5/22/04	2045	515	R	8	0	0	0	0	
12	Methow	5/23/04	2050	530	R	8	0	0	0	0	
13	Methow	5/24/04	2100	504	R	8	0	0	0	0	
14	Methow	5/25/04	2105	530	R	8	0	0	0	0	
15	Methow	5/26/04	2250	530	R	8	5	10	3	0	
16	Methow	5/27/04	2145	515	R	8	0	0	0	0	
17	Methow	5/28/04	2110	440	R	8	3	0	0	0	
18	Methow	5/29/04	2130	515	R	8	0	2	0	0	
19	Methow	5/30/04	2115	515	R	8	3	0	0	0	
20	Methow	5/31/04	2110	600	R	8	3	0	0	0	
21	Methow	6/1/04	2100	500	R	8	0	0	0	0	
22	Methow	6/2/04	2145	530	R	8	2	2	0	0	
23	Methow	6/3/04	2115	445	R	8	1	1	0	0	
24	Methow	6/4/04	2115	515	R	8	0	0	0	0	
25	Methow	6/5/04	2145	540	R	8	0	0	0	0	
26	Methow	6/6/04	2110	515	R	8	1	0	0	0	
27	Methow	6/7/04	2115	505	R	8	1	0	0	0	
28	Methow	6/8/04	2245	500	R	8	0	1	1	0	
29	Methow	6/9/04	2330	755	R	8	3	2	0	0	
30	Methow	6/10/04	2245	540	R	8	0	0	0	0	
31	Methow	6/11/04	2030	1015	R	8	2	1	0	0	

What's getting in the way?

- Semantic differences across datasets
  - What is being observed (exactly which "entity"?)
  - What is being measured (which "characteristic"?)
  - How it is being measured ("standard" & "protocol"?)
  - what is the "context"?
  - What are the relationships among "measurements"?
- How are the "observations" comparable?

Partial Solutions

– 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

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, 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. The mixtures also occurred in different geographic locations (i.e. Dean Creek on southern Sapelo Island, Marsh Landing on southwestern Sapelo Island, and Shell Hammock, which is near the University of Georgia Marine Institute). In May 1996, 16 1mx1m plots were placed within each plant mixture and alternate plots were assigned to control and fertilization treatments. Pelletized fertilizer (29% N, 3% P, 4% K) was broadcast into fertilization treatment plots by hand at the rate of 60g/m^2 every 2 weeks. 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 mass, and

Partial Solutions

– Structural Metadata

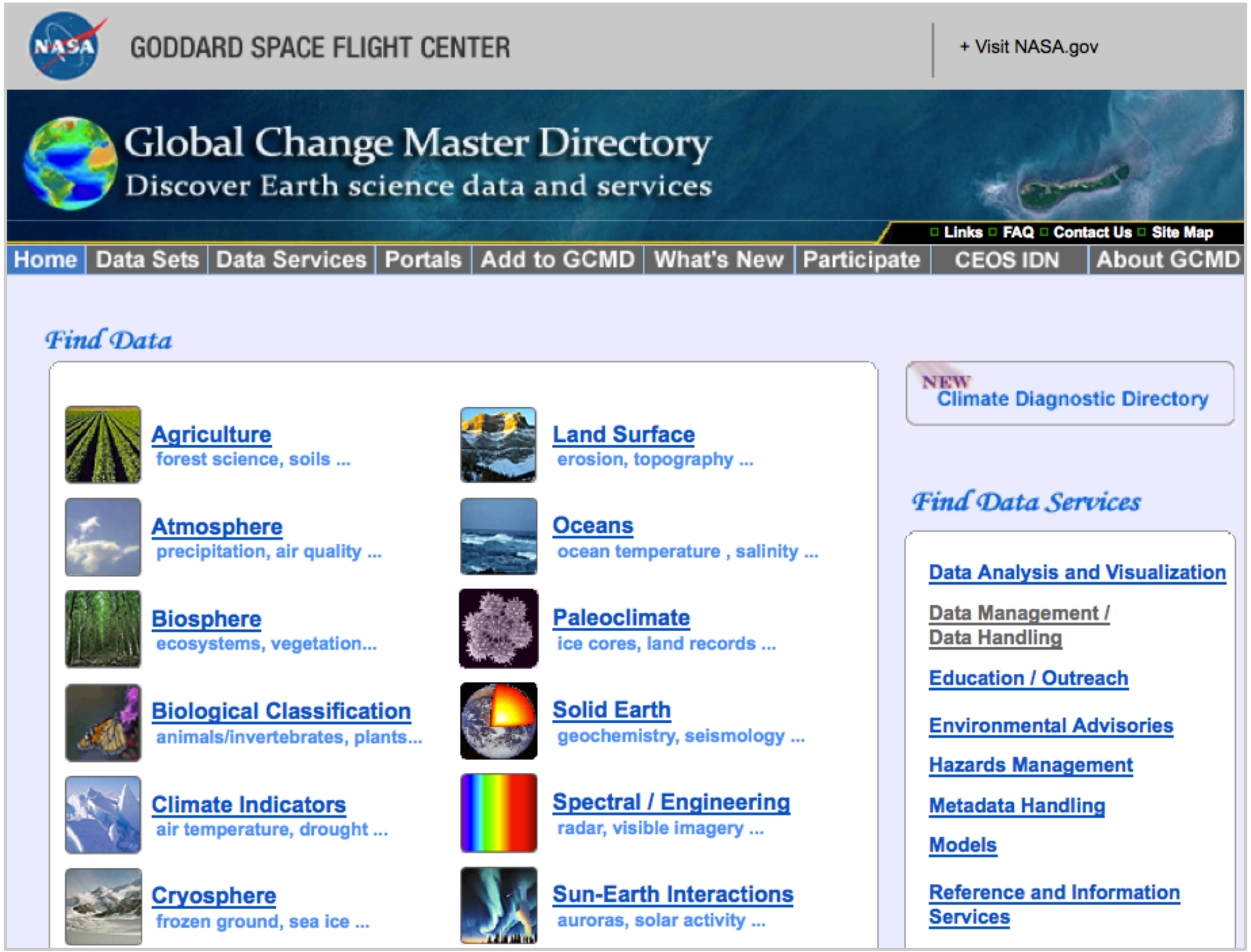
- Number of columns
- Name (informal "meaning") of columns
- Allowable values (e.g., dates, times, integers for catch data)

	A	B	C	D	I	J	K	L	M	N	O	P
1												
2			**Tucannon Smolt**									
3			Pulled									
4			Debris: L=1, M=2, H=3		**Wild Chinook**				**Hatchery Chinook**			
5	start	start	end	end	Wild				Hatchery			
6	date	time	Date	time	Spr.	Mort.	Fall	Mort	Blue VIE	Mort.	Purple VIE	Mort.
7	10/8/07	1015	10/9/07	830	0	0	0	0	0	0	0	0
8	10/9/07	831	10/10/07	830	1	0	0	0	0	0	0	0
9	10/10/07	831	10/11/07	915	0	0	0	0	0	0	0	0
10	10/11/07	916	10/12/07	930	0	0	0	0	0	0	0	0
11	10/12/07	931	10/15/07	1000	2	0	0	0	0	0	0	0
12	10/15/07	1001	10/16/07	1200	0	0	0	0	0	0	0	0
13	10/16/07	1201	10/17/07	930	1	0	0	0	0	0	0	0
14	10/17/07	931	10/18/07	1015	6	0	0	0	0	0	0	0
15	10/18/07	1016	10/19/07	1130	5	0	0	0	0	0	0	0
16	10/19/07	1131	10/20/07	1245	3	1	0	0	0	0	0	0
17	10/20/07	1246	10/21/07	1230	2	0	0	0	0	0	0	0
18	10/21/07	1231	10/22/07	1100	9	0	0	0	0	0	0	0

Partial Solutions

- Controlled Vocabularies
  - Carefully selected list of words and phrases
  - Agreed upon within a community
  - Used as “look up tables” in databases
  - Reduce ambiguity, increase search precision
  - Example: GCMD keyword thesaurus

GCMD keyword thesaurus



Federated data schema approach

- Standardize the data schema
  - Critical to collaboration
  - “Early-binding approach”
  - Recode each group’s data to a common format
  - Juvenile Migration Exchange (JMX) Database
  - NCEAS working group

Semantic approach

- Use a common “scientific observations” model
- Annotate existing datasets
- “Late-binding approach”
- Ingest, align and convert data on-the-fly

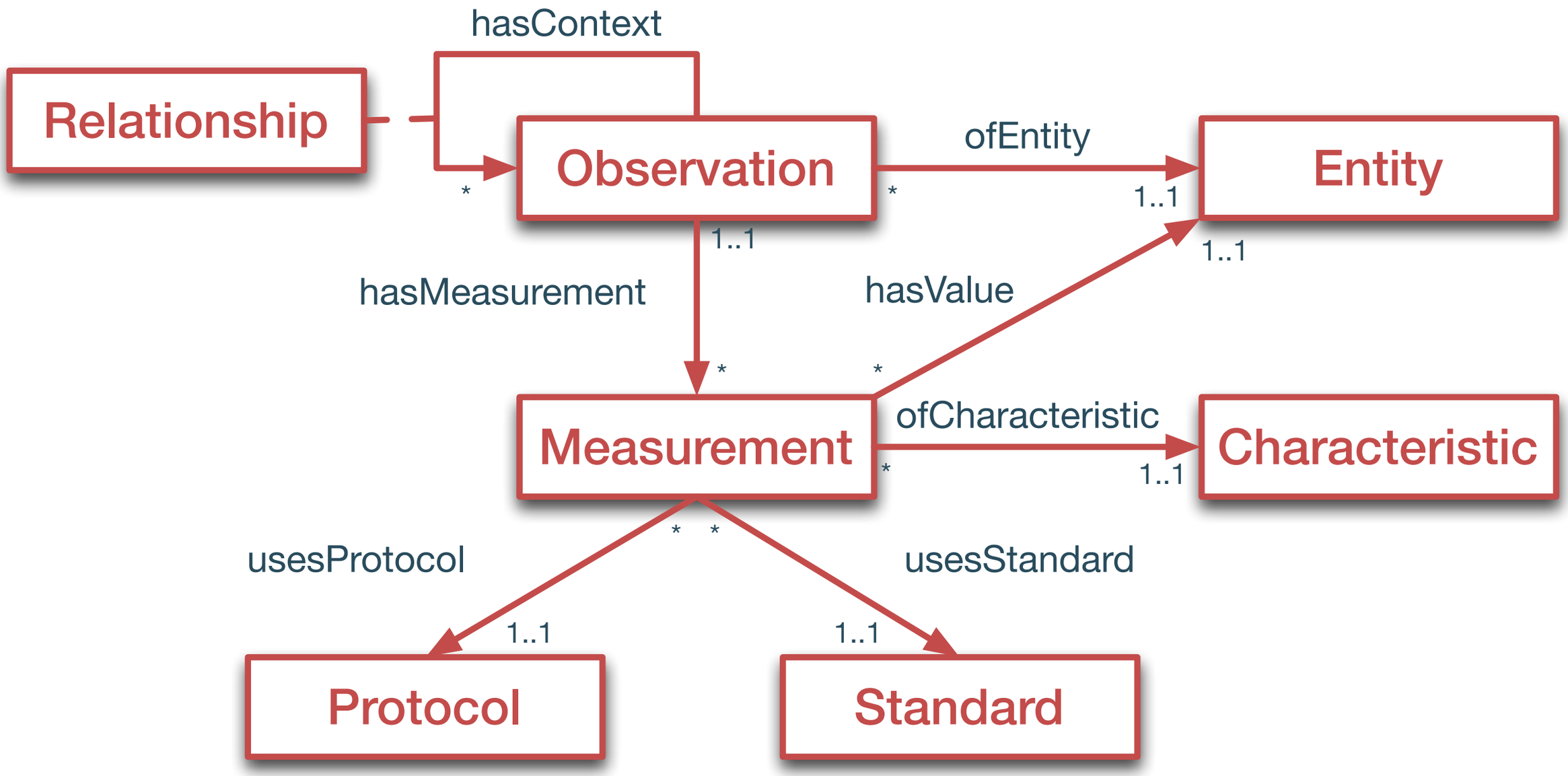
Ontologies

- A model of formally representing knowledge
- A "specification of a shared conceptualization"
- Rigorously defined concepts and their relationships
- Allows programmatic inference and reasoning
- Example:
  - Catch == PopulationSample
    - SteelheadCatch == SteelheadPopulationSample
    - SteelheadPopulationSample != CohoPopulationSample

OBOE Scientific Observations

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

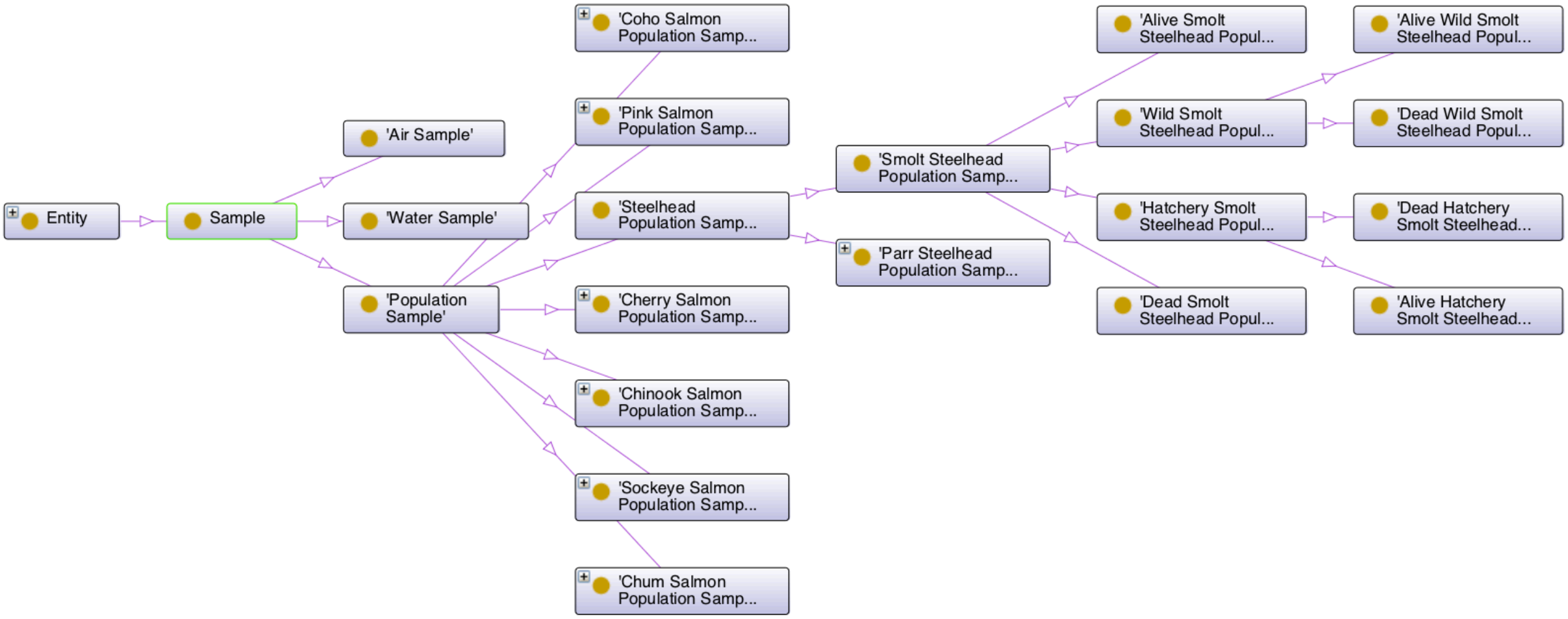
Core OBOE model



Salmon Ontology

- Thing
  - Characteristic
  - 'Characteristic Qualifier'
  - Entity
    - 'Administrative Feature'
    - 'Anatomical Entity'
    - Container
    - 'Ecological Community'
    - 'Ecological Habitat'
    - Instrument
    - Organism
      - Adult
      - Algae
      - 'Dead Organism'
      - Egg
      - Fish
        - 'Adult Stage Fish'
        - Alevin ≡ 'Alevin Stage Fish'
        - 'Alevin Stage Fish' ≡ Alevin
        - 'Egg Stage Fish'
        - Parr ≡ 'Parr Stage Fish'
        - 'Parr Stage Fish' ≡ Parr
        - Salmonid ≡ Salmonidae
        - Salmonidae ≡ Salmonid
          - Onchorhynchus
            - 'Onchorhynchus kisutch' ≡ 'Coho Salmon'
            - 'Oncorhynchus aguabonita' ≡ 'Golden Trout'
            - 'Oncorhynchus clarkii clarkii' ≡ 'Cutthroat Trout'
            - 'Oncorhynchus gorbuscha' ≡ 'Pink Salmon'
            - 'Oncorhynchus keta' ≡ 'Chum Salmon'
            - 'Oncorhynchus masou' ≡ 'Cherry Salmon'
            - 'Oncorhynchus mykiss' ≡ Steelhead ≡ 'Rainbow Trout'
            - 'Oncorhynchus nerka' ≡ 'Sockeye Salmon'
            - 'Oncorhynchus tshawytscha' ≡ 'King Salmon' ≡ 'Chinook Salmon'

Salmon Ontology

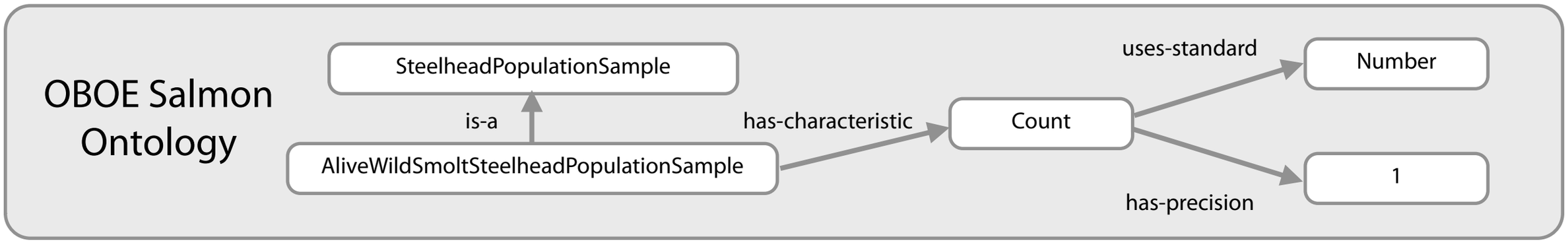


Salmon Ontology

- Full ontology available at:

  http://ecoinformatics.org/oboe-ext/salmon.1.0/oboe-salmon.owl

Linking data to the ontology



Structural Metadata

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

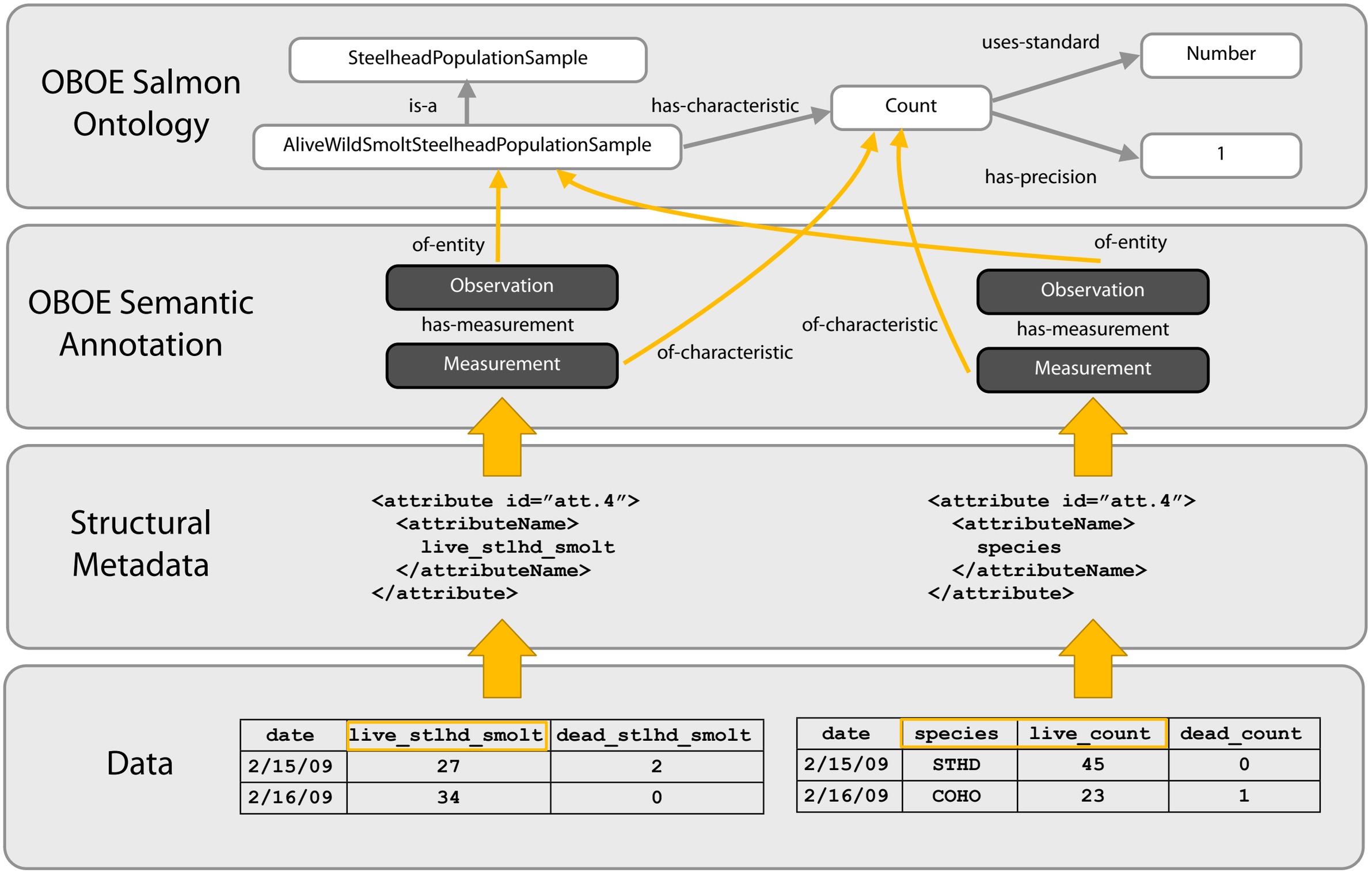
```
<attribute id="att.4">
  <attributeName>
    species
  </attributeName>
</attribute>
```

Data

date	live_stlhd_smolt	dead_stlhd_smolt
2/15/09	27	2
2/16/09	34	0

date	species	live_count	dead_count
2/15/09	STHD	45	0
2/16/09	COHO	23	1

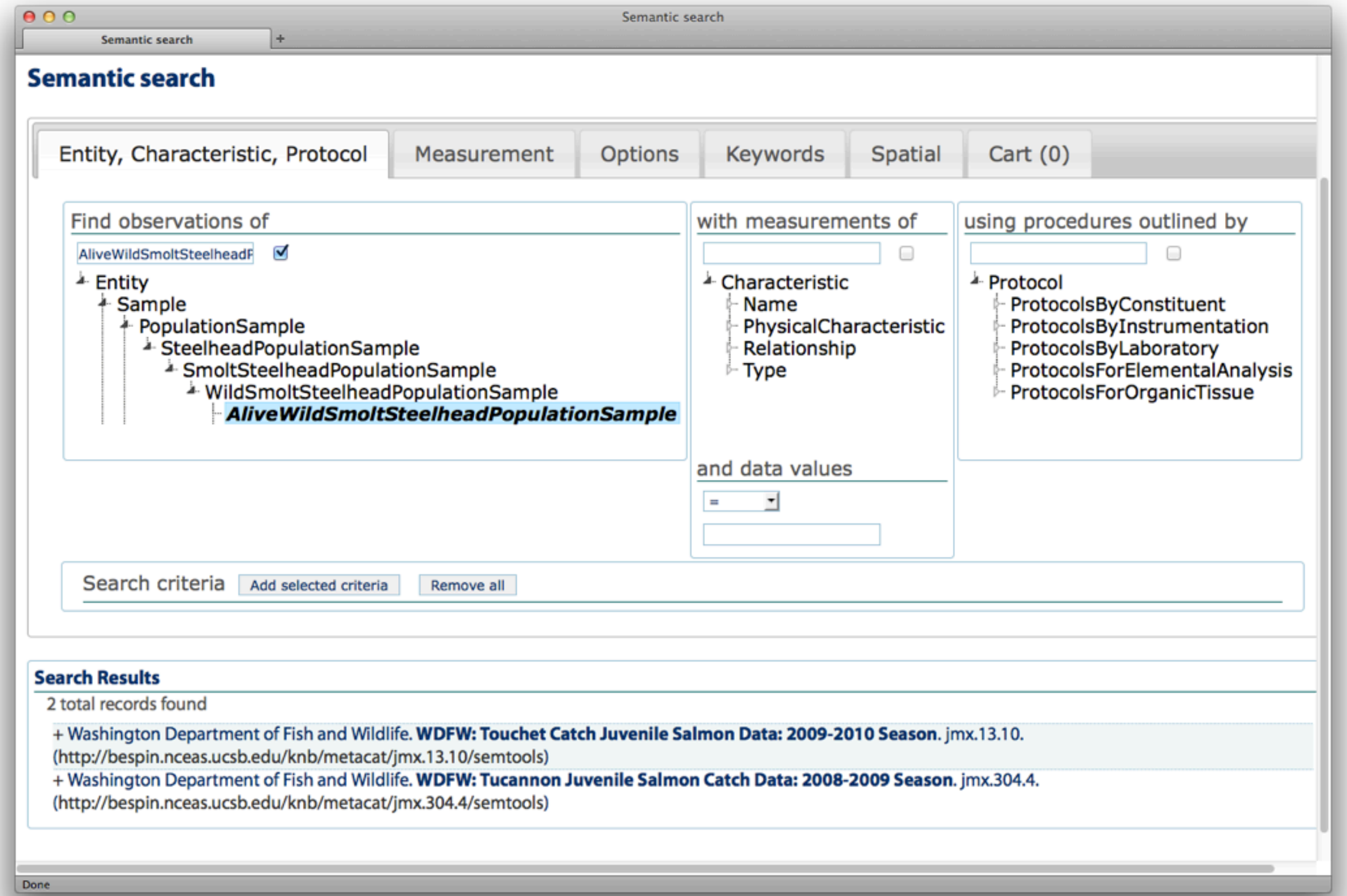
Linking data to the ontology



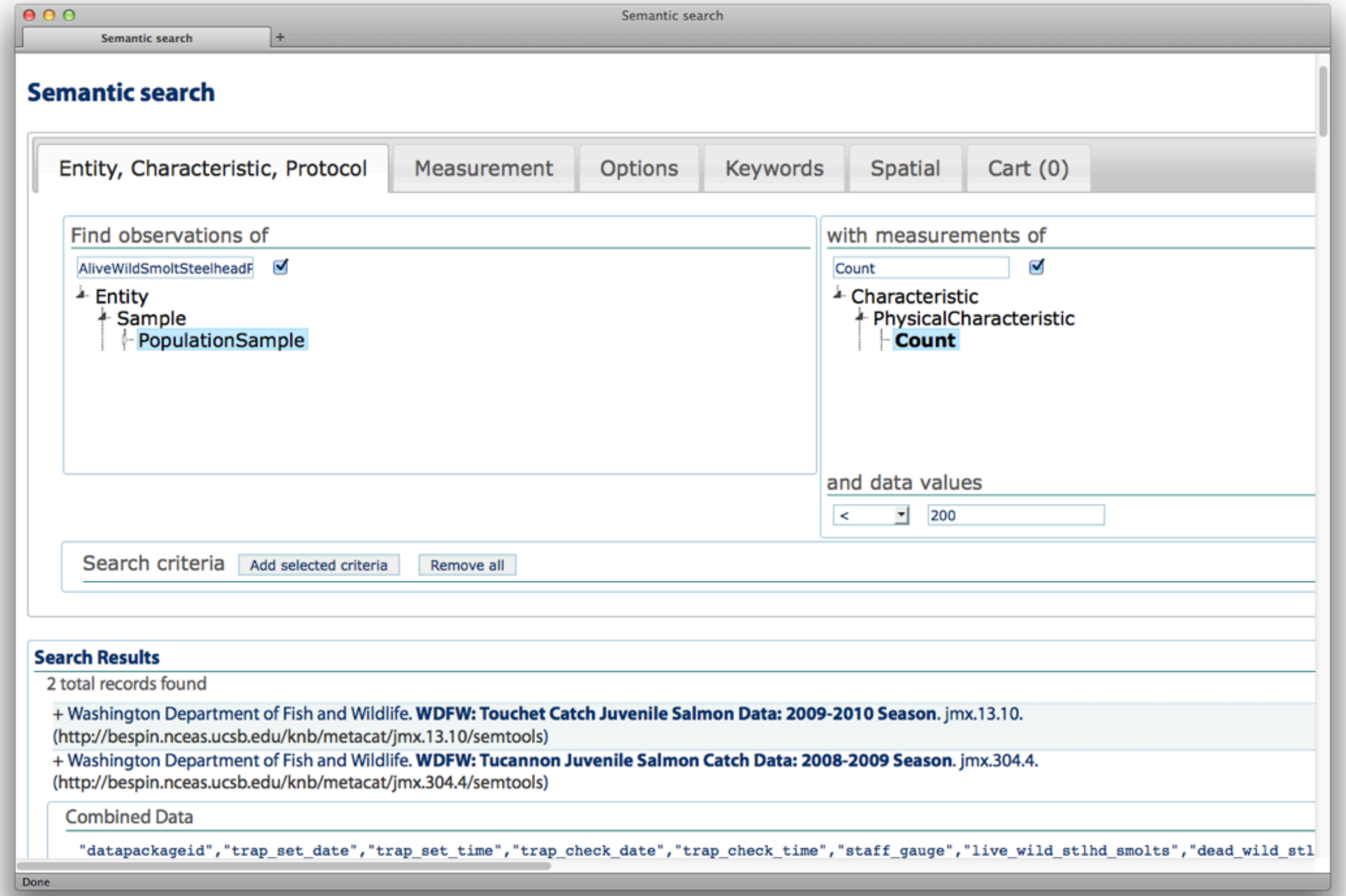
date	live_stlhd_smolt	dead_stlhd_smolt
2/15/09	27	2
2/16/09	34	0

date	species	live_count	dead_count
2/15/09	STHD	45	0
2/16/09	COHO	23	1

Search Example: “steelhead”



Search Example: +Characteristic



Combined Search Results

```
"datapackageid","trap_set_date","trap_set_time","trap_check_date","trap_check_time","staff_gauge","live_wild_stlhd_smolts","dead_wild_stlh
"jmx.304.4","16-Oct-2009 00:00:00","06-Jul-0002 12:00:00","18-Oct-2009 00:00:00","01-Jan-0001 14:00:00","20.0","0.0","0.0","0.0","0.0","0.
"jmx.304.4","18-Oct-2009 00:00:00","01-Jan-0001 14:00:00","19-Oct-2009 00:00:00","01-Jan-0001 15:00:00","20.0","0.0","0.0","0.0","0.0","0.
"jmx.304.4","19-Oct-2009 00:00:00","01-Jan-0001 15:00:00","20-Oct-2009 00:00:00","03-May-0001 23:00:00","20.0","0.0","0.0","0.0","0.0","0.
"jmx.304.4","20-Oct-2009 00:00:00","03-Jun-0001 23:00:00","21-Oct-2009 00:00:00","03-Jan-0001 22:00:00","20.0","31.0","0.0","0.0","0.0","0
"jmx.304.4","21-Oct-2009 00:00:00","03-Jan-0001 22:00:00","22-Oct-2009 00:00:00","03-Mar-0001 13:00:00","20.0","32.0","0.0","0.0","0.0","0
"jmx.304.4","22-Oct-2009 00:00:00","03-Apr-0001 13:00:00","23-Oct-2009 00:00:00","01-Jan-0001 13:00:00","20.0","5.0","0.0","0.0","0.0","0.
"jmx.304.4","23-Oct-2009 00:00:00","01-Jan-0001 13:00:00","26-Oct-2009 00:00:00","03-Mar-0001 12:00:00","20.0","0.0","0.0","0.0","0.0","0.
"jmx.304.4","26-Oct-2009 00:00:00","03-Apr-0001 12:00:00","27-Oct-2009 00:00:00","04-Jan-0001 01:00:00","23.0","0.0","0.0","0.0","0.0","0.
"jmx.304.4","27-Oct-2009 00:00:00","04-Jan-0001 01:00:00","28-Oct-2009 00:00:00","04-Jan-0001 01:00:00","","43.0","0.0","0.0","1.0","0.0",
"jmx.304.4","28-Oct-2009 00:00:00","04-Jan-0001 01:00:00","29-Oct-2009 00:00:00","01-Jan-0001 14:00:00","21.0","16.0","0.0","0.0","0.0","0
"jmx.304.4","29-Oct-2009 00:00:00","01-Jan-0001 14:00:00","30-Oct-2009 00:00:00","01-Jan-0001 10:00:00","21.0","17.0","0.0","0.0","0.0","0
"jmx.304.4","30-Oct-2009 00:00:00","01-Jan-0001 10:00:00","02-Nov-2009 00:00:00","01-Jan-0001 10:00:00","22.0","6.0","0.0","0.0","0.0","0.
"jmx.304.4","02-Nov-2009 00:00:00","01-Jan-0001 10:00:00","03-Nov-2009 00:00:00","04-Jan-0001 08:00:00","21.0","7.0","0.0","0.0","0.0","0.
"jmx.304.4","03-Nov-2009 00:00:00","04-Jan-0001 08:00:00","04-Nov-2009 00:00:00","04-Jan-0001 08:00:00","21.0","5.0","0.0","0.0","0.0","0.
"jmx.304.4","04-Nov-2009 00:00:00","04-Jan-0001 08:00:00","05-Nov-2009 00:00:00","01-Jan-0001 10:00:00","21.0","21.0","0.0","0.0","0.0","0
"jmx.304.4","05-Nov-2009 00:00:00","01-Jan-0001 10:00:00","06-Nov-2009 00:00:00","04-Jan-0001 21:00:00","22.0","7.0","0.0","0.0","0.0","0.
"jmx.304.4","06-Nov-2009 00:00:00","04-Jan-0001 21:00:00","09-Nov-2009 00:00:00","04-Jan-0001 08:00:00","22.0","5.0","0.0","0.0","0.0","0.
"jmx.304.4","09-Nov-2009 00:00:00","04-Jan-0001 08:00:00","10-Nov-2009 00:00:00","04-Jan-0001 08:00:00","22.0","2.0","0.0","0.0","0.0","0.
"jmx.13.10","23-Dec-2009 00:00:00","04-Jan-0001 11:00:00","24-Dec-2009 00:00:00","01-Jan-0001 10:00:00","25.0","16.0","0.0","0.0","0.0","0
"jmx.13.10","24-Dec-2009 00:00:00","01-Jan-0001 10:00:00","28-Dec-2009 00:00:00","01-Jan-0001 10:00:00","23.0","0.0","0.0","0.0","0.0","0.
"jmx.13.10","28-Dec-2009 00:00:00","01-Jan-0001 10:00:00","29-Dec-2009 00:00:00","04-Jan-0001 21:00:00","23.0","0.0","0.0","0.0","0.0","0.
"jmx.13.10","29-Dec-2009 00:00:00","04-Jan-0001 21:00:00","30-Dec-2009 00:00:00","01-Jan-0001 10:00:00","22.0","0.0","0.0","0.0","0.0","0.
"jmx.13.10","30-Dec-2009 00:00:00","01-Jan-0001 10:00:00","31-Dec-2009 00:00:00","03-Mar-0001 10:00:00","22.0","0.0","0.0","0.0","0.0","0.
"jmx.13.10","31-Dec-2009 00:00:00","03-Apr-0001 10:00:00","04-Jan-2010 00:00:00","01-Jan-0001 11:00:00","24.0","0.0","0.0","0.0","0.0","0.
"jmx.13.10","04-Jan-2010 00:00:00","01-Jan-0001 11:00:00","05-Jan-2010 00:00:00","01-Jan-0001 13:00:00","32.5","48.0","0.0","1.0","1.0","0
"jmx.13.10","05-Jan-2010 00:00:00","01-Jan-0001 13:00:00","06-Jan-2010 00:00:00","01-Jan-0001 11:00:00","36.0","0.0","0.0","0.0","0.0","0.
"jmx.13.10","06-Jan-2010 00:00:00","01-Jan-0001 11:00:00","07-Jan-2010 00:00:00","01-Jan-0001 10:00:00","32.0","176.0","0.0","0.0","0.0","
"jmx.13.10","07-Jan-2010 00:00:00","01-Jan-0001 10:00:00","08-Jan-2010 00:00:00","01-Jan-0001 11:00:00","31.0","100.0","0.0","0.0","0.0","
"jmx.13.10","08-Jan-2010 00:00:00","01-Jan-0001 11:00:00","11-Jan-2010 00:00:00","01-Jan-0001 11:00:00","26.0","15.0","0.0","0.0","0.0","0
"jmx.13.10","11-Jan-2010 00:00:00","01-Jan-0001 11:00:00","12-Jan-2010 00:00:00","01-Jan-0001 10:00:00","27.0","34.0","0.0","0.0","0.0","0
"jmx.13.10","12-Jan-2010 00:00:00","01-Jan-0001 10:00:00","13-Jan-2010 00:00:00","05-Jun-0002 10:00:00","27.0","0.0","0.0","0.0","0.0","0.
"jmx.13.10","13-Jan-2010 00:00:00","06-Jul-0002 10:00:00","14-Jan-2010 00:00:00","05-Jun-0002 10:00:00","30.0","5.0","0.0","0.0","0.0","0.
"jmx.13.10","14-Jan-2010 00:00:00","06-Jul-0002 10:00:00","15-Jan-2010 00:00:00","01-Jan-0001 10:00:00","29.0","0.0","0.0","0.0","0.0","0.
"jmx.13.10","15-Jan-2010 00:00:00","01-Jan-0001 10:00:00","19-Jan-2010 00:00:00","03-Mar-0001 10:00:00","27.0","0.0","0.0","0.0","0.0","0.
"jmx.13.10","19-Jan-2010 00:00:00","03-Apr-0001 10:00:00","20-Jan-2010 00:00:00","01-Jan-0001 11:00:00","26.0","0.0","0.0","0.0","0.0","0.
"jmx.13.10","20-Jan-2010 00:00:00","01-Jan-0001 11:00:00","21-Jan-2010 00:00:00","01-Jan-0001 12:00:00","25.5","0.0","0.0","0.0","0.0","0.
"jmx.13.10","21-Jan-2010 00:00:00","01-Jan-0001 12:00:00","22-Jan-2010 00:00:00","04-Jan-0001 08:00:00","26.0","0.0","0.0","0.0","0.0","0.
"jmx.13.10","22-Jan-2010 00:00:00","04-Jan-0001 08:00:00","28-Jan-2010 00:00:00","03-Jan-0001 22:00:00","24.0","1.0","0.0","0.0","0.0","0.
"jmx.13.10","28-Jan-2010 00:00:00","03-Jan-0001 22:00:00","29-Jan-2010 00:00:00","01-Jan-0001 13:00:00","25.5","0.0","0.0","0.0","0.0","0.
"jmx.13.10","29-Jan-2010 00:00:00","01-Jan-0001 13:00:00","01-Feb-2010 00:00:00","01-Jan-0001 10:00:00","26.0","0.0","0.0","0.0","0.0","0.
```

Future Possibilities

- As a compliment to the JMX database
  1. Annotate the master database structures
  2. Annotate data from agencies, universities, etc.
  3. Use the ontology query capabilities as a transformation engine to produce JMX-compliant data
- *Annotated databases can be repurposed for other synthetic studies

Take Home Points

- Ontology development and use
  - Must be a community-based effort
  - Start at the data sources and build up
  - Involves:
    - Scientists with Domain Expertise
    - Researchers in Knowledge Representation
    - Software Developers / Data Managers
- Can compliment “early-binding” standardization efforts
- Generalized data integration is a phenomenally challenging problem for synthesis applications

Acknowledgments

- This material is based upon work supported by:
- Ecotrust
- 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