Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Resource Discovery for Extreme Scale Collaboration

JESSE WEAVER, ALAN CHAPPELL, WILLIAM SMITH, SUMIT PUROHIT

Pacific Northwest National Laboratory
Richland, WA

PETER FOX, PATRICK WEST, BENNO LEE

Tetherless World Constellation, Rensselaer Polytechnic Institute
Troy, NY

September 23, 2014

IR#: PNNL-SA-105496

Data Sharing Practices circa 2011

Of 1,329 scientists surveyed…

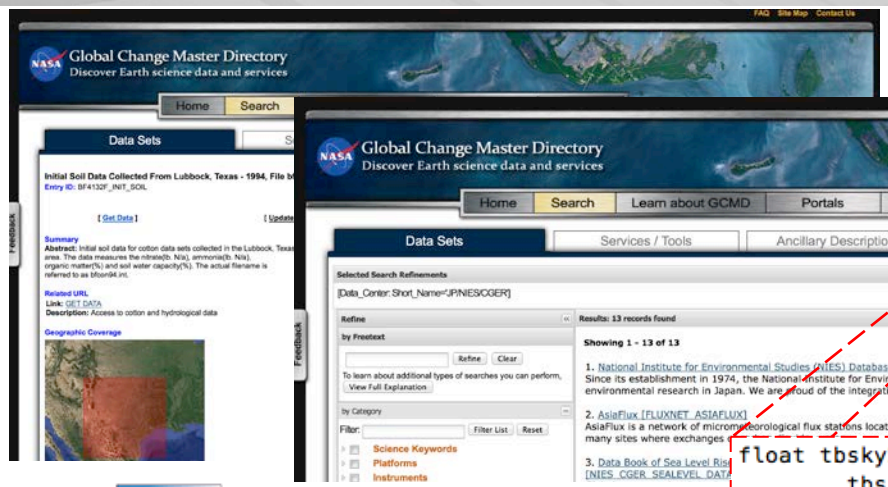
- "Nearly two thirds … agreed that lack of access to data generated by other researchers or institutions is a major impediment to progress in science." [Tenopir et al, 2011]

- "Nearly one third of the respondents chose not to answer whether they make their data available to others. Of those who did respond, 46% reported they do not make their data electronically available to others." [Tenopir et al, 2011]

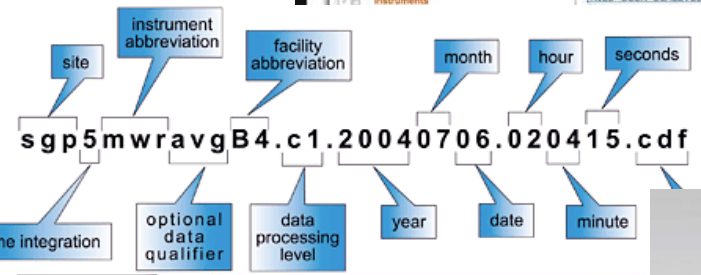
- "…, most of the respondents (85%) are interested in using other researchers' datasets, if those datasets are easily accessible." [Tenopir et al, 2011]

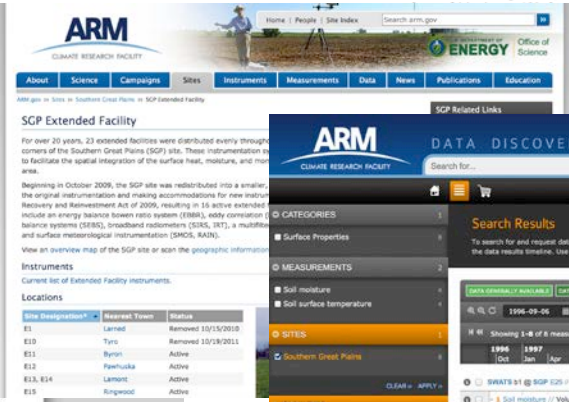
[Tenopir et al, 2011] Carol Tenopir, Suzie Allard, Kimberly L. Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. Data shring by scientists: practices and perceptions. PLoS One, 6(6), 2011.

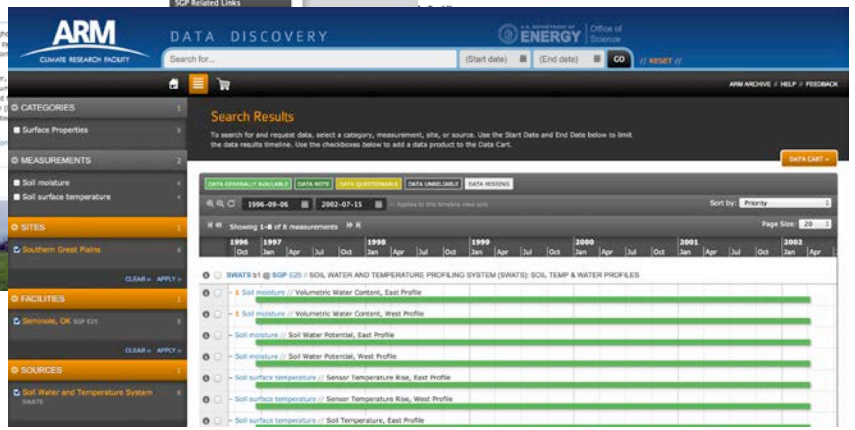
Example Metadata Sources



```
float tbsky23(time) ;
        tbsky23:long_name = "Mean 23.8 GHz sky brightness temperature" ;
        tbsky23:units = "K" ;
```





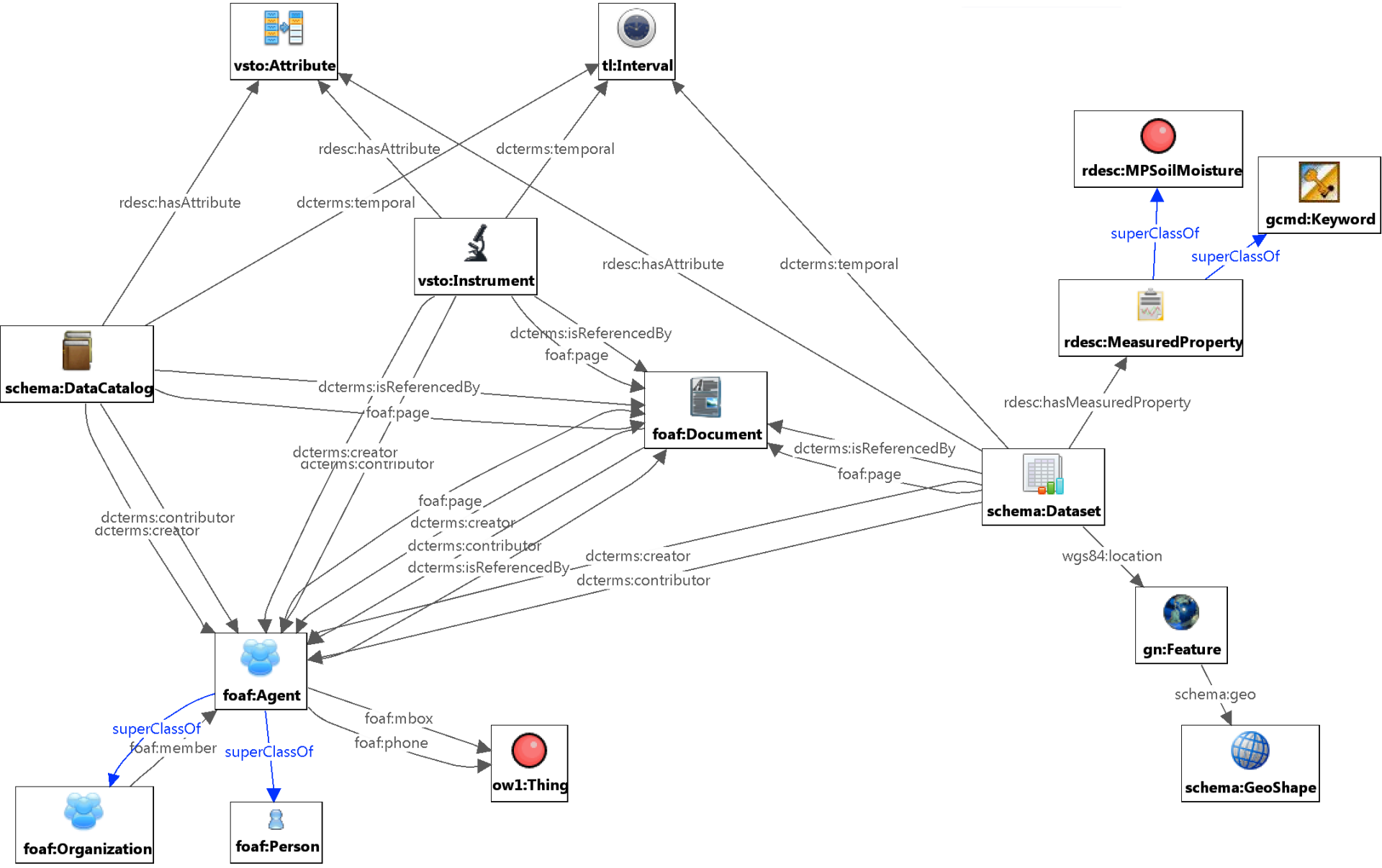


Lessons from CS Domains

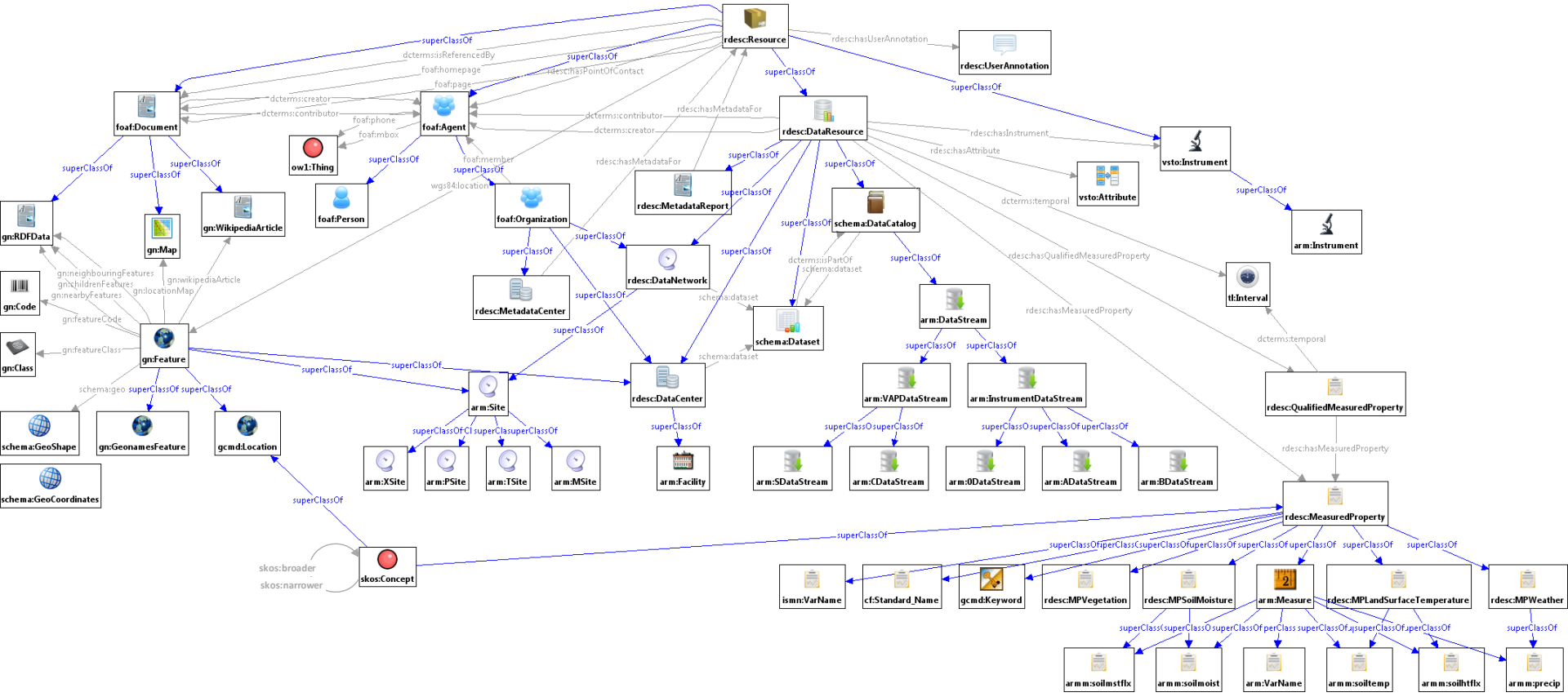
- Databases.
  - Getting data from disk is expensive (or data is “far away”).
  - So we want to minimize page fetches from disk.
  - Solution is to build indexes in memory so you fetch only relevant pages.
  - Analogously, we need to build “indexes” that are “closer” (local) than the data, so that we don’t have to wander on the Web.
  - But what should we index? Depends on the queries.

- Web search.
  - Are PageRank and similar algorithms physically executed on the Web?
  - No! You first crawl whatever (meta)data you need to rank the pages.
  - In other words, leave the data where it is.
  - Collect the *metadata* and use it to build indexes.

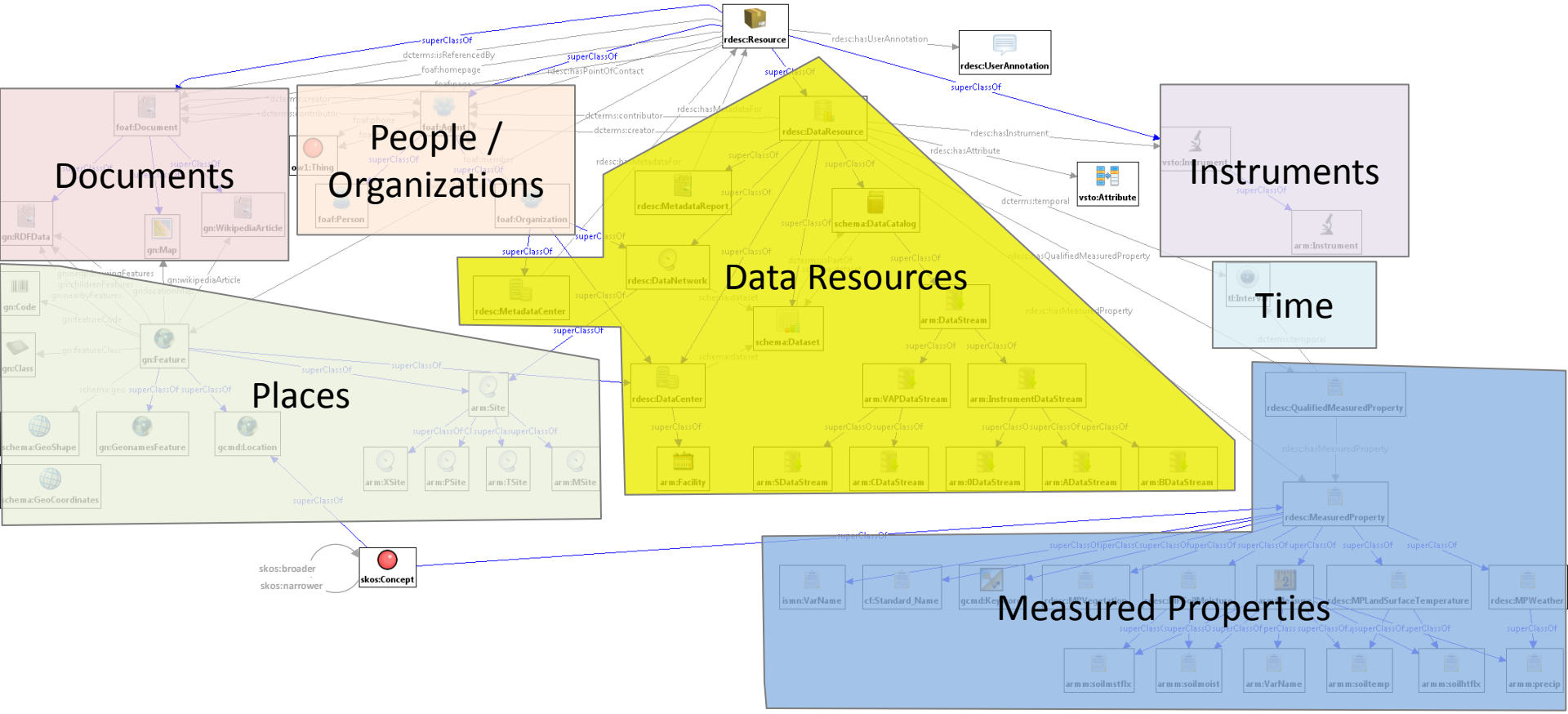
Initial Ontology



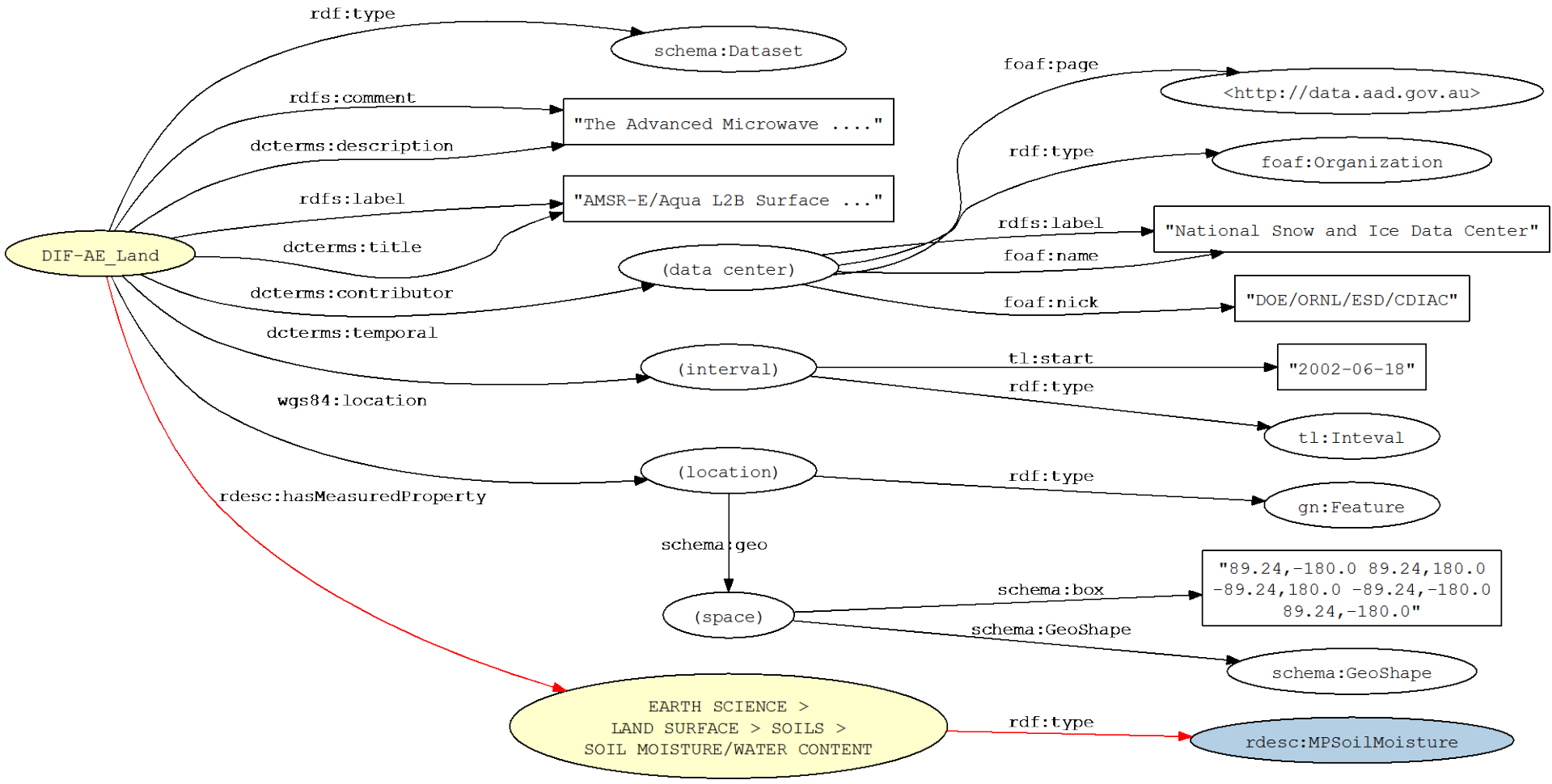
Today's Ontology



Today's Ontology

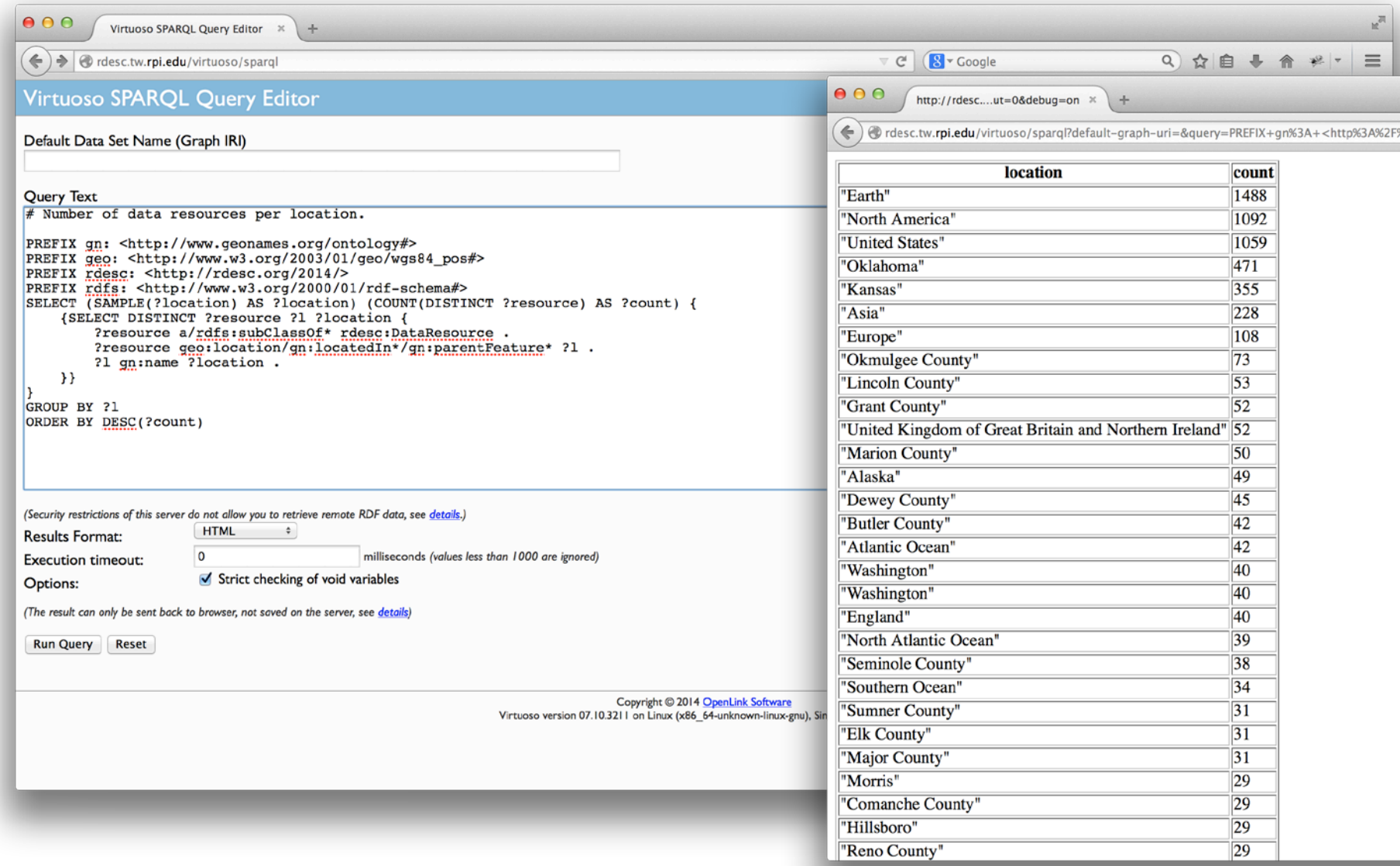


Example Metadata



Thanks to Gerald John M. Manipon of NASA for providing access to initial RDF descriptions of GCMD datasets.

Augmentation with Linked Open Data



location	count
"Earth"	1488
"North America"	1092
"United States"	1059
"Oklahoma"	471
"Kansas"	355
"Asia"	228
"Europe"	108
"Okmulgee County"	73
"Lincoln County"	53
"Grant County"	52
"United Kingdom of Great Britain and Northern Ireland"	52
"Marion County"	50
"Alaska"	49
"Dewey County"	45
"Butler County"	42
"Atlantic Ocean"	42
"Washington"	40
"Washington"	40
"England"	40
"North Atlantic Ocean"	39
"Seminole County"	38
"Southern Ocean"	34
"Sumner County"	31
"Elk County"	31
"Major County"	31
"Morris"	29
"Comanche County"	29
"Hillsboro"	29
"Reno County"	29

Analytic Queries (as opposed to lookup)

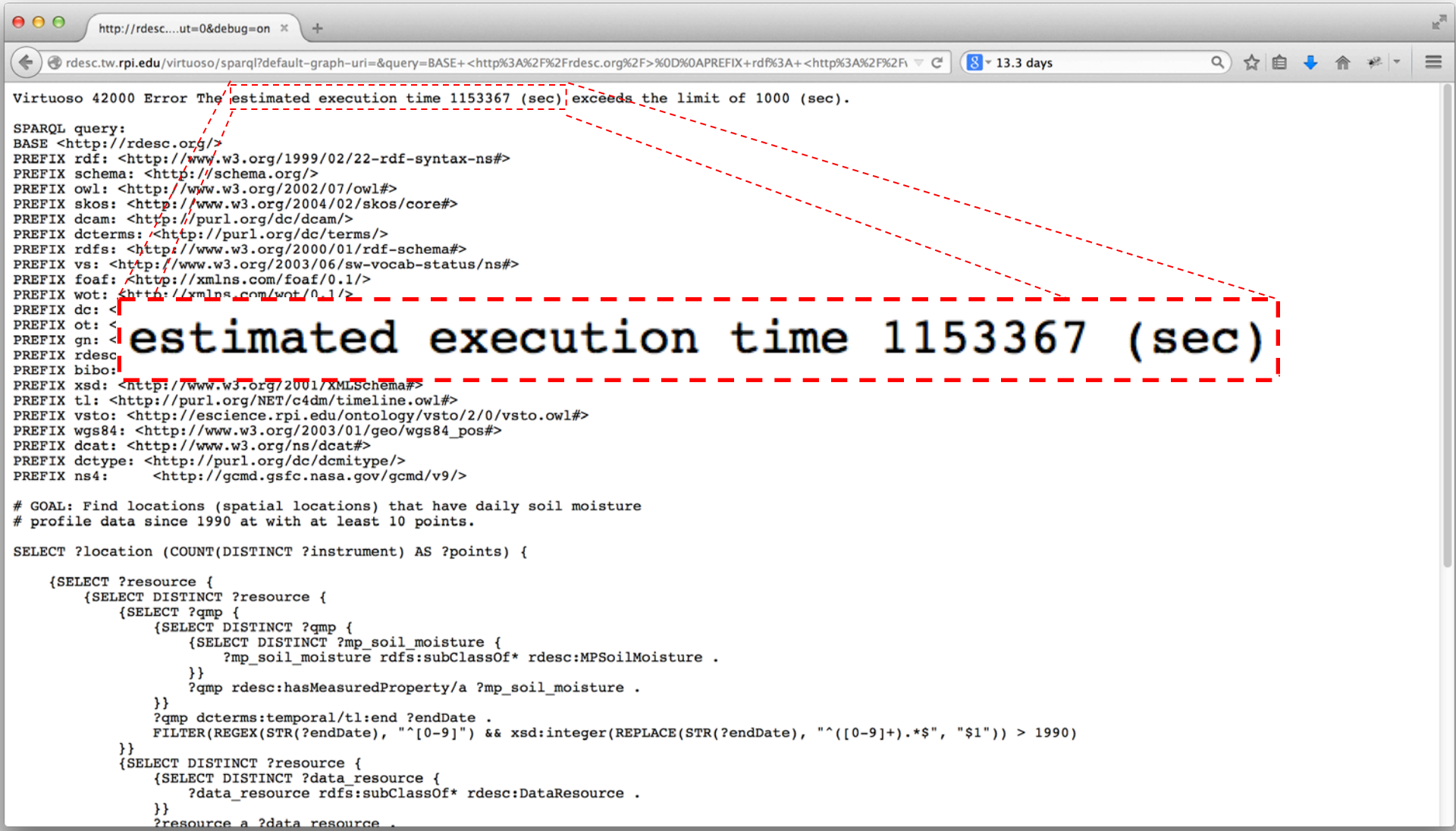
Find soil moisture profile information for the Walnut Gulch (aka Walnut Creed) watershed that provides daily data for the time period after 1990 with at least 10 points.

Find locations that have daily soil moisture profile data since 1990 with at least 10 points.

Find locations that have daily soil moisture profile and surface temperatures (at 2 meters) since 1990 with at least 10 points.

Find locations that have hourly weather data with at least 10 locations since 1990. Would then like to be able to quickly access the vegetation information around the instruments.

Analytic Query on VOS 7



Analytic Query with GEMS

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

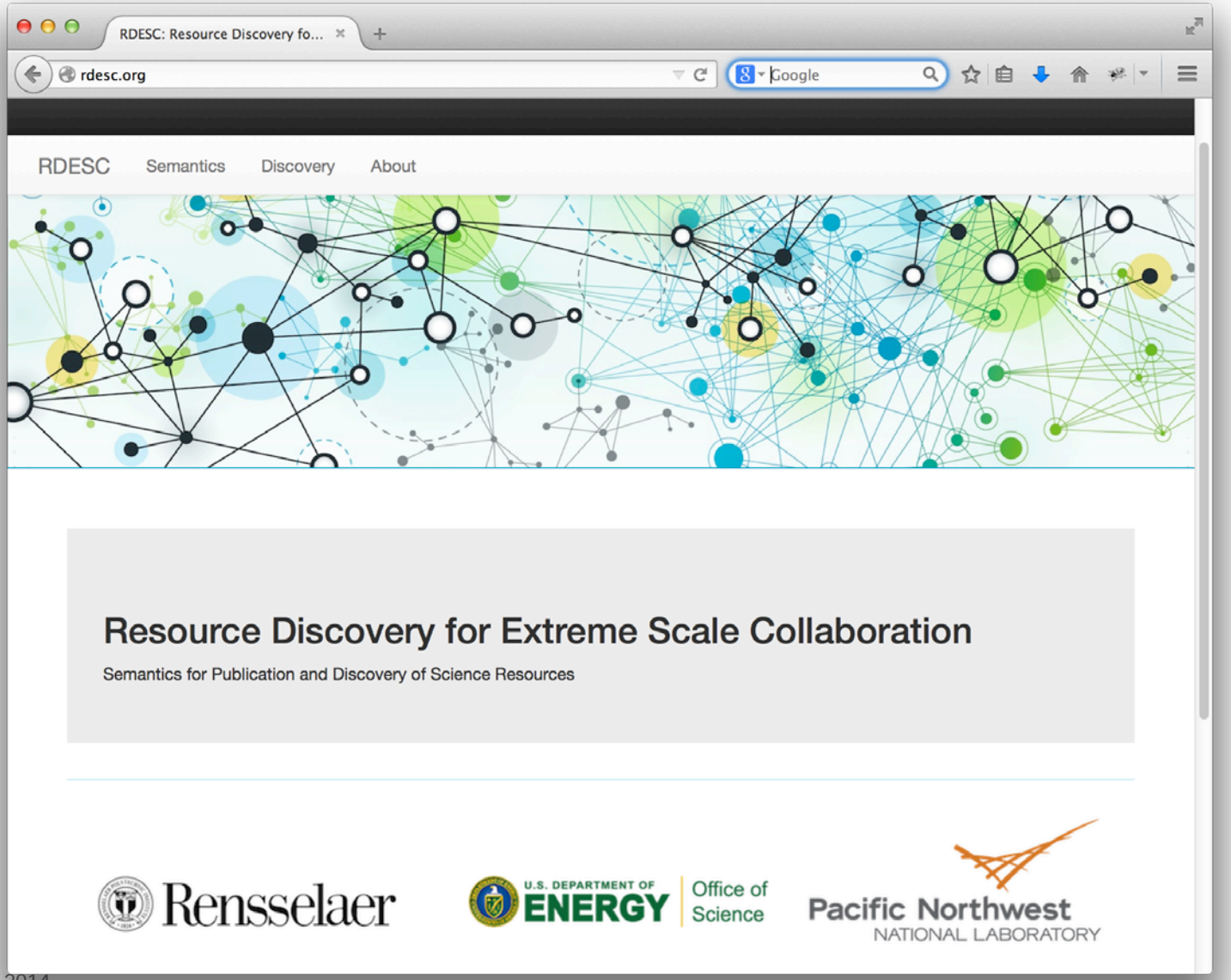
	#results	8 nodes	16 nodes	32 nodes
Ingest		1218	580	406
Query Fig.3	76	0.492	1.74	0.724
Query Fig.5	1	0.585	0.706	0.690
Query Fig.9	0	5.29	4.93	2.04
Query Fig.10	68	15.9	9.38	8.16

Table 1: Average ingest and query times (in seconds) over 10 runs, on fixed RDESC dataset of nearly 1.4B triples, varying number of nodes

[Weaver et al, 2014]

[Weaver et al, 2014] Jesse Weaver, Vito Giovanni Castellana, Alessandro Morari, Antonino Tumeo, Sumit Purohit, Alan Chappell, David Haglin, Oreste Villa, Sutanay Choudhury, Karen Schuchardt, and John Feo. Toward a Data Scalable Solution for Facilitating Discovery of Science Resources. Parallel Computing, 2014. (in press)

UI Demo


RDESC: Resource Discovery fo...
rdesc.org
Google
RDESC
Semantics
Discovery
About
Resource Discovery for Extreme Scale Collaboration
Semantics for Publication and Discovery of Science Resources
Rensselaer
U.S. DEPARTMENT OF ENERGY
Office of Science
Pacific Northwest
NATIONAL LABORATORY

What question does your research motivate you to ask now?

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- What questions would domain scientists ask?
- Certainly questions will differ between domains.
- This information will better inform us as to what kind of metadata needs to be collected and what ontologies need to be developed to support them.
- What level of reasoning will be needed to support easy analytical queries? if any at all?