

Managing observation semantics in CUAHSI Hydrologic Information System

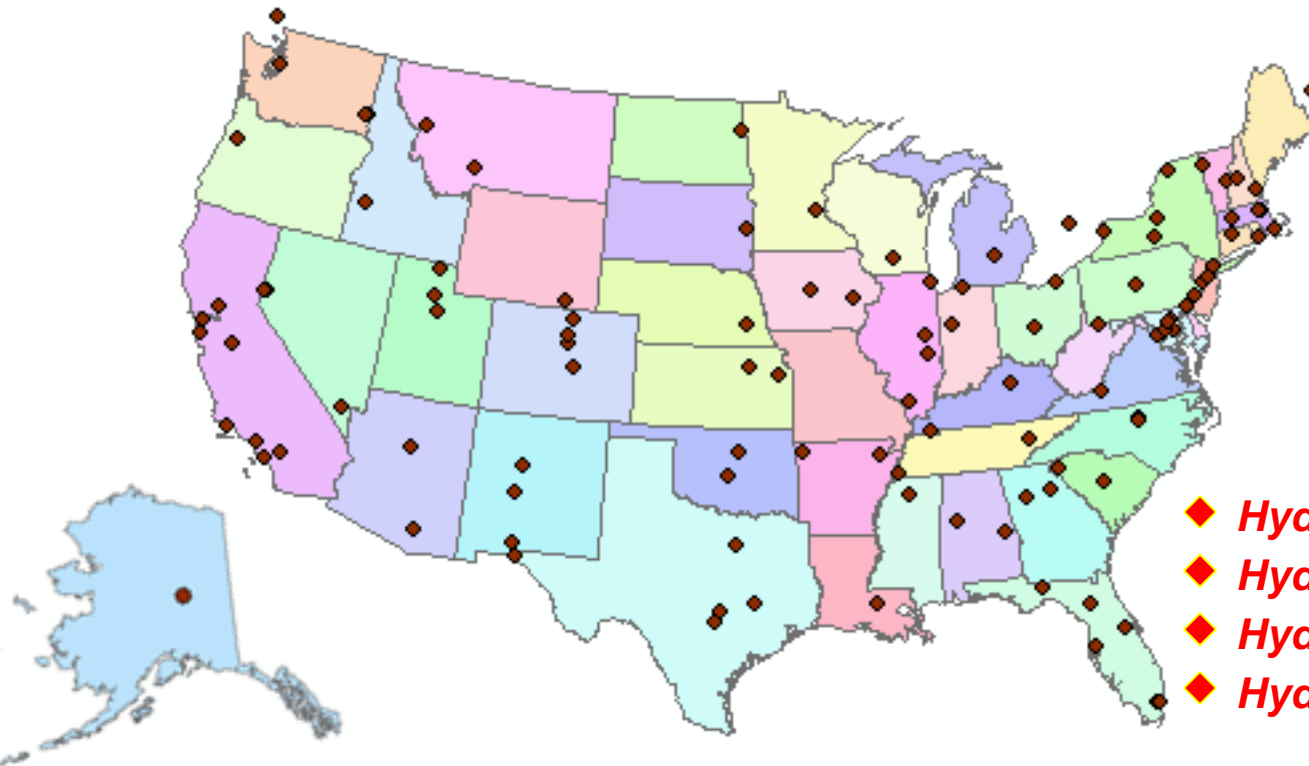


Ilya Zaslavsky

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.



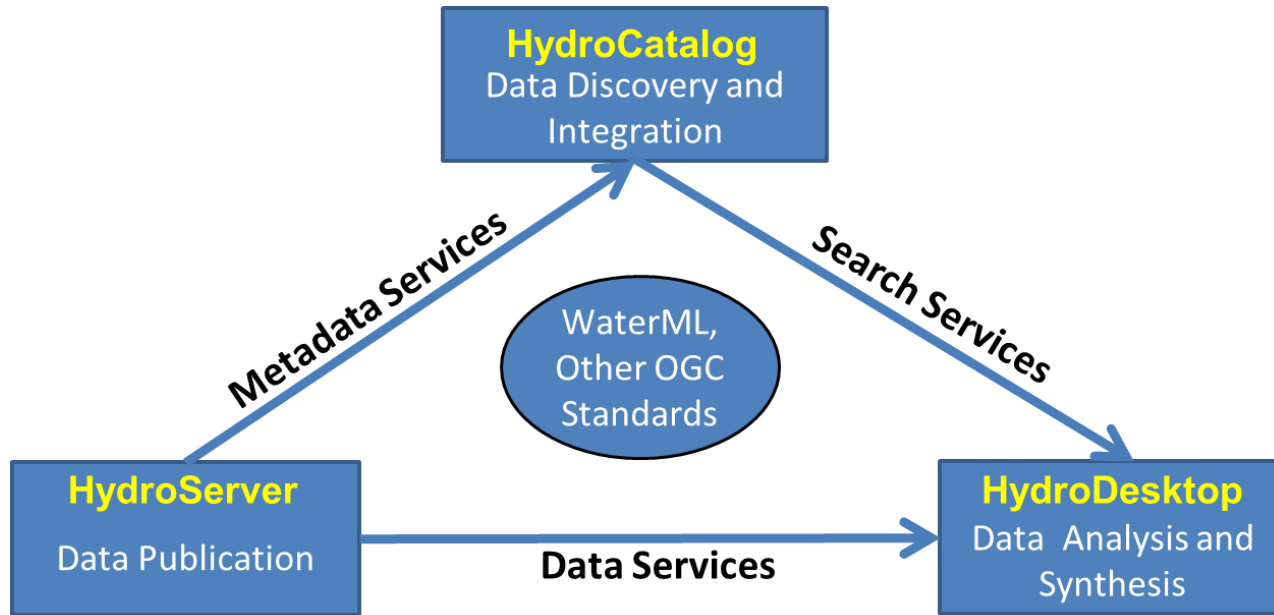
CUAHSI
universities allied for water research

<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



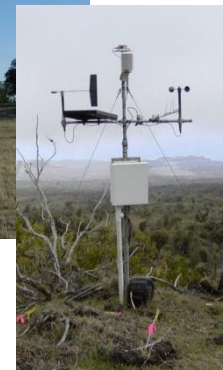
Soil water



Rainfall & Snow



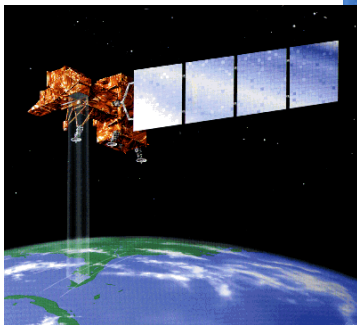
Meteorology



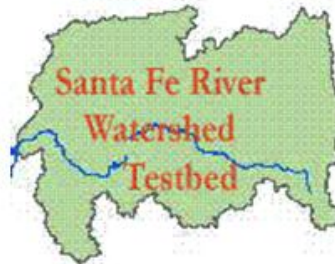
Modeling



Remote sensing



Sources of Observations Data



Observations Data Model (ODM)



Streamflow

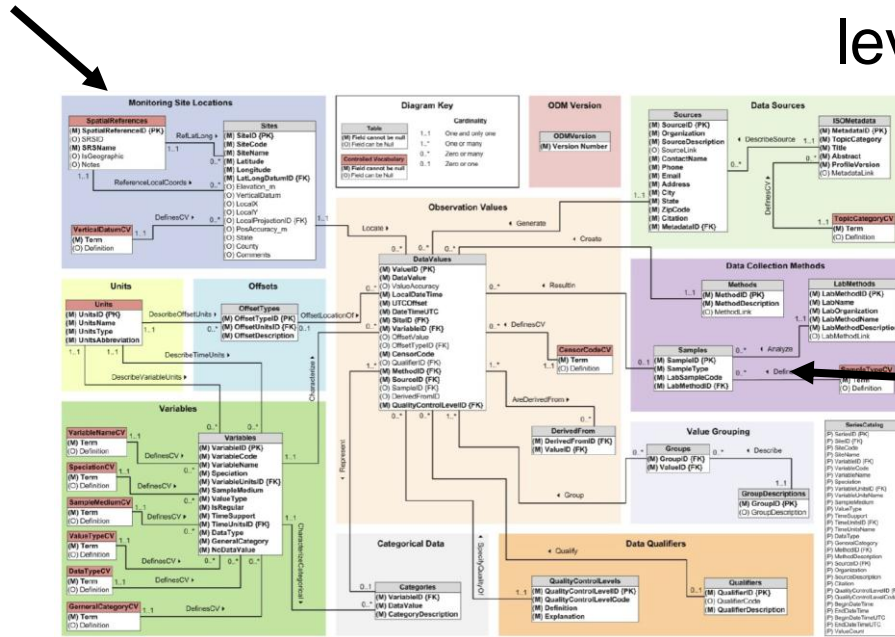
Groundwater levels



Precipitation & Climate



Water Quality



Soil moisture data



Flux tower data



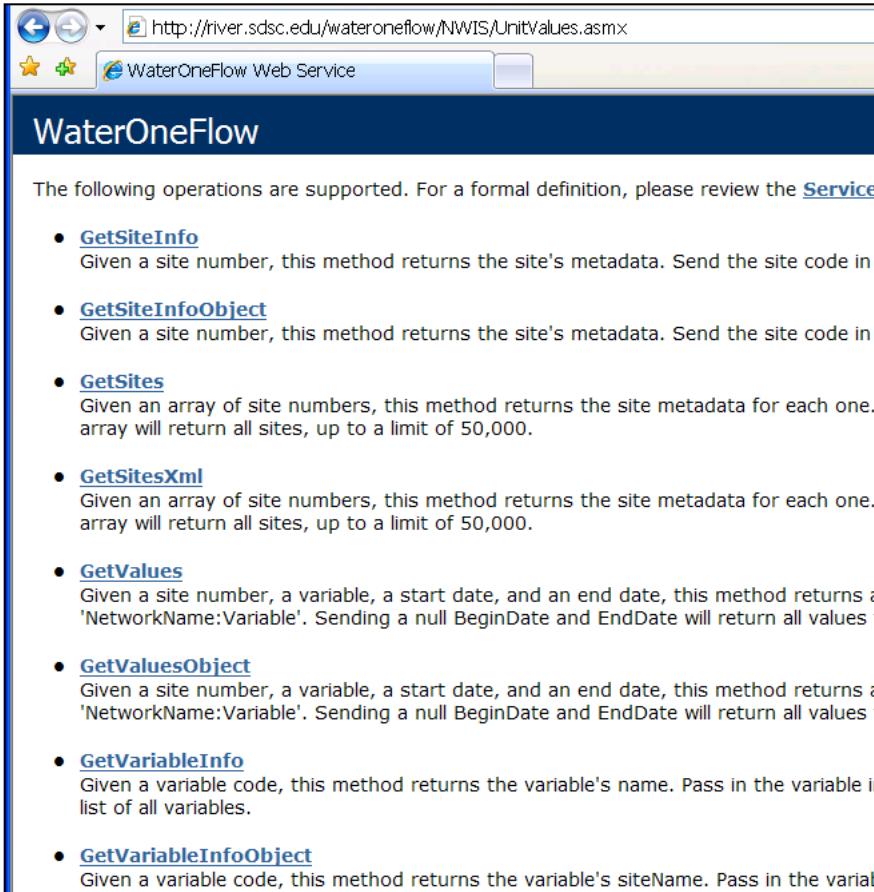
- A **relational database** at the single observation level
- Metadata for **unambiguous interpretation**
- Traceable heritage from **raw** measurements to **usable** information
- Promote **syntactic** and **semantic** consistency
- **Cross dimension** retrieval and analysis

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**



The screenshot shows a web browser window with the URL `http://river.sdsc.edu/wateroneflow/NWIS/UnitValues.aspx`. The page title is "WaterOneFlow Web Service". The main heading is "WaterOneFlow". Below the heading, it states: "The following operations are supported. For a formal definition, please review the [Service](#)".

- **GetSiteInfo**
Given a site number, this method returns the site's metadata. Send the site code in
- **GetSiteInfoObject**
Given a site number, this method returns the site's metadata. Send the site code in
- **GetSites**
Given an array of site numbers, this method returns the site metadata for each one. array will return all sites, up to a limit of 50,000.
- **GetSitesXml**
Given an array of site numbers, this method returns the site metadata for each one. array will return all sites, up to a limit of 50,000.
- **GetValues**
Given a site number, a variable, a start date, and an end date, this method returns a 'NetworkName:Variable'. Sending a null BeginDate and EndDate will return all values
- **GetValuesObject**
Given a site number, a variable, a start date, and an end date, this method returns a 'NetworkName:Variable'. Sending a null BeginDate and EndDate will return all values
- **GetVariableInfo**
Given a variable code, this method returns the variable's name. Pass in the variable list of all variables.
- **GetVariableInfoObject**
Given a variable code, this method returns the variable's siteName. Pass in the variat

```
<timeSeries>
- <sourceInfo xsi:type="SiteInfoType">
  <siteName>Colorado Rv at Austin, TX</siteName>
  <siteCode network="NWIS" siteID="4619631">08158000</siteCode>
- <geoLocation>
  - <geogLocation xsi:type="LatLonPointType" srs="EPSG4326">
    <latitude>30.24465429</latitude>
    <longitude>-97.694448</longitude>
  </geogLocation>
</geoLocation>
</sourceInfo>
- <variable>
  <variableCode vocabulary="NWIS" default="true" variableCode="08158000">08158000</variableCode>
  <variableName>Discharge, cubic feet per second</variableName>
  <units unitsAbbreviation="cfs" unitsCode="35">cubic feet per second</units>
</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>
```

International Standardization of WaterML

Hydrology Domain Working Group

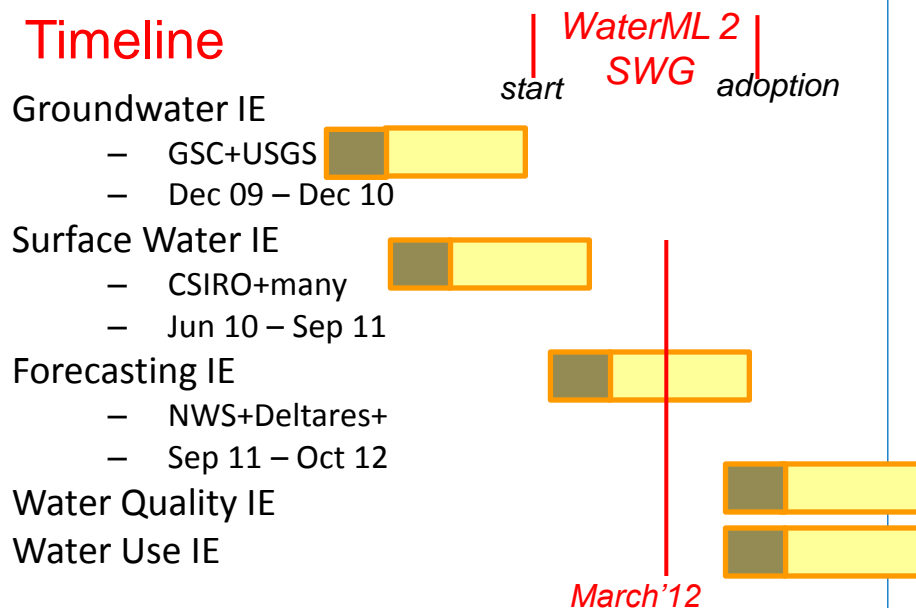
- working on WaterML 2.0
- organizing Interoperability Experiments focused on different sub-domains of water
- towards an agreed upon feature model, observation model, semantics and service stack



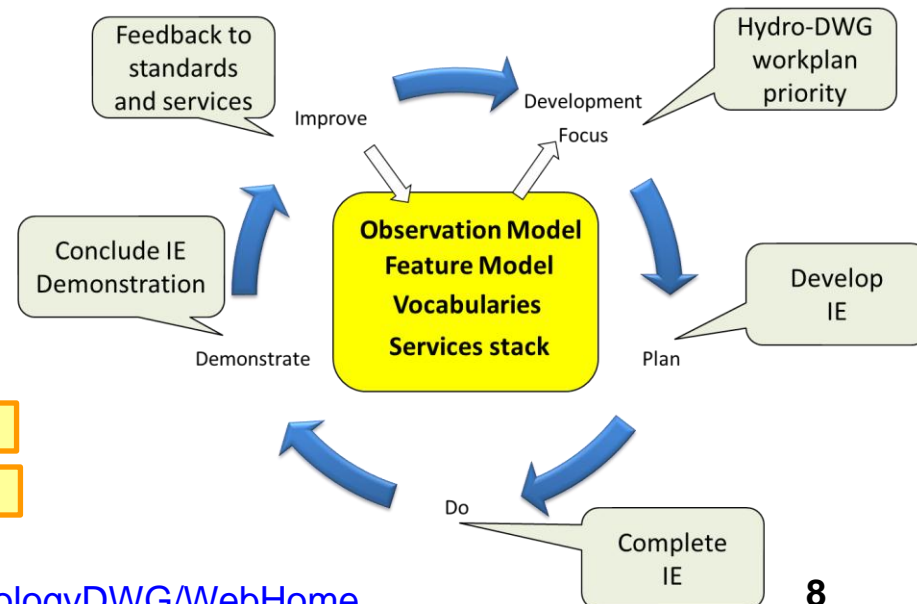
World Meteorological Organization
Working together in weather, climate and water



Timeline

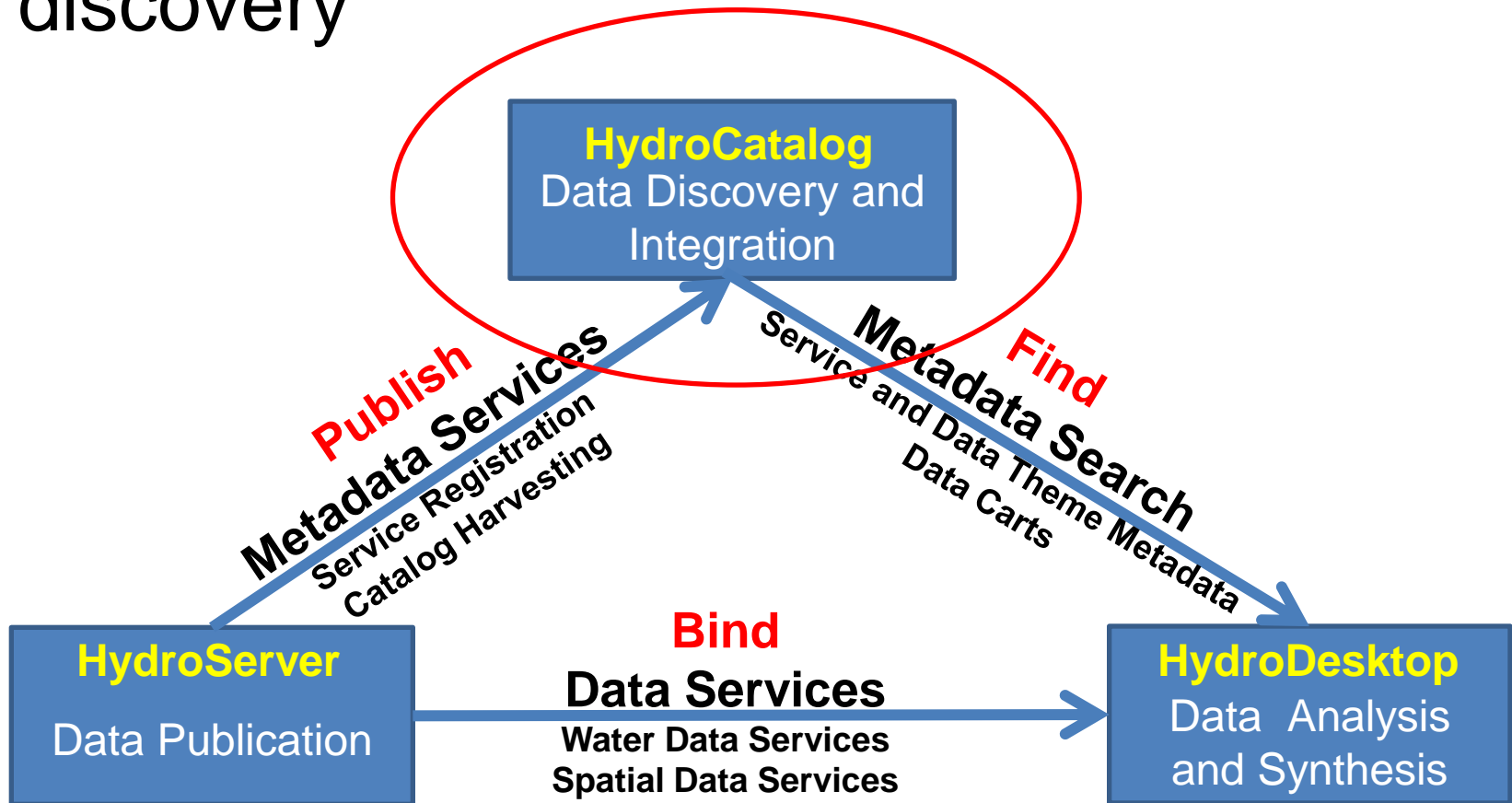


Iterative Development

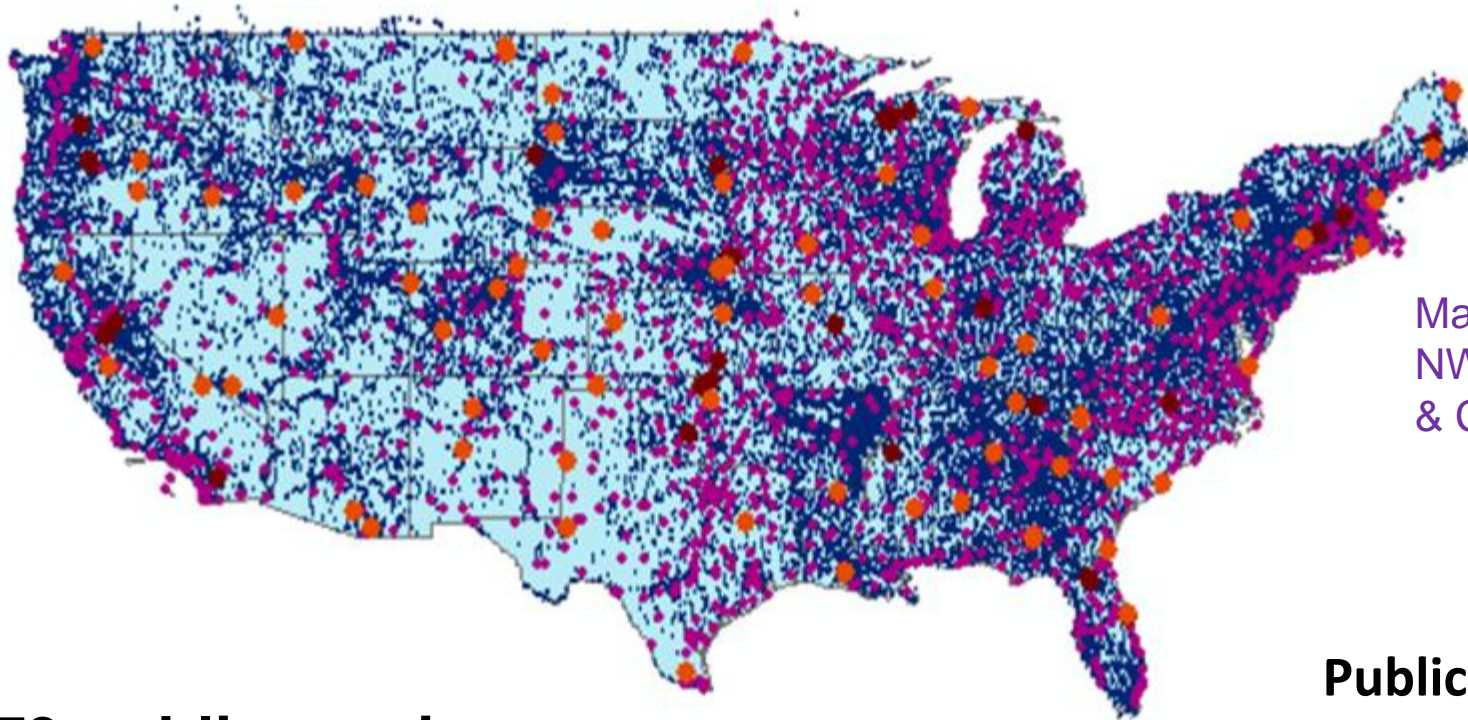


HIS Central - HydroCatalog

- 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

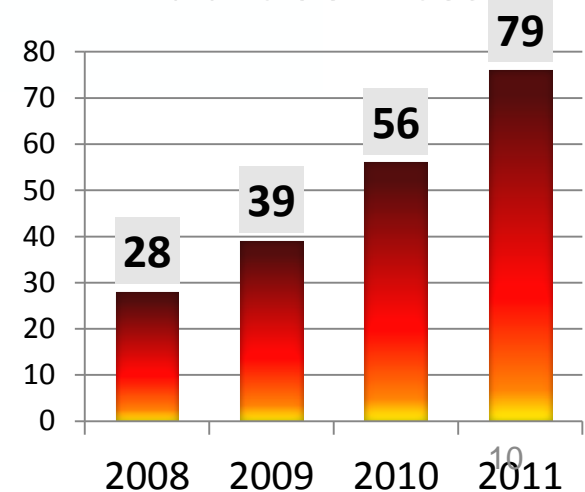
25.1 million series

Referencing 100+ billion data values

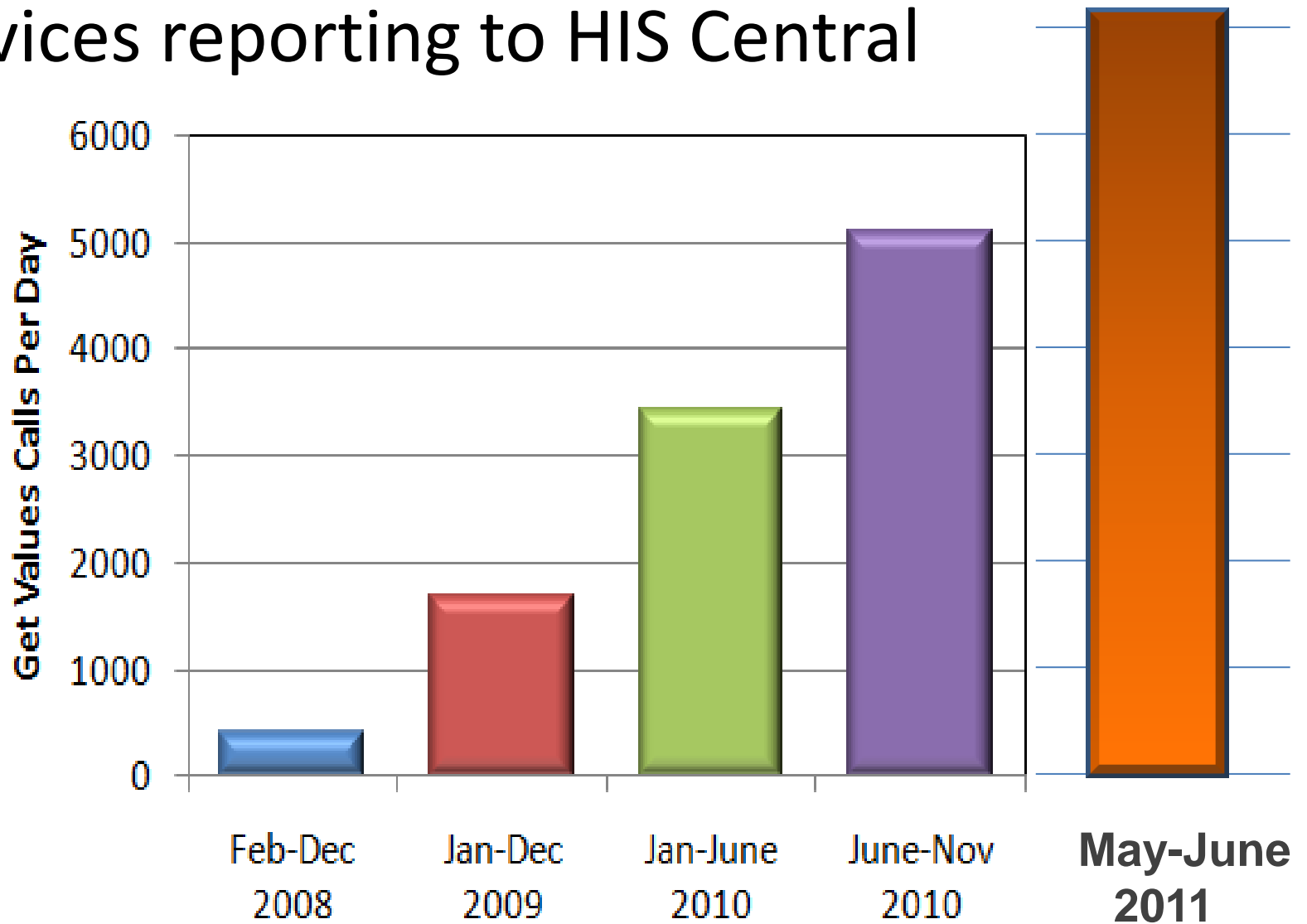
*Available via HISCentral
discovery services*

Available via GetValue requests

Public Services



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</siteCode>
- <geoLocation>
  - <geogLocation xsi:type="LatLonPointType" srs="EPSG:4326">
    <latitude>30.24465429</latitude>
    <longitude>-97.694448</longitude>
  </geogLocation>
</geoLocation>
</sourceInfo>
<variable>
  <variableCode vocabulary="NWIS" default="true" variableCode="08158000">08158000</variableCode>
  <variableName>Discharge, cubic feet per second</variableName>
  <units unitsAbbreviation="cfs" unitsCode="35">cubic feet per second</units>
</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>
</values>
</timeSeries>
```

- 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

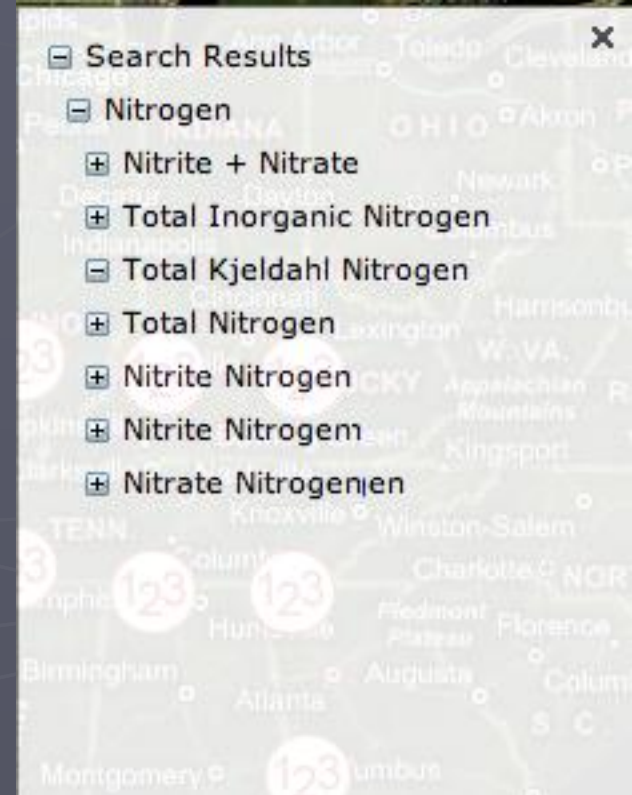
Managing Varying Semantics

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: Dissolved 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**

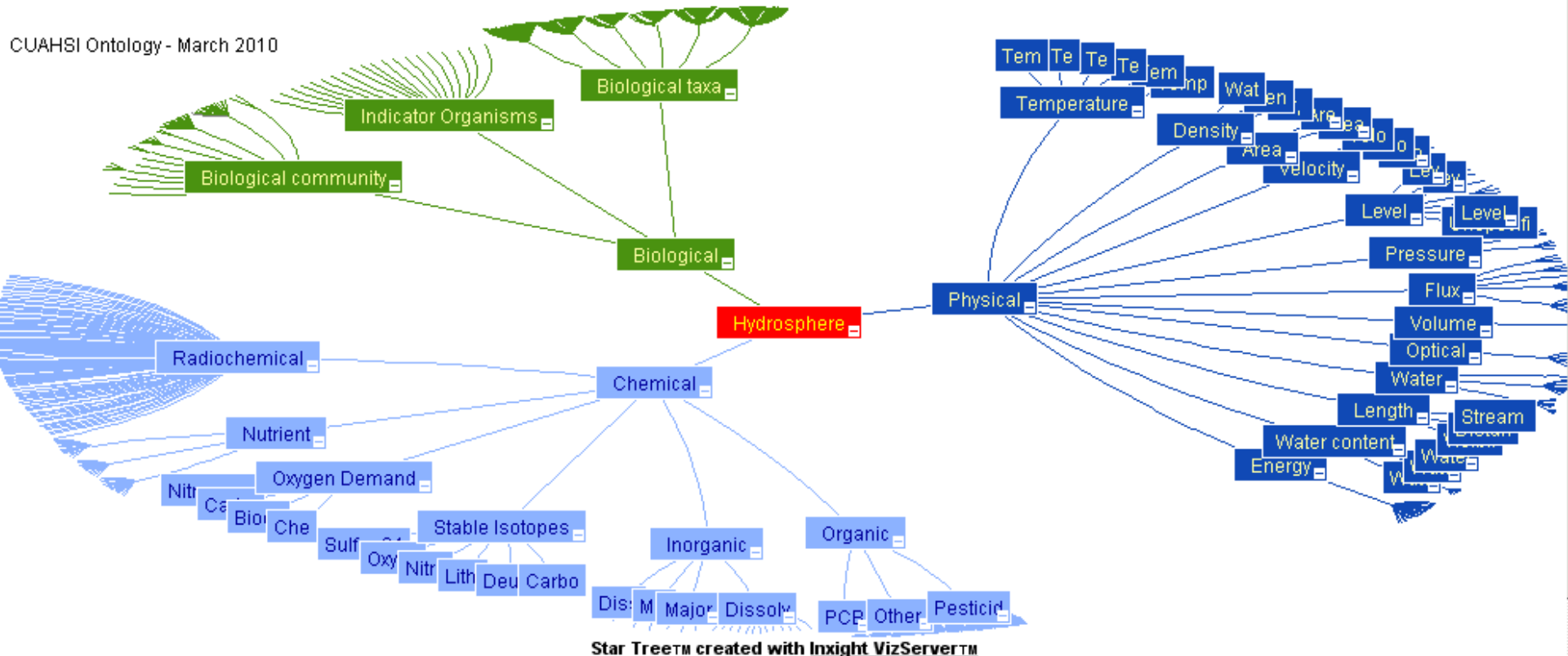
Parameter have multiple meanings, for example “stage”, i.e. a water level measurement versus an art performance venue

- **Synonymy**

‘Total Kjeldahl Nitrogen’ vs. ‘Ammonia+Organic Nitrogen’, or
‘Stream Gauge’ ↔ ‘Stream Stage’ ↔ ‘Gauge Height’ ↔ ‘Gauge’

Hydrologic Concept Hierarchy

CUAHSI Ontology - March 2010



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

Tagging variables in submitted datasets

The screenshot displays the CUAHSI HIS Central Data Annotation Tool interface. The main area shows a hierarchical ontology tree for 'CUAHSI Ontology - March 2010'. The tree is structured as follows:

- Major
 - Major, non-metals
 - Major, non-metals (sub-category)
 - Major, metals (sub-category)
 - Major, metals
 - Major, metals (sub-category)
 - Major, bulk properties
 - Major, bulk properties (sub-category)
 - Major, metals (sub-category)
- Minor
 - Minor, non-metals
 - Minor, non-metals (sub-category)
 - Minor, metals
 - Minor, metals (sub-category)
- Dissolved Solids
 - Solids, total dis
 - Solids, fixed dissolved
- Inorganic
 - Inorganic (sub-category)
- Organic
 - Organic (sub-category)
- Dissolved Gas
 - Dissolved Gas (sub-category)
- PCBs
 - PCBs (sub-category)

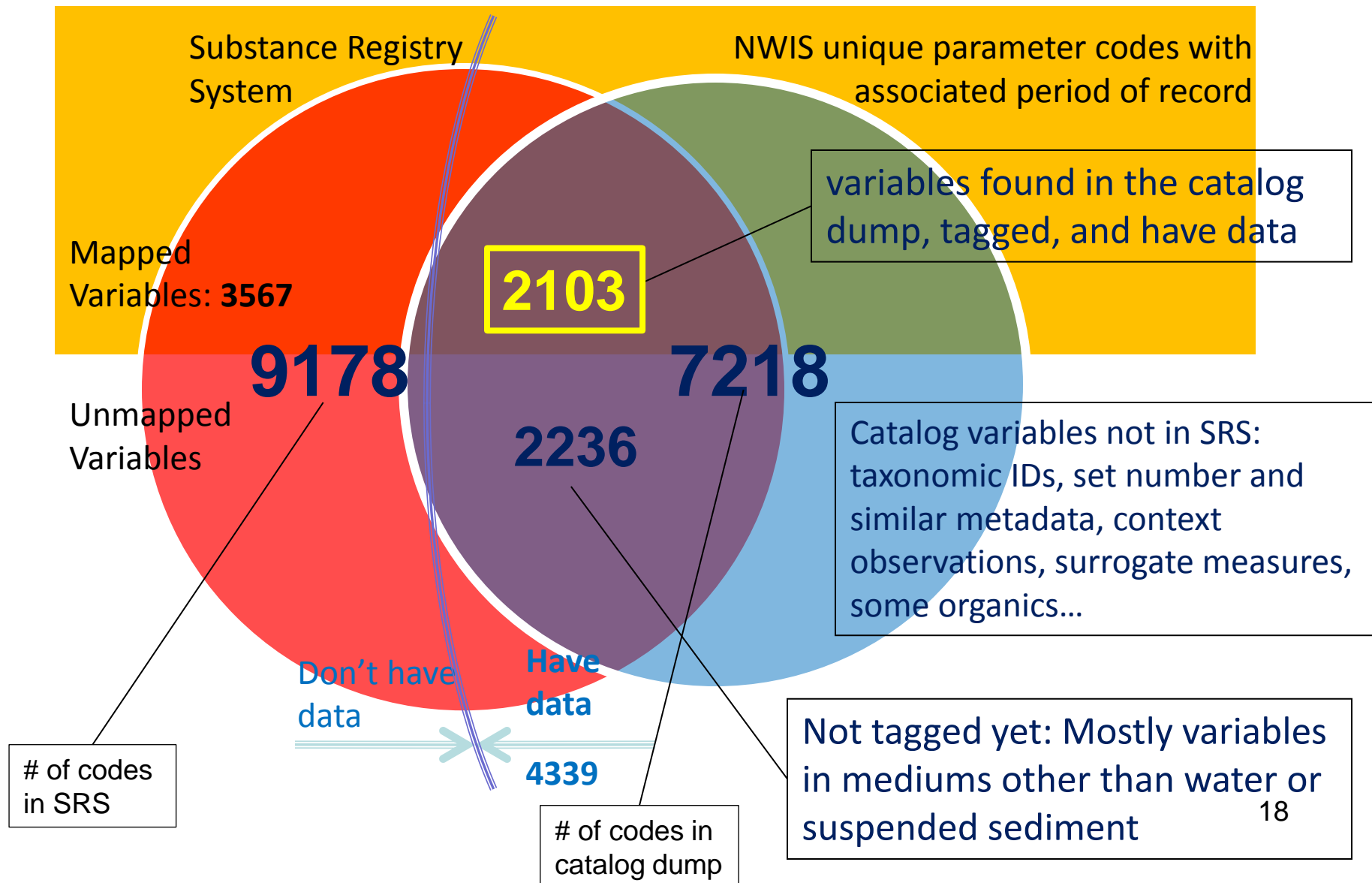
Below the tree, there is a search bar and a 'Search' button. The bottom section of the interface contains a table of variables and their associated metadata.

Variable Name	Code	Medium	select
sulfate	czo_boulder:so4_	surface water	select
silicon	czo_boulder:si	surface water	select
cation sum	czo_boulder:cation_sum	surface water	select
anion sum	czo_boulder:anion_sum	surface water	select
charge balance	czo_boulder:charge_balance	surface water	select

At the bottom of the table, there are page numbers: 1 2 3 4 5 6 7.

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

conceptName	conceptID	variableID	variableIDold	variableName
nitriteNitrogen	45	32	NWIS:00613	Nitrite, water, filtered, milligrams per liter as nitrogen
		33	NWIS:00615	Nitrite, water, unfiltered, milligrams per liter as nitrogen
		34	NWIS:00616	Nitrite, bed sediment, total, dry weight, milligrams per kilogram as nitrogen
		35	NWIS:62954	Nitrite, solids, dry weight, micrograms per gram as nitrogen
		36	NWIS:71855	Nitrite, water, unfiltered, milligrams per liter
		37	NWIS:71856	Nitrite, water, filtered, milligrams per liter
		38	NWIS:76009	Nitrite, suspended sediment, total, milligrams per liter as nitrogen
		39	NWIS:91012	Nitrite, water, filtered, tons of nitrogen per day
		40	NWIS:99116	Nitrite, water, filtered, field, milligrams per liter as nitrogen
		41	NWIS:99125	Nitrite, water, unfiltered, field, milligrams per liter as nitrogen
		62	EPA:17115-1	Nitrogen, Nitrite (NO2) as N
		63	EPA:17115-2	Nitrogen, Nitrite (NO2) as N
		64	EPA:335-1	Nitrogen, Nitrite (NO2) as NO2
		65	EPA:335-2	Nitrogen, Nitrite (NO2) as NO2
		127	CIMS:NO2F	NITRITE NITROGEN AS N (FILTERED SAMPLE)
		131	CIMS:NO2W	NITRITE NITROGEN AS N (WHOLE SAMPLE)
		134	CIMS:NO3F	NITRATE NITROGEN AS N (FILTERED SAMPLE)
		136	CIMS:NO3W	NITRATE NITROGEN AS N (WHOLE SAMPLE)
		223	TCEQ:00615	NITRITENITROGEN,TOTAL(MG/LASN)
		226	TCEQ:00616	NITRITENITROGEN,BOTTOMDEPOS.(MG/KG-NDRYWT)
234	TCEQ:00618	NITRATENITROGEN,DISSOLVED(MG/LASN)		
246	TCEQ:00620	NITRATENITROGEN,TOTAL(MG/LASN)		
273	TCEQ:00621	NITRATENITROGEN,BOTTOMDEPOS.(MG/KG-NDRYWT)		

Catalog methods: GetMappedVariables; GetSearchableConcepts; GetSeries; GetOntologyTree...

Community management of vocabularies

CUAHSI Lexicon

Category:Watqualcompound:r

Basic

Name: Watqualcompound:nobleGases
 type: http://www.w3.org/2002/07/owl#Class
 subClassOf: Category:navigation.chemicalParameters
 label: Noble Gases

Advanced

last Modified Date: 2009-11-4

Facts about Watqualcompound:nobleGases

Label: Watqualcompound:nobleGases and Noble Gases +
 ModifiedDate: 2009-11-4 +
 SubClassOf: Navigation.chemicalParameters -
 Type: Http://www.w3.org/2002/07/owl -

Semantic Wiki

Concept Editor

Inorganic

- Dissolved Gas
 - Hydrogen
 - Hydrogen sulfide
 - Argon
 - Krypton
 - Xenon
 - Nitrogen, gas
 - Nitrous oxide
 - Sulfur dioxide
 - Oxygen, dissolved
 - Oxygen, dissolved percent of s
 - Carbon dioxide
 - Chlorine
 - Fluorine
- Dissolved Solids
 - Solids, fixed dissolved
 - Solids, total dissolved
- Major
 - Major, metals
 - Major, non-metals
 - Major, bulk properties
- Minor
 - Minor, metals
 - Aluminum
 - Barium

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 your 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 Jeff Horsburgh David Tarboton
 CUAHSI Utah State University Utah State University
 jarrigo@cuahsi.org jeff.horsburgh@usu.edu 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 populate 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 in 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