CUAHSI Profile of OGC services for Time Series

By David Maidment, Ilya Zaslavsky, and David Valentine,

Presented at the US Water Summit, OGC Technical Committee Meeting,
Boulder, CO, 20 September 2011

Introduction

The Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) is an organization representing more than 120 US universities that is supported by the National Science Foundation to advance hydrologic science through development of a Hydrologic Information System (HIS). This project has been operating since 2004, and it defined a language, WaterML, or Water Markup Language, to convey water observation time series through the internet.

The backbone of the HIS service-oriented architecture design (Figure 1) is a set of standard web service application programming interfaces that define interactions between hydrologic data publication platform (HydroServer), the data cataloguing and discovery system (HydroCatalog) and client applications, such as HydroDesktop. The key standards used in the current operational version of HIS are WaterML 1 and WaterOneFlow services. These specifications have been designed to unify hydrologic data discovery and access for academic data sources that store data in CUAHSI Observations Data Model (ODM) and large federal and state repositories (e.g. maintained by USGS, EPA, NCDC) that follow their own storage, metadata and access conventions. To date, CUAHSI HIS has incorporated 70 observation networks from government and academic publishers, with over 2 million observation locations, 18,000 variables, 23.3 million time series, referencing 5.2 billion data points. The system is constantly expanding as new networks are registered.

To establish a higher level of compatibility between a wider group of water data sources, at the national and international scales, and to take advantage of multiple third-party software applications, the CUAHSI HIS services oriented architecture is now transforming its key interfaces to be compatible with OGC standards. Another key advantage of this transition is that the Open Geospatial Consortium provides transparent and community-accepted procedures and protocols for governing standards development. In 2008, CUAHSI approached OGC and the World Meteorological Organization (WMO), and these two organizations subsequently established an OGC/WMO Hydrology Domain Working Group to advance open standards development for hydrologic data. The mission of this international group of experts in standards for water data and related fields is to examine existing standards, develop standardization priorities, coordinate development of specifications, organize their testing in a series of interoperability experiments, and lead the standards to community adoption. This group has coordinated development of a new version, WaterML2, which is being proposed to OGC and WMO as an international standard for exchange of hydrologic time series information.

CUAHSI established a Concept Development Study with the OGC to examine how OGC standards could be used in this larger context of hydrologic data interoperability at the national and international levels (Bermudez and Arctur, 2011). In this study, four functions were specified: publication, cataloging, discovery, and accessing of time series and a CUAHSI Profile of OGC standards were identified to support these functions. The present summary document describes the minimal requirements for a data source to become part of hydrologic data sharing within OGC-based CUAHSI HIS infrastructure. This is a summary of a larger engineering document arising from the Concept Development Study describing migration of CUAHSI HIS components to standard OGC encodings and service interfaces.

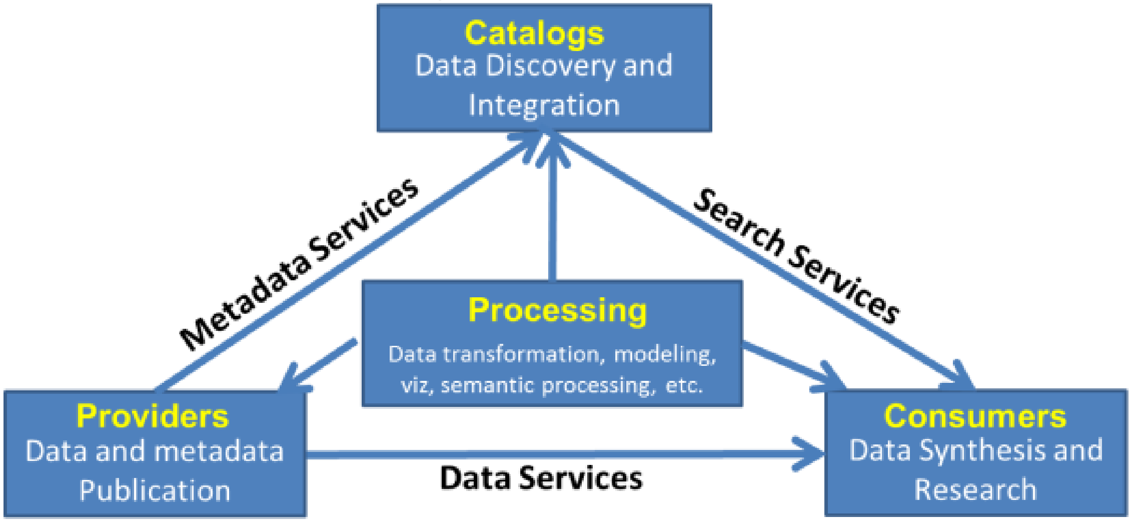
Catalogs
Data Discovery and Integration
Metadata Services
Search Services
Processing
Data transformation, modeling, viz, semantic processing, etc.
Providers
Data and metadata Publication
Data Services
Consumers
Data Synthesis and Research

Figure 1. The HIS design follows the standard "publish-find-bind" pattern, and integrates providers of data, catalog and processing services, and data synthesis and research clients.

Minimal Requirements

In order to utilize hydrologic data assembled by the CUAHSI HIS community system, and publish additional data into the system, a data provider needs to:

1. Describe the purpose of the water data collection, provide standard collection-level metadata, and specify a contact person to follow up with as needed.
2. Provide a list of observation data series from a data source, in a standard format. This is based on CUAHSI HIS experience setting a data series (location-variable-time-units-method) as the best granule for hydrologic data discovery.
3. Describe a standard method to access the data.
4. Provide mappings of variables to a standard set of parameter terms, to enable parameter-based search across multiple sources.
5. Ensure compatibility with established controlled terminology (desirable), to enable easier data interpretation and additionally search patterns.

Service Stack

The CUAHSI HIS project has demonstrated that a comprehensive description of a set of hydrologic time series can be represented by a "stack" of three OGC services presenting a catalog, metadata and data, respectively. The *catalog* is conveyed using the OGC Catalog Services for the Web (CSW), which provides standard information about the data provider and indexes one or more underlying web feature services. The *metadata* are conveyed using an OGC web feature service (WFS) with one record per time series, that record the "who" (publisher of the data), "what" (the variable being described), "where" (spatial location of the sampling site), and "when" (begin date and end date of the record). The metadata record also contains the url of the service call to obtain the time series, and if necessary the parameters for a SOAP call to obtain the time series. The *data* are conveyed using a time series web service, currently GetValues request of CUAHSI Water One Flow services which returns information in WaterML 1 format, and later using GetObservations method of a OGC Sensor Observations Service using OGC WaterML2, which is the of profile of OGC Observations and Measures This services stack is completely free-standing and can be published by any water data provider, large or small, using public, open source standards for which existing code and tools are available.

The services stack is shown in Figure 2. A client queries a CSW provider for service metadata. The client then parses the returned service records, identifies appropriate WFS services (based on spatial coverage and attribute keywords), and queries the WFS for observation series of interest, using an extended set of query filters. The client can then display the observation series on a map as WFS records. The user selects the series of interest (from a map display, from a table of series, or by applying additional filters), letting the client formulate SOS requests against the data services. The client will receive WaterML 2 time series as the output, including references to features of interest. Feature information is retrieved from WFS with site information. The time series data from the SOS/WaterML 2 response, and the feature information from the WFS/GML response, can be utilized by the client and stored in the clients' data model.

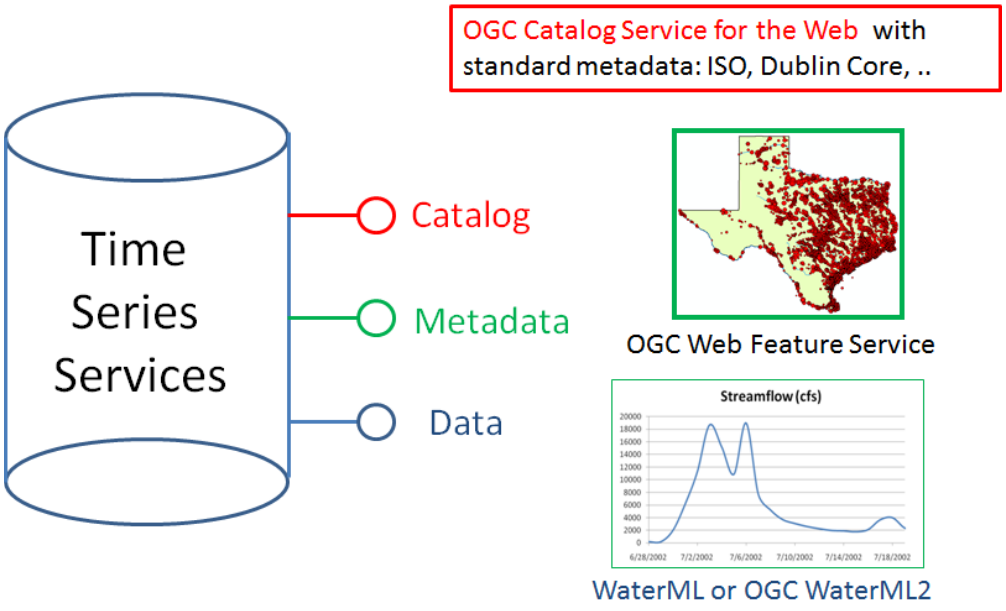

Figure 2. Services in an OGC-compliant hydrologic information system

This information model is very flexible. It can be used to catalog and describe the time series produced by a single organization, in which case all the records in the WFS metadata refer to the same information provider. It can also be used to describe a *theme*, which means a normalized and curated collection of time series of a particular kind, such as for streamflow, water temperature or bacteria, where the actual observations data may come from multiple information providers. A key point is that a single CSW address can index many WFS metadata services – at the HIS Central facility at the San Diego Supercomputer Center, approximately seventy WFS metadata descriptions of academic and national WaterML services are indexed by a single CSW address, and similarly at the Center for Research in Water Resources of the University of Texas at Austin, a single CSW address indexes many time series data sets compiled by state and local agencies in Texas.

Semantic Mediation

The services stack described is suitable for a water agency or institution such as a university to publish its water observations data. It provides *syntactic mediation*, or consistency of *format* and access mechanism, among the data sources in the sense that all are then using the same web services functions, and the XML language structures of the responses to those functions are defined in a standardized and consistent way. There is, however, a further need for *semantic mediation*, or consistency of *meaning*, among the various information sources. For example, the USGS uses the parameter code 00060 to identify streamflow discharge data, and 00065 to identify stage height data. Other organizations have different identifying numbers or names for the same data types.

One approach for addressing this issue is a *hydrologic ontology*, which is a hierarchical collection of concepts for describing hydrologic information. CUAHSI's hydrologic parameter ontology is divided into physical, chemical and biological terms, and is arranged in a tree structure of successively refined branch concepts until reaching *leaf concept*s. The central concept is named *hydrosphere*, and the streamflow discharge leaf concept is reached on the *physical* branch, the *flux* sub-branch, and the *discharge, streamflow* leaf concept. Similarly, to reach stage height data, the concept pathway is hydrosphere – physical – level – gage height, stream. Each concept pathway from the central concept to a leaf concept is unique. A single leaf concept can appear as the end point of several pathways in the event that one concept has several useful meanings. An associated table of synonyms allows for searching for information using terms not contained in the formal ontology. The CUAHSI Hydrologic Ontology tables can be seen at: http://his.cuahsi.org/ontologyfiles.html

The hydrologic ontology is incorporated within the services stack shown in Figure 2 by having an attribute field in the Metadata Web Feature Service that contains the leaf concept corresponding to the variable published by a particular organization. Using this approach, the CUAHSI HIS project has demonstrated that it is possible to search across multiple services and obtain comparable information from a distributed metadata and data system as would be obtained if all the information had been centralized in a single location.

Community Vocabularies

The HIS stack provides a controlled terminology system for controlled and common vocabulary terminology (http://his.cuahsi.org/mastercvreg.html). We encourage individuals and organizations to adopt and extend common terminology, by submitting additional terms to the CUAHSI master vocabulary registry. In the OGC stack, keyword and terminology lists can be exposed through service descriptions (GetCapabilities). The combination of service descriptions and community-managed controlled terminology will simplify discovery and access for water data from multiple organizations.

Water Themes

The combination of syntactic and semantic mediation means that a client application can search across information from multiple water information providers and derive a consistent result. Another way of presenting such information is by using *water themes*, in other words, to classify water information into particular packaged information types as is done in the GIS field. This approach has been adopted by the Texas Water Development Board, who are publishing WaterML web services to the water information collected by four large state water agencies in Texas in a system called Water Data for Texas. In this instance a particular set of variables are chosen and the association of how each agency describes this variable to a centralized standard variable list is made manually. This allows for quality control of what information is included in the theme, and, if necessary, provision for conversion of units so that all the information in the theme is presented in consistent units of measurement. The establishment of water themes is a less general access mechanism than the hydrologic ontology, but it provides a simplified mechanism for packaging curated, normalized and quality controlled information of a particular kind across a set of water service provider organizations.

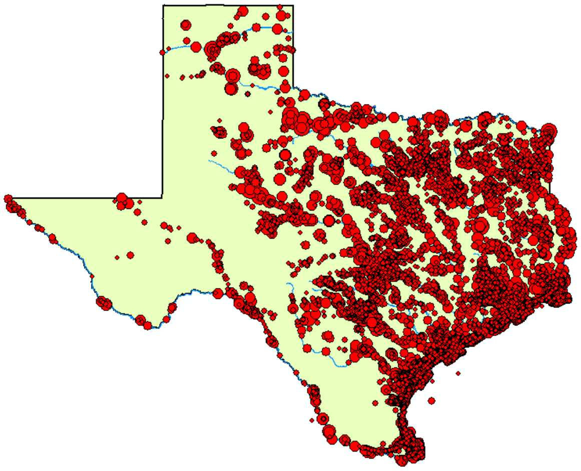

Figure 3. Water temperature theme from Water Data for Texas.

Advantages of an open standards-based infrastructure for water data

The advantages of an open standards-based infrastructure for water data include:

- Support of multiple client applications, desktop (e.g. CUAHSI HydroDesktop) or web-based, developed by multiple distributed teams;
- Reliance on mature OGC service standards and encodings: externally governed, community-tested in multiple settings, with detailed documentation and validation suites – instead of proprietary vendor formats;
- Ability to publish water data using commercial off the shelf software implementing the OGC standards: CSW, WFS and SOS;
- Ability to federate across multiple distributed and separately managed catalogs of water data;
- Architecture and implementation flexibility, and ability to tune system performance to different data publication and access scenarios.
    - For example, CUAHSI HIS maintains a centralized series catalog at the San Diego Supercomputer Center, which is regularly synchronized with distributed services at data publishers, to support consistent metadata services and fast cross-network data discovery.
    - In the current CUAHSI HIS, data are not centrally cached. However, this may change in the future to improve performance, enable queries based on data values, support data validation, and manage data themes.
    - The data can be harvested and cached by desktop clients, to improve performance and enable cross-service and cross-series queries based on data values.
- Community-managed semantics/vocabularies (in near future), which adds the ability to publish local vocabularies into the system and integrate them with extant semantic search facilities.

Conclusions

The CUAHSI profile of OGC services has been developed through seven years of research and has been proven by being applied to seventy water data services registered at the San Diego Supercomputer Center, and to comparable data services compiled at the University of Texas for state and local data water access in Texas.   It describes a stand-alone set of services which rely on OGC standards baseline and community agreement about syntactic and semantic description of hydrologic data. These services comprise a catalog, metadata and data access, and enable a water organization to publish their time series information in a form that can be indexed and searched in a comparable way across a set of water organizations in a region, at a local, state, national or international scale.

Reference

Bermudez, L. and D. Arctur (Eds) (2011), Water Information Services, Concept Development Study, OGC Engineering Report, OGO Report No. OGC 11-013r6, 12 July 2011. Obtainable from http://portal.opengeospatial.org/files/?artifact_id=44834