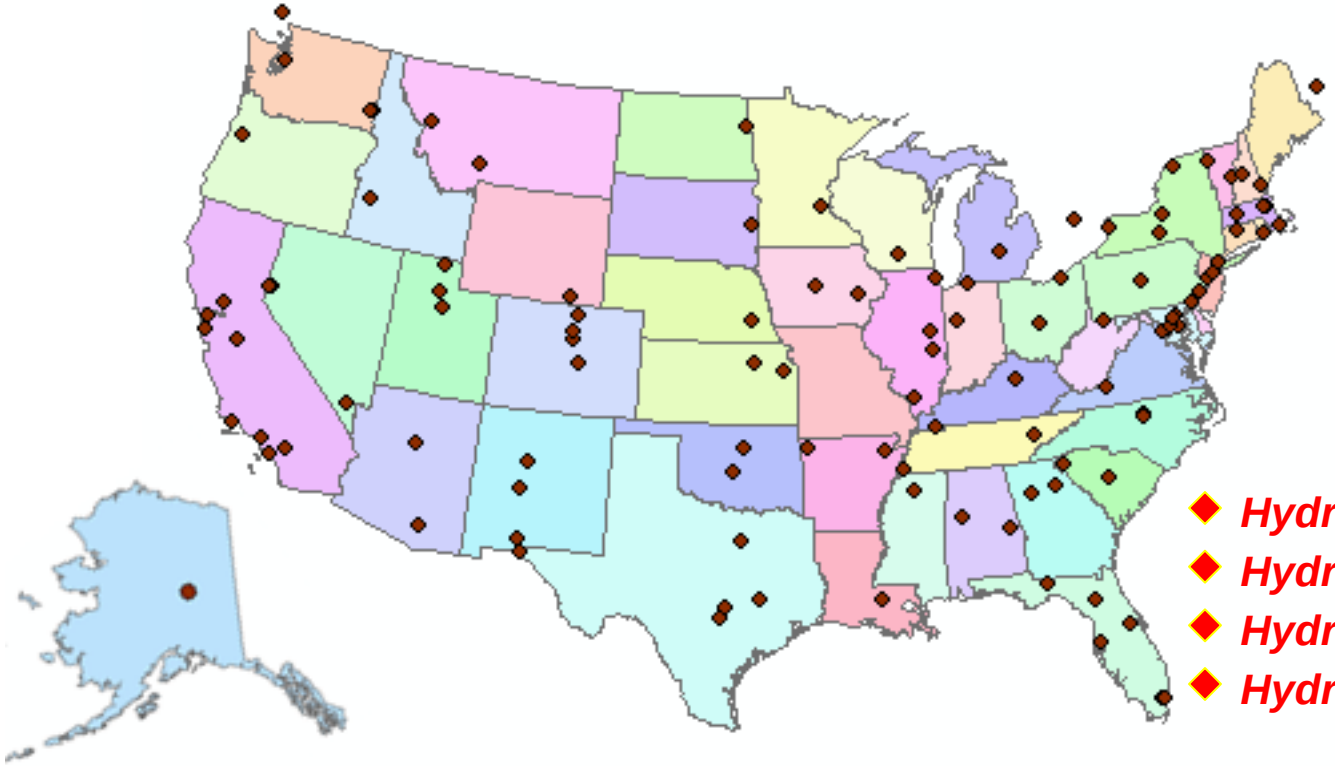
Managing observation semantics in CUAHSI Hydrologic Information System

Ilya Zaslavsky

Spatial Information Systems Lab
San Diego Supercomputer Center
UCSD

Consortium of Universities for the Advancement of Hydrologic Science, Inc.



~125 US Universities

- ***Hydrologic Information System***
- ***Hydrologic Measurement Facility***
- ***Hydrologic Modeling***
- ***Hydrologic Education Outreach***

An organization representing more than one hundred United States universities, receives support from the National Science Foundation to develop infrastructure and services for the advancement of hydrologic science and education in the U.S.

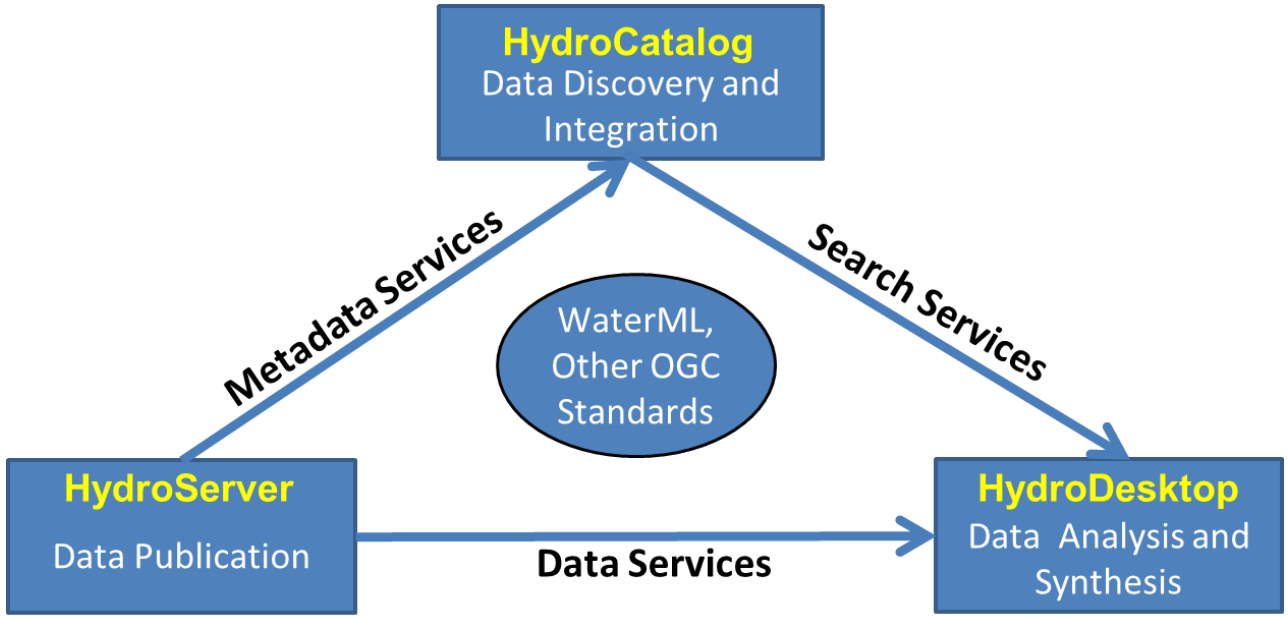
CUAHSI
universities allied for water research

http://www.cuahsi.org/

What is the CUAHSI HIS?

UT-Austin, SDSC/UCSD, Utah State U, Idaho State U, Drexel U, U of So. Carolina
PI: D. R. Maidment (UT-Austin)

http://his.cuahsi.org



CUAHSI HIS: NSF support through 2012 (GEO)

Partners:

Academic: hydrologic observatories at universities, CZO…

Government: USGS, EPA, NCDC, NWS, state and local

Commercial: Microsoft, ESRI, Kisters

Standardization: OGC, WMO (Hydrology Domain WG); adopted by USGS, NCDC, Army Corps of Eng

An online distributed system to support the sharing of hydrologic data from multiple repositories and databases via standard water data service protocols; software for data publication, discovery, access and integration.

Water Data

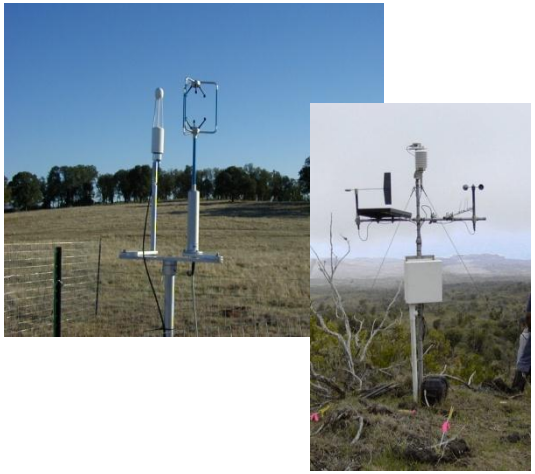
Water quantity and quality

Soil water

Rainfall & Snow

Modeling

Meteorology



Remote sensing

Sources of Observations Data

The University of Montana, Missoula

Reynolds Creek Experimental Watershed

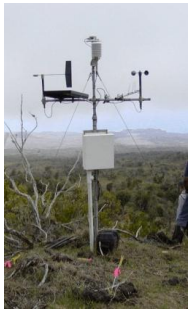
Dry Creek Experimental Watershed

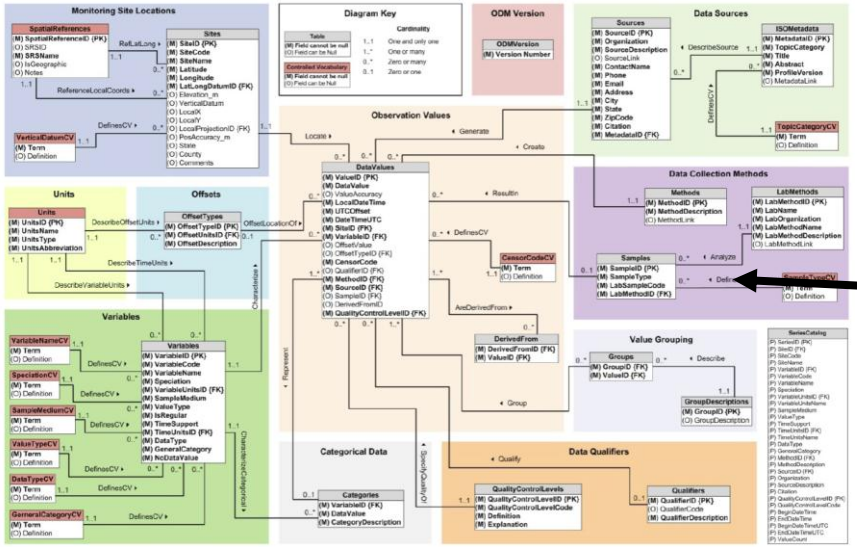
Observations Data Model (ODM)

Streamflow

Groundwater levels

Precipitation & Climate





Soil moisture data

Water Quality

Flux tower data

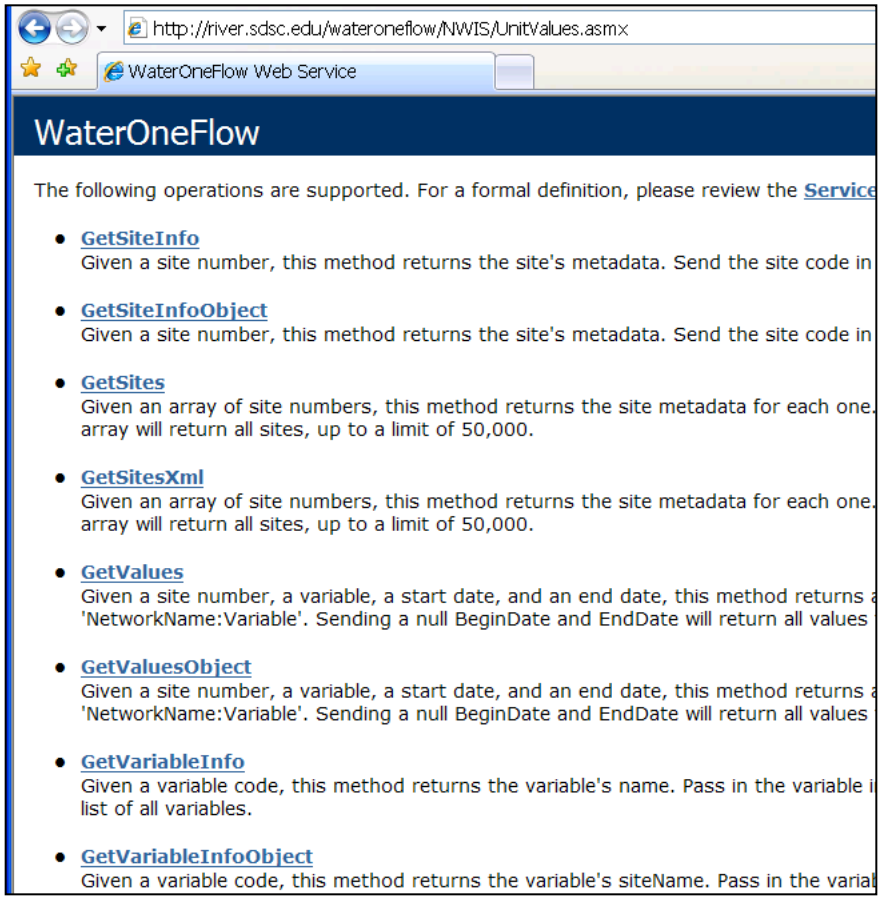
- A relational database at the single observation level
- Metadata for unambiguous interpretation
- Traceable heritage from raw measurements to usable information
- Promote syntactic and semantic consistency
- Cross dimension retrieval and analysis

Horsburgh, J. S., D. G. Tarboton, D. R. Maidment, and I. Zaslavsky (2008), A relational model for environmental and water resources data, *Water Resources Research*, 44, W05406, doi:10.1029/2007WR006392.

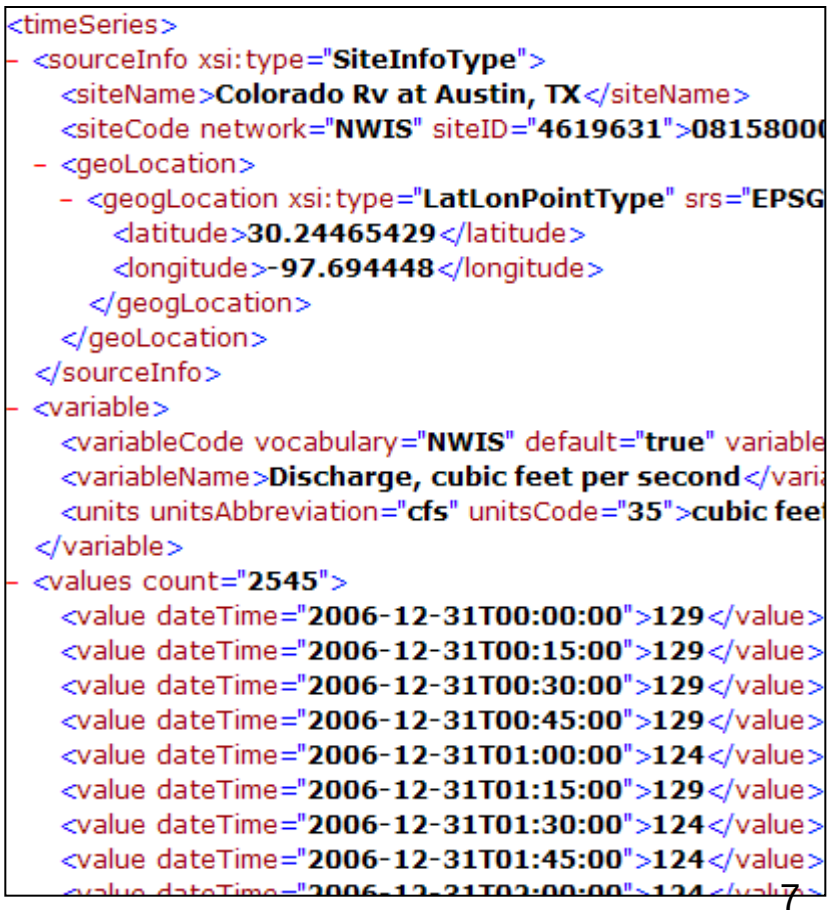
WaterML and WaterOneFlow

WaterML is an XML language for communicating water data
WaterOneFlow is a set of web services based on WaterML

- Set of query functions



- Returns data in WaterML



International Standardization of WaterML

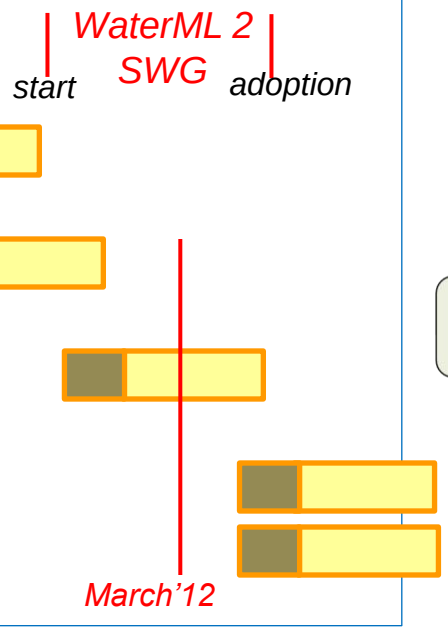
Hydrology Domain Working Group

- working on WaterML 2.0
- organizing Interoperability Experiments focused on different sub-domains of water
- towards an agreed upon feature model, observation model, semantics and service stack

World Meteorological Organization
Working together in weather, climate and water

OGC®
Open Geospatial Consortium, Inc.

Timeline



Groundwater IE
- GSC+USGS
- Dec 09 – Dec 10

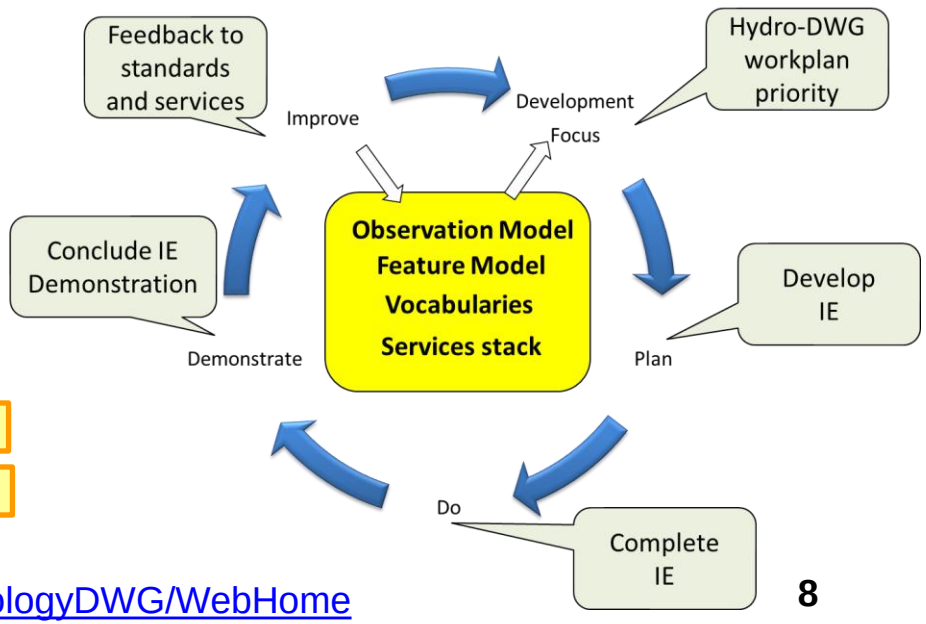
Surface Water IE
- CSIRO+many
- Jun 10 – Sep 11

Forecasting IE
- NWS+Deltares+
- Sep 11 – Oct 12

Water Quality IE

Water Use IE

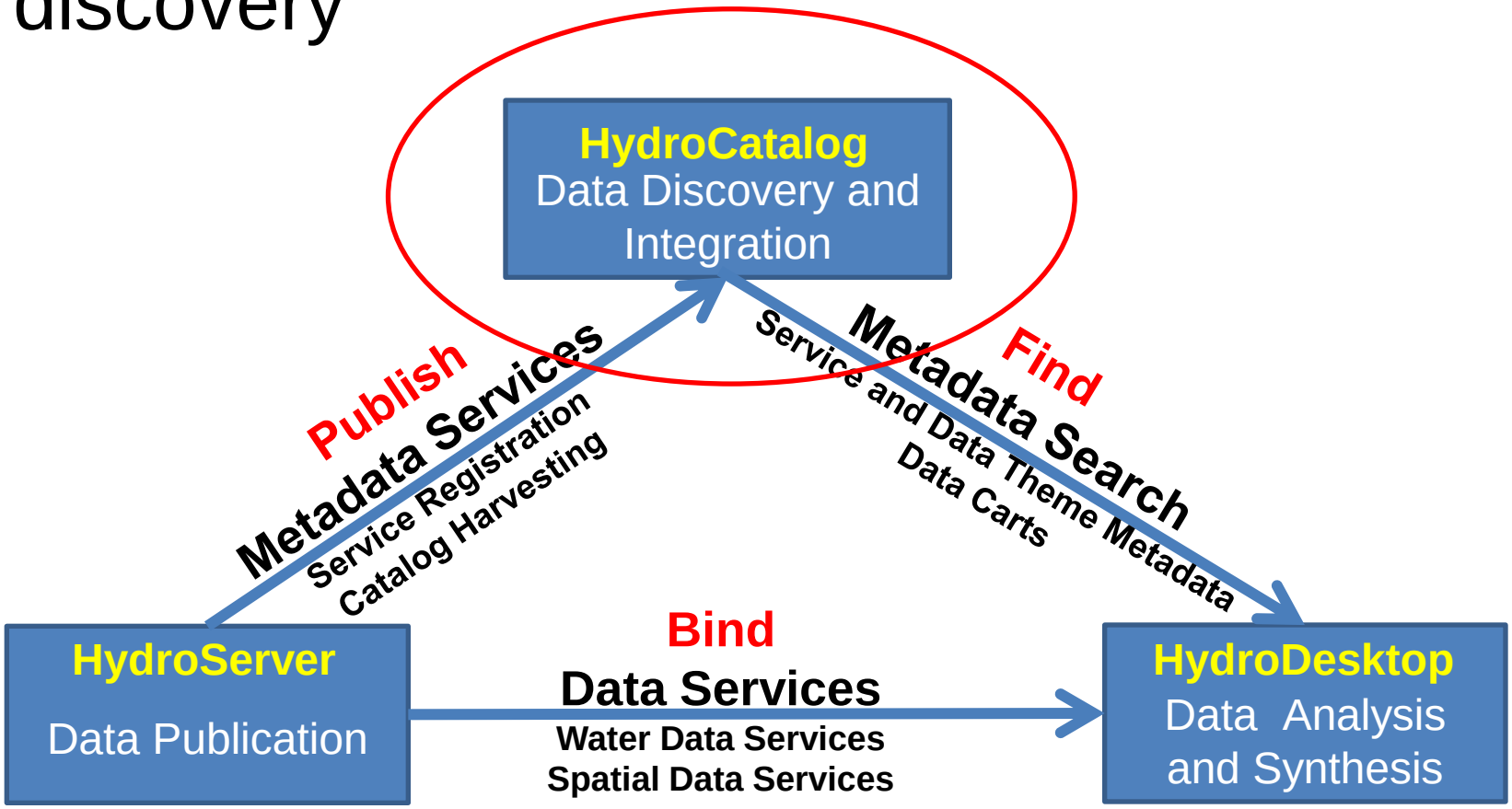
Iterative Development



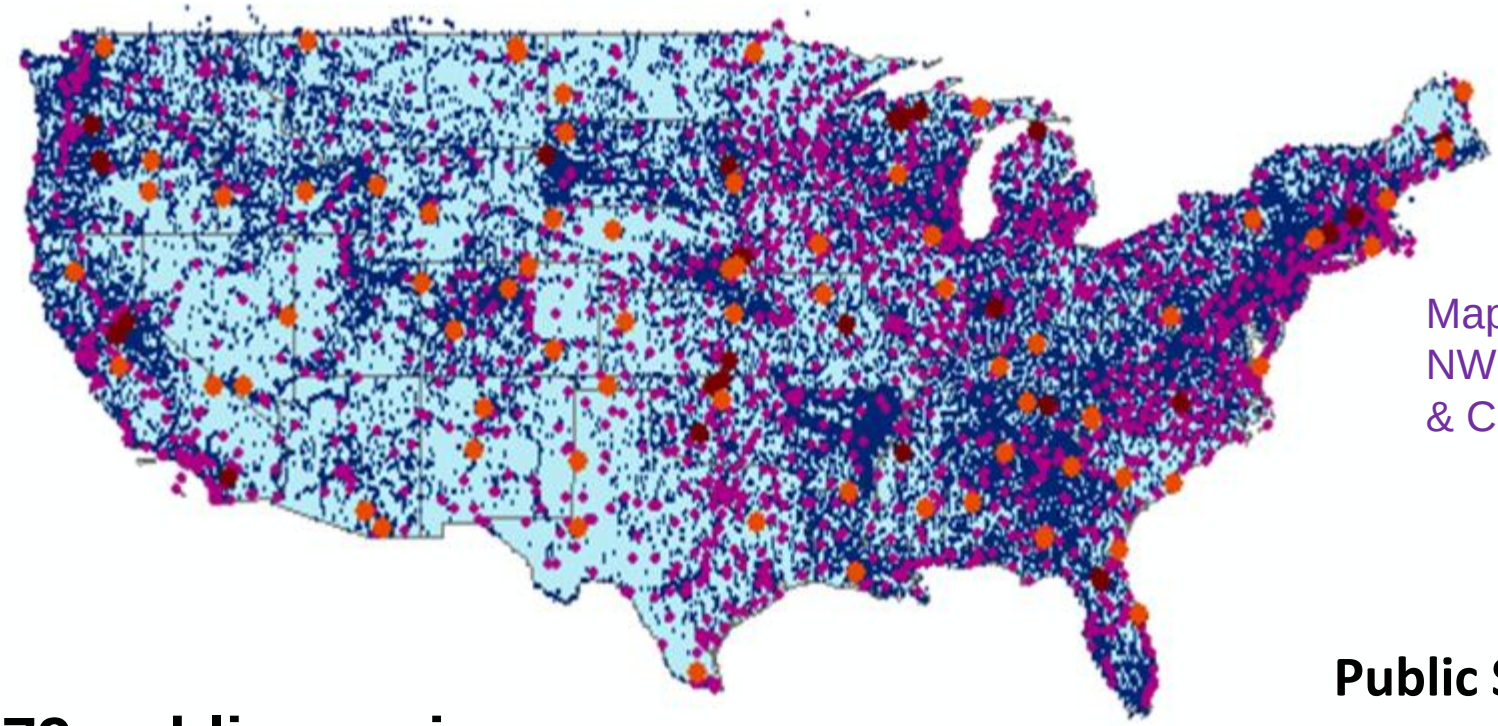
http://external.opengis.org/twiki_public/bin/view/HydrologyDWG/WebHome

HIS Central - HydroCatalog

- Central metadata catalog supporting data discovery



HIS Central Catalog

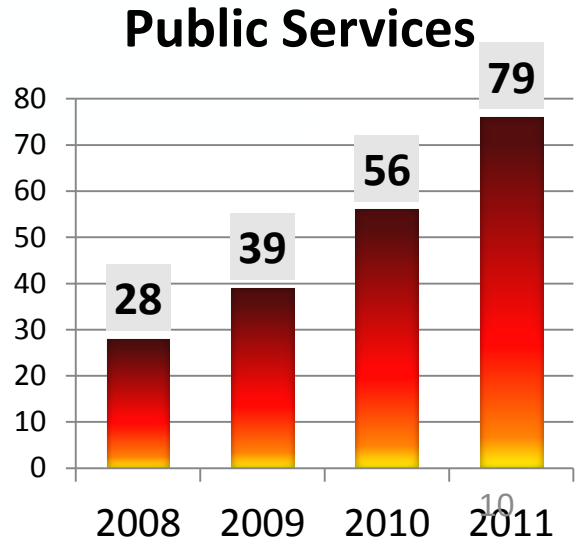


Map integrating NWIS, STORET, & Climatic Sites

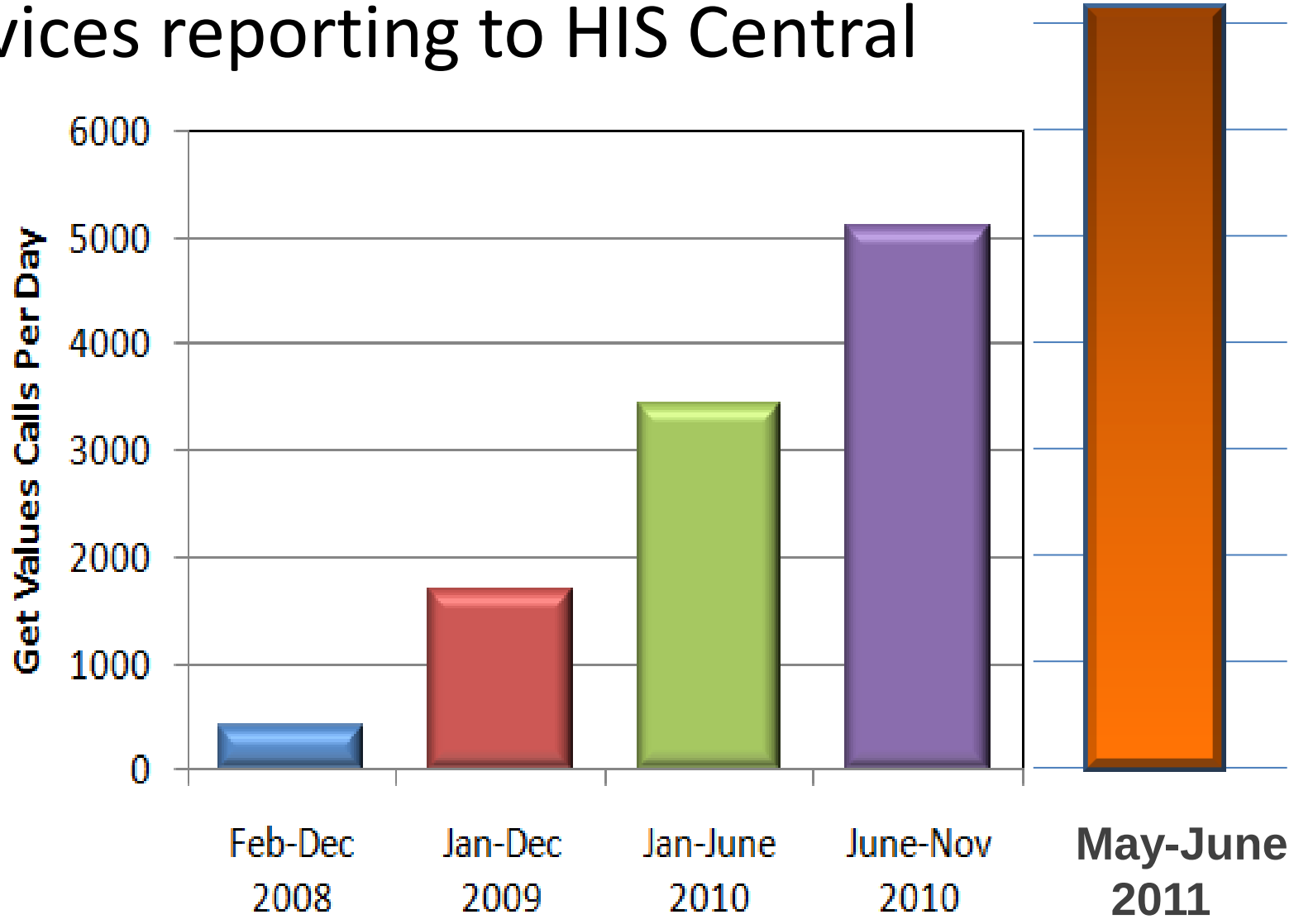
79 public services
16,000+ variables
2.31+ million sites
25.1 million series
Referencing 100+ billion data values

Available via HISCentral discovery services

Available via GetValues requests



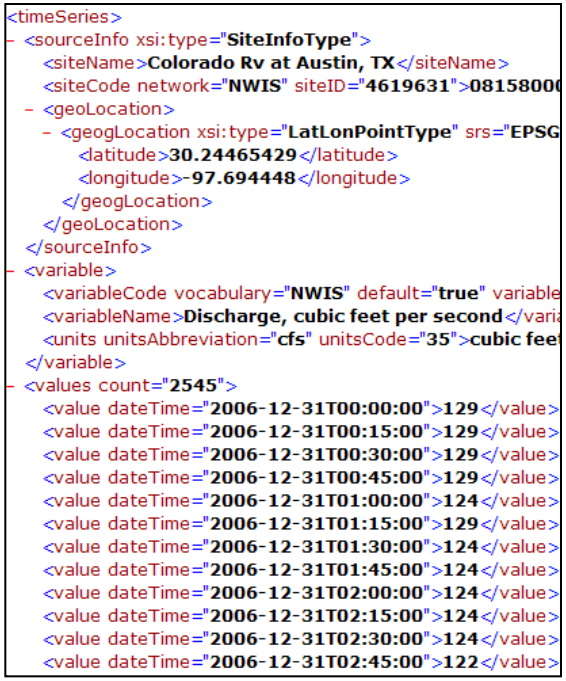
Growth in GetValues calls for all services reporting to HIS Central



Average for 2011: **13,569/day**

Data Heterogeneity

- Syntactic mediation
  - Heterogeneity of format
  - Use WaterML to get data into the same format



```
<timeSeries>
- <sourceInfo xsi:type="SiteInfoType">
    <siteName>Colorado Rv at Austin, TX</siteName>
    <siteCode network="NWIS" siteID="4619631">08158000
  - <geoLocation>
    - <geogLocation xsi:type="LatLonPointType" srs="EPSG
        <latitude>30.24465429</latitude>
        <longitude>-97.694448</longitude>
      </geogLocation>
    </geoLocation>
  </sourceInfo>
- <variable>
    <variableCode vocabulary="NWIS" default="true" variable
    <variableName>Discharge, cubic feet per second</vari
    <units unitsAbbreviation="cfs" unitsCode="35">cubic fee
  </variable>
- <values count="2545">
    <value dateTime="2006-12-31T00:00:00">129</value>
    <value dateTime="2006-12-31T00:15:00">129</value>
    <value dateTime="2006-12-31T00:30:00">129</value>
    <value dateTime="2006-12-31T00:45:00">129</value>
    <value dateTime="2006-12-31T01:00:00">124</value>
    <value dateTime="2006-12-31T01:15:00">129</value>
    <value dateTime="2006-12-31T01:30:00">124</value>
    <value dateTime="2006-12-31T01:45:00">124</value>
    <value dateTime="2006-12-31T02:00:00">124</value>
    <value dateTime="2006-12-31T02:15:00">124</value>
    <value dateTime="2006-12-31T02:30:00">124</value>
    <value dateTime="2006-12-31T02:45:00">122</value>
```

- Semantic mediation
  - Heterogeneity of meaning
  - Each water data source uses its own vocabulary
  - Match these up with a common controlled vocabulary
  - Make standard scientific data queries and have these automatically translated into specific queries on each data source

Managing Varying Semantics

In measurement units…

acre feet	acre-feet
micrograms per kilogram	micrograms per kilgram
FTU	NTU
mho	Siemens
ppm	mg/kg

In parameter names…

Nitrogen: **e.g. NWIS parameter # 625 is labeled 'ammonia + organic nitrogen', Kjeldahl method is used for determination but not mentioned in parameter description. In STORET this parameter is referred to as Kjeldahl Nitrogen.**

And: *Dissloved oxygen*

Semantics of Hydrologic Variables

- USGS: "parameters"
  - Over 18,000 terms; names overloaded (name, method, units, medium) – but inconsistent, often implied
  - E.g. "Calcium, water, unfiltered, recoverable, milligrams per liter"
- EPA: "characteristics"
  - Derive from SRS, aligned with Chemical Abstract Service registry
  - E.g. "calcium"
- NCDC: "elements"
  - 4-letter abbreviations (e.g. MXRH = max. relative humidity)
- CF standard names:
  - E.g. runoff_amount_excluding_baseflow
- OGC O&M: "observed properties"
- CUAHSI HIS: "variables"
  - In real databases, lots of proxy terms ("voltage")

Semantic Mediation

- **Hyponymy**

  Parameter "Groundwater level", "Stream stage", "Reservoir level" versus "Water level", which Water Level?

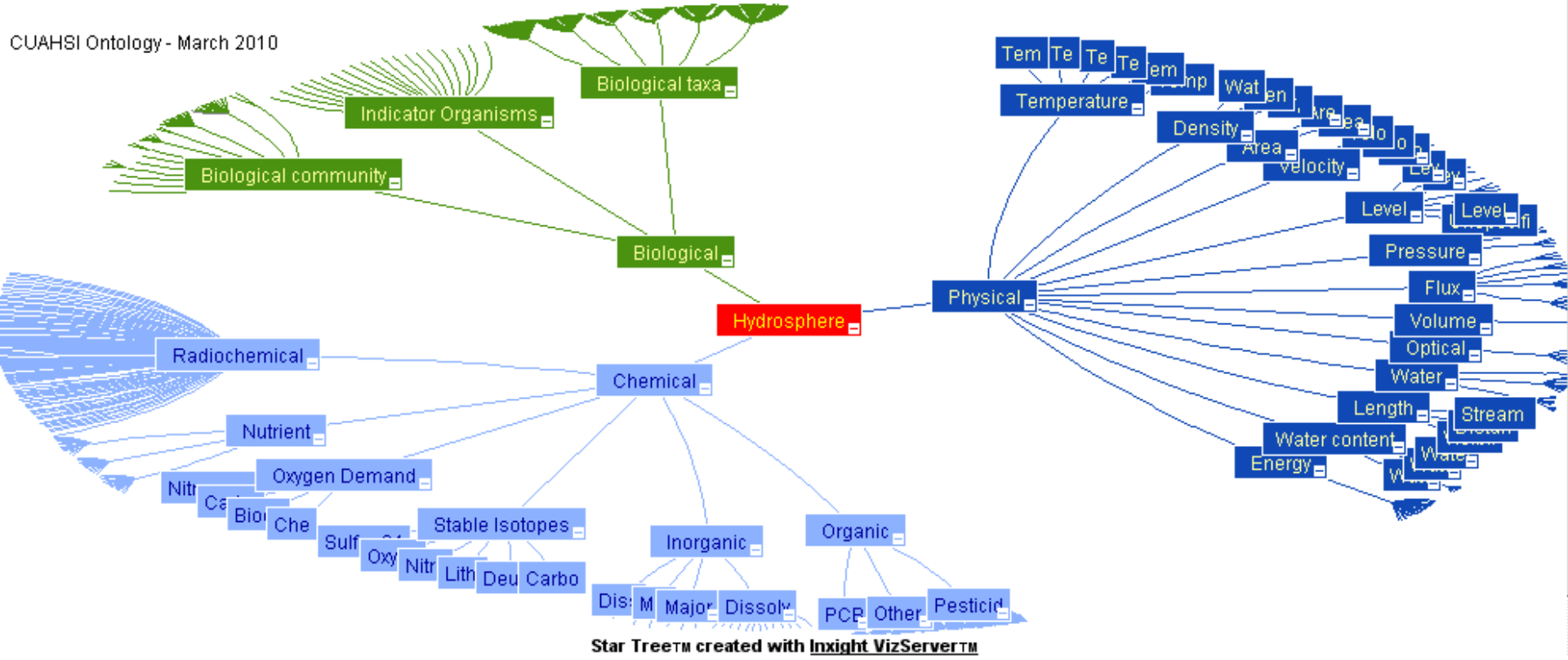
- **Polysemy**

  Parameter have multiple meanings, for example "stage", i.e. a water level measurement versus an art performance venue

- **Synonymy**

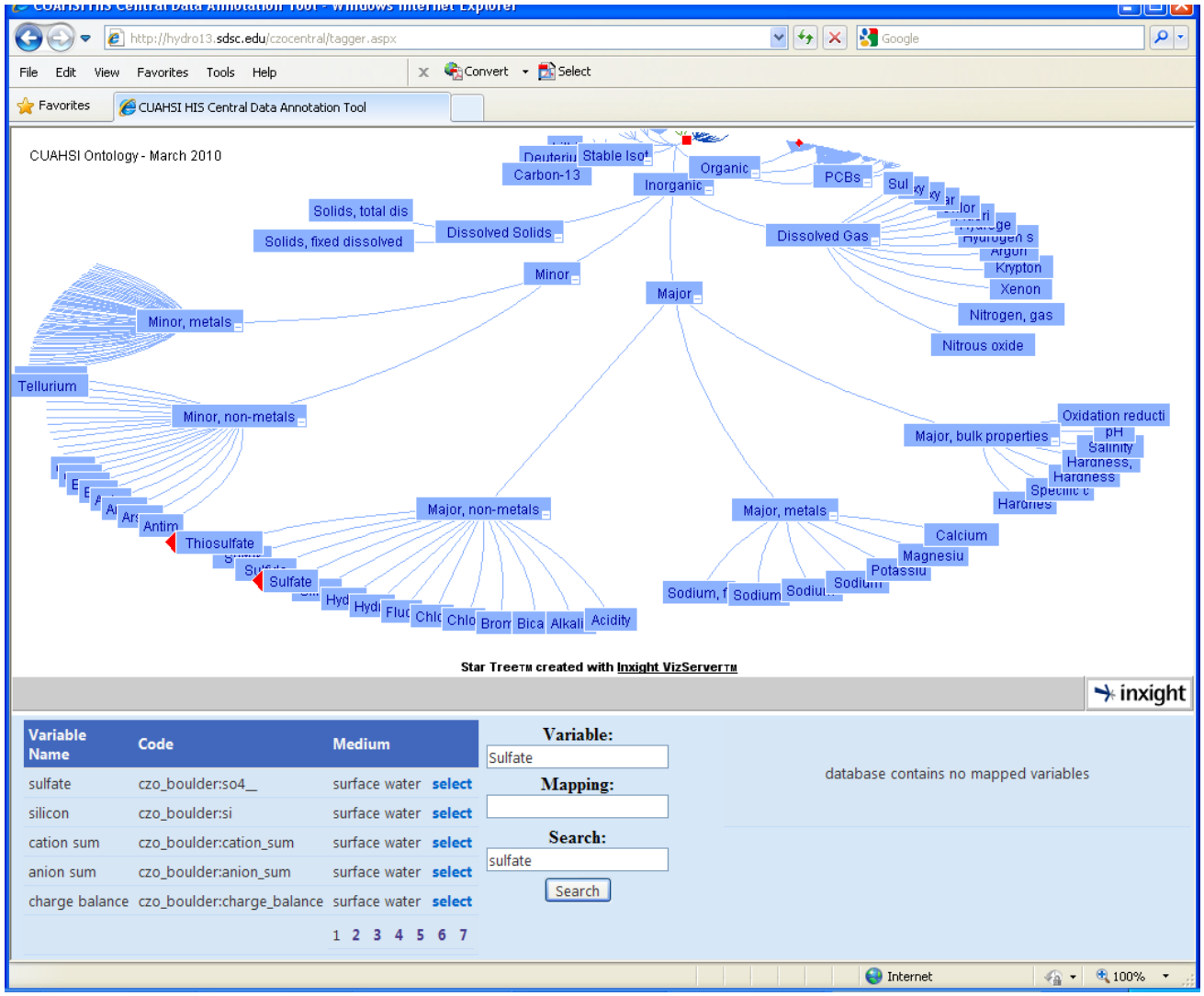
  'Total Kjeldahl Nitrogen' vs. 'Ammonia+Organic Nitrogen', or 'Stream Gauge' ⇔'Stream Stage' ⇔ 'Gauge Height' ⇔ 'Gauge'

Hydrologic Concept Hierarchy



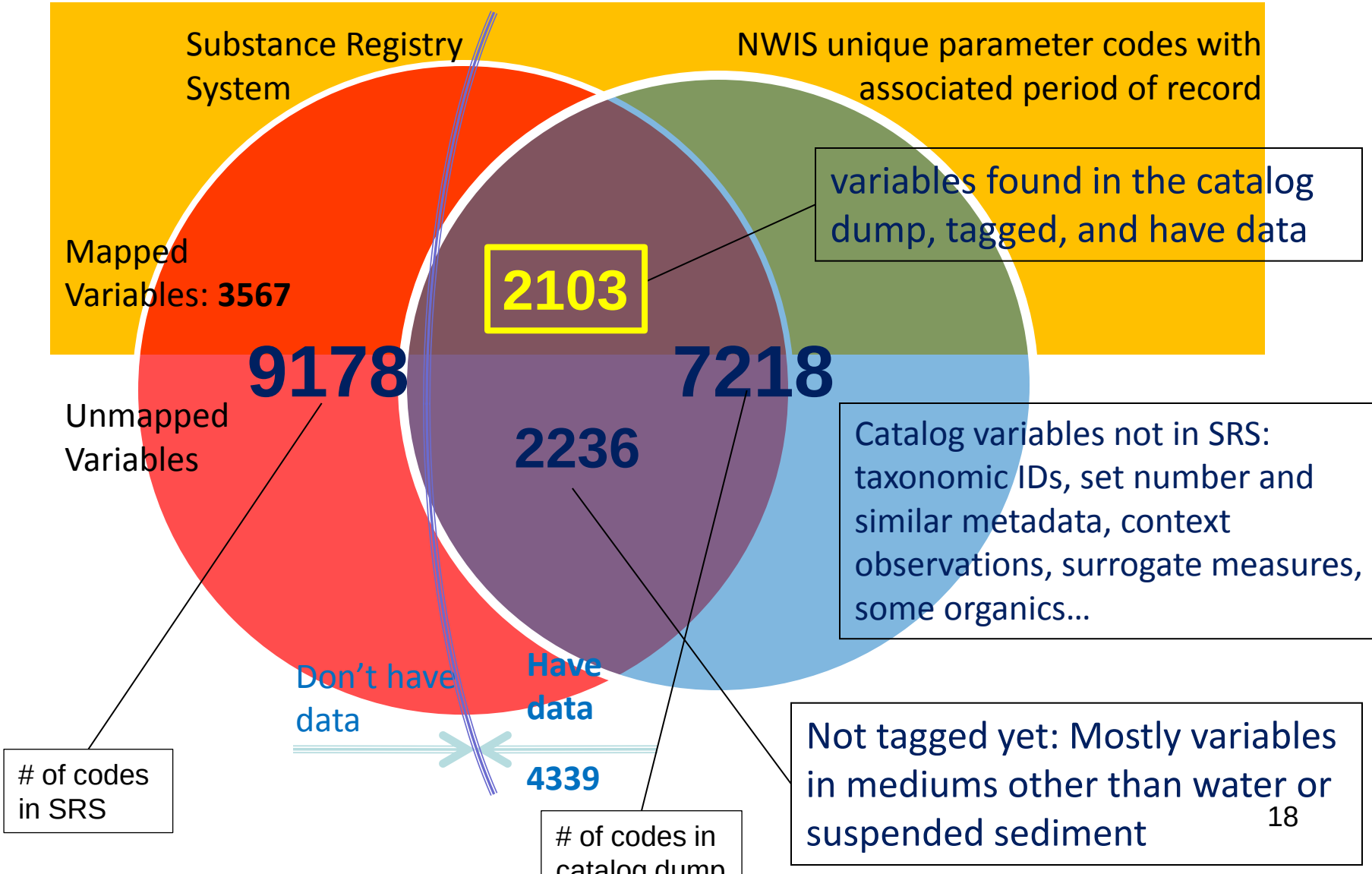
http://hiscentral.cuahsi.org/startree.aspx

Tagging variables in submitted datasets

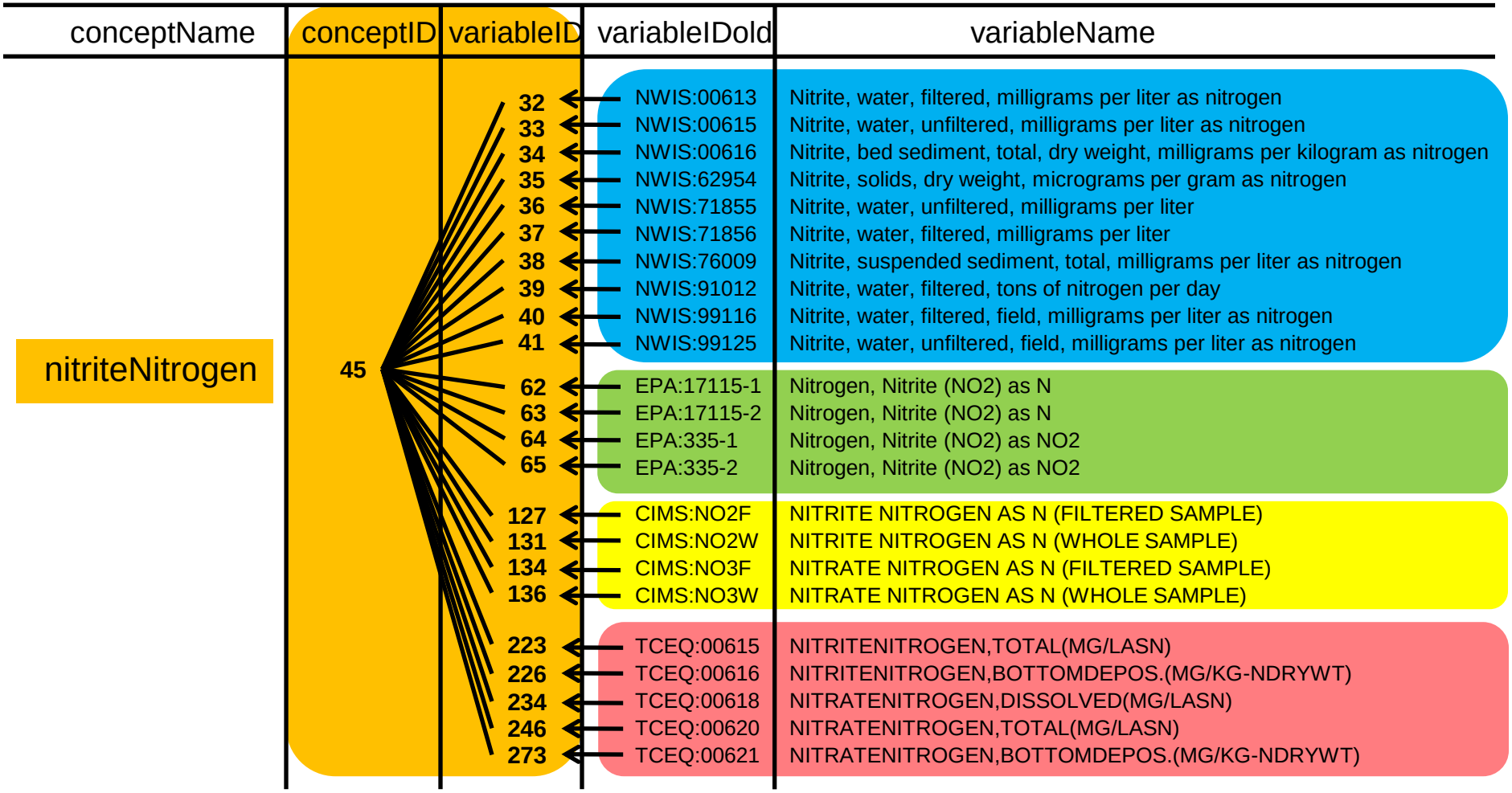


Time series can be discovered by keywords, once variables are associated with concepts in hydrologic ontology. The tagger application is available as part of HIS Web Service Registry

Aligning semantic hierarchy, SRS, and NWIS catalog dump



Concept-based search given concept-variable maps

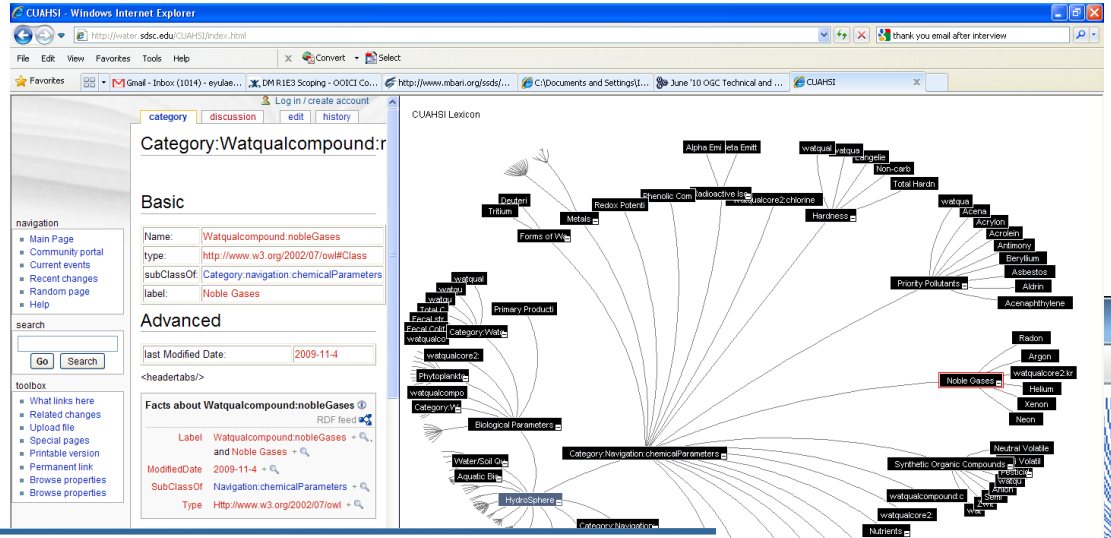


Catalog methods: GetMappedVariables; GetSearchableConcepts; GetSeries; GetOntologyTree…

Community management of vocabularies

Semantic Wiki

Concept Editor



Master Controlled Vocabulary Registry for ODM 1.1

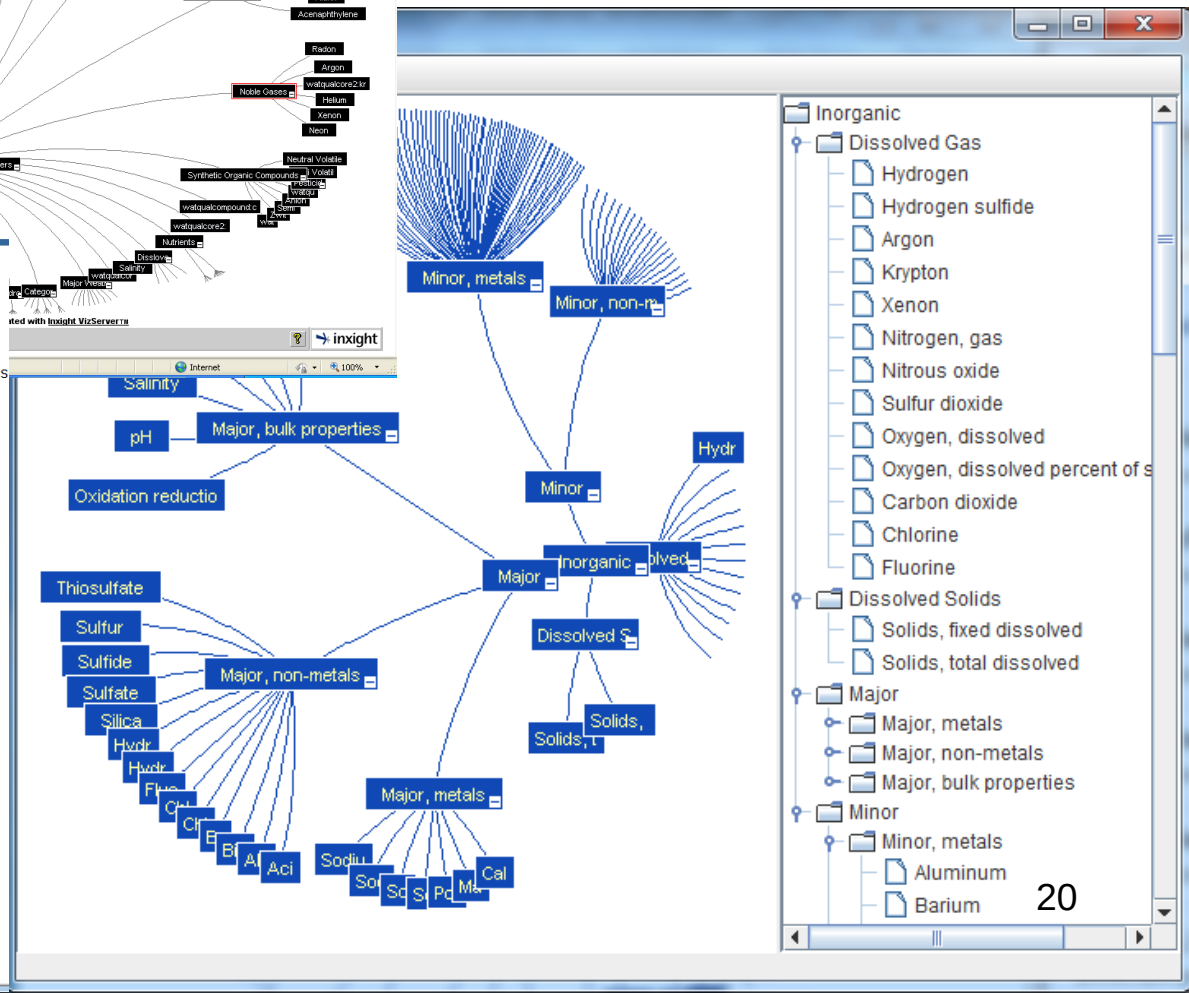
CUAHSI's ODM 1.1 has several controlled vocabulary tables. This web page has been developed to promote consistency between different instances of the ODM through a moderated system for changing the master controlled vocabularies. This web page displays the master controlled vocabulary entries and allows users to request additions or changes to these. Users may then update their ODM controlled vocabulary tables from this master set using ODM tools as described below.

Changes that you request will be forwarded to the master controlled vocabulary moderators who will attend to requests as promptly as possible. When you submit a request, you should receive an email verifying that you request has been received. When your request is approved, you should also receive an email confirmation. If you have a request that cannot be accommodated on this website, please contact the master controlled vocabulary moderators:

Jennifer Arrigo CUAHSI jarrigo@cuahsi.org
Jeff Horsburgh Utah State University jeff.horsburgh@usu.edu
David Tarboton Utah State University david.tarboton@usu.edu

Controlled Vocabularies:

CensorCodeCV:	Used to populate the CensorCode field of the DataValues table
DataTypeCV:	Used to populate the DataType field of the Variables table
GeneralCategoryCV:	Used to populate the GeneralCategory field in the Variables table
SampleMediumCV:	Used to populate the SampleMedium field in the Variables table
SampleTypeCV:	Used to populate the SampleType field in the Samples table
SiteTypeCV:	Used to populate the SiteType field in the Sites table
SpatialReferences:	Defines the coordinate systems used in the Sites table
SpeciationCV:	Used to poputate the Speciation field in the Variables table
TopicCategoryCV:	Used to populate the TopicCategory field in the ISOMetadata table
Units:	Defines the units used in the Variables and Offset types tables
ValueTypeCV:	Used to populate the ValueType field in the Variables table
VariableNameCV:	Used to populate the VariableName field in the Variables table
VerticalDatumCV:	Used to populate the VerticalDatum field im the Sites table



Some Lessons Learned

- Well defined and narrow use cases to demonstrate benefits of semantic approaches
  - “.. ontologies are the tail, not the dog”
- Having explicit vocabularies (classifiers) is a must in a distributed system; community shall be included in the development and evolution of vocabularies
- It is critical to capture and evolve domain knowledge in a form that the community is comfortable with
- Transition from implicit domain knowledge to explicit encoding requires community consensus - and an organization to manage the consensus