

How to use Water Data to Produce Knowledge: Data Sharing with the CUAHSI Water Data Center

Jon Pollak

The Consortium of Universities for the Advancement of Hydrologic
Science, Inc. (CUAHSI)

August 20, 2014

Lower Mekong Initiative International Workshop



Agenda and Goals

- Background and introduction on CUAHSI
- The CUAHSI Water Data Center
 - The Hydrologic Information System
- Challenges

I hope you will come away with a feeling that there are existing tools that can help this region develop cyberinfrastructure for sharing data

CUAHSI

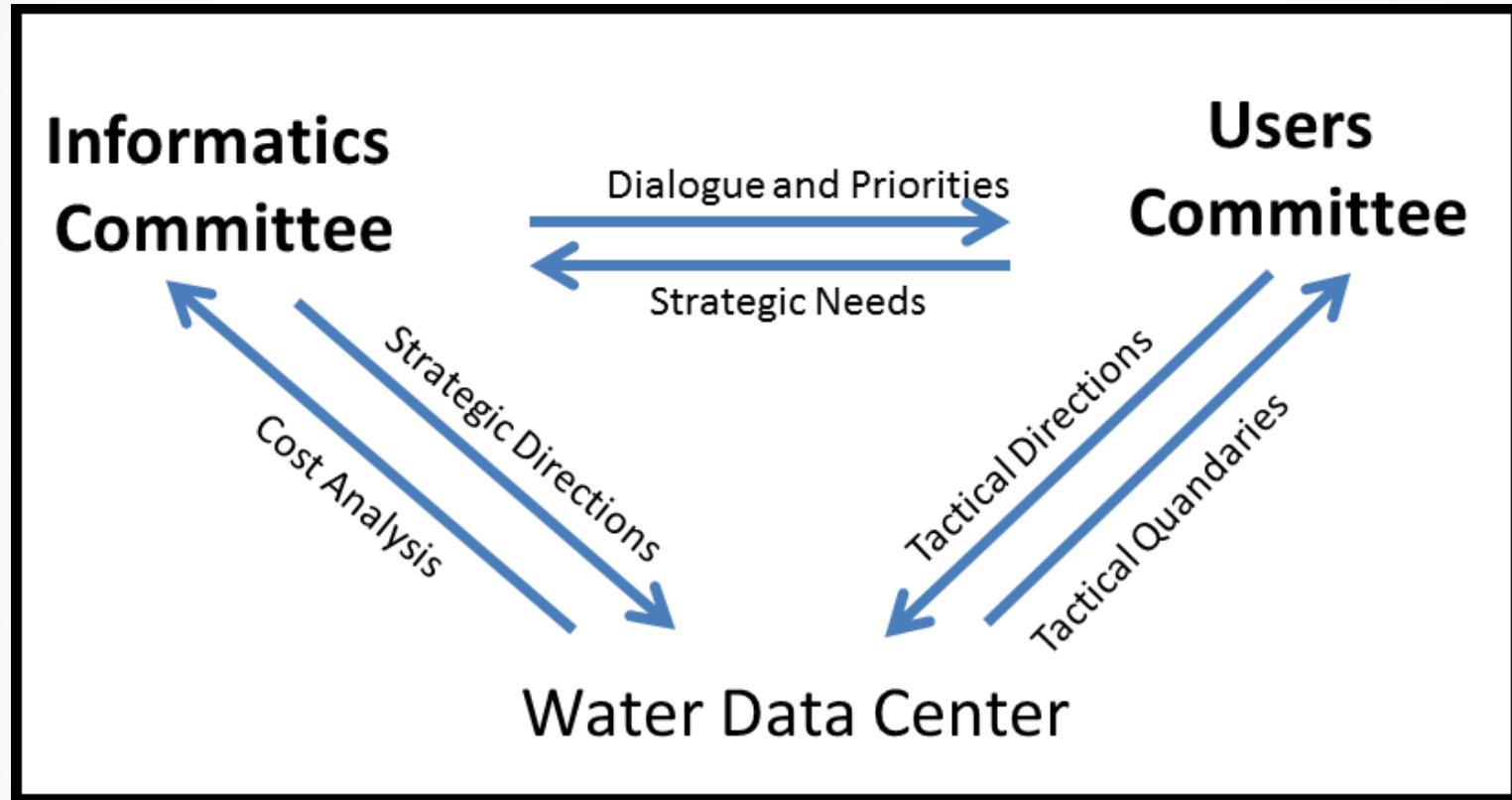


- Consortium of Universities for the Advancement of Hydrologic Science, Inc.
- Established in 2001
- Over 100 members
- Funding primarily by the U.S. National Science Foundation
- Largest program is a Water Data Center (WDC) that focuses on sharing and archiving of academic research data

Mission of CUAHSI WDC

- Provide data services for the academic community
 - Data Management/Archive
 - Data Access
- To define how to achieve this mission, CUAHSI uses a community governance structure that directs staff:
 - Board of Directors
 - Standing Committee on Informatics
 - WDC Users Committee

Governance of CUAHSI



- A Board of Directors oversee all activity of the consortium
- Two committees provide guidance for the Water Data Center

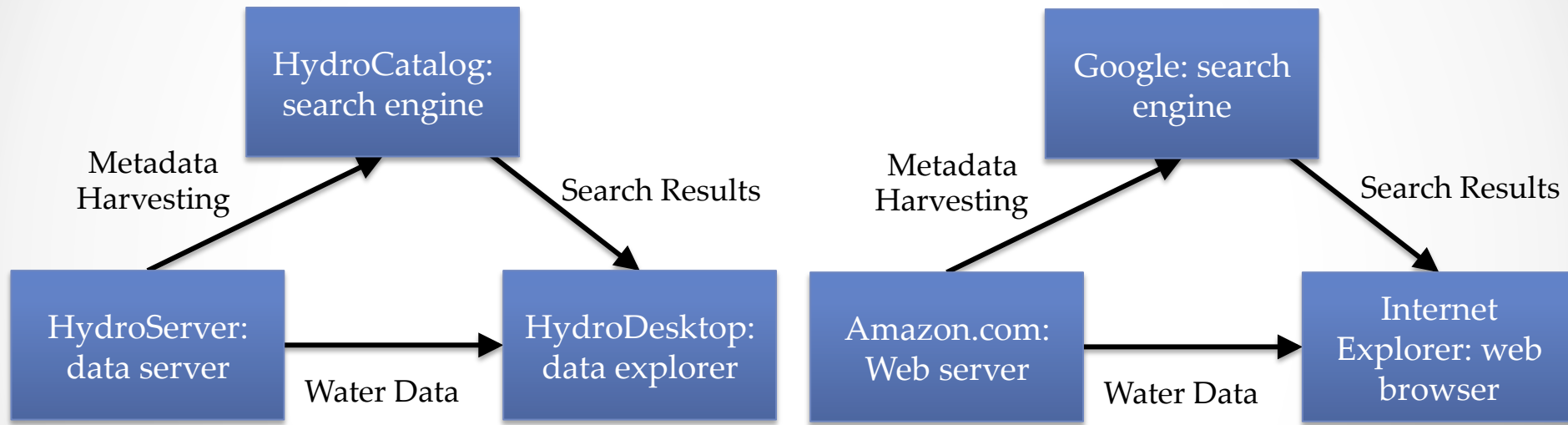
Why Share Data?

- Data re-use
- Data validation and verification
- In the U.S., Data Management Plans are a requirement for those receiving grant money from the National Science Foundation
 - Most data must eventually become public

Who is using our services and software?

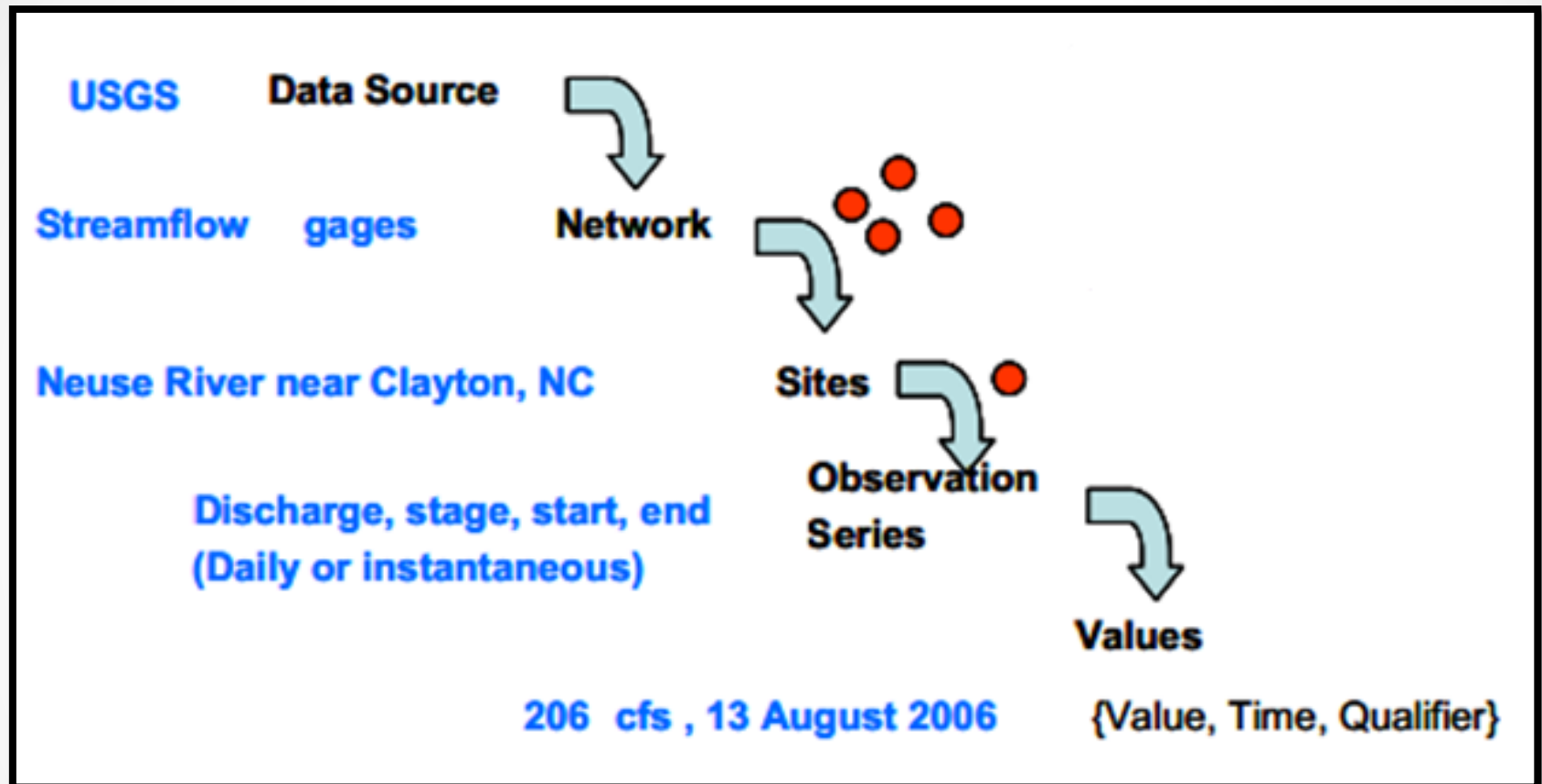
- U.S. National Science Foundation-funded research
 - Shale Network → Research Coordination Network evaluating impacts of hydraulic fracturing on water resources
 - Critical Zone Observatories
 - Global Lakes Ecological Observatory Network (GLEON)
- U.S. Local government and watershed associations
- Italian Environmental Protection Agency (ISPRA)

The Hydrologic Information System (HIS)



“Services-Oriented Architecture” ...A web service is software that transmits data over the internet

Time Series Data



A unique time series is defined by:

- **Variable:** What is being measured
- **Site:** Location where measurement is made
- **Method:** How the measurement is made
- **Source:** Who made the measurement
- **Quality Control:** Has the data been reviewed for errors and/or modified?

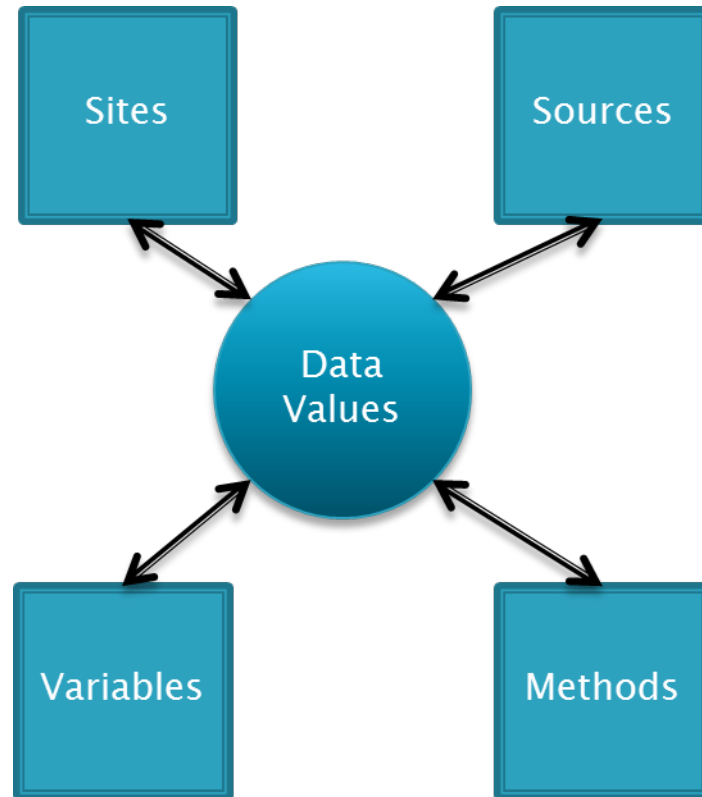
Standards

- Information Model
- Semantics
- WaterML



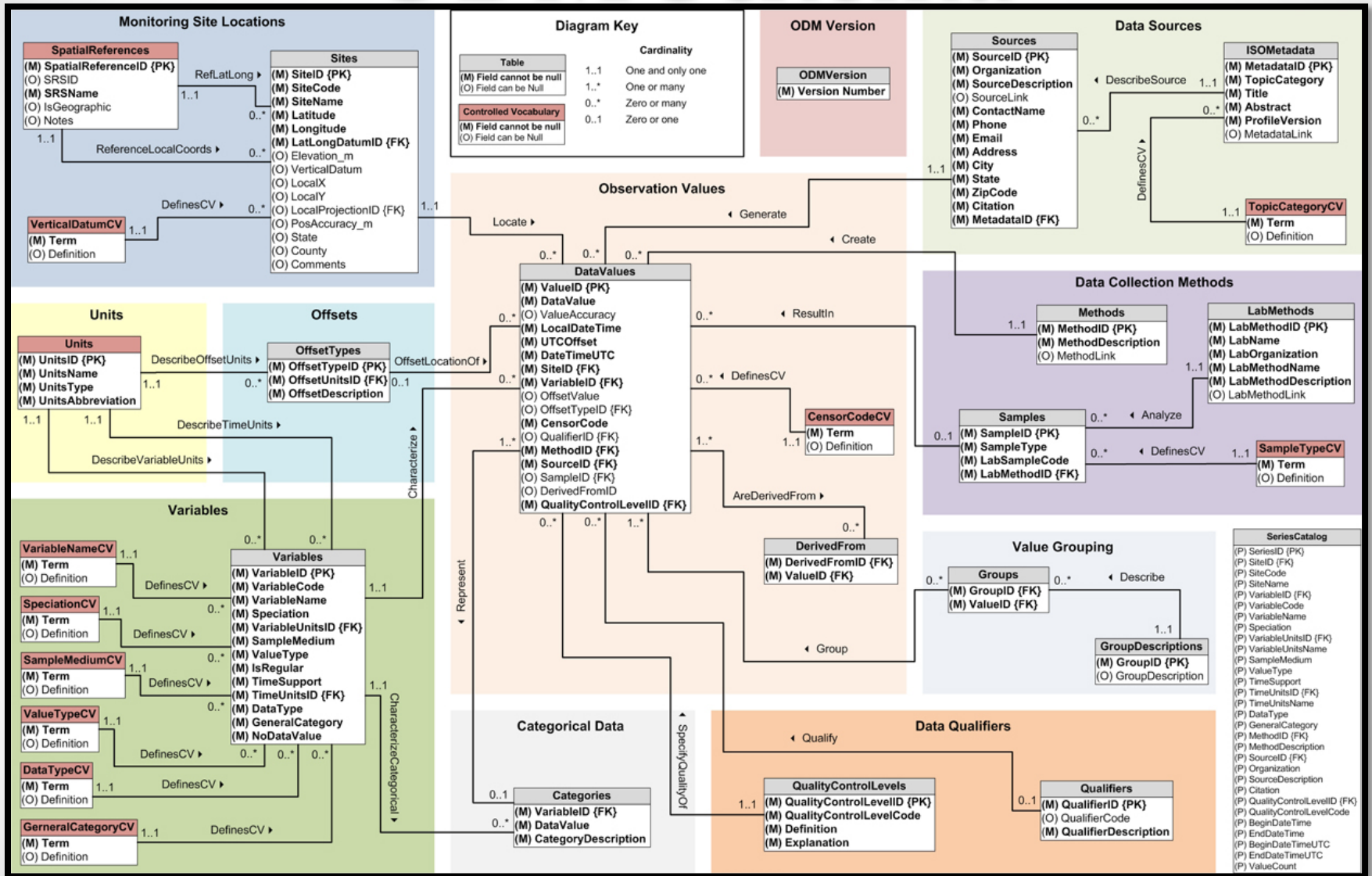
Our Information Model:

The Observations Data Model (ODM)



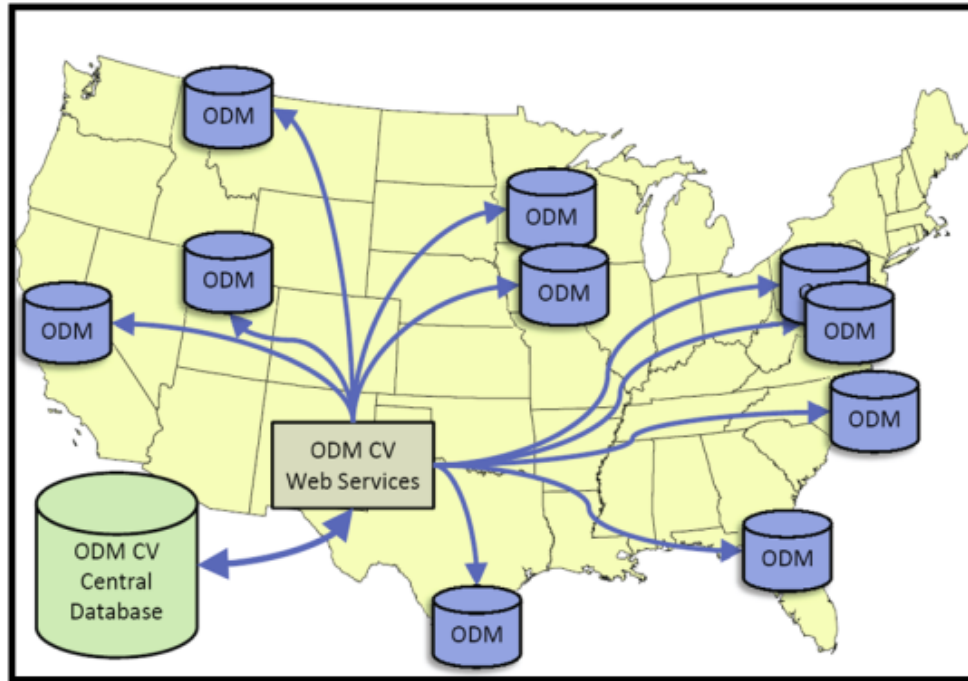
Above: Generalization of ODM1

ODM Schema



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

Controlled Vocabularies



Variable Name

Investigator 1:	"Temperature, water"
Investigator 2:	"Water Temperature"
Investigator 3:	"Temperature"
Investigator 4:	"Temp."

Controlled Term

...
Sunshine duration
Temperature
Turbidity
...

WaterML

- XML for transmitting time series water data
- Becoming a global standard
 - Adopted by Open Geospatial Consortium (OGC)
 - Under consideration by the World Meteorological Organization (WMO)

```
<timeSeries>
- <sourceInfo xsi:type="SiteInfoType">
  <siteName>Colorado Rv at Austin, TX</siteName>
  <siteCode network="NWIS" siteID="4619631">0815800</siteCode>
- <geoLocation>
  - <geogLocation xsi:type="LatLonPointType" srs="EPSG" >
    <latitude>30.24465429</latitude>
    <longitude>-97.694448</longitude>
  </geogLocation>
</geoLocation>
</sourceInfo>
- <variable>
  <variableCode vocabulary="NWIS" default="true" variableCode="0815800" >0815800</variableCode>
  <variableName>Discharge, cubic feet per second</variableName>
  <units unitsAbbreviation="cfs" unitsCode="35">cubic feet</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>
```

This is the standard that makes the HIS interoperable with other systems!

Above: Screenshot of WaterML1

Cloud Data Storage

- The CUAHSI Water Data Center is a “virtual data center”
- We own very little infrastructure relative to the amount of data in our catalog
- Employing cloud technology
 - Scalability and reliability
 - Cost is calculated based on usage
 - We have chosen to use the Microsoft Azure Cloud

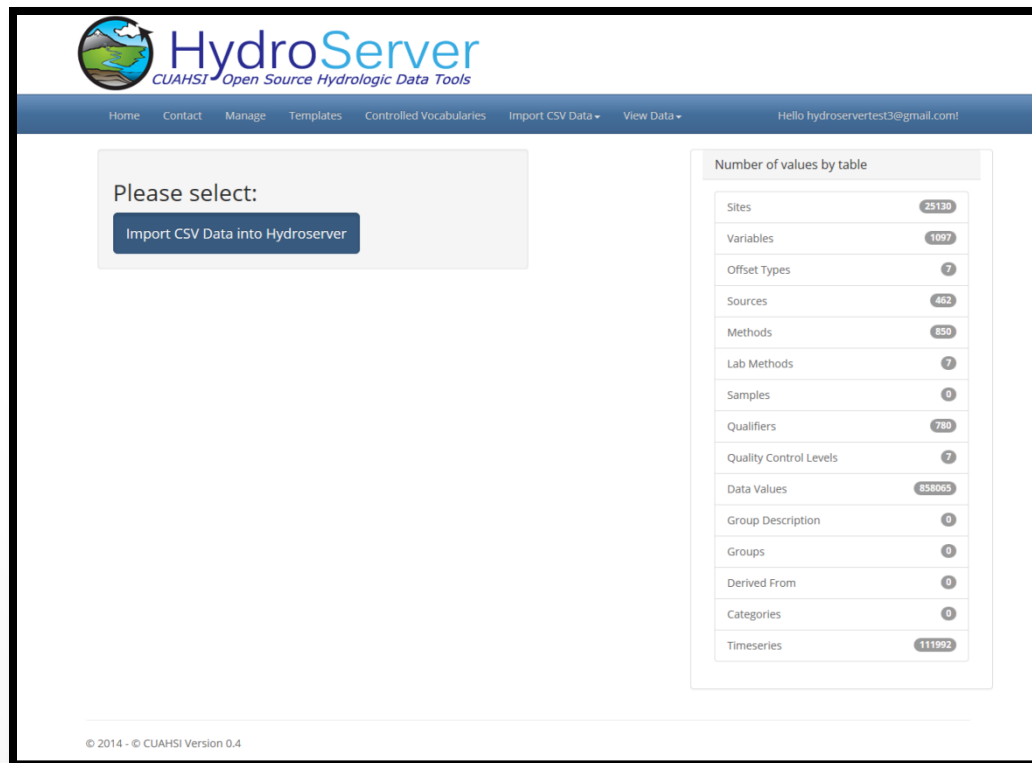


Data Access

The screenshot displays the CUAHSI HydroDesktop software interface. The title bar reads "CUAHSI HydroDesktop - World Map.dspix*". The main menu includes File, Map, Search, Table, Graph, Edit, HydroR, and Help. The toolbar contains icons for Current View, Select Features, Draw Rectangle, Search Area, Keyword, Time Range, Data Sources, Search, Show Attribute Table, Show Map Popups, Download Settings, and Download Selected. The Search field is set to "All", and the Time Range is from 7/19/2013 to 7/29/2013. The Legend on the left shows Map Layers with checked options for Rivers, Online Basemap, Lakes, and Countries. The map displays a topographic view of North Vietnam, with major cities like Hanoi, Haiphong, and Nam Dinh labeled. A blue line represents a river network, and a magnifying glass icon is positioned over the Phu Ly area. The status bar at the bottom shows the current coordinates: Longitude: 106°01'49"E, Latitude: 20°34'59"N, and "No Layer Selected".

Data Publication and Archive

- HydroServer: Our software for publishing water data
 - Includes: Database, web service, data management tools
- We have deployed this software to the Azure Cloud



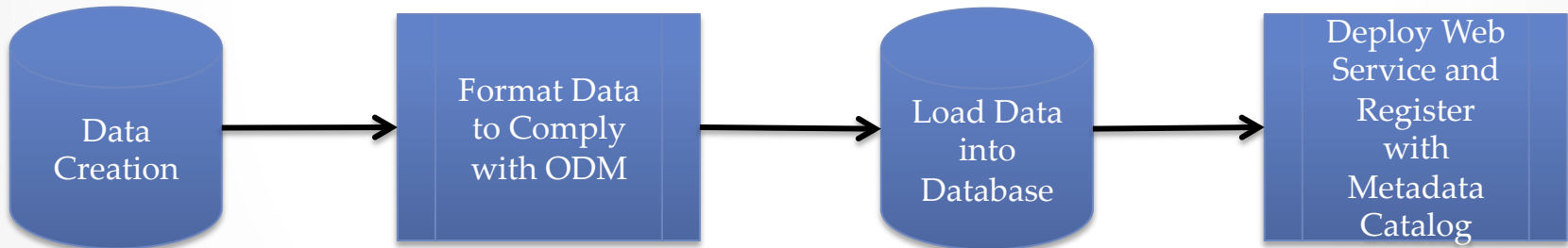
The screenshot displays the HydroServer web interface. At the top, there is a navigation bar with the following links: Home, Contact, Manage, Templates, Controlled Vocabularies, Import CSV Data, and View Data. The user is logged in as 'Hello hydroservertest3@gmail.com'. The main content area is divided into two sections. On the left, there is a 'Please select:' prompt with a button labeled 'Import CSV Data into Hydroserver'. On the right, there is a table titled 'Number of values by table' which lists various data tables and their corresponding values.

Number of values by table	
Sites	25120
Variables	1097
Offset Types	7
Sources	462
Methods	830
Lab Methods	7
Samples	0
Qualifiers	780
Quality Control Levels	7
Data Values	838003
Group Description	0
Groups	0
Derived From	0
Categories	0
Timeseries	111992

© 2014 - © CUAHSI Version 0.4

The WDC Provides...

- Adherence to Standards
- Cloud Storage
- We are funded to deal with issues of long term data persistence



Above: Workflow for publishing and archiving data with the WDC.

HIS Central

hiscentral.cuahsi.org/pub_network.aspx?n=52

 **CUAHSI HIS**
Sharing hydrologic data

Home All Data Services [Login](#) [Register](#)

Little Bear River Experimental Watershed, Northern Utah, USA

Utah Water Research Laboratory, Utah State University

LittleBearRiver

Utah State UNIVERSITY

WaterML Service:
http://icewater.usu.edu/littlebearriver/cuahsi_1_1.asmx?WSDL

WFS Service:
<http://hiscentral.cuahsi.org/WFS/52/cuahsi.wfs?request=getCapabilities>

Contact: Jeff Horsburgh
jeff.horsburgh@usu.edu
435-797-2946

Service Statistics:			
Sites:	16	Geographic Extent:	41.71847
Variables:	61		-111.9464 -111.7993
Values:	23288662		41.49541

Last Harvested on 11/30/2013 12:15:50 PM
(updated weekly, assumed current)

Abstract

Utah State University is conducting continuous monitoring within the Little Bear River watershed of northern Utah, USA to investigate the use of surrogate measures such as turbidity in creating high frequency load estimates for constituents that cannot be measured continuously.

Keywords:

Discharge, Water Quality, Pollutant Loads, Continuous Data, Surrogate Measures, Oxygen Dynamics, Hydrochemical Response

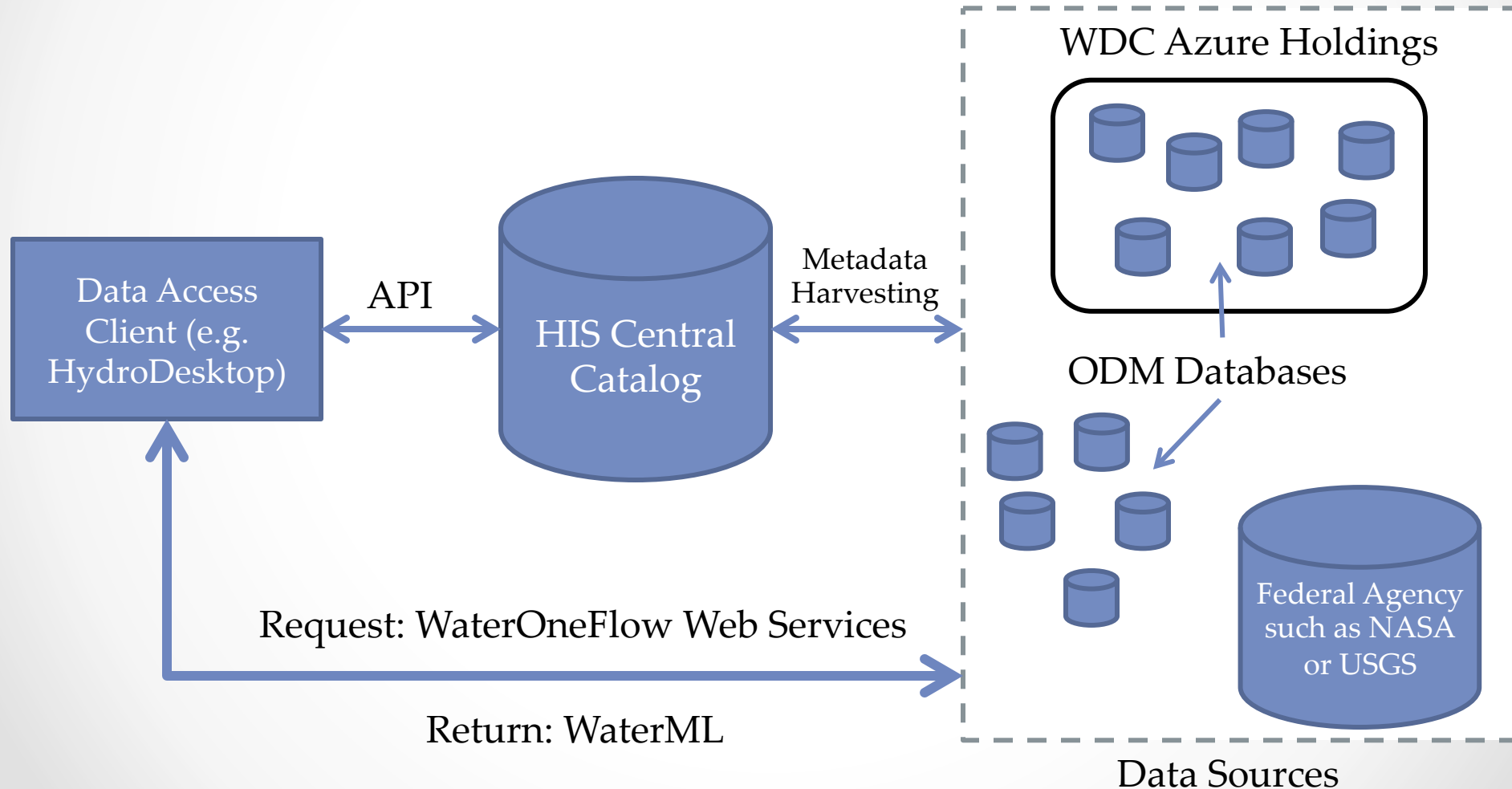


Citation

Horsburgh, J. S., D. K. Stevens, D. G. Tarboton, N. O. Mesner, A. Spackman Jones, and S. Gurrero (2009) Monitoring data collected within the Little Bear River Experimental Watershed, Utah, USA, Utah State University.

- Metadata database and software that facilitates search and discovery
- Public webpage for each data source
- Deployed in the Azure Cloud

A Distributed Ecosystem



Open Data Access & Open Source Development

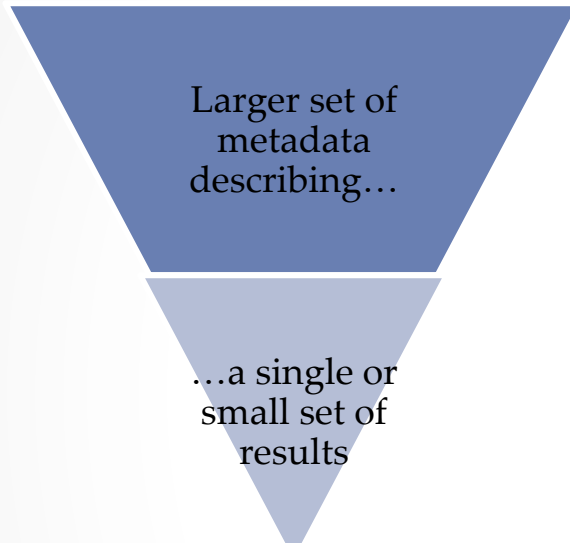
- All data published with the WDC is openly available
- All software developed is posted online
 - www.github.com/cuahsi
 - www.hydrodesktop.org

Data and Metadata Standardization

- Training researchers to follow best practices
 - Providing tools to ease this burden
- Developing and maintaining approaches to cross-domain data discovery such as controlled vocabularies

Continuous Data versus Discrete Data

Discrete

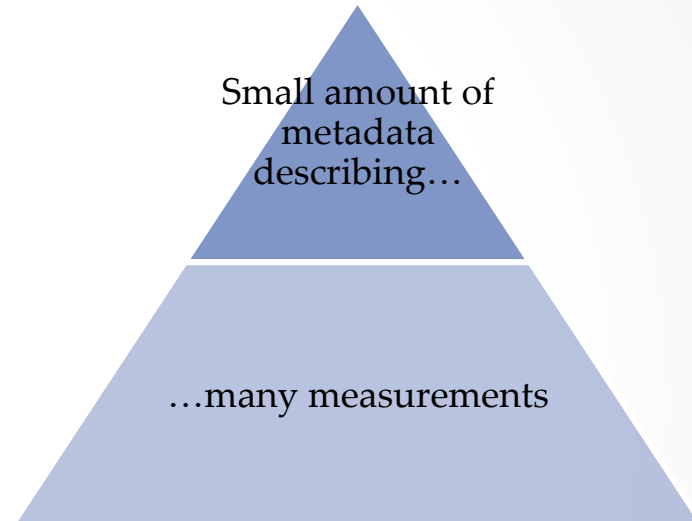


Discrete Monitoring

- A **sample** is taken and sent to a lab for further analysis
- Typically a one-time event that can be repeated as needed



Continuous



Continuous Monitoring

- A **sensor** is used to record a continuous stream of data about 1 particular analyte or a small set of analytes (i.e. flow, dissolved oxygen, pH, etc).
- Values are reported at set intervals (i.e. every 15 minutes, 1 hour, etc.)



Data Storage and Retrieval

- What does one need to consider?
 - **High reliability:** the data source can't stop working.
 - **High persistence:** when you put something there, it should stay there.
 - **High availability:** the data source must handle multiple simultaneous uses.
 - **Quick recovery:** in the unlikely event of a failure, data must become available again quickly.

Water Data Presents Unique Challenges

- 4000+ identifiable kinds of information (and growing).
- Interdisciplinarity and data reusability: data is used by several distinct disciplines.
- Dynamic data: much data is real-time.
- Data use for different purposes:
 - Scientific inquiry
 - Decision support
 - Education and outreach

Some Conclusions

- There are many existing challenges for sharing water data
 - Cultural
 - Technological
- There are existing (free!) tools and standards that could provide a starting point for sharing data within the Lower Mekong Region

Thank you!



- CUAHSI Website: www.cuahsi.org
- My email: jpollak@cuahsi.org
- CUAHSI WDC Github: www.github.com/cuahsi

