Toward Integrated Environmental Modeling Using Research Data Infrastructures

Jeffery S. Horsburgh

Civil and Environmental Engineering
Utah State University

UtahState University

Utah Water Research Laboratory
UWRL

The Motivators

(from the perspective of an Environmental Engineer interested in water)

- Uncertain climate, land use change, population growth, will alter hydrology, water availability, and water quality
- We need better information to manage water resources
  - But the questions are now transdiciplinary
  - The next generation of models will look very different
- Significant challenges in integrated modeling
  - Coupling models/model components from different disciplines
  - Matching data to required model inputs and outputs at spatial and temporal scales relevant to decision making

Research Data Infrastructures

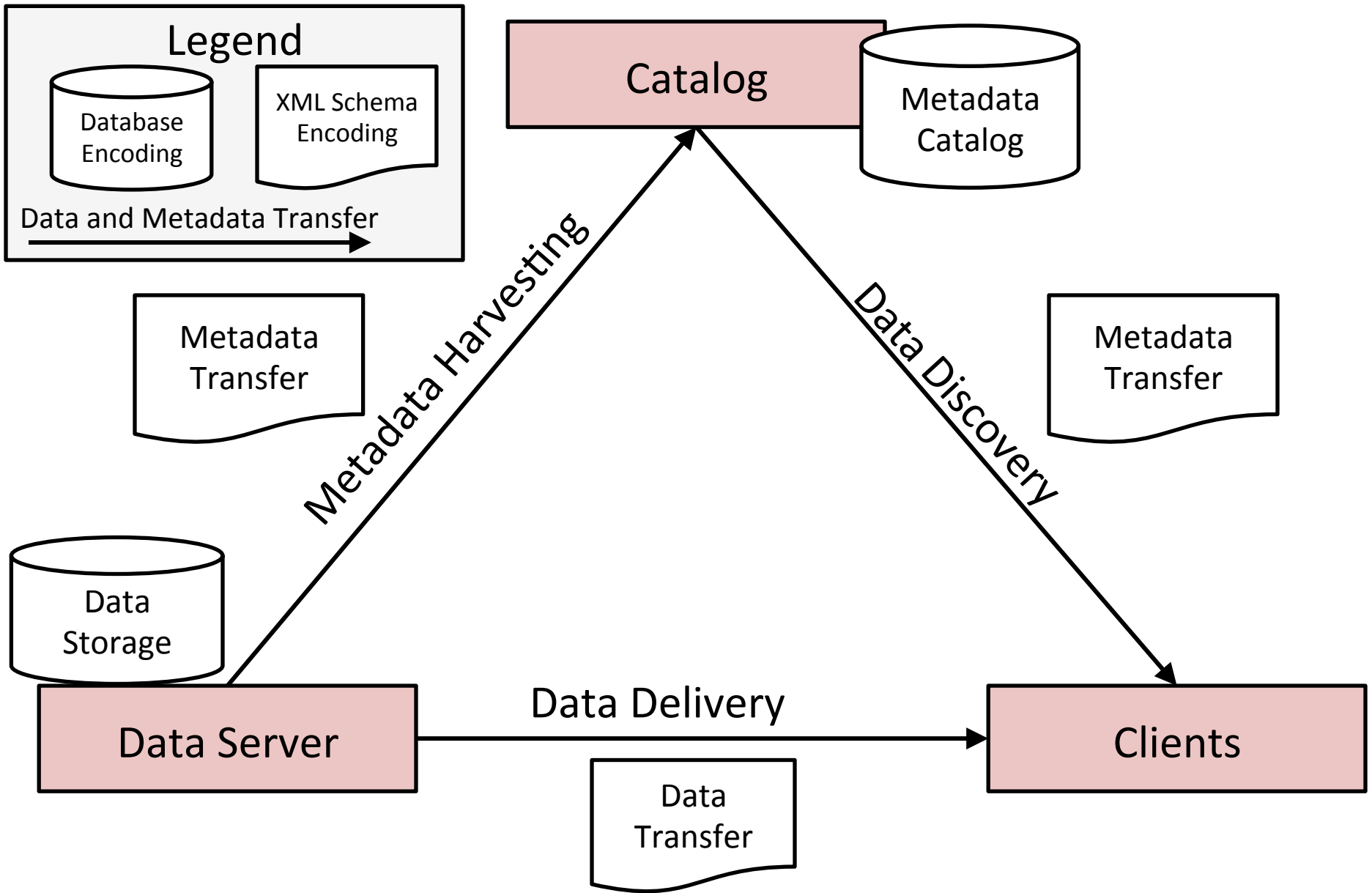
- Research data infrastructures in the U.S. are still disciplinary:
  - Hydrology Community: CUAHSI Hydrologic Information System and Water Data Center
  - Ecological Community: DataONE, Knowledge Network for Biocomplexity (KNB)
  - Critical Zone Community: Critical Zone Observatory Integrated Data System (CZOData)
  - Ocean Community: Integrated Ocean Observing System (IOOS)
  - …

And, it's not uncommon to find hydrologic data in all of these systems!

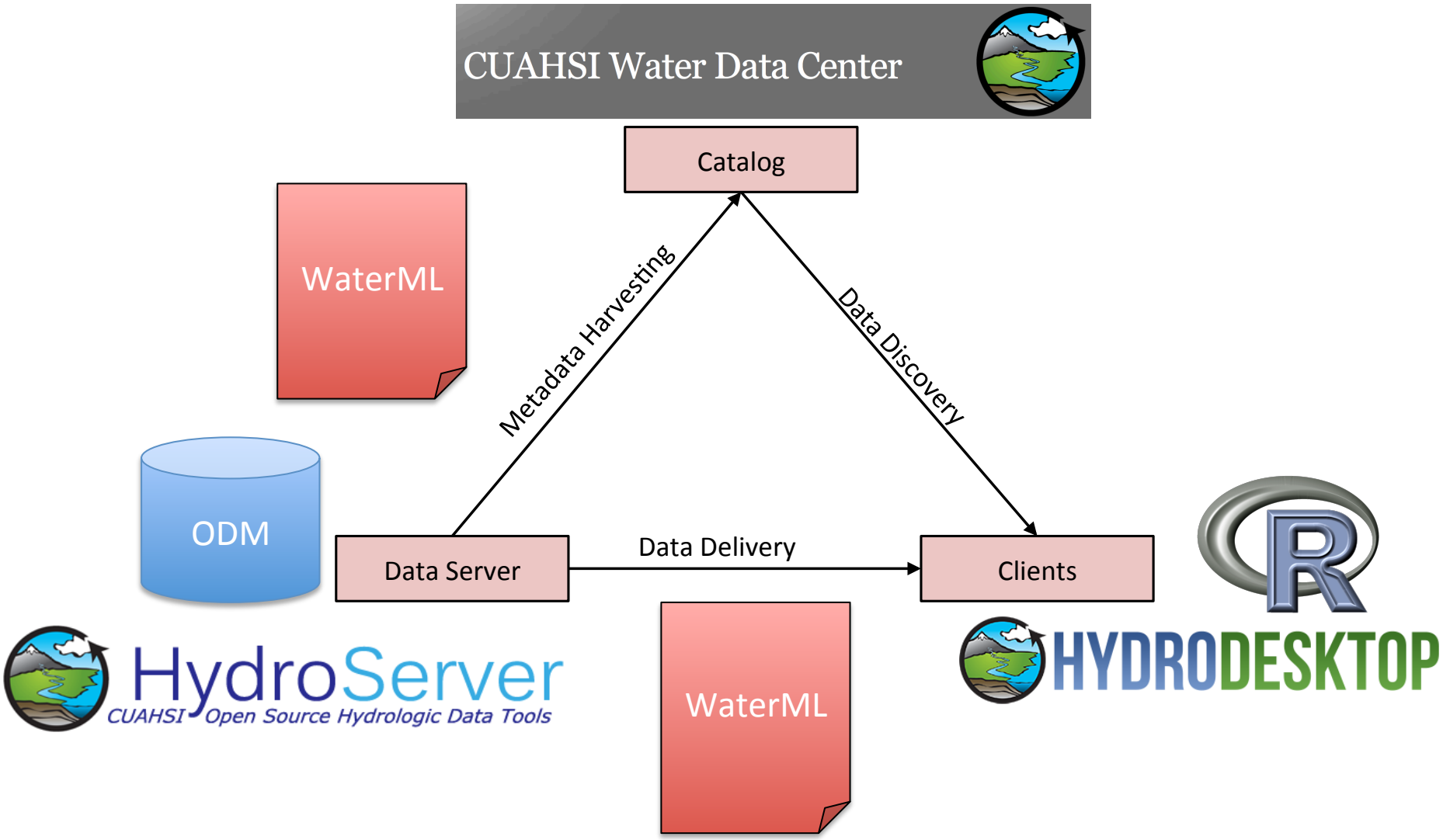
Research Data Infrastructures

- Built using principles of service-oriented architecture (SOA)
- Rely on standard data encodings – but domain specific
- In some cases rely on standard semantics – but not across systems and domains
- Focused on publishing or sharing data on the Internet
  - Common functions: metadata harvest, catalog, search/discovery, data access

Services Oriented Architecture



CUAHSI
HIS
Sharing hydrologic data



DataONE

Coordinating Nodes

Multiple metadata standards

EML

ISO 19115

Dublin Core

FGDC

...

Catalog

ONEMercury

Metadata Harvesting

Data Discovery

Data Server

Data Delivery

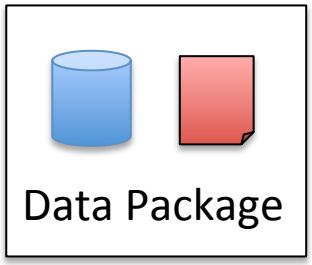
Clients

ONE R

Kepler

Member Nodes

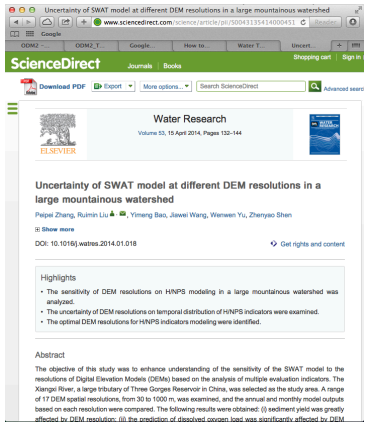
No restrictions on data format(s)



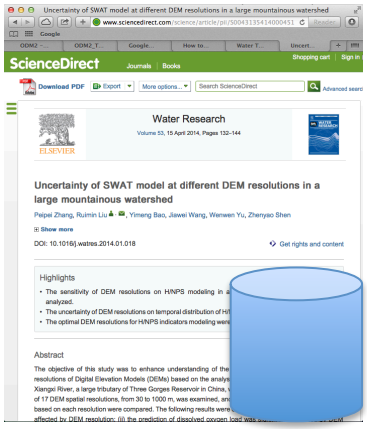
Investigator Toolkit

python

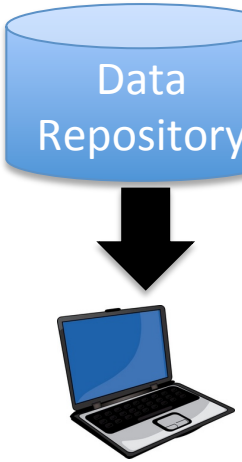
Reproducible Science Versus Integrated Modeling



Search for model application papers



Paper cites data source

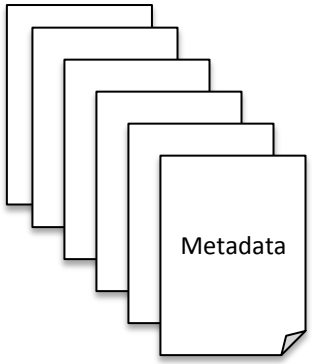


Download data and run the model

Reproducible Science Versus Integrated Modeling

Data Repository X

Query the data repository for "Streamflow"



Query returns 1000 metadata results

IEDA
INTEGRATED EARTH DATA APPLICATIONS

EarthChem

figshare
credit for all your research

WQP
Water Quality Data

CZO | CRITICAL ZONE OBSERVATORIES
U.S. NSF NATIONAL PROGRAM

DataONE
Data Observation Network for Earth

CUAHSI
HIS
Sharing hydrologic data

USGS
science for a changing world
National Water Information System

Reproducible Science Versus Integrated Modeling

- Sharing and publication is important for reproducible science
  - Supported well by existing research data infrastructures
- But, the next scientist/modeler looking for data to perform a different analysis
  - Has trouble finding what they need
  - Data are site/study specific and "outsiders" struggle to determine if the data are appropriate for their use
  - Many models require data at spatial and temporal scales not found in research data infrastructures

Challenges

- It takes a “data savvy” investigator considerable effort to discover and access datasets from multiple repositories for a synthetic analysis
- There is a shortage of “data savvy” investigators
- There is currently a mismatch between the functionality of research data infrastructures and what is required for integrated modeling

A Hypothetical Query

I am interested in modeling the effects of aquatic nutrient concentrations on stream metabolism.

"Show me locations where high frequency stream discharge, water temperature, and dissolved oxygen data have been collected in third order streams for which samples of nitrogen and phosphorus have also been collected and that are within one mile of a weather station that measures solar radiation for the same time period."

Spatial Context

"Show me **locations** where high frequency stream discharge, water temperature, and dissolved oxygen data have been collected in **third order streams** for which samples of nitrogen and phosphorus have been collected and that are **within one mile of a weather station** that measures solar radiation for the same time period."

Temporal Context

"Show me locations where **high frequency** stream discharge, water temperature, and dissolved oxygen data have been collected in third order streams for which samples of nitrogen and phosphorus have been collected and that are within one mile of a weather station that measures solar radiation **for the same time period**."

Measured Variables

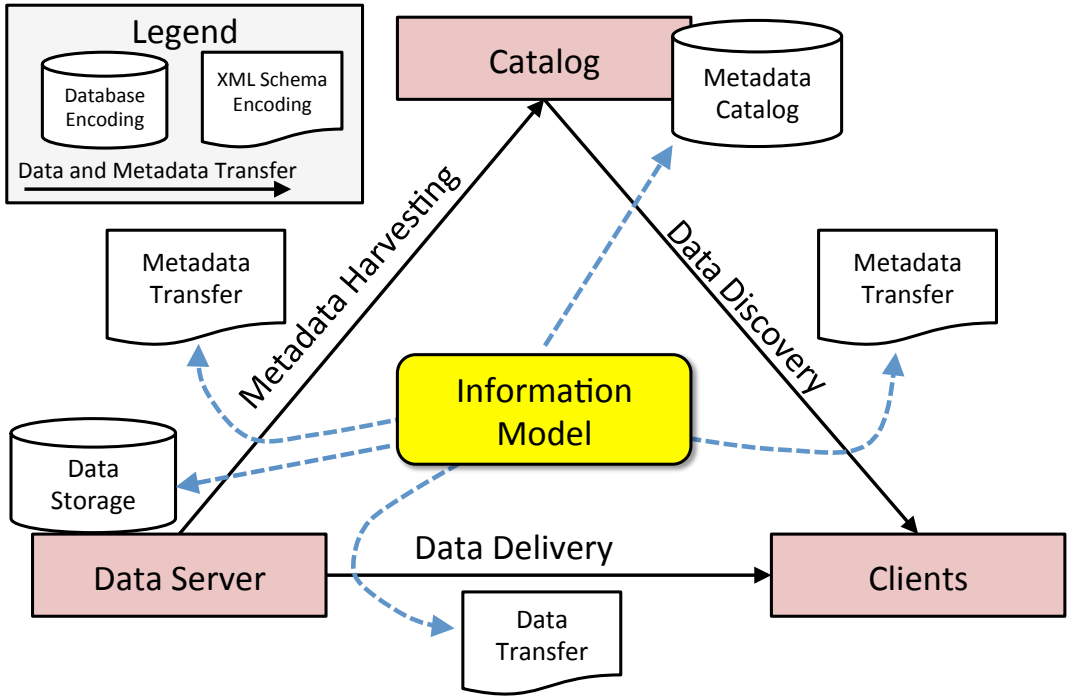
"Show me locations where high frequency **stream discharge, water temperature, and dissolved oxygen data** have been collected in third order streams for which samples of **nitrogen and phosphorus** have been collected and that are within one mile of a weather station that measures **solar radiation** for the same time period."

Toward Integrated Environmental Modeling Using Research Data Infrastructures

- A vision for next generation research data infrastructures
  - An integrated view of the best available "data" that holistically represent the earth's hydrologic and environmental systems
  - Seamless presentation of and access to data across existing earth science data repositories
- What is needed:
  - Common information modeling
  - Linking data to the geo-environment
  - Capturing the knowledge content of data
  - Integration with government and agency repositories

Common Information Modeling

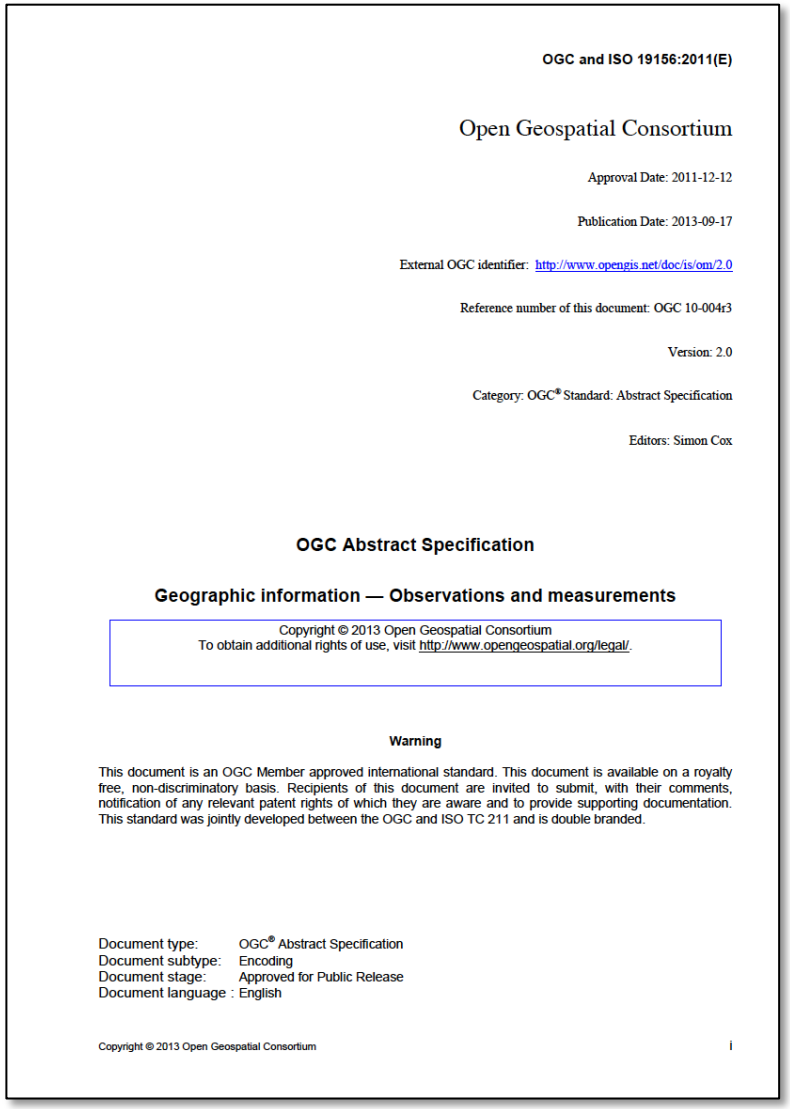
- Challenges:
  - Geoscience cyberinfrastructures represent common informational elements inconsistently
  - Opportunities for standardized representation of **spatial**, **temporal**, and **measured variable** contexts
  - Semantic and syntactic heterogeneity are major hurdles!



<u>Information Model</u>: Agreement about information needed to describe observational data

Observation Information Models

OGC's Observations & Measurements



OGC and ISO 19156:2011(E)

Open Geospatial Consortium

Approval Date: 2011-12-12

Publication Date: 2013-09-17

External OGC identifier: http://www.opengis.net/doc/is/om/2.0

Reference number of this document: OGC 10-004r3

Version: 2.0

Category: OGC® Standard: Abstract Specification

Editors: Simon Cox

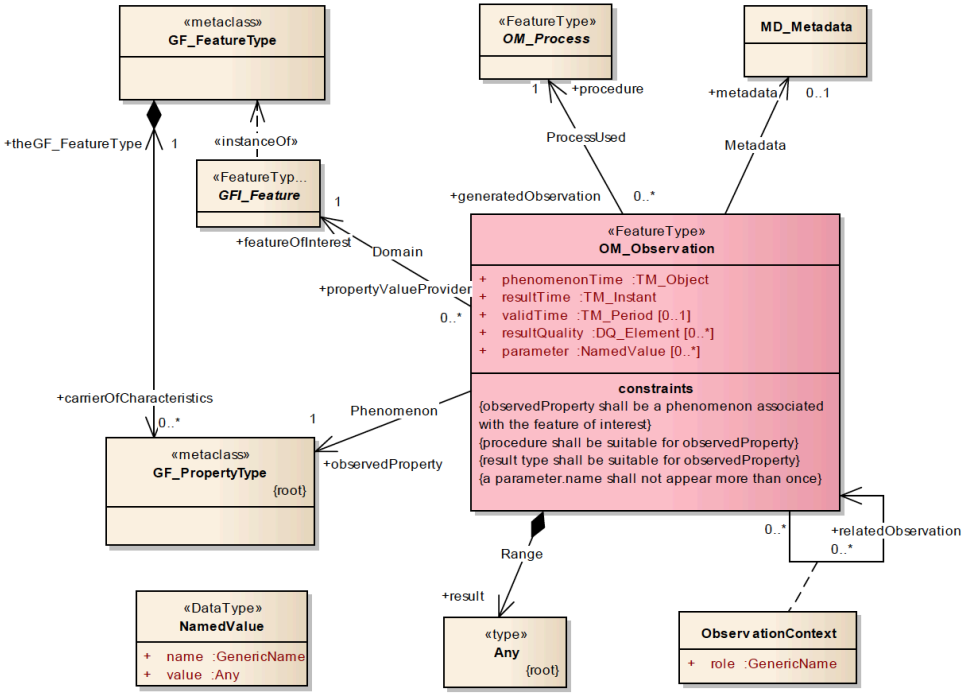
OGC Abstract Specification

Geographic information — Observations and measurements

Warning

This document is an OGC Member approved international standard. This document is available on a royalty free, non-discriminatory basis. Recipients of this document are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation. This standard was jointly developed between the OGC and ISO TC 211 and is double branded.

Document type: OGC® Abstract Specification
Document subtype: Encoding
Document stage: Approved for Public Release
Document language : English



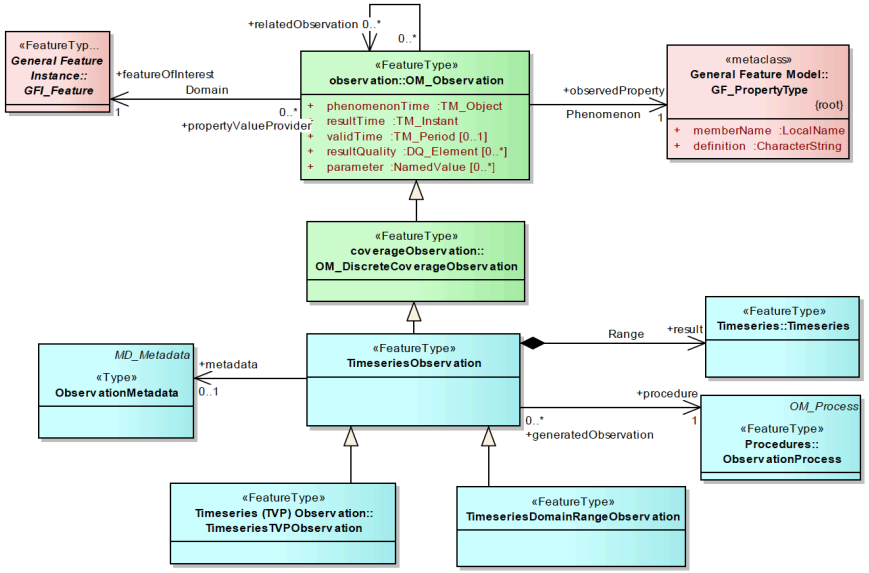
Observation

- Sampling Feature
- Feature of Interest
- Procedure
- Phenomenon/observed property
- Result

In Practice: Profile!

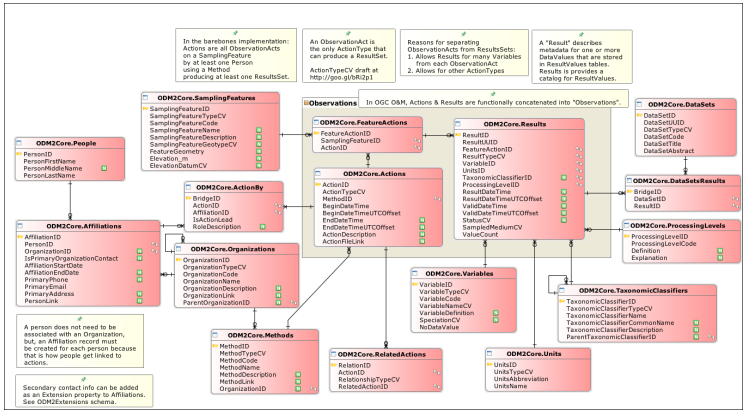
WaterML 2.0

- International standard for time series of water observations data



ODM2

- Additional result types beyond time series
- Water quality and solid earth samples, sections, transects, profiles



Common Semantics for Earth Observations

CUAHSI HIS | EarthChem | CZOData | IOOS

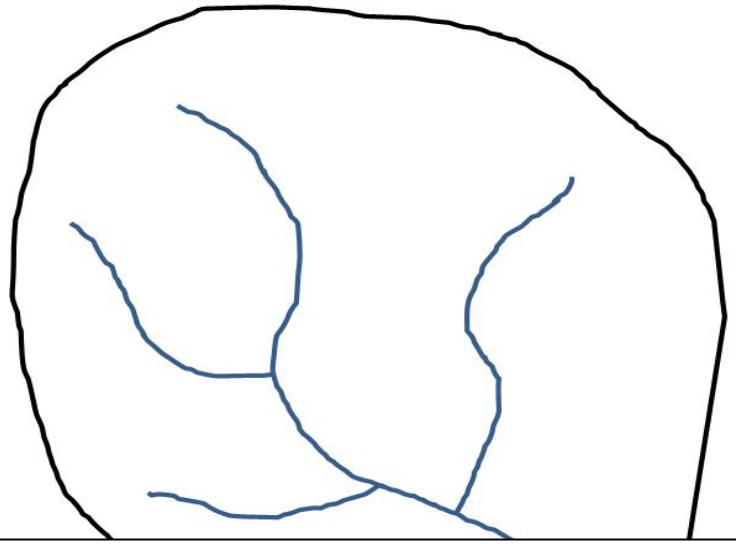
Linking Data to the Geo-Environment
(aka – Spatial Context)

- Feature of Interest – a real world feature whose properties are observed
- Geospatial support – the area or volume to which an observation applies
- Geospatial context – the relationships between a feature of interest and its surrounding features

Linking Data to the Geo-Environment

An Example from Hydrology

Geographic Features



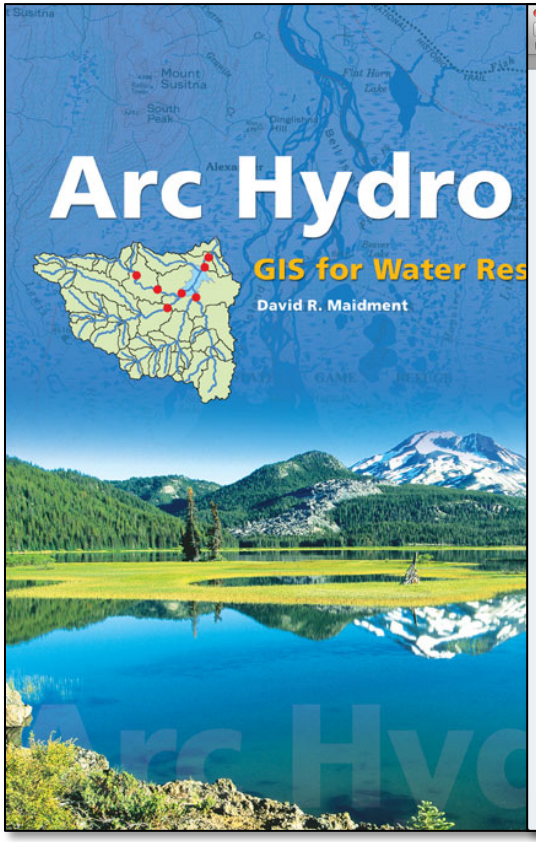
Observations			
1	1/1/2012	Streamflow	100 cfs
2	1/2/2012	Streamflow	101 cfs
3	1/3/2012	Streamflow	102 cfs

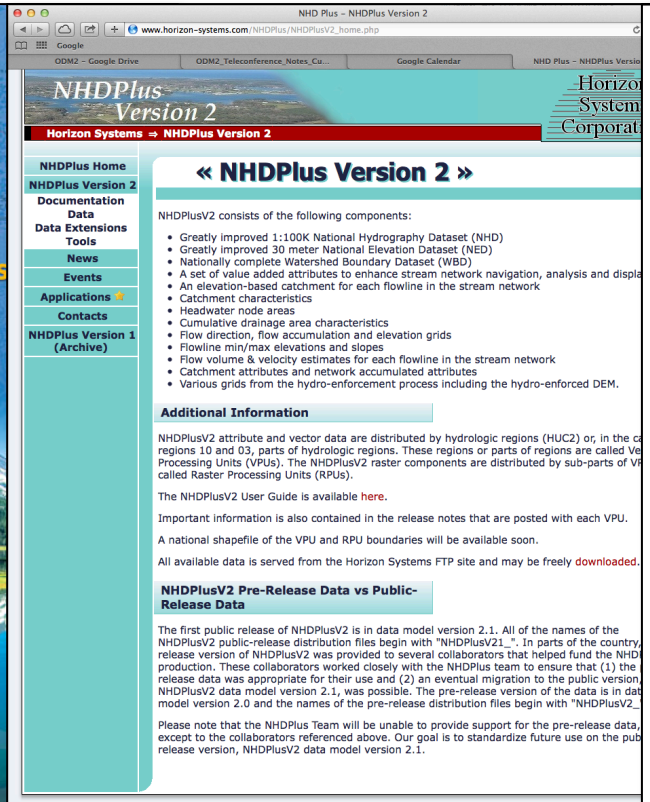
Sampling Feature	
1	Stream Gage

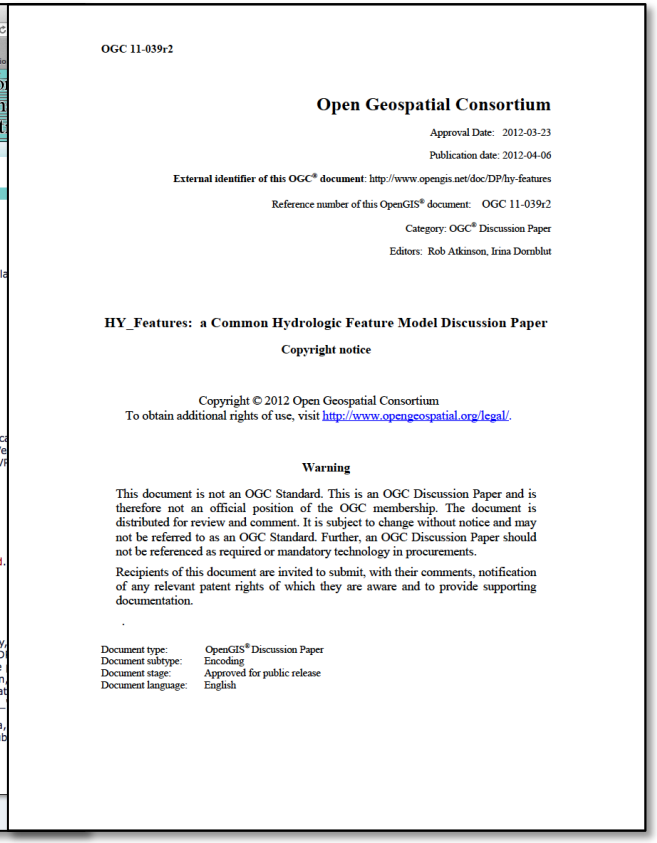
Feature of Interest		
1	Cross Section	Point
2	Stream	Line
3	Catchment	Polygon

Currently, data users must determine that the gage lies on a particular river, measures the outflow of a particular catchment, is downstream of another gage, shares its location with a water quality monitoring site, and is located near a weather station.

Linking Data to the Geo-Environment






OGC 11-039r2

Open Geospatial Consortium

Approval Date: 2012-03-23

Publication date: 2012-04-06

External identifier of this OGC® document: http://www.opengis.net/doc/DP/hy-features

Reference number of this OpenGIS® document: OGC 11-039r2

Category: OGC® Discussion Paper

Editors: Rob Atkinson, Irina Dornblut

HY_Features: a Common Hydrologic Feature Model Discussion Paper

Copyright notice

Copyright © 2012 Open Geospatial Consortium
To obtain additional rights of use, visit http://www.opengeospatial.org/legal/.

Warning

This document is not an OGC Standard. This is an OGC Discussion Paper and is therefore not an official position of the OGC membership. The document is distributed for review and comment. It is subject to change without notice and may not be referred to as an OGC Standard. Further, an OGC Discussion Paper should not be referenced as required or mandatory technology in procurements.

Recipients of this document are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

Document type: OpenGIS® Discussion Paper
Document subtype: Encoding
Document stage: Approved for public release
Document language: English

Much of this work has been done already!

Capturing the Knowledge Content of Data

- What do the data mean?
- How have they been used?
- What conclusions have been drawn?
- What are appropriate uses?

Linking data to papers is great, but only a small fraction of data end up in papers!

Capturing the Knowledge Content of Data

Data Annotation

Comments

Amber Jones 9 minutes ago

These data characterize diurnal fluctuations in temperature in the Little Bear River at its terminus just upstream of Cutler Reservoir. At its lower end, the hydrology of the Little Bear River has been highly modified, with streamflows dominated by releases from Hyrum Reservoir and with several agricultural diversions and return flows. These temperature observations reflect a high degree of human modification within what was typically a spring snowmelt dominated system.

Link | Reply

Liked by 0 +1

Tony Melcher 6 minutes ago

I used these data in a fish habitat assessment for the Little Bear River. They demonstrate the diurnal variability in water temperature, and I was able to use them to assess potential acute and chronic effects of elevated stream temperatures on salmonid fish species.

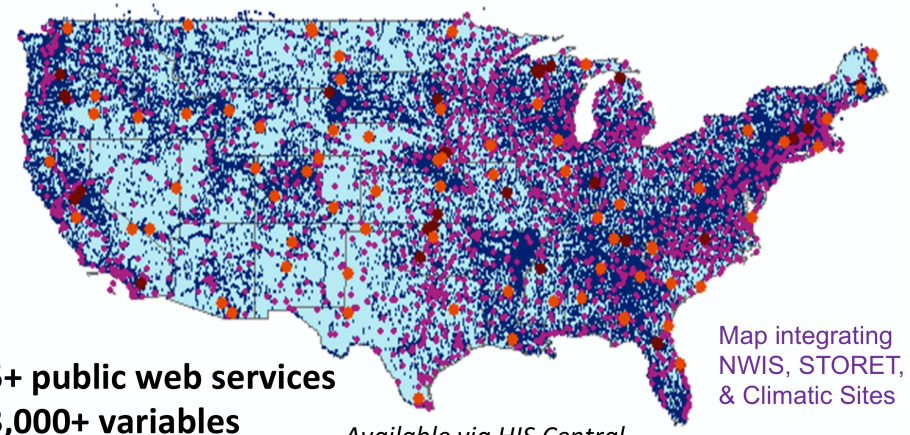
Link | Reply

Liked by 0 +1

Integration with Government Agency Data Repositories

- Many models require data at spatial and temporal scales not found in research data infrastructures
- For some science domains – there is far more data in agency repositories

HIS Central HydroCatalog Content



Map integrating NWIS, STORET, & Climatic Sites

65+ public web services
13,000+ variables
1.96+ million sites

Available via HIS Central discovery services

23.3 million observation time series
Referencing 5+ billion data values

Available via GetValues requests

Metadata for most services are harvested weekly

Summary

- Speeding innovations in synthesis and modeling in the geosciences will require:
  - New CI techniques and tools
  - A work force capable of developing and using these tools
- Enhancing the discoverability, accessibility, and use of data in research repositories for modeling:
  - Common information modeling – standards promote interoperability
  - Linking data to the geo-environment for spatial context
  - Capturing the knowledge content of data

Questions?

Jeffery S. Horsburgh
Civil and Environmental Engineering
Utah State University
jeff.horsburgh@usu.edu

Utah State University

Utah Water Research Laboratory
UWRL