CUAHSI

universities allied for water research

CUAHSI HYDROLOGIC INFORMATION SYSTEM 2010 STATUS REPORT

David G. Tarboton[1], David Maidment[2], Ilya Zaslavsky[3], Daniel P. Ames[4], Jon Goodall[5], Jeffery S. Horsburgh[1]

1. Utah Water Research Laboratory, Utah State University
2. Principal Investigator, Center for Research in Water Resources, The University of Texas at Austin
3. San Diego Supercomputer Center, University of California at San Diego
4. Department of Geosciences, Idaho State University
5. Department of Civil and Environmental Engineering, University of South Carolina

November 13, 2010

Distribution

Acknowledgment

Funding for this document was provided by National Science Foundation under Grant No. EAR-0622374. The report has benefited from the contributions of the entire CUAHSI Hydrologic Information System team, the CUAHSI program office and collaborators at ESRI (Jack Dangermond, Dean Djokic, Zichuan Ye) and Drexel (Michael Piasecki). The contributions of partners who provide data using CUAHSI HIS technology are invaluable. These include the USGS (David Briar), NCDC (Richard Baldwin), Critical Zone Observatories (Mark Williams), ICEWATER data network (http://icewater.inra.org/), the National Science Foundation EPSCoR program (award number EPS-0814387) and NWS West Gulf River Forecast Center (implemented in collaboration with John McEnery, UT Arlington).

Table of Contents

INTRODUCTION

The Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) (http://www.cuahsi.org) is an organization representing 122 US universities, which is supported by the Earth Sciences Division of the National Science Foundation to develop infrastructure and services to advance hydrologic science in the nation's universities. One component of CUAHSI's activity, also funded by the National Science Foundation, is a Hydrologic Information System (HIS) project, which is developing infrastructure and services to improve access to hydrologic data.

The overall goals of this project are:

- **Data Access** – providing better access to a large volume of high quality hydrologic data;
- **Hydrologic Observatories** – storing and synthesizing hydrologic data for a region;
- **Hydrologic Science** – supporting science by providing a stronger hydrologic information infrastructure;
- **Hydrologic Education** – bringing more hydrologic data into the classroom.

The purpose of this report is to summarize the present status of the HIS project and to set priorities and present a plan for work to be done in the remaining year of the project. The overarching goal for the final year of work will be to operationalize as many components of the CUAHSI HIS system as possible by the end of 2011 when the present funding for the project ends. This will involve moving from research and development versions of HIS software to release versions that can be used and maintained by the hydrology community with the support of the CUAHSI program office.

The HIS Project is carried out by a multi-university team at five universities:

- **University of Texas at Austin** – David Maidment, Tim Whiteaker, Eric Hersh, James Seppi, Jingqi Dong, Fernando Salas, Harish Sangireddy
- **San Diego Supercomputer Center** – Ilya Zaslavsky, David Valentine, Tom Whitenack, Matt Rodriguez
- **Utah State University** – David Tarboton, Jeff Horsburgh, Kim Schreuders, Stephanie Reeder, Edward Wai Tsui, Ravichand Vegiraju, Ketan Patil
- **Idaho State University** – Dan Ames, Ted Dunsford, Jiří Kadlec, Yang Cao, Dinesh Grover
- **University of South Carolina** – Jon Goodall, Anthony Castronova

In addition, Yoori Choi serves as user support specialist at the CUAHSI Program Office. Richard Hooper, President and CEO of CUAHSI, is PI on a separately funded, but related project for ontology development.

ARCHITECTURE OF HIS

The concept of a Hydrologic Information System can be viewed in multiple ways: 1) as a way of publishing hydrologic data in a uniform way; 2) as a way of discovering and accessing remote water information archives in a uniform way; and 3) as a way of displaying, synthesizing and analyzing water information and exporting it to other analysis and modeling systems. By definition, a system is an array of connected components, and in the instance of the CUAHSI HIS, the components are defined as software applications that store, access and index hydrologic information. The connections among them are established by *web services*, which are automated functions that enable one computer to make appropriate requests of another computer and receive responses through the internet. In this sense, the HIS team and its partners have created a *services-oriented architecture* for water information. Josuttis (2007) defines a SERVICES-ORIENTED ARCHITECTURE as "a concept that applies to large, distributed information systems that have many owners, are complex and heterogeneous, and have considerable legacies from the way their various components have developed in the past." This definition certainly applies to the water resources field, which has thousands of agencies and individuals who collect and archive water information in their own way.

Another way of thinking about a Hydrologic Information System is by analogy with a Geographic Information System (GIS). Tomlinson (2003) states that "a GIS stores spatial data with logically-linked attribute information in a GIS storage database where analytical functions are controlled interactively by a human operator to generate the needed information products." This definition implies that all the information has been harvested and stored in a local database and is then available for analysis and interpretation. However, unlike GIS where the data are static and change little through time, a hydrologic information system is representing phenomena that are inherently dynamic and vary greatly through time.

These two concepts, (1) the services oriented architecture; and (2) the desktop hydrologic information system serve to provide the overarching vision of the system that we are developing (Figure 1). To date we have prototype functionality for all three components of the services oriented architecture and data transmission formats for the data exchanges between them, all of which are elaborated on in this report.

In terms of the desktop hydrologic information system, we have developed a prototype desktop application that combines the analysis of GIS, modeling and observations. It downloads, stores and operates on the information on a local desktop computer. Our present implementation is still under active development so these are continually being refined. It also has not yet developed the capability to integrate weather, climate and remote sensing data. Our investigation has demonstrated however, that same triangle of web services among servers, catalogs and users shown in Figure 1 can also be applied to weather and climate data and to remote sensing data, with the adjustment that instead of describing a variable by a time series observed at a point location for an interval of time, we are describing a variable by a grid of values observed over a region for an interval of time.

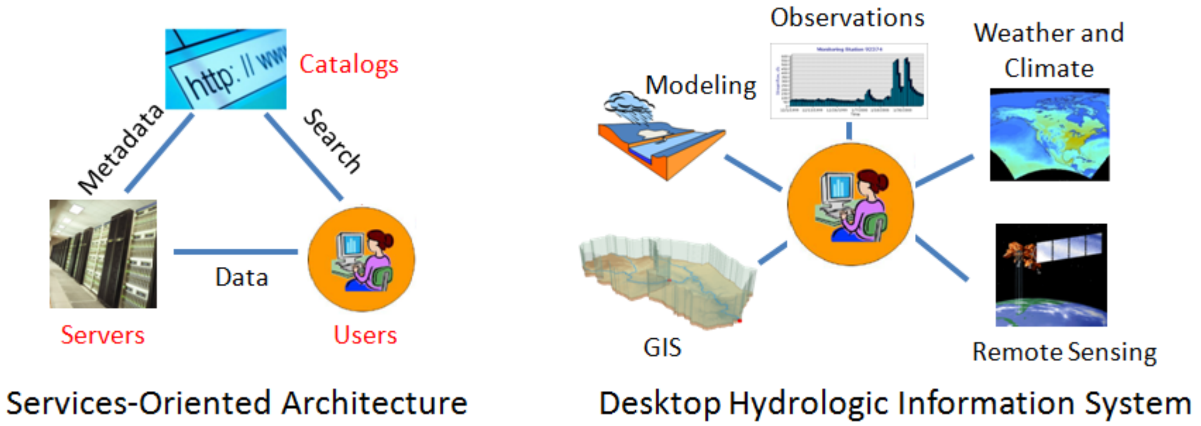


Figure 1. Hydrologic Information System Overarching Vision

The HIS services-oriented architecture is comprised of three classes of functionality: 1) data publication (HydroServer), 2) data cataloging (HydroCatalog), and 3) data discovery, access and analysis (HydroDesktop) (Figure 2). This functionality follows the general paradigm of the Internet. HydroServer publishes data similar to the way Internet web servers publish content. HydroDesktop consumes data published from HydroServer, similar to the way web browsers consume Internet content. HydroCatalog supports data discovery based on indexed metadata similar to the way search engines support the discovery of Internet content. Within the HIS, each of the three components shown in Figure 2 either publish or consume information via web services.

There are three categories of web services in the CUAHSI HIS architecture:

- Data Services – which convey the actual data.
- Metadata Services – which convey metadata about specific collections or series of data.
- Search Services – which enable search, discovery, and selection of data and convey metadata required for accessing data using data services.

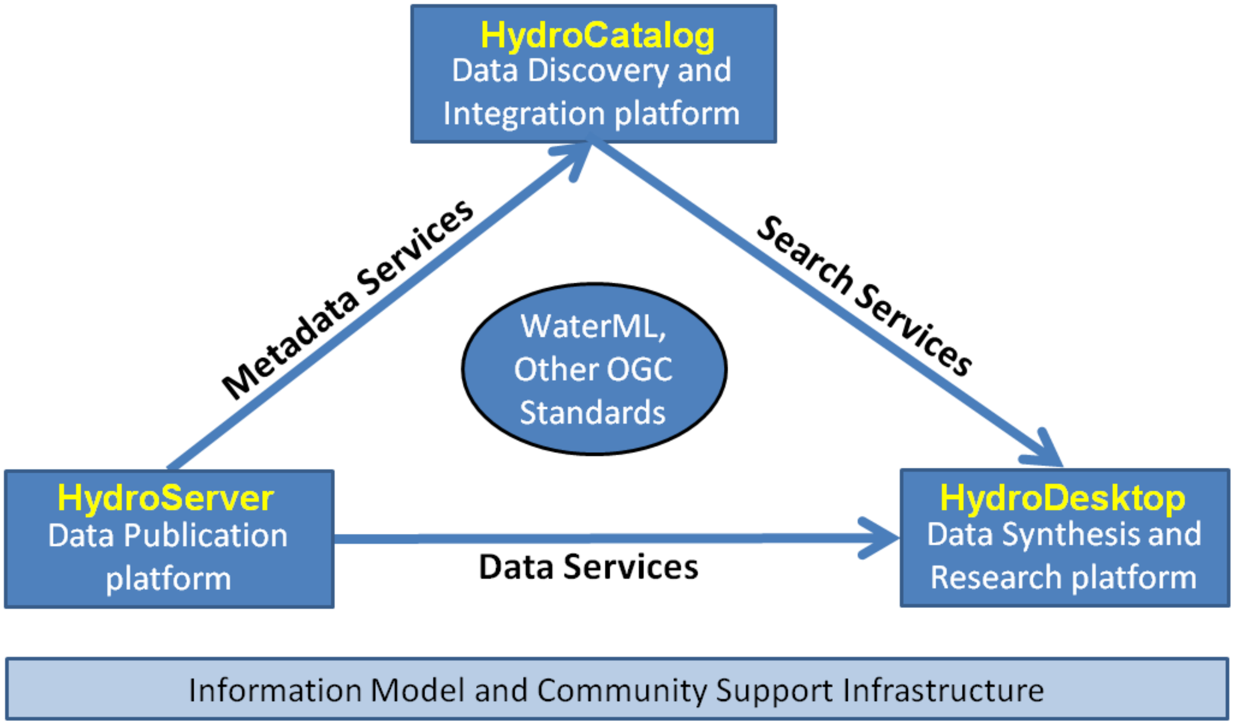

HydroCatalog
Data Discovery and Integration platform
Metadata Services
Search Services
WaterML, Other OGC Standards
HydroServer
Data Publication platform
Data Services
HydroDesktop
Data Synthesis and Research platform
Information Model and Community Support Infrastructure

Figure 2. Components of CUAHSI HIS Services Oriented Architecture

The formats for transmission of information between these systems and the interfaces that enable the communication between them (the connecting arrows in Figure 2) are critical to the functioning of the system. CUAHSI HIS has developed WaterML, an XML based language for transmitting observation data via web services (Zaslavsky et al., 2007). CUAHSI HIS also relies on other established standards (e.g. World Wide Web Consortium Simple Object Access Protocol and Open Geospatial Consortium Geographic Markup Language) for transmission of information between the three primary components.

At the base of Figure 2 is the information model and community support infrastructure upon which the system is founded. The information model is the conceptual model used to organize and define sufficient metadata about hydrologic observations for them to be unambiguously interpreted and used. Within HydroServer, it is encoded as a relational database structure within the Observations Data Model (ODM) (Horsburgh et al., 2008) and the HydroServer Capabilities database to ensure that data and metadata are stored together. It also serves as the conceptual basis for WaterML to ensure that data and associated metadata are transmitted with fidelity when data are downloaded. HydroDesktop implements the information model within its data repository database ensuring that local copies of data retrieved from a server maintain their original context. ODM includes a number of controlled vocabularies for metadata such as units, variable names, sample media etc., where semantic consistency in describing observations is important. The information model also includes an Ontology used to represent a hierarchy of concepts that categorize the variables being measured. The ontology has been developed to support concept based search. The Ontology and Controlled Vocabulary components of the information model have been developed to provide semantic consistency of the terms used in metadata and to support search and discovery based on this semantics. A web site collects and manages community additions and edits to controlled vocabulary content. A similar site is planned for ontology content. These web sites represent community infrastructure to allow dynamic growth of this content while encouraging semantic consistency across the user community.

The architecture shown in Figure 2 has evolved as a general approach for sharing hydrologic observations data. The HIS project has developed functional prototypes of each of the components, but the intention of the HIS is for the system to be general and open to allow the participation of others, much like the Internet is general and open. For example, the HydroServer software stack is not the only entry point for data publishers. Anyone can publish data using web services that deliver data in WaterML format and thus have their data become part of this system. In fact, the United States Geological Survey (USGS) and the National Climatic Data Center (NCDC) have already adopted WaterML for publication of some of their data and have programmed web services that support some of the HydroServer functionality from their systems. The USGS daily and instantaneous value services (http://waterservices.usgs.gov/rest/USGS-DV-Service.html and http://waterservices.usgs.gov/rest/WOF-IV-Service.html) provide data encoded as WaterML. Similarly, NCDC serves data in WaterML format for some of their climate data online datasets (http://www7.ncdc.noaa.gov/rest/). It is through broad uptake of the services oriented architecture of the HIS based on existing and emerging standards that this system will become sustainable.

Similarly the HydroCatalog and HydroDesktop functionality is not limited to the software we have developed. The definition of standard functionality for transmission of information to and from a catalog provider enables others to establish their own catalogs to serve different purposes or communities. In some cases the community may benefit from multiple competing catalogs (much as everyone benefits from the competition for better search functionality between Internet search engines like Google, Yahoo and Bing). HydroDesktop is our prototype client for consumption of web service based hydrologic data, but this does not preclude others from establishing their own (perhaps competing) clients (just as there is competition among internet browsers).

Reliance on independently developed and governed standards is one of the key elements of project sustainability beyond the current funding cycle. Other components of sustainability that comprise the community architecture are:

- Interacting with the community of CUAHSI HIS adopters and users
- Cultivating an open software development model (including infrastructure to support distributed code management, code reviews and refactoring, unit and user interface testing, automated builds) and encouraging contributions from developers outside the project team
- Education and dissemination effort (seminars, workshops, presentations, class exercises, tutorials, learning modules)

- Maintaining a solid operational foundation of the system (high availability data discovery system, hardware and service monitoring and reporting, service testing and validation)
- Engagement with key, long-standing government, university and industry groups, capable of contributing to the system and data development and maintenance beyond the funding cycle (federal and state agencies, libraries, leading companies such as ESRI and Kisters)
- Extending CUAHSI HIS technology in several NSF-supported research and cyberinfrastructure projects

Functionality of HydroServer

HydroServer is envisioned to be a self-contained, complete hydrologic data and metadata publication system that permits data publishers to control their own data while still being part of a distributed national/international system allowing universal access to the data (Horsburgh et al., 2010). HydroServer is targeted at investigators who are collecting data within research watersheds or observatories, although the software is general and can be used by anyone who wants to share hydrologic observations. The HydroServer software stack relies on the protocols and standards established by the HIS project and consists of a number of software applications that are being developed and managed as open source software using an open source code repository (http://hydroserver.codeplex.com).

An important principle has emerged from our work on publishing hydrologic data. **HydroServer functionality should support complete description of the data and metadata**. We refer to this as the self describing principle and this stems from the fact that the person or organization creating the data is generally best suited to provide metadata, and should have control over data publication.

HydroServer (Figure 3) supports publication of both point observations data stored in one or more ODM databases (Horsburgh et al., 2008) and published using WaterOneFlow web services and geospatial data published using OGC Web services from ArcGIS Server.

Each HydroServer has a Capabilities Database that catalogs metadata about the regions for which data have been published (e.g., an experimental watershed or observatory) and the list of services that have been published on the server for each region. These three web services comprise the service interface. A suite of tools to load, edit and assist with the management of ODM data has been developed. By being implemented using Microsoft SQL Server database software, tools that Microsoft provides are also available for ODM data loading and editing. A configuration tool has been built that provides an interface for defining the contents of the Capabilities Database. The Capabilities Web Service includes methods that return, in XML format, the list of regions for which data have been published, the published point observations data services, and the list of published spatial data services, along with appropriate metadata for each. By doing so, all of the capabilities of the HydroServer are published in an XML format that can be discovered by registration and cataloging services (HydroCatalog), making a HydroServer self-describing. A program can discover all of the capabilities of the HydroServer simply by calling the Capabilities Web service, and metadata can be registered and harvested automatically. The ODM Tools suite and capabilities configuration tool comprise the data manager interface. Finally, a suite of data presentation and visualization tools has been created for HydroServer. The suite includes the HydroServer Website, the Time Series Analyst, and the HydroServer Map Website. These provide a public browser accessible graphical user interface to the data holdings of the HydroServer.

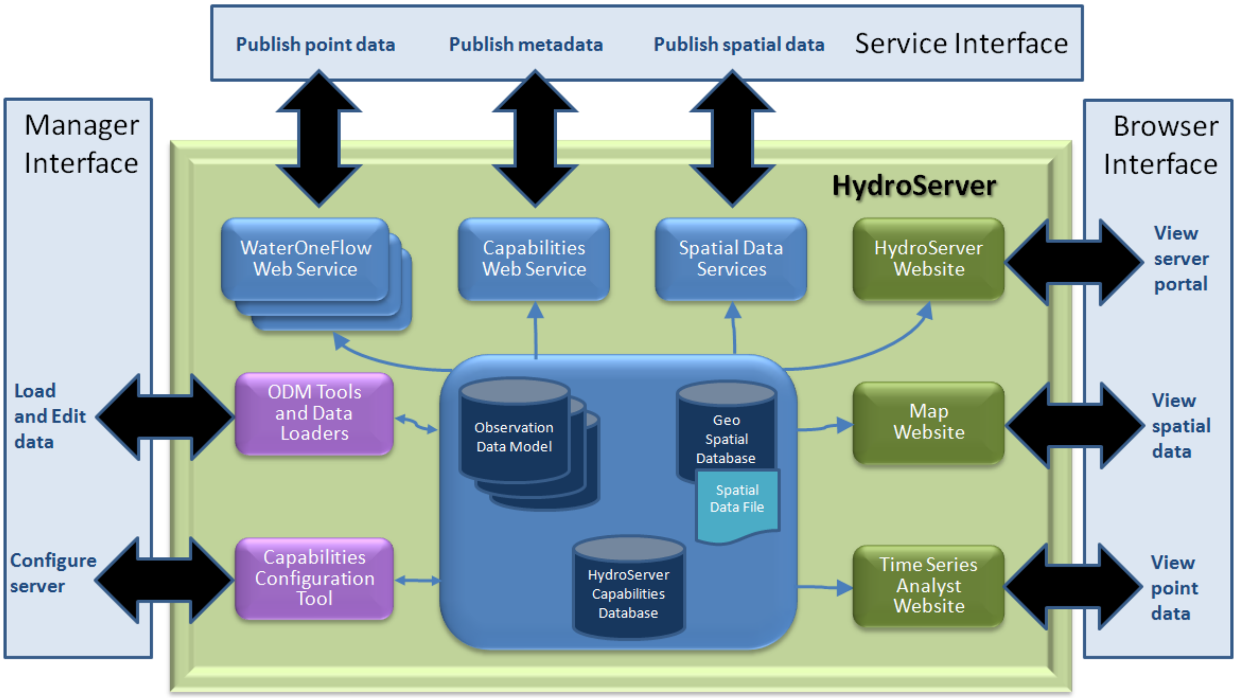


Figure 3. HydroServer Architecture and Functionality

HydroServer relies on commercial off the shelf software (e.g., Microsoft Windows, .Net Framework, Internet Information Services, and SQL Server database software as well as ESRI's ArcGIS Server software) for general purpose functionality and includes the following specific components developed by the HIS team:

- *Observations Data Model (ODM)* – a standard relational database model for storing hydrologic point observations
- *ODM Tools* – a software application for managing data stored in an ODM database
- *ODM Data Loader* – a software application for loading table-based data into an ODM database
- *ODM Streaming Data Loader* – a software application for loading streaming data from environmental sensors into an ODM database
- *ODM Web Data Loader* – a web application for loading table-based data into an ODM database
- *WaterOneFlow Web Services* – a web application for publishing hydrologic observations stored in an ODM database on the Internet in WaterML format
- *Capabilities Database, Capabilities Database Configuration Tool and Capabilities Web Services* – a database and software tools for publishing the capabilities of a HydroServer on the Internet
- *HydroServer Web Site* – a web application that provides simple web browser based access to the data resources hosted on a HydroServer
- *HydroServer Time Series Analyst* – a web application for visualization of point observations published on a HydroServer
- *HydroServer Map Web Application* – a web application for visualization of spatial datasets published on a HydroServer

FUNCTIONALITY OF HYDROCATALOG

HydroCatalog is the discovery part of the system linking data publishers and application clients in CUAHSI HIS. Data discovery across multiple observational data services is enabled by a centralized Metadata Catalog Database, which contains descriptions of the datasets hosted on the many federated data servers on which data are published. HydroCatalog interfaces with data publishers through its web sites, interfaces with WaterOneFlow web services, and interfaces with desktop clients through search and ontology web services (Figure 4).

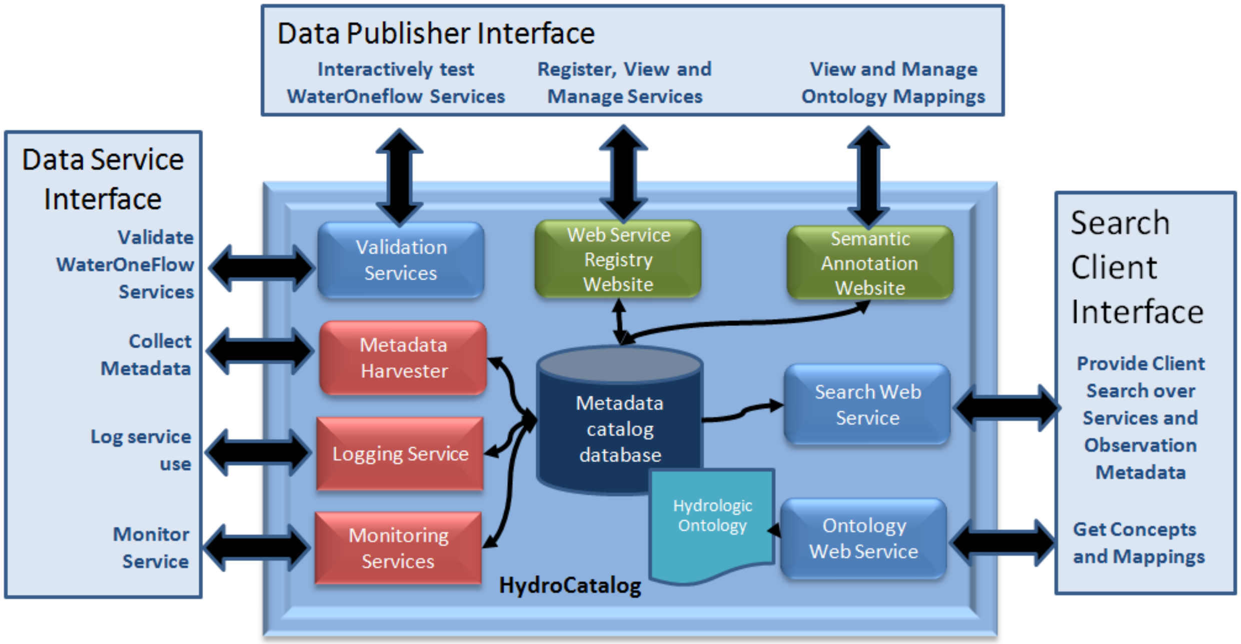

Data Publisher Interface
Interactively test WaterOneflow Services
Register, View and Manage Services
View and Manage Ontology Mappings
Data Service Interface
Validate WaterOneFlow Services
Collect Metadata
Log service use
Monitor Service
Validation Services
Web Service Registry Website
Semantic Annotation Website
Metadata Harvester
Logging Service
Monitoring Services
Metadata catalog database
Hydrologic Ontology
Search Web Service
Ontology Web Service
HydroCatalog
Search Client Interface
Provide Client Search over Services and Observation Metadata
Get Concepts and Mappings

Figure 4. HydroCatalog Architecture and Functionality

HydroCatalog supports discovery of data by keywords, which represent concepts in the CUAHSI ontology and a collection of their synonyms. Search functionality requires that variable names in registered services are associated with terms at the leaf nodes of this hierarchy. Data publishers first register their WaterOneFlow web services with the HydroCatalog Web Service Registry, providing a service level metadata description that includes a title, abstract, publisher and contact information, spatial and temporal extent, recommended citation, and links to organization web sites. Data publishers have the opportunity to validate their WaterOneFlow services by testing them using the Validation Services. Registration of a service triggers the Metadata Harvester to harvest the metadata from the web service and store it in the metadata catalog database. Once the metadata is stored in the database, data publishers can use the tagging application on the Semantic Annotation Website to map their variables to terms in the hydrologic ontology. The ontology can be visualized on part of the Semantic Annotation website (currently at http://hiscentral.cuahsi.org/startree.aspx).

Once tagging is complete, the metadata are discoverable through the Search and Ontology Web Service. The metadata harvester is capable of doing periodic metadata harvests for each of the registered WaterOneFlow web services to ensure that the metadata catalog database is kept up to date. A Logging Service records use information on WaterOneFlow services that report use back to HydroCatalog. The Monitoring Service periodically accesses registered WaterOneFlow services to monitor their status so that breaks in service may be identified and rectified, or services that go offline be de-listed (after first attempting to work with their owners to reinstate them).

The Search and Ontology Web Service that exposes the contents of the metadata catalog database includes a number of web service methods that enable programmers who are developing client software applications to execute spatial, temporal, and semantic searches on the catalog to discover relevant data. HydroCatalog searches are mediated across all sources of data in the catalog, resulting in lists of data series that match search criteria. Search results contain all of the information necessary to retrieve data in WaterML format from the data server on which the data are hosted, and client applications that use the HydroCatalog search services (e.g., HydroDesktop) can use the information contained within the search results to retrieve the data on demand.

Functionality of HydroDesktop

HydroDesktop is a free and open source Desktop Hydrologic Information System (Figure 5) that helps users discover, use, manage, analyze and model hydrologic data. HydroDesktop complements the CUAHSI HIS services oriented architecture for the publication of hydrologic data, and at the same time represents the third critical component in this services oriented architecture, the client that users use to search and access data.

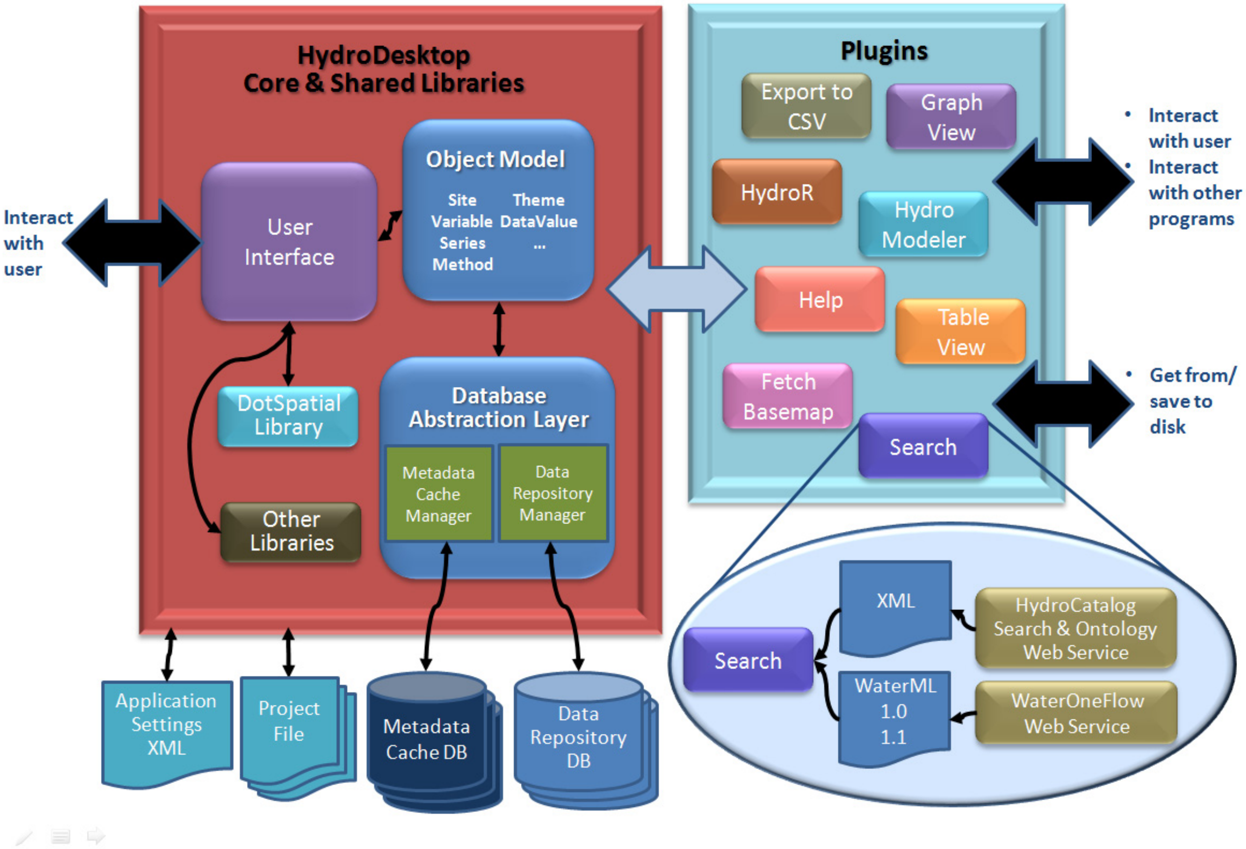


Figure 5. HydroDesktop Architecture and Functionality

The Geographic Information System (GIS) components of HydroDesktop are built from the open source DotSpatial library, while the time series components use HIS web services. The result is a spatially-enabled system for downloading observational data describing our water environment. The architecture of HydroDesktop (Figure 5) is structured to take advantage of centralized cataloging functionality (from HydroCatalog) as well as distributed data (from HydroServers).

The DotSpatial/MapWindow engine used by HydroDesktop provides geographic visualization capability. HydroDesktop uses a plugin architecture, and plugins provide attribute table visualization, editing, map printing, projections, symbology, and base maps. HydroDesktop Plugins also support searching for, downloading, viewing, graphing, editing, exporting, printing, and modeling with time series data. The search plugin allows search by area, time range, variables, key words, and server. Like HydroServer, HydroDesktop is open source software developed using an open source code repository (http://hydrodesktop.codeplex.com).

The HIS architecture supports the concept of the development many different client tools. Through the project lifespan a number of such tools have been prototyped and made available as examples of web-services consumption on various software platforms. These have included a Microsoft Excel spreadsheet, HydroExcel, that can connect directly to and download data from specific HydroServers; a web application (HydroSeek) for searching for data on HIS Central; and various extensions for Matlab, ArcGIS and other software tools.

The HydroDesktop client was envisioned as a free and open source desktop software application that supports many of the features demonstrated in the prototype clients noted above including search and discovery using HIS Central, download of data from specific HydroServers, basic GIS and time series data visualization and analysis, export and import of spatial and temporal data, and hydrologic modeling using data retrieved from the HIS network. While not intended to be an all-encompassing data analysis software application (e.g., a complete GIS or statistical/modeling system), HydroDesktop has been developed with an extensible plugin architecture that creates opportunities for incorporating such capabilities either by custom programming of new capabilities specifically for HydroDesktop or through interfacing (loosely or tightly as shown below) with third party software applications.

Version 1.0 Functional Specifications – The HydroDesktop Version 1.0 Functional Specification Document available on the documentation page of the HydroDesktop web site (http://hydrodesktop.codeplex.com) describes the following key functions as central to version 1.0 of the software:

- Data Discovery
  - Data Discovery Using the HIS Central Metadata Catalog
  - Data Discovery Directly From WaterOneFlow Web Services
  - Data Discovery for Thematic Datasets
  - Processing of Search Results
- Data Download
  - Downloading Observational Data
  - Downloading GIS Datasets
- Data Visualization
  - Visualization and Analysis of Spatial Data
  - Visualization and Analysis of Observational Data
- Data Import and Export
  - Importing and Exporting Spatial Datasets
  - Importing and Exporting Observational Datasets
- Project Workspace

    - Saving and opening sets of data
    - Saving and loading project and plugin settings
- Plug-in Interface
    - Extending the software through plugins built by the HIS team
    - Extending the software through plugins built by third parties

Following is a brief description of how these capabilities have been implemented within the 1.0 version of HydroDesktop.

Search and Discovery Functions – At the heart of HydroDesktop is the capability to search for, discover, download, visualize and export data from the HIS network. Search and discovery is primarily achieved through a search plugin that allows a user to search based on:

- Area – The user must select a polygon on the map from one of the default data layers (counties, states, major watersheds) or from a polygon layer added by the user. Alternatively the user can draw a box on the map to identify a search area.
- Key Words – The user can optionally specify a set of key words related to observed variables to be used in the search query. Key words can be found by browsing a tree-view control or by typing key words in a search box. If no key words are selected then the query defaults to all variables.
- HydroServers – The user can optionally specify specific HydroServers or HIS services to include in the query. If none are specified then all known services are included in the search.
- Time Range – The user can optionally specify a time range for the data search by indicating a start and stop date which bound the time period of interest.

The user creates the search and executes it. This results in the creation of a “search results” layer showing all points on the map where data series were found. The user then selects series of interest from the map and executes a data download function which retrieves all of the data to the local computer database.

Time Series Data Visualization, Editing, and Export Functions – Once data have been downloaded into the HydroDesktop database, they can be immediately viewed graphically or tabularly through a “Graph View” plugin and a “Table View” plugin respectively. Both of these tools were developed by the HIS team specifically for HydroDesktop and extend functionality also available in the HydroServer Time Series Analyst. Graph visualization includes the ability to view time series, probability, histogram, and box-and-whisker plots that are extensively customizable and can be exported as graphic files for use in reports or other documents. The Table View plugin allows the user to view the data in parallel (time stamps and values for all series shown in parallel columns of the table) or in series (all data and metadata with metadata in columns and time stamps with values in rows.) This plugin also allows the user to export the data to a comma separated values (CSV) file. The “Edit View” plugin enables users to modify existing time series and derive new time series from existing data (e.g., derive daily average values from 15 minute or hourly data).

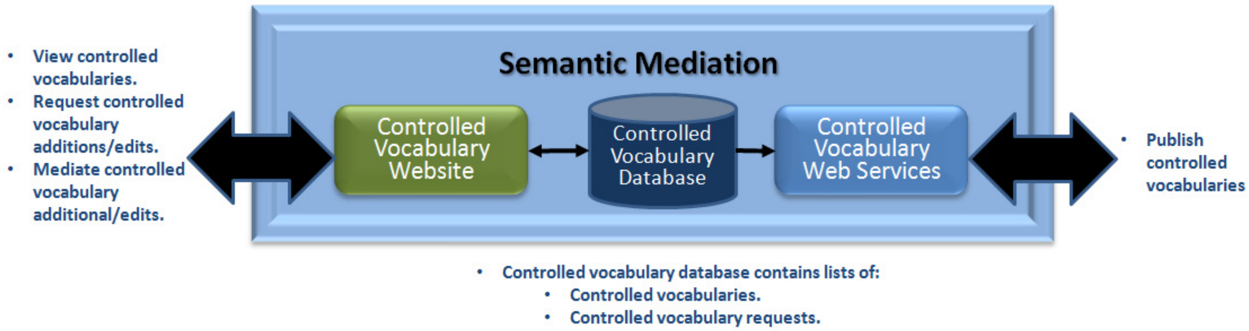
Geographic Information Systems (GIS) Functions – HydroDesktop is built largely upon the open source MapWindow GIS software development framework Application Programmer Interface (API) called DotSpatial (see http://www.dotspatial.org/). This “tightly integrated” GIS functionality supplies HydroDesktop with a large (and growing) number of geospatial data visualization and analysis capabilities. The open source MapWindow GIS project was started at Utah State University in 2001 and has grown to be one of the most widely used open source GIS software applications with over 7,000 downloads per month globally. The most recent version of MapWindow includes the DotSpatial programmer API as its core GIS library and both projects (MapWindow and DotSpatial) have been used as models for how to create an open source development community around HydroDesktop. Indeed, by using the same plugin interface (defined in DotSpatial and used in MapWindow 6), HydroDesktop is cross compatible with plugins and extensions developed for the MapWindow software by third party developers. It is expected that this will help add to the long term sustainability of the HydroDesktop software project.

Statistical Analysis and Modeling Functions – Through its plugin interface, HydroDesktop has been extended to support extensive statistical analysis and modeling capabilities. Recognizing the cost prohibitive challenges and associated massive software development effort that would be required to build custom statistical analysis and modeling capabilities natively into the HydroDesktop application, HIS project team members made the decision early in the project to provide such capabilities through loose (or tight) coupling with 3rd party software applications. Specifically two unique and very powerful plugins have been constructed for HydroDesktop and are included with the default software installation following this paradigm. The first is a plugin called HydroModeler that wraps the European Union-developed OpenMI modeling framework. OpenMI (see www.openmi.org) defines a model interoperability interface that allows hydrologic and other time-step based models to interact with each other – passing data between models – as needed to simulate complex natural systems. The HydroModeler plugin to HydroDesktop provides a complete implementation of the OpenMI specification and specifically allows modelers to integrate HIS derived datasets into their models. The second 3rd party software which has been wrapped in the HydroDesktop plugin environment is the statistical software, "R". R is an extremely powerful script/command line based open source statistical analysis software tool based on the same scripting language used in the popular proprietary "S-Plus" software. The HydroR plugin provides an R scripting and execution environment directly within HydroDesktop, thereby extending the statistical analysis capabilities of HydroDesktop immensely.

Information Model and Community Infrastructure

The organization and representation of data is at the heart of much of what HIS is. ODM (Horsburgh et al., 2008), WaterML (Zaslavsky et al., 2007) and the HIS Ontology (Beran and Piasecki, 2009; Piasecki and Beran, 2009) were developed early in the project and provide the conceptual and logical foundation for much of the current functionality of the HIS System. As a mechanism for data organization and storage, ODM adopted an atomic model using a relational database star schema to annotate individual observations with the metadata required for observations to be unambiguously interpreted. This has the advantage that relational database methods can be used to extract data organized into collections along different dimensions, such as for example a snapshot in time across multiple points in space, versus time series at a point, and relational database management systems are very good at efficiently handling data loading and retrieving tasks. As a mechanism for transmitting observations over the Internet, WaterML adopted a series model for the delivery of observations as time series, which has resulted in much of the current functionality of the HIS being focused on time series. The key is that ODM and WaterML share a common information model. Although they are organized differently and are used for different purposes, they contain the same information.

The semantics of ODM are held in a series of controlled vocabulary tables that are centrally mediated (Horsburgh et al., 2009). A master list of approved ODM controlled vocabulary terms is maintained within a central controlled vocabulary database (Figure 6). This central repository represents a community vocabulary for describing environmental and water resources data in that it was developed by researchers working within the Hydrologic Science community. It is dynamic and growing; users can add new terms or edit existing terms by using the functionality available through the HIS website (http://his.cuahsi.org/mastercvreg/cv11.aspx). If a data manager cannot find an appropriate term to describe data that is being added to an ODM database, he or she can navigate to the HIS website and use an online form to request addition of an appropriate term to the master controlled vocabulary. Once a new term is accepted, it becomes part of the master database. This system serves the purpose of encouraging the use of consistent terminology by all users of HIS.


Semantic Mediation
Controlled Vocabulary Website
Controlled Vocabulary Database
Controlled Vocabulary Web Services

- View controlled vocabularies.
- Request controlled vocabulary additions/edits.
- Mediate controlled vocabulary additional/edits.

- Publish controlled vocabularies

- Controlled vocabulary database contains lists of:
  - Controlled vocabularies.
  - Controlled vocabulary requests.

Figure 6. Central Semantic Mediation Architecture and Functionality

DATA PUBLICATION MODELS

In our experience to date with the services-oriented architecture described above, three patterns of implementation for data publication have emerged.

- HydroServer-based data services
- Water agency data services (e.g. from USGS, EPA, NWS)
- Hosted data services

In the *HydroServer* approach a water research center or a water agency creates their own data server that stores observations data in an ODM database and publishes them through WaterOneFlow web services. This pattern is implemented at more than a dozen universities in the US for publishing water research data, and at some water agencies, such as the Texas Water Development Board, which is in the process of applying CUAHSI technology to publish the main state level water databases in Texas. Another example of this approach is being implemented at the University of Texas at Arlington to which the National Weather Service's West Gulf River Forecast Center has supplied all of its history from 1995 to the present of hourly and daily Multisensor Precipitation Estimate data (mainly derived from Nexrad measurements). These data are stored in an ODM database and published using WaterOneFlow web services as precipitation time series indexed to points on a regular array mesh, like a set of virtual rain gages distributed over the landscape. The size of this precipitation database is 5TB.

The *water agency data services* approach (Figure 7a) has been used where a water agency has an existing water data archive and wishes to retain its current structure. In this instance, as exemplified by the US Geological Survey, the agency programs a customized WaterOneFlow web service to provide access to its water observations data in the WaterML language and supplies HIS Central with a data dump of its observations metadata. A variant of this approach is used by the EPA to provide access to its STORage and RETrieval (STORET) water quality data – this information is published as a different kind of web service within the Water Quality Exchange (WQX) framework using the WQX XML schema, which CUAHSI translates to become WaterML. EPA periodically provides a dump of the entire STORET database to HIS Central from which the data series metadata is extracted for inclusion in the CUAHSI HIS Central metadata catalog.

In a *hosted data service* (Figure 7b) the data publisher creates data files of observations data and conveys them to a CUAHSI data repository where the information is stored in an ODM datbase and served using WaterOneFlow web services based on the WaterML language. The data publisher relies on the support services of CUAHSI to host their data. Such a service is presently offered by the San Diego Supercomputer Center (SDSC). One variant on this pattern is used by the six NSF Critical Zone Observatories (CZO) to centralize their observations data at the SDSC. In the CZO model data is published on the CZO websites in a pre-specified ASCII file format. Each CZO runs their own internal data management system that after internal quality control and release procedures publishes these ASCII files on the CZO project web sites. The new files are regularly retrieved into the SDSC hosting application, validated against controlled vocabularies, ingested into an ODM database, and WaterOneFlow web services are automatically updated and publish the data in WaterML. This approach is convenient for water scientists who want to store, maintain, and display their own data however they choose, and who want also to have the data published at a central location so that it can more readily be synthesized and compared with similar information measured elsewhere. This form of hosted data service is referred to as "hold and serve" because CUAHSI actually holds the data. Another variation on hosted data service is referred to as a "flow through" data service. In this case an agency or organization is maintaining, managing, and actively updating the authoritative copy of their data, but in a form that is not CUAHSI compatible. A flow through data service establishes a mapping from the 3rd party data format to WaterML and implements this as an on demand function that, when triggered by a WaterOneFlow web service request, retrieves the data from the 3rd party source, reformats it as WaterML and transmits to the calling program. The retrieval of data from the 3rd party data source may be via "scraping" a website. Such a procedure was used with the early implementations of USGS NWIS web services and is currently used with a prototype SNOTEL web service. This approach is fragile because changes in the 3rd party website may break scraping functionality, so this approach is best implemented in cooperation with the 3rd party provider. This approach is useful where the 3rd party provider wants to maintain control over the authoritative copy of their data and does not want, or have the capacity to publish using the CUAHSI HIS model.

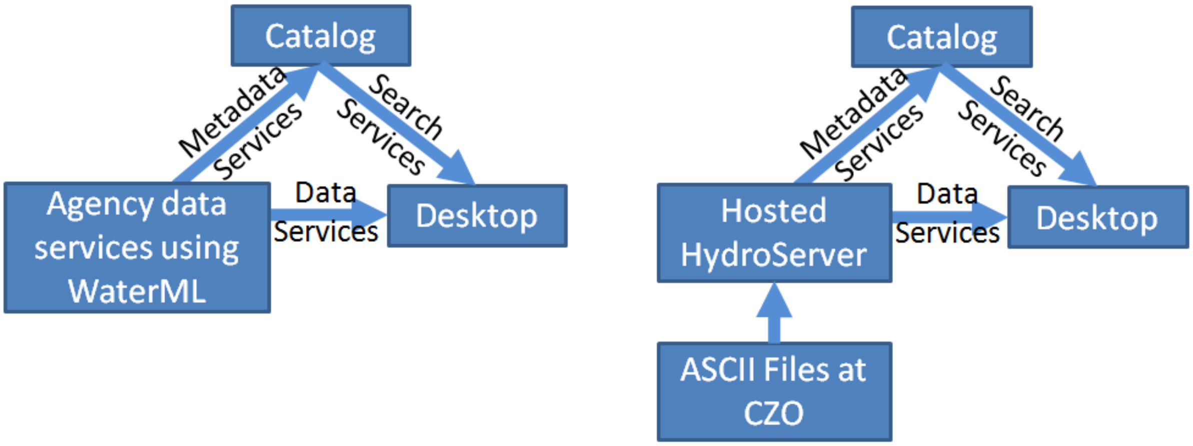


Figure 7. Architectures of agency and hosted data service publication models

ROLES AND RESPONSIBILITIES

Sustaining HIS requires ongoing operation and maintenance of the cyberinfrastructure it represents. This section identifies the roles and responsibilities that need to be filled to sustain this operation, regardless of who or what organization fulfills each role.

DATA PUBLISHER

Three modes of data publication were outlined above. The HydroServer approach requires obtaining hardware and configuring software on a data server. Details on how to do this are given below. The water agency data services approach requires the agency to develop their own capability to publish water data and metadata services using WaterML. The strategy behind this approach has been established but ongoing expertise is needed to continue to work with agencies and provide guidance on this. Hosted data services start with the same requirements as the HydroServer approach, but differ in that they need to be filled by a central (e.g., CUAHSI) facility. Additionally, capability to receive data from publishers, support multiple (possibly virtual) servers and provide the expertise to run this system and support users is required. In the remainder of this section we first summarize some of the key responsibilities of a data publisher using HydroServer, focusing on the expertise required to carry them out. Detailed instructions on publishing data using HydroServer are given on the web page: http://his.cuahsi.org/hydroserver.html. We then discuss the approach established for working with agencies to establish water data services, and lastly identify additional roles required (e.g. at a CUAHSI central facility) to support hosted data services.

HYDROSERVER APPROACH

Establish HydroServer Hardware
Setting up a HydroServer requires one or more physical servers on which the software stack will be installed. Data publishers who wish to use the HydroServer software must either purchase a physical server or procure access to and space within existing server infrastructure to host the HydroServer software and applications. HydroServer can be implemented on a single server, although it is quite flexible in that it can be deployed across multiple machines in the case where server administrators divide labor among multiple machines (e.g., they have a separate web server, database server, GIS server, etc.). For approximately $5,000 - $8,000, a server can be purchased that is adequate for implementing a HydroServer. Approximate minimum system requirements include: dual or quad core processor, at least 4 – 8 GB of RAM, 500 GB or more of hard disk space, and a 1 GB network adapter.

There are a several significant challenges accompanying the setup of physical hardware for hosting a HydroServer. First, a HydroServer must be a web server and so must have a dedicated internet connection and IP address. Additionally, the server must be located within an organization’s network in a location where it can be accessed by the rest of the Internet using standard internet ports and protocols. Where existing server infrastructure is used, data publishers must ensure that they have appropriate access. Another significant challenge is that hardware eventually needs to be replaced. Organizations that create HydroServers should have a plan for migrating data and services to new hardware if and when the need arises. The vast heterogeneity in Information Technology (IT) environments at different universities and organizations can complicate HydroServer setup. Because of this, acquisition and set up of appropriate hardware for hosting a HydroServer may require assistance from an Information Technology (IT) professional.

Acquire and Install HydroServer Software

A complete HydroServer installation requires the following commercial off the shelf software:

- Microsoft Windows 2008 Server
- Microsoft Internet Information Services (IIS)
- Microsoft .Net Framework Version 3.5
- Microsoft SQL Server 2008
- ESRI ArcGIS Server 9.3.1 for .Net, Enterprise Advanced

HIS supplied software available from http://his.cuahsi.org/hydroserver.html includes:

- Observations Data Model (ODM)
- ODM Tools
- ODM Data Loader
- ODM Streaming Data Loader
- WaterOneFlow Web Services
- Capabilities Database, Capabilities Database Configuration Tool and Capabilities Web Services
- HydroServer Web Site
- HydroServer Time Series Analyst
- HydroServer Map Web Application

Installing the required commercial and HIS supplied software requires IT expertise. A server preconfigured with Microsoft Windows Server, IIS, .Net Framework, and SQL Server can be purchased, however the cost of the server increases significantly if these components are purchased pre-installed. Installation of ArcGIS Server must be completed by the HydroServer administrator. Detailed software manuals for each HIS supplied software component are provided via the HIS website. Although these manuals contain step-by-step instructions for installing and configuring each HIS software component, expertise is needed in software installation, deployment and configuration of web applications, and database creation and management.

Obtaining and maintaining licenses for the commercial software listed above can be a barrier for adoption of HydroServer. A free version of SQL Server, called SQL Server Express, is available for download and will work with all of the HIS tools. Alternative spatial data servers can be substituted for ArcGIS server for publishing spatial data services, however the HydroServer map application currently requires ArcGIS server. Many universities can qualify for Microsoft's Academic Alliance program which enables academic departments to pay a nominal annual fee for access to licenses for the Windows operating systems, SQL Server, and Microsoft's code development environments. IIS is part of the Windows operating system, and the .Net Framework can be downloaded for free.

Establish a Domain for HydroServer

The HydroServer web applications require that an Internet domain be set up for a HydroServer (i.e., the Internet address at which the applications will be located, e.g. http://icewater.usu.edu). The process for establishing a domain is described in general in HydroServer documentation, but a HydroServer Data Publisher needs to work with the IT professionals within their organization, as this process is specific to each organization that hosts a

HydroServer. Because of this, setting up a HydroServer requires personnel with sufficient computer system administration expertise to work with an IT professional.

Organize and Load Data to be Published

Organizing data to be published on an a HydroServer and loading it into the appropriate data structures for publication can be the most time consuming and difficult tasks in setting up a HydroServer. These tasks require one or more individuals with the expertise to not only familiarize themselves with and understand the scientific meaning of the data to be published, but also to become sufficiently familiar with relational database concepts and data manipulation and reorganization techniques and technologies so that they can understand ODM and load data. A basic knowledge of spreadsheets, relational databases, SQL Server database management, and the use of file sharing and operating systems is needed. Metadata needs to be gathered and written and choices made as to how to represent the data in ODM, mapping variables and concepts onto the ODM semantics. Data needs to be formatted into one of the formats that can be accommodated by the ODM data loaders. Where GIS data is to be published using the Web Map application the data analyst will need GIS skills, specifically with respect to ArcGIS Server.

Detailed documentation of ODM and its use is available at http://his.cuahsi.org/odmdatabases.html, and documentation describing how to publish spatial datasets on a HydroServer is available. However, the time and expertise required to organize and load data into appropriate structures on a HydrServer remain as significant impediments for adoption of HydroServer. This is partially because it is rare to find hydrology domain scientists who already have the necessary IT skills to accomplish the task or to find data or information scientists who have the necessary hydrology domain skills to understand the data. There exists a significant opportunity to train students in hydrology and engineering so that they will have these skills.

Register Services with HydroCatalog

We encourage that services hosted on a HydroServer that publish data and metadata be registered with a HydroCatalog to ensure that they can be discovered within the larger context of the HIS. Currently this requires filling out a form on a registration website at HIS Central. During the registration process, service level metadata describing the characteristics and contents of the service must be supplied by a data manager. No special expertise is required, however the person registering the service must be familiar with the contents of the service in order to specify accurate metadata.

Maintain HydroServer

A HydroServer Data Publisher is responsible for ensuring that their HydroServer is adequately supported physically (e.g. sufficient power and cooling and proper server hardware environmental conditions). Additionally, they are responsible for establishing a back-up plan for their server that ensures that their databases and applications are adequately protected from a system or disk failure. Protection from viruses and outside attack is also the responsibility of the server owner. This may require some expertise from an IT professional to ensure that a server is adequately secure (according to the organization's IT policies) and that a viable backup and recovery plan is established.

Contribute to HydroServer Development

The HIS supplied HydroServer software is being managed in an open source code repository at http://hydroserver.codeplex.com. At that website, the source code for each component can be accessed, there is a discussion forum for posting questions and getting answers, and there is an Issue Tracker for posting bug reports and feature request related to the HydroServer software. Depending on the level of expertise of a HydroServer administrator, we encourage participation in the forums, posting of bugs and issues, as well as contributions and enhancements to the source code of each of the HydroServer components.

Agency Approach

In working with agencies to have them set up water data services a key requirement is specification of the format and functionality required to publish water data services. The standardization efforts through the OGC Hydrolology Domain Working Group are critical in this regard and capability and expertise to continue these efforts is required to sustain the growth in agency uptake of the CUAHSI HIS model. Standardization is also critical to catalyze the engagement of industry partners such as ESRI and Kisters and further extend the adoption of the system. From working with US federal government agencies (primarily the USGS and NCDC) in establishing agency water data services, the following steps have been developed to guide the process and interactions.

1. Establish an agreement with the agency
2. Identify the scope of the service
3. Translate the semantics of the service to WaterML
4. Include agency personnel in OGC/WMO Hydrology Domain Working Group where appropriate
5. Develop a first draft of the web service
6. Perform unit testing, over a series of validation cases
7. Establish an agency metadata service. This may be at the agency or hosted by CUAHSI (e.g. at SDSC) based on transmission of the metadata to CUAHSI.
8. Establish procedures for metadata service updates - updating the server metadata database if hosted at CUAHSI
9. Document and register the water data service with a HydroCatalog
10. Review and test the service together with the agency, for possible approval as “operational”

Additional roles to support hosted data services

SDSC has established prototype functionality for hosting data services in a centralized system with redundancy, failover management and load balancing to support continuous robust operation (Figure 8). This facility is referred to as HIS Central and combines data hosting and serving, prototype HydroCatalog functionality and centralized aspects of HIS program operations.

As currently implemented, HIS Central relies on two pairs of servers (web servers and mirrored database servers) housed in two different buildings on the UCSD campus. Server hardware and software are continuously monitored with R-U-On infrastructure (Figure 4). Hardware or process malfunction alerts are delivered to developers’ emails and - for a small group of HydroCatalog managers at SDSC – over SMS. A comprehensive reporting system (available at http://water.sdsc.edu/reports) provides access to the catalog content and usage statistics. It currently supports 20+ tabular and graphical reports.

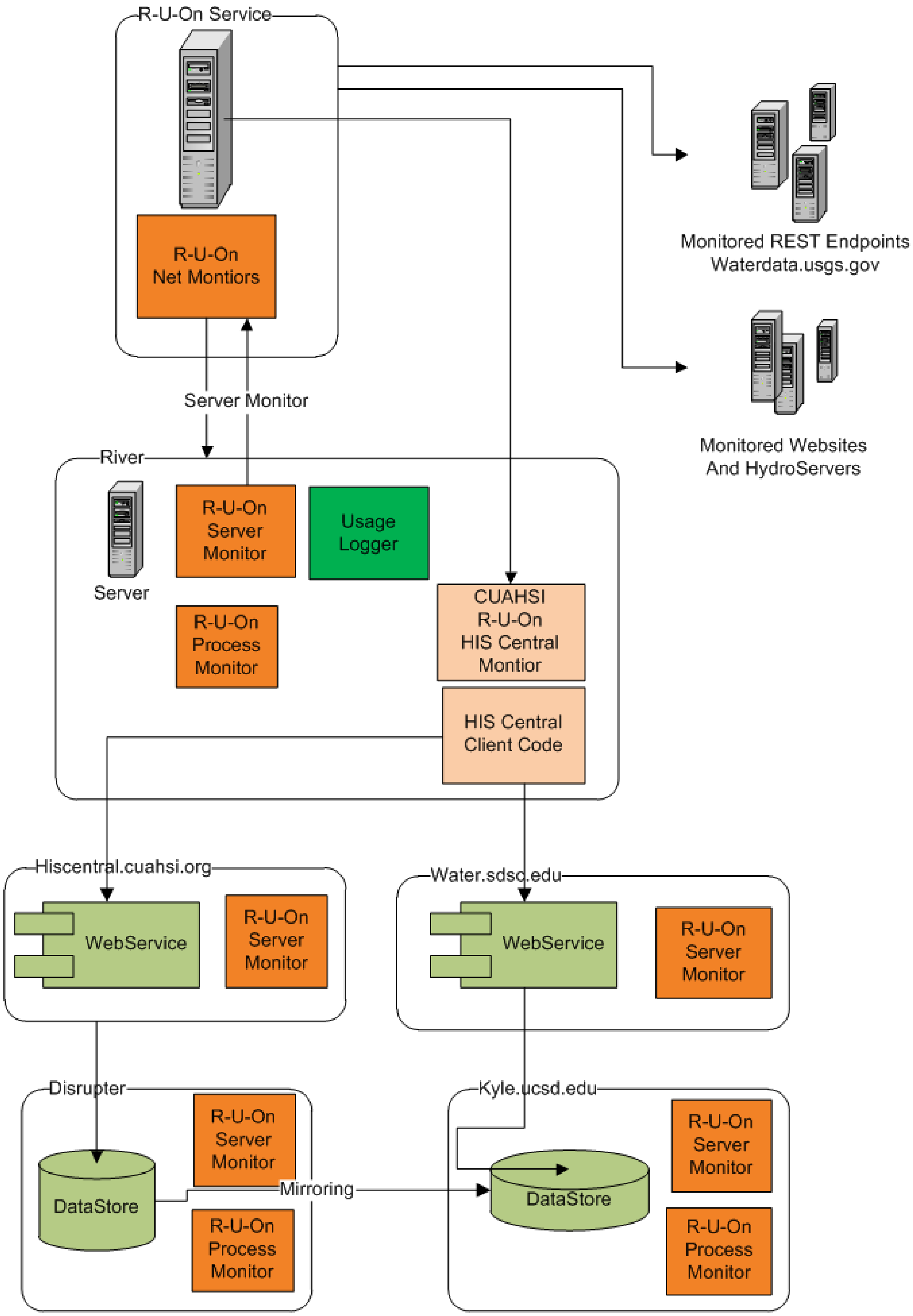


Figure 8. The current physical setup of HIS Facility at SDSC and the monitoring infrastructure

HIS Central Operations

While HIS is largely a federated system, centralized functionality is required to keep it coordinated and sustained. This functionality includes:

- Semantic mediation of controlled vocabularies and the ontology
- Managing HydroDesktop and HydroServer CodePlex source code repositories
- Engaging and supporting users through training, mailing lists, project web sites, wikis, etc.
- Operation of a HydroCatalog
- HydroServer hosting services
- Monitoring all aspects of the system and logging and reporting use
- Keeping HIS infrastructure up-to-date with respect to standards for hydrologic data exchange
- Support of data management in NSF projects
- Establishing policies and user agreements

Semantic Mediation System

The CUAHSI HIS Semantic Mediation System includes both the controlled vocabularies used by ODM and the hydrology domain ontology used to support data discovery in HydroCatalog. Maintaining these systems requires a web programmer and a database programmer who can see to the upkeep of the custom software and hardware that supports these systems, as well as implementing any changes or fixes that are required. Operation of the system for moderating the ODM control vocabulary content requires one or more moderators familiar with ODM controlled vocabularies and the science behind hydrologic information and measurements so that correct decisions on units and semantics are made. Maintaining the ontology will have similar requirements. In both cases, comprehensive knowledge of HydroServer data publication is required so that the context for semantic choices can be appreciated. Additional skills are required for maintaining CUAHSI HIS ontology, including understanding of Substance Registry System and other components of the parameter ontology, ability to update concept tables and generate concept tree visualizations, and manage and update mappings between concepts and parameters.

HydroDesktop and HydroServer CodePlex Repository Management

HydroDesktop and HydroServer software development is managed using the CodePlex repository. A coordinator for each project is required to sustain these efforts. Coordinators must have deep understanding of the structure of the code and the programming skills to design, code, compile, and release the applications. Since code may be contributed by different developers with different skill levels and varying understanding of the overall system, code reviews, code refactoring, unit testing, automated builds, and management of code trunk and branches are required. Additionally, the coordinators must be able to track and respond to bug reports and feature requests made via the CodePlex sites, as well as contribute substantively to the discussion forums in support of the community of developers who are contributing to these systems.

User support

User support requires comprehensive knowledge of the HIS System to the level that user questions can either be answered directly or directed to the person who is responsible for a given component of the system and that has the expertise to identify and resolve problems. This implies that multiple people will be involved in supporting users – a user support specialist who may be able to address many common issues, questions, and problems, as well as others who are more familiar with the in-depth workings of the HIS System components. User support will also require competence in interacting with users via diverse media (telephone, email, social networking, online forums) and the capability to select and set up appropriate media for the community. User support will include both data consumers and data publishers and will require conducting workshops and training sessions for both groups.

Operation of a HydroCatalog

Operation of a HydroCatalog requires implementing the hardware and software that supports HydroCatalog, as well as expertise in its use and configuration. Skills in web programming, systems administration (maintaining redundant setup, load balancing, etc.) and database management (indexing, partitioning, synchronization) are required. In addition, HydroCatalog management requires understanding the nuances of data from different sources, an ability to map them to the common information model, an ability to administrate and monitor metadata harvesting and data ingestion, and the ability to assist with the semantic tagging of variables.

HydroServer Hosting Services

Establishing a HydroServer for data publication may be beyond the ability or desire of many users. A HydroServer hosting service at HIS Central could fulfill the data publication needs for such users. The skills required for accomplishing this are the same as those required for a HydroServer data publisher. Additional skills include an ability to obtain organizational and institutional support for the hosting infrastructure, which will likely require a higher level of systems configuration and administration skills and capability to manage multiple, possibly virtual or cloud based servers. Several types of data hosting have been explored under this role: 1) hosting entire virtual HydroServers for users who prefer to maintain and manage a dedicated server but either cannot afford a physical server or do not have the institutional infrastructure to host a server; 2) an ODM/water data repository hosting for users who prefer to delegate data and service management to HIS Central; and 3) backups and archiving of data to HIS Central for redundancy.

Monitoring

HIS Central maintains a monitoring and reporting infrastructure. When new resources are added to HIS (services, catalogs, web sites, etc.), the monitoring system must be updated to include respective system monitors such as automated content validation routines, service availability monitoring, and service use and data access monitoring. This system must be maintained by an individual that is capable of specifying appropriate alert levels for system failures and addressees to be notified if troubleshooting is required. Additionally, data collected by the monitoring system must be compiled and made available so that service use and performance statistics can be compiled for reporting purposes.

Keeping HIS Infrastructure Up-to-Date with Standards

Reliance on community standards for service definitions and data exchange is a key component of sustained system operation. Understanding the existing and emerging standards in hydrology and other scientific domains, ability to make informed choices with respect to infrastructure updates for better standards compliance, developing and updating water data services as the standards progress, and communicating implementation experience and change requests to standards working groups are the skills required to ensure that the HIS infrastructure is kept up to date with respect to standards. Additionally, opportunities for collaboration with other cyberinfrastructure projects will need to be sought out to ensure that HIS is interoperable with other data publication, preservation, and synthesis efforts.

Supporting Hydrologic Data Management in NSF Funded Projects

Since October 2010, NSF requires data management plans to accompany all submitted proposals. In response to this, CUAHSI HIS prepared a statement of support for hydrologic data management plans. Multiple data management support requests have been directed to the project team since the Water Sustainability and Climate Request for Proposals was released, wherein HIS was recommended as a potential data management solution. This role will require CUAHSI to provide leadership to the hydrology community by assessing individual project needs and suggesting appropriate data management approaches for hydrologic data generated by projects. It will also require CUAHSI HIS to explore long-term data preservation options, which may be done in collaboration with other NSF funded cyberinfrastructure projects such as those funded by the DataNet program.

Establishing Policies and User Agreements

A governance system is needed to oversee the operation of any HIS facilities or programs to ensure that they are consistent with the mission of CUAHSI, and to set appropriate policies and software and data licensing and use agreements. CUAHSI's mission includes improving access to data, information and models (http://www.cuahsi.org/docs/stratplan/strat-plan-20101006.pdf). A governance system may include a standing committee comprised of members from academia and agency or industry partners. The role and membership of the CUAHSI Informatics standing committee is described at http://www.cuahsi.org/stdcomm-info.html.

Categories of Users and Features Provided by HIS

HIS targets three categories of users: researchers, educators and students, primarily at CUAHSI member universities, but also in the general water community at large. What should a researcher be able to do with HIS? What should an educator be able to do with HIS? What should a student be able to do with HIS? Answering these questions highlights the use cases that motivated creation of the system and motivate the need for ongoing efforts to make operational and further develop the system.

For a researcher, CUAHSI HIS should improve their ability to do the following:

- Get the data they need (if it exists) in a format that is unambiguous and easy to use. Establishing a water data services oriented architecture that provides data that is syntactically and semantically consistent with sufficient metadata to facilitate unambiguous interpretation strives to improve this capability.
- Discover what data of interest exists in an area of interest. The HIS information model includes search capabilities using ontology concepts that are linked to data series, supporting semantic search on concepts. GIS functionality supports spatial search functionality.
- Synthesize data from multiple sources. The water data services oriented architecture provides the capability to discover and integrate data from multiple sources.
- Combine data from other disciplines into a cross disciplinary analysis. The class of data that the HIS information model can support is quite broad and has been used for hydrologic, biologic, ocean and atmospheric data. The OGC standards approach, and in particular the WaterML standard, which is being harmonized with the OGC Observations and Measurements standard,is intended to ensure cross disciplinary interoperability of HIS data.

- Manipulate and analyze data. The HIS desktop application HydroDesktop is conceived as a general purpose framework to support the acquisition, manipulation and analysis of hydrologic data. In today's software ecosystem one tool cannot fit all purposes, rather the key is to develop interoperability between components. The HydroDesktop plugin capability endeavors to enable this. Already interoperability with the R statistical package, and modeling capability using OpenMI has been demonstrated.
- Manage the data being collected in an experimental watershed. This is a big challenge. Experimental watersheds typically gather a vast heterogeneous array of measurements that need to be organized so that they are accessible to the teams of scientists involved. Software and hardware is needed for getting data from the field to the grid. HydroServer has been developed to fill this capacity and the capability to ingest data streamed in from sensors and store and organize this data has been demonstrated. There are needs to add to this data management capability, keeping better track of laboratory sample measurements and managing access to the data so that it may be quality controlled and held private for first publication analyses by the researchers collecting it before being opened to the general public.
- Integrate and harmonize the data collected from multiple experimental watersheds so as to facilitate network scale analyses. This is another big challenge. Networks of observatories such as the Long Term Ecological Research (LTER) sites, Critical Zone Observatories (CZO) and Water, Sustainability and Climate (WSC) observatories are increasing. Both CZO and WSC are new NSF programs created in the last 4 years. The diversity of research across these observatories makes achieving integration and consistency at the data level difficult, yet that is often exactly what is required to do the transformative big science that these observatories strive to enable. The network of CZO's is building some network level data management capability taking advantage of HIS web services, semantic mediation and HydroServer technology.
- Advance the science of information technology in many domain science areas. There is the opportunity for fundamental computer and information science research centered on the HIS information model and architecture. The open source code repository (CodePlex) used for system development opens the door to engagement with this community.

For an educator CUAHSI HIS should improve their ability to do the following:

- Provide greater access to data in the classroom. HydroDesktop has already been used in an online inter institution class by the PI and co-PI (Maidment and Tarboton). This also involved a non HIS instructor (Irmak) at the University of Nebraska. Students used HydroDesktop to search for and acquire the data needed to gain an understanding of streamflow in their home state. Once capability such as HydroR and HydroModeler are better developed, these could support classroom modeling and analysis exercises and enrich the learning experience in advanced settings.
- Publish data collected by students in class learning experiences to encourage collaboration and collective learning. To enable this, it is important to streamline and simplify the systems for publishing data in HydroServer, which at present can be quite intimidating. Nevertheless Maidment and Tarboton have had student projects in their online class that have established and published data using HIS capability.
- Train the next generation of engineers and scientists to better use Information Technology by providing education on the information science aspects of HIS. The development of HIS has involved a large number of students. Students can be exposed to data management procedures and gain the necessary skills to support their own research. Additionally, The CodePlex open development environment provides a powerful capability for students to contribute new functionality to HIS at their appropriate skill level and learn the skills of integrated system development.

Students are in many instances researchers or participants in a class with educators so much of the functionality and features listed above applies directly to students too. Additionally HIS should improve students' ability to:

- Engage in independent learning and curiosity driven discovery. The wealth of data potentially at the fingertips of curious students opens tremendous possibilities. This is enhanced by the real time nature of much of the data so that students can explore hydrologic events (e.g., floods) as, or shortly after, they occur and become more informed participants in related societal response. To be successful in filling this role, free and easily accessible are essential features of HIS software. Platform independence is another key features. HydroDesktop, although quite sophisticated can be quickly downloaded and installed on run of the mill PC's and laptops. A Mac version is under development. The inclusion of online basemap information provides rich visual and information content that provides important context for curiosity driven discovery.

Current Status

Overall Architecture

One way to summarize the current status as regards the overall architecture of HIS is to say that a very large prototype of a services-oriented architecture for water observations data in the United States has been developed and demonstrated. The insights gained from building this prototype have revealed that the necessary functionality of this SERVICES-ORIENTED ARCHITECTURE can be performed by a conceptual design that consists of search services, metadata services, and data services. This conceptual design can be physically implemented using a number of different technologies, as demonstrated by the existing prototype and including OGC Web Services, or in ArcGIS.com map services. To achieve *semantic mediation*, or unifying of concepts in this information set, the observations variables in all data services must be associated with concepts in a common ontology, which CUAHSI has established, largely by building on existing work on semantic mediation between the USGS and EPA over the past several years. To achieve *syntactic mediation*, or uniformity of data format, the metadata and data responses must use a common specification, in the case of water observations metadata as a standardized set of map attributes where each observations series forms one feature symbolized by its point location in space, and in the case of the water data by use of a formally specified water markup language, WaterML.

Overall, the uptake of CUAHSI HIS has been satisfying. Some of the highlights include:

- USGS and NCDC publishing some data using HIS WaterML
- OGC Hydrology Domain Working Group evaluating WaterML as OGC standard
- ESRI using CUAHSI model in ArcGIS.com GIS data collaboration portal (Dangermond and Maidment, 2010)
- Kisters WISKI support for WaterML data publication
- Australia Bureau of Meteorology Water Accounting System has adopted aspects of HIS
- NWS West Gulf River Forecast Center Multi-sensor Precipitation Estimate published from ODM using WaterML

There are also implementations of CUAHSI HIS software components that others have developed for other platforms. ODM has been ported to MySQL and PostgreSQL at UC-Berkeley and UC-Davis for use in the Keck HydroWatch project (http://sensor.berkeley.edu/aboutDB.html). Besides USGS and NCDC, CUAHSI HIS water data services have been developed independently in Php (at Baltimore Hydrologic Observatory) and Java (at Phoenix

LTER). Various clients of web services have been also created, including a Google Earth based interface developed at CSIRO Australia (which is now an open source Codeplex project at http://his2kml.codeplex.com/).

Information Model and Community Infrastructure

Current distributable HIS software uses the ODM and WaterML 1.1 versions. The physical implementations of ODM, WaterML, HydroDesktop Database and HydroServer Capabilities database, while generally consistent are not precisely aligned in their information content and the need for some harmonization and generalization has been identified. There is also a need to include additional metadata required to support modeling. This includes information related to unit conversion factors, unit dimensions, multi-dimensional geometries, and provenance for storing multiple model runs. Some of these features are included in one or more of the information models used in the HIS, but again harmonization and generalization are required. We are looking to the Open Modeling Interface information model as our primary means for supporting modeling linking and coupling, therefore an important goal is to ensure interoperability between the HIS and OpenMI information models.

Moving towards a sustainable infrastructure for hydrologic data requires alignment of the data exchange model with community data exchange schemas and service interface standards, such as developed by the Open Geospatial Consortium. This alignment will let HIS: a) better engage government partners in hydrologic data sharing, as they are more likely to follow approved standards in publishing their data services, and b) use independently developed commercial software implementing the standards. Therefore, it will lower code development and maintenance efforts within the project, letting the HIS team focus on fundamental issues of hydrologic data interoperability and on operational data support of the hydrologic research community.

To this end, the project team has been actively participating in the Open Geospatial Consortium, working on closer alignment of CUAHSI HIS service model with OGC standards. The OGC/WMO Hydrology Domain Working Group (DWG) focuses on harmonizing existing encodings of hydrologic data and developing an agreed community model for hydrologic observations, hydrologic features, vocabularies, and service stack. One of the key outcomes of this effort is WaterML 2.0, a proposed standard for exchanging hydrologic time series, developed as a specialization of the OGC Observations and Measurement specification. The WaterML 2.0 schema will be presented at the upcoming OGC Technical Committee meeting and enter a Request for Comments period, preparing it for being voted in as an OGC standard. As part of the Hydrology DWG work, CUAHSI HIS team members contributed to WaterML, 2.0 development, participated in OGC Interoperability Experiments (Ground Water IE and Surface Water IE), and organized Hydrology DWG meetings in the U.S. Drawing from this work, the CUAHSI information models will be aligned with OGC models, and feedback on information model requirements for hydrologic analysis and modeling will be provided to OGC.

In addition to WaterML, several OGC specifications are directly relevant to the use cases supported by the CUAHSI HIS service-oriented infrastructure:

- Catalog Services for the Web (CSW): service interface specification defining how metadata about services, datasets and similar information objects can be published, searched and connected with
- Web Feature Service (WFS): service interface defining how geographic features can be exchanged between servers and clients
- Sensor Observation Service (SOS): service interface defining how observational data (from sensors or other measuring devices or procedures/algorithms) can accessed

- Geography Markup Language (GML): XML grammar for expressing geographic features, normally carried over WFS services
- Observations and Measurements (O&M) encoding standard defining an abstract model and XML encoding for observations, normally carries over SOS services.

In recent pilot work, a combination of these service interfaces and encoding schemas has been shown to provide some capability for hydrologic data discovery and retrieval, although there are still unresolved questions about the degree to which existing HydroCatalog search capability can be supported, as well as debate as to whether search is better implemented as a centralized or desktop function.

The HIS project team is currently evaluating the use of WFS to publish data series metadata and is working on a specification describing the precise set of time series metadata fields to use in such an approach. This specification is being developed in consultation with USGS and ESRI. Besides the metadata fields normally relayed via WaterOneFlow's GetSites, GetSiteInfo and GetVariableInfo requests, these agency-hosted WFS services will include associations between measured parameters and concepts defined in the EPA/USGS Substance Registry System. This is intended to enable semantics-based discovery of time series records directly in agency catalogs, eliminating the need for harvesting and tagging them to the central catalog.

Several issues will need to be addressed if this new approach is to be adopted: a) managing large WFS services (e.g. the expected number of records in the USGS WFS service we are discussing may exceed 18 million), b) exposing parameter and concept information from WFS service in the CSW registry, sufficient to select registered services for further analysis. The team is currently experimenting with potential solutions to these challenges, ensuring that existing functionality, reliability and robustness of catalog search are not compromised in the process.

HydroServer

The functionality described in the functionality section above is all complete and available to users from the HIS website. Authentication and Access Control have been identified by the hydrology community as a major missing capability, as has as the need for simple web based data loading functionality. We are currently working on a prototype for a web data loader that is nearing completion and enables users to load data into an ODM database that is hosted on a HydroServer. We have also completed scoping and preliminary design for an authentication and access control system for HydroServer.

Although HydroServer is a viable option for sharing hydrologic data, there is currently no model for long term archival and curation of hosted datasets. Data hosted on a HydroServer will remain available as long as the HydroServer is running, but if a HydroServer is shut down or cannot be maintained it isn't clear what can or should happen to the data. Because of this, it is clear that a community data repository is needed that can accept data on a permanent basis from individuals or organizations that either do not have the expertise or cannot afford to set up or maintain a HydroServer on their own.

HydroCatalog

As described above, HydroCatalog provides capability for registering and searching academic water data services, and keeping time series metadata for multiple remote services up-to-date. It also enables search over multiple agency catalogs; however keeping the agency catalog information up-to-date and available for concept-based search has been hampered by the laborious procedures of ingesting agency catalog dumps and semantic tagging of parameters. The reliance on custom code in HIS Central implementations of HydroCatalog has also been identified as a critical limitation that is a concern for sustainability. As a result an effort is underway to transition HydroCatalog functionality to a standards based framework relying on OGC standards and ESRI Geoportal catalog services. Work is required to ensure that functionality is not lost in this transition. A new HydroCatalog approach will be initially piloted for selected federal agency sources: work is ongoing with USGS to have NWIS catalog information served via a standards-based mechanism using CSW. At the same time, we will create additional WFS service endpoints for registered academic water data services, to make them compatible with the new model.

HydroDesktop

HydroDesktop is undergoing rapid development with 7 releases since March 25, 2010. It has been used in a GIS in Water Resources class. Significant new search, basemap, modeling and analysis functionality have been added. This includes major refactoring to simplify the user interface based on feedback received at the 2009 AGU Fall Meeting. The new interface is more intuitive with a simple set of tabbed “ribbon controls” at the top of the main form, providing easy access to map, graph, table, modeling, and other views. We have also simplified and more tightly integrated the search module providing default values for most search parameters and saving search results as a standard shapefile that can be opened in third party GIS software. The HydroDesktop plugin interface has been streamlined such that third party developers can now more easily develop plugins that only extend the GIS aspects of the software or fully take advantage of the local HIS database. These plugin improvements had the effect of speeding up the user interface when doing simple tasks such as switching between views. These functionality improvements were accompanied by over 2000 code commits by over 25 registered developers on the code management system (http://hydrodesktop.codeplex.com). HydroDesktop has been downloaded over 12,800 times from this repository.

While the software continues to evolve on a fairly rapid basis, we have recognized the need to provide a degree of consistency and usability to the software through a well defined release management plan. One of the key elements of that plan which was recently introduced was the introduction of the use of a “branch” and “trunk” approach to managing the code base. This is a fairly standard practice amongst software development teams and allows the developers to create a release as a “branch” of the software code base which remains somewhat stable – now new functions are allowed to be introduced to a branch, only bug fixes. Parallel to the release branch, the source code “trunk” continues to move forward with new functions, features and bug fixes in preparation of another official release. Currently we are working on an active branch, version 1.1, and an active trunk which is moving toward the functional specifications of version 1.2.

HydroModeler is a plug-in application that extends the core HydroDesktop application to support hydrologic modeling. HydroModeler makes use of the Open Modeling Interface (OpenMI) standard and OpenMI Association Technical Committee (OATC) Software Development Kit (SDK) to provide a “plug-and-play" modeling framework within HydroDesktop. By adopting the OpenMI standard and by leveraging core HydroDesktop functionality for data management and visualization, HydroModeler is able to focus on the more specific task of running integrated model configurations. To support the use of the OpenMI for creating process-level hydrologic modeling components, we designed and implemented a SDK called the Simple Model Wrapper (SMW) (Castronova and Goodall, 2010). We have conducted tests of the SMW and shown that it can be used to simulate rainfall/runoff as a set of interlinked model components (Castronova and Goodall, In Revision). We are now using the SMW to create a library of hydrologic model components, and documenting the process of creating new model components in order to encourage community participation. As more models are refactored to follow the OpenMI standard, we anticipate that HydroModeler will become an increasingly powerful application for hydrologic and water resource modeling.

DATA PUBLICATION

As of November 2010, the HydroCatalog hosted information about 58 public services (over 100 services total, including services registered at HIS training workshops or intentionally designated as "private" by service publishers). The services provide information about 18,000+ hydrologic variables measured at over 1.96 million sites and organized into 23.3 million time series which are searchable using the HIS Central services. The services reference a total of 5.1 billion data points available via standard GetValues requests.

The current content of the federal agency services is presented in the following table:

Network Name	Site Count	Value Count	Earliest Observation	Notes
NWISDV	31852	303843342	01/01/1861	WaterML-compliant GetValues service from NWIS, catalog ingested
EPA	236094	78076394	01/11/1900	SOAP wrapper over WQX services, catalog ingested
NWISUV	11758	84530304	60 DAYS	WaterML-compliant GetValues Service, catalog ingested
NCDC ISH	24770	3000000*	1/1/2005	WaterML-compliant GetValues service from NCDC
NWISIID	375831	86485762	9/1/1867	SOAP wrapper over NWIS web site, catalog ingested
NWISGW	833681	8491383	1/1/1800	SOAP wrapper over NWIS web site, catalog ingested
RIVERGAGES	2206	263591424	1/1/2000	WaterML compliant REST services from Army Corps of Engineers

(*) Estimated

In addition to the federal agencies, the Central HIS catalog holds metadata from several state-wide and regional repositories (e.g. Chesapeake Bay Information Management System, regional NWS repositories, precipitation data from NADP, MPE and HydroNEXRAD, Florida, Texas and Vermont state agencies) and multiple academic sources. This large variety of observation networks, with different spatial and temporal patterns of data collection and different update frequencies allowed us to tune HydroCatalog into a comprehensive and robust system that can support a variety of client applications and search trajectories.

The number of data retrievals recorded by the system has been increasing steadily (Figure 9).

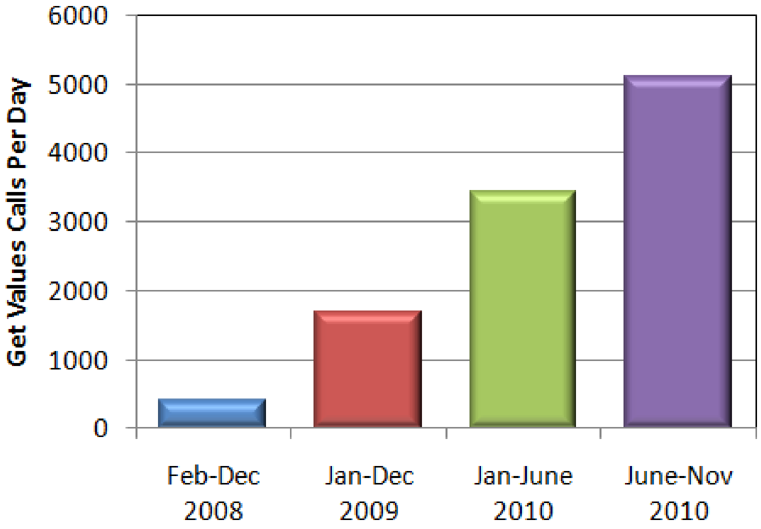


Figure 9. Growth in GetValues calls for all services reporting to HIS Central

2011 Plan

Our strategic goal for 2011 is to produce HIS version 2.0. We are presently at HIS version 1.1. HIS 2.0 has been conceived as a standards based refactoring of the architecture to make it more sustainable. A proposed approach for doing this is given in a draft white paper prepared in consultation with ESRI and USGS partners at a recent working meeting (Maidment et al., 2010). Planned activities for 2011 include:

(1) Develop a complete functional specification for standards based web services for sharing water observations data. This includes, but is more than, the WaterML 2.0 specification being developed through the OGC process because it needs to specify and standardize the function calls involved in publishing water data. This specification, possibly in the form of a Best Practices document approved through the OGC Hydrology Domain Working Group, is required to promote consistency across multiple agencies and organizations publishing water data using the CUAHSI HIS model.

(2) Develop a strategy for searching for weather and climate data grids and remote sensing grids that is compatible with what we are currently doing in the search for time series using a single CUAHSI Ontology.

(3) Evaluate the extent to which the current HIS software stack can be transitioned into a standards based paradigm and what the trade-offs are in doing so.

(4) Develop a complete specification for HydroCatalog interfaces so that HydroCatalogs can be implemented in other locations (e.g. Texas) and water data agencies (e.g. Data.Gov or GeoData.Gov). This will involve defining and agreeing upon the precise functionality that HydroCatalog supports. This functionality should ideally support single URL service registration and harvesting based on HydroServer metadata services.

(5) By the end of 2011, all HydroServers to be self describing supporting their own data and metadata. Metadata housed in a HydroCatalog at HIS Central to be derived from HydroServer metadata web services as the primary source and used to support integrated search.

(6) Work with USGS, EPA, NCDC and other federal agencies to encourage them to stand up the HIS Metadata-Data services that are required for their metadata to be complete and harvestable by one or more HydroCatalogs. Focus in particular on USGS and EPA so that we can avoid having local metadata files for their services at HIS Central and have a fully distributed metadata system operational for their information.

(7) Demonstrate using the Texas HIS data compilations how a state-based data system constructed according to CUAHSI specifications can be integrated with federal and local data to provide more complete and integrated water observations information coverage at the state level. Work with ESRI to show how a gallery of water observations metadata maps with attached time series information can be posted on ArcGIS.com and made publicly available and searchable as map services.

(8) Continue to develop HydroDesktop and harden it to get out the bugs. Release an operational version to the general CUAHSI community. Release a version that runs on the Macintosh through the mono framework. Create class exercises for hydrology class that show how to use HydroDesktop in a teaching setting. Provide web-based training seminars so people can learn how to use HydroDesktop and the HIS services architecture.

(9) Extend HydroDesktop so that it can search across weather and climate grids and remote sensing information and download needed datasets into a folder within HydroDesktop. Be able to display such information in HydroDesktop. The goal is for HydroDesktop to be able to access the full content of HydroServer geospatial services (i.e. OGC WFS, WMS, WCS) as well as common grid and NetCDF data formats.

(10) Evolve the core database model within HydroDesktop so that it is fully OpenMI compliant and have a library of OpenMI compliant models and model components that can run within the HydroDesktop Hydromodeler extension.

(11) Extend HydroServer so that it supports Authentication and Access Control functionality that users have frequently asked for. Implement a Web Data Loader for ODM databases hosted on a HydroServer to ease data loading. Support on HydroServer the storage and publication of ontology mapping for variables within ODM databases. Support standards based metadata services based on WaterML 2.0 in coordination with transitioning other HIS functionality to this approach. Create REST endpoints for HydroServer services.

(12) Have an HIS User Conference in 2011 to present all this to the CUAHSI community. We are evaluating the merits of either late Spring (May) or Fall (October) for doing this.

Specification of complete metadata for HydroServer functionality will necessitate some revisions to the common conceptual information model that is used in ODM, WaterML, HydroDesktop Database and HydroServer Capabilities database to better integrate the ontology structure in this model as well as new informational elements requested by the community. This will involve coordination with the new NSF –funded Hydrologic Ontology project to develop a system for community management of the hydrology domain ontology. We will need to establish a systematic procedure for upgrading components of the HIS to implement changes made to the common information model.

Future Funding

Efforts are under way to secure funding for HIS development to continue once the current funding ends. In planning and seeking future funding, the following considerations need to be addressed:

- How is it best to provide ongoing production level support for HIS? CUAHSI HIS has been identified in some NSF program solicitations as a solution for data archiving requirements and all NSF proposals in the future are required to address data management. HIS is actively supported by CUAHSI through a full time User Support Specialist, periodic user training workshops, and promotion at national meetings such as AGU. How should this evolve and become sustainable?
- What should be the interface between the operational system and innovative development? The present project has defined the framework and provided a large scale prototype for a services oriented architecture hydrologic data publication and sharing system. As this moves in to operational use there are efforts to continue to innovate. A proposal has been developed for expansion of HIS in interactive data access and model and tool sharing through an online, collaborative environment referred to as CUAHSI Online. CUAHSI Online is envisioned as a web-based collaboration portal that will enable scientists to easily discover and access data and models, retrieve them to their desktop or perform analyses in a high performance computing environment and thereby enhance research, education and application of hydrologic knowledge. CUAHSI Online will add a new class of functionality to HIS that builds on the concepts of social networking to enable simplified collaborative data, model and tool sharing.
- How should contributions from multiple participants with separate funding be coordinated and sustained? CUAHSI HIS has grown bigger than one funded NSF project. There are now multiple funded projects involving HIS team members and others in the community that use HIS contributions in some way shape or form. A challenge is how to develop a collaboration mechanism that coordinates these contributions into a coherent system.

References

Beran, B. and M. Piasecki, (2009), "Engineering new paths to water data," Computers & Geosciences, 35(4): 753-760, http://dx.doi.org/10.1016/j.cageo.2008.02.017.

Castronova, A. M. and J. L. Goodall, (2010), "A generic approach for developing process-level hydrologic modeling components," Environmental Modelling & Software, 25(7): 819-825, http://dx.doi.org/10.1016/j.envsoft.2010.01.003.

Castronova, A. M. and J. L. Goodall, (In Revision), "Simulating watersheds using loosely integrated model components," Journal of Hydrologic Engineering.

Dangermond, J. and D. Maidment, (2010), "Integrating Water Resources Information Using GIS and the web," 2010 AWRA Spring Specialty Conference Geographic Information Systems (GIS) and Water Resources VI, Orlando Florida, American Water Resources Association, Middleburg, Virginia, TPS-10-1, http://www.awra.org/orlando2010/doc/awrakeynote.pdf.

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.

Horsburgh, J. S., D. G. Tarboton, M. Piasecki, D. R. Maidment, I. Zaslavsky, D. Valentine and T.Whitenack, (2009), "An integrated system for publishing environmental observations data," Environmental Modelling & Software, 24(8): 879-888, http://dx.doi.org/10.1016/j.envsoft.2009.01.002.

Horsburgh, J. S., D. G. Tarboton, K. A. T. Schreuders, D. R. Maidment, I. Zaslavsky and D. Valentine, (2010), "Hydroserver: A Platform for Publishing Space-Time Hydrologic Datasets," 2010 AWRA Spring Specialty Conference Geographic Information Systems (GIS) and Water Resources VI, Orlando Florida, American

Water Resources Association, Middleburg, Virginia, TPS-10-1, http://www.awra.org/orlando2010/doc/abs/JefferyHorsburgh_7cb420e3_6602.pdf.

Josuttis, N. M., (2007), SOA in practice - the art of distributed system design, O'Reilly Press, Sebastapol, CA, 324 p.

Maidment, D., T. Whiteaker, J. Seppi, F. Salas, H. Sangireddy, I. Zaslavsky and D. Valentine, (2010), "Sharing Water Observations Data Using Web Services," Draft paper.

Piasecki, M. and B. Beran, (2009), "A semantic annotation tool for hydrologic sciences," Earth Science Informatics, 2(3): 157-168, http://dx.doi.org/10.1007/s12145-009-0031-x

Tomlinson, R., (2003), Thinking about GIS, ESRI Press, Redlands CA, 283 p.

Zaslavsky, I., D. Valentine and T. Whiteaker, (2007), "CUAHSI WaterML," OGC 07-041r1, Open Geospatial Consortium Discussion Paper, http://portal.opengeospatial.org/files/?artifact_id=21743.