Data Interoperability in the Hydrologic Sciences

The CUAHSI Hydrologic Information System

David G Tarboton[1], David Maidment[2], Ilya Zaslavsky[3], Dan Ames[4], Jon Goodall[5], Richard P Hooper[6], Jeff Horsburgh[1], David Valentine[3], Tim Whiteaker[2], Kim Schreuders[1]

[1] Utah State University
[2] University of Texas at Austin
[3] San Diego Supercomputer Center
[4] Idaho State University
[5] University of South Carolina
[6] Consortium of Universities for the Advancement of Hydrologic Science, Inc.

dtarb@usu.edu, maidment@mail.utexas.edu, zaslavsk@sdsc.edu, dan.ames@isu.edu, goodall@cec.sc.edu, RHooper@cuahsi.org, jeff.horsburgh@usu.edu, valentin@sdsc.edu, twhit@mail.utexas.edu, kim.schreuders@usu.edu

***Abstract*—The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) has been established to promote research infrastructure that advances Hydrologic Sciences. Hydrologic Information Systems (HIS) are part of this infrastructure. Hydrologic information is collected by many individuals and organizations in government and academia for many purposes, including general monitoring of the condition of the water environment and specific investigations of hydrologic processes and environments. This paper describes HIS capability developed to promote data sharing and interoperability in the Hydrologic Sciences with the ultimate goal of enabling hydrologic analyses that integrate data from multiple sources. The CUAHSI HIS is an internet based system to support the sharing of hydrologic data. It is comprised of hydrologic databases and servers connected through web services as well as software for data publication, discovery and access. The system that has been developed provides new opportunities for the water research community to approach the management, publication, and analysis of their data systematically. The system's flexibility in storing and enabling public access to similarly formatted data and metadata has created a community data resource from public and academic data that might otherwise have been confined to the private files of agencies or individual investigators. HIS provides an analysis environment for the integration of data from multiple sources and serves as a prototype for the infrastructure to support a network of large scale environmental observatories or research watersheds.**

Keywords—Hydrologic Information System; Web services; Data Model; Hydrology

I. Introduction

The advancement of hydrologic science is critically dependent on the assembly and synthesis of hydrologic data. The Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) is an organization representing 135 universities and affiliated organizations, funded by the US National Science Foundation, to develop community infrastructure and services to advance hydrologic science. This paper describes the CUAHSI Hydrologic Information System (HIS), a community information systems technology project to improve access to hydrologic data.

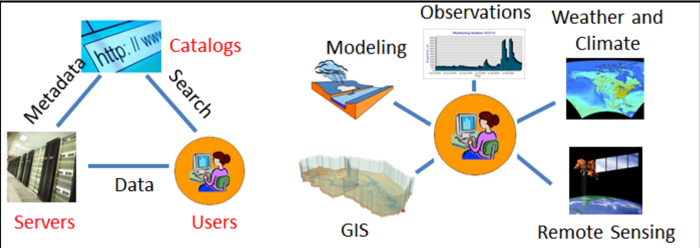
The CUAHSI HIS project [1, 2] has as a goal the development of standards, systems, and software to enhance access to and interoperability among water data from multiple sources. We have built a prototype system centered on a services-oriented architecture [3] that defines the interfaces between system components for publishing, cataloging and accessing hydrologic data and a desktop hydrologic information system that supports the integration and analysis of hydrologic data retrieved from multiple sources.

II. Architecture

Two concepts, (1) the services oriented architecture; and (2) the desktop hydrologic information system underlie the architecture of the system that we are developing (Fig. 1).

The HIS services-oriented architecture can be viewed as: 1) a way of publishing hydrologic data in a uniform way; 2) a way of discovering and accessing remote water information archives in a uniform way; and 3) a way of displaying, synthesizing and analyzing water information and exporting it to other analysis and modeling systems. The connections among components are established by web services.

The concept of HIS desktop application software is somewhat analogous to Geographic Information System (GIS) desktop software that supports storage and analysis of logically linked data [4]. Our implementation, "HydroDesktop" provides an analysis environment within which data from multiple sources can be discovered, accessed and integrated.



a) Services-Oriented Architecture b) Desktop Hydrologic Information System

Figure 1. Hydrologic Information System Overarching Vision.

We have developed prototype functionality for all three components of the services oriented architecture and data transmission formats for the data exchanges between them. 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 and has not yet developed the capability to integrate weather, climate and remote sensing data illustrated in Fig. 1, but does synthesize GIS, point observations and time series and modeling.

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) (Fig. 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.

The components shown in Fig. 2 either publish or consume information via the following categories of web services:

- 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.

The formats for transmission of information between these systems and the interfaces that enable the communication between them (the connecting arrows in Fig. 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 [5]. The web services are referred to as WaterOneFlow web services. CUAHSI HIS also relies on other established standards such as World Wide Web Consortium Simple Object Access Protocol (SOAP) and Open Geospatial Consortium (OGC) Geographic Markup Language (GML) for transmission of information between the three primary components.

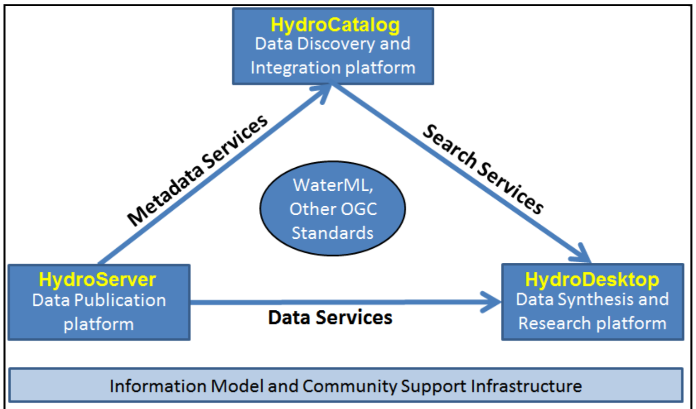

HydroCatalog
Data Discovery and Integration platform
Metadata Services
Search Services
WaterML, Other OGC Standards
HydroServer
Data Publication platform
Data Services
HydroDesktop
Data Synthesis and Research platform
Information Model and Community Support Infrastructure

Figure 2. Components of CUAHSI HIS Services Oriented Architecture.

At the base of Fig. 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 using the Observations Data Model (ODM) [6] relational database and the HydroServer Capabilities Database to ensure that data and metadata are stored together. The information model 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 a defined 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 these semantics. A web site collects and manages community additions and edits to controlled vocabulary content to allow dynamic growth of this content while encouraging semantic consistency across the user community.

The architecture shown in Fig. 2 has evolved as an approach for sharing hydrologic observations data that is general and open to allow broad participation. 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. 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. HydroDesktop is our prototype client for consumption of web service based hydrologic data, but this does not preclude others from establishing their own clients.

III. 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 [7]. 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).

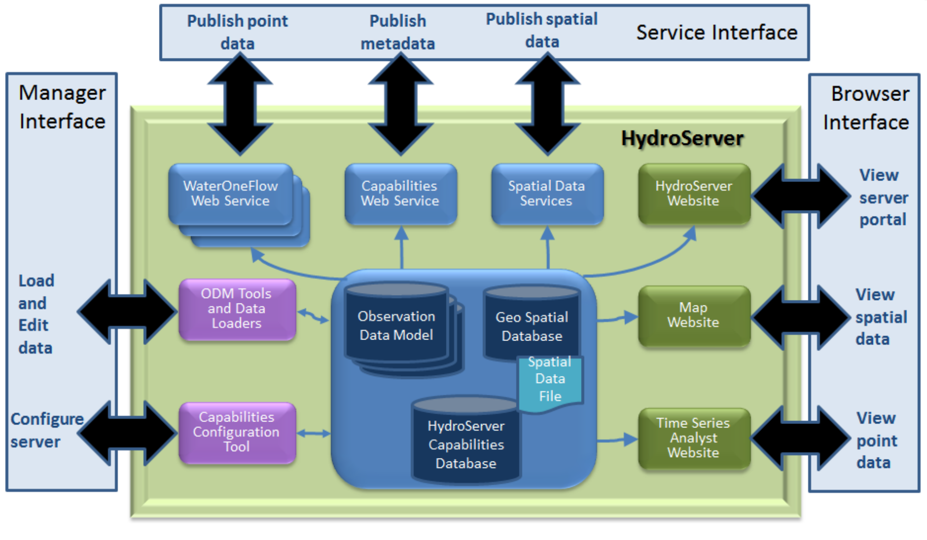


Figure 3. HydroServer Architecture and Functionality

An important principle that has emerged from our work on HydroServer , is that server 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. A catalog should not be required to aquire or generate additional metadata when supporting the discovery of data from a HydroServer.

HydroServer (Fig. 3) supports publication of both point observations data stored in one or more ODM databases [6] 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 data and web services it publishes. 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 can be discovered and metadata harvested automatically by registration and cataloging services (HydroCatalog), making a HydroServer self-describing. 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. A configuration tool has been built that provides an interface for defining the contents of the Capabilities Database. The ODM Tools suite and capabilities configuration tool comprise the data manager interface.

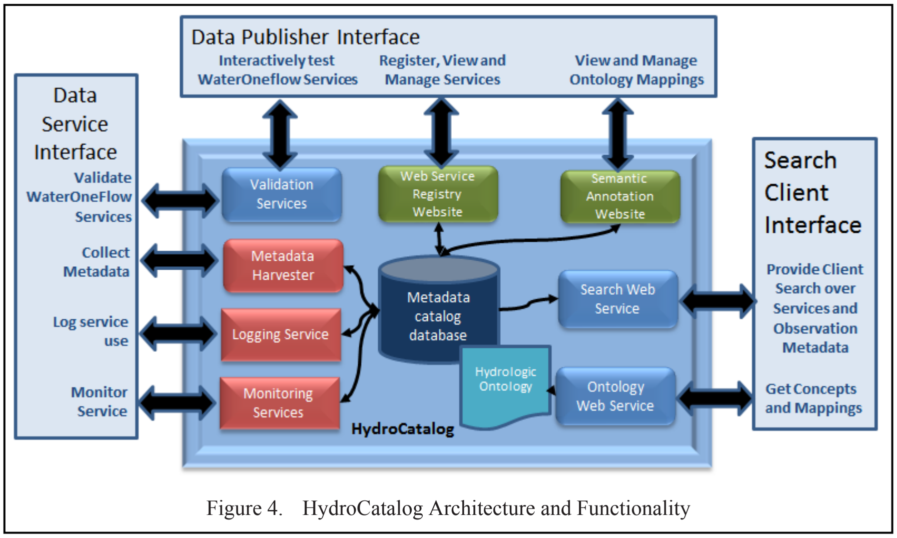
Finally, a suite of data presentation and visualization tools has been created for HydroServer. The suite includes the HydroServer Website, the Time Series Analyst, and the HydroServer Map Website. These provide a public browser accessible graphical user interface to the data holdings of the HydroServer.

IV. HYDROCATALOG

HydroCatalog is the discovery component of the system linking data publishers and application clients. Data discovery across multiple 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 (Fig. 4).

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 nodes of this hierarchy. Data publishers first register their WaterOneFlow web services with the HydroCatalog Web Service Registry. 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 does 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



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

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 spatial, temporal, and semantic searches across all sources of data in the catalog. 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. HydroCatalog software is open source software managed at (http://hydrocatalog.codeplex.com).

V. HydroDesktop

HydroDesktop is a free and open source Desktop Hydrologic Information System (Fig. 5) that helps users discover, use, manage, analyze and model hydrologic data.

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 (Fig. 5) is structured to take advantage of centralized cataloging functionality from HydroCatalog as well as distributed data from HydroServers.

The DotSpatial project (http://dotspatial.codeplex.com/) has been under development by members of the HIS team as well as an international open source volunteer community and members of the MapWindow project (see mapwindow.org) since April 2010. Since it's first release, DotSpatial has been downloaded over 40,000 times and it currently receives approximately 200 downloads per day by user-developers exploring free and open source alternatives for GIS enabled custom software targeting the Microsoft Windows operating system.

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

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.

Once data have been downloaded into the HydroDesktop database, they can be immediately viewed graphically or

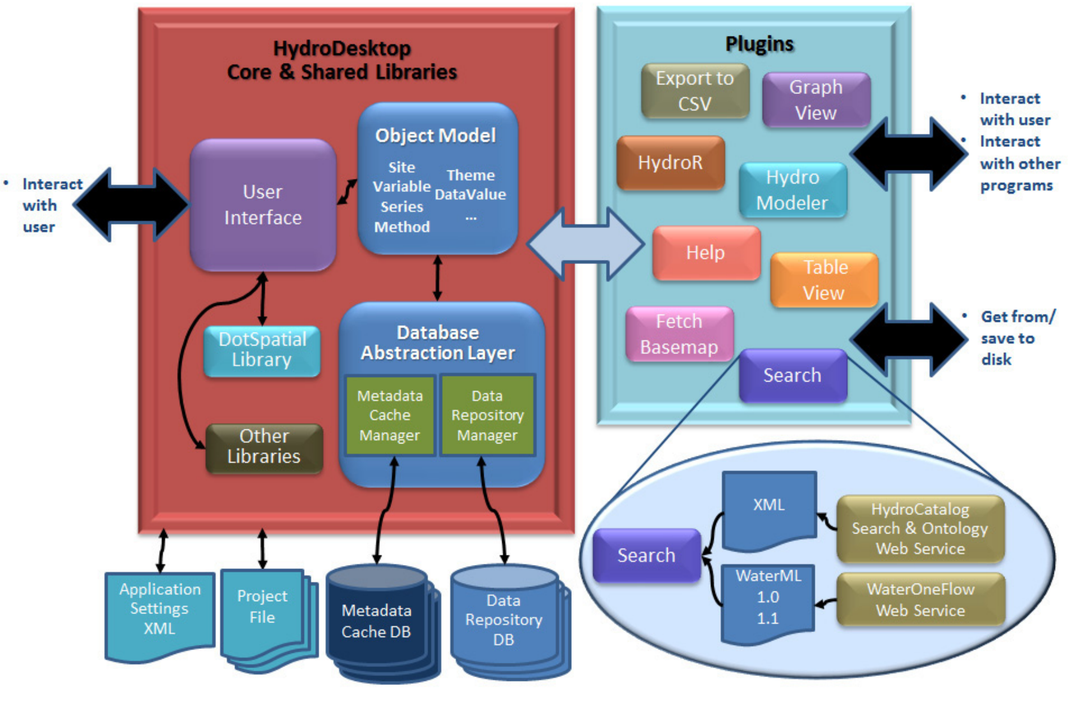


Figure 5.   HydroDesktop Architecture and Functionality

tabularly through a “Graph View” plugin and a “Table View” plugin respectively. 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 tabular form and export the data to a comma separated values (CSV) file. The “Edit View” plugin enables editing.

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 coupling with 3rd party software applications. Specifically two unique and very powerful plugins have been constructed for HydroDesktop. The first is a plugin called HydroModeler that leverages the OpenMI modeling framework developed under European Union funding. 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 an implementation of the 1.4 version of the OpenMI standard and specifically allows modelers to read HIS derived datasets into their models and write model outputs back into the HydroDesktop database.

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. Fig. 6 illustrates the HydroDesktop interface highlighting the integration of data from multiple sources and combining, map, graph and search capability.

VI. Use and Community Support

Table 1 summarizes the data available and its recorded use from instances of HydroServer registered with the CUAHSI HydroCatalog at SDSC. There is also use of the open source software that is downloaded by others and not registered here for which we do not have data. Standard HydroServer refers to installations, typically at universities, that have used the HydroServer software stack we have developed to publish data. Hybrid HydroServer refers to large existing federal datasets that the HIS project has wrapped with a WaterOneFlow web service.

The United States Geological Survey (USGS) and the National Climatic Data Center (NCDC) have 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.

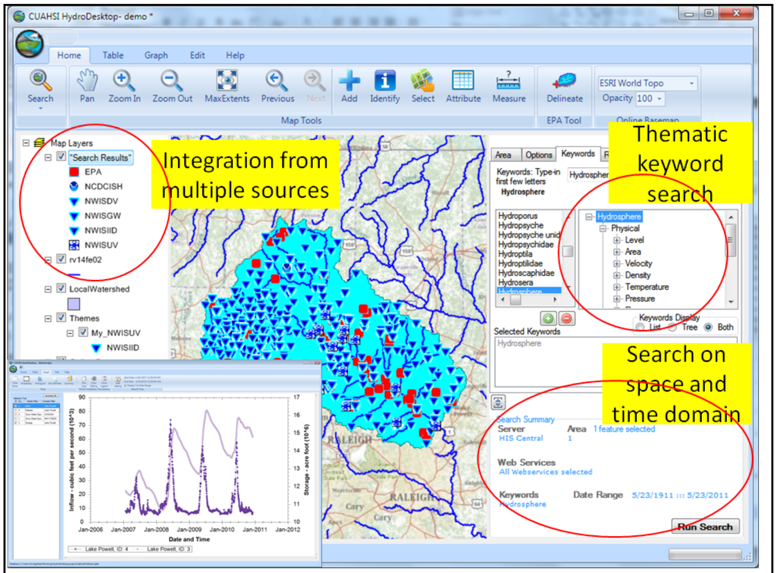


Figure 6. HydroDesktop Interface Illustration

TABLE I. CUAHSI HydroServer Use Data

	Standard HydroServer	Hybrid HydroServer
Number of registered WaterOneFlow data services	66	6
Number of sites	462,992	1,490,113
Number of variables	5,978	6,892
Number of data values	>4 billion	>0.9 billion
Number of GetValues requests 7/1/2009-6/30/2010	46,055	64,810[b]
Number of GetValues requests 7/1/2010-6/30/2011	571,560[a]	43,723[b]

a. 435,762 of these are from the new West Gulf River Forecast Center NEXRAD precipitation data service that started in the latter year.

b. These are dominated by USGS NWIS Unit Values requests that dropped off when services to obtain this data directly from the USGS became available.

Reliance on independently developed and governed standards is one of the key elements of project sustainability. Other considerations that support sustainability 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

Development of HIS is done under the auspices of CUAHSI with 135 member organizations (mostly university), which sets policies such as software licensing, data publication and data use agreements. CUAHSI is advised by its Informatics Standing Committee that provides user and community input on priorities and needs necessary to support the academic research community.

VII. Conclusions

There is a fundamental need within the hydrologic and environmental engineering communities for new, scientific methods to organize and utilize observational data that overcome the syntactic and semantic heterogeneity in data from different experimental sites and sources and that allow data collectors to publish their observations so that they can easily be accessed and interpreted by others. The tools and partnerships that CUAHSI HIS has developed provide: (1) **Data Storage** in an Observations Data Model (ODM) and publication through HydroServer; (2) **Data Access** through internet-based WaterOneFlow web services using a consistent data language, called WaterML from HydroDesktop; (3) **Data Discovery** through a National Water Metadata Catalog and thematic keyword search system at HydroCatalog and (4) **Integrated Modeling and Analysis** within HydroDesktop. These functions support a high level of interoperability for hydrologic data. Beyond technical aspects, HIS has also focused on scientific, organizational, and infrastructure aspects of hydrologic data integration, which represent an important part of its contribution – in particular building partnerships with major federal and state agencies to incorporate their data into the system and ingrate with data provided by multiple academic partners.

The HIS is a federated system linking data from multiple providers. As such, data availability and quality does depend on it being maintained by the provider. The ODM data model provides capability to document data sources, methods and quality controls, but there is no filter on data quality that may be published using HIS technology. In this respect the system is also like much other information on the internet, buyer beware, user's need to assess for themselves the suitability of data for a particular purpose. As with broken links on the internet, when servers go down data becomes unavailable. The system does enable the capability for institutions to establish data centers to store data that is critical to them and CUAHSI is working to establish such a long term data center to archive community data.

The combination of HIS capabilities creates a common window on water observations data for the United States unlike any that has existed before, and is also extensible worldwide. This system represents new opportunities for the water research community to approach the management, publication, and analysis of their data systematically. The system's flexibility in storing and enabling public access to similarly formatted data and metadata has created a community data resource from public and academic data that might otherwise have been confined to the private files of agencies or individual investigators. HydroDesktop provides an analysis environment for the integration of data from multiple sources and serves as a prototype for the infrastructure to support a network of large scale environmental observatories or research watersheds.

For more information about the CUAHSI HIS and access to the tools and code, all freely distributed and open source, under the Berkeley Software Distribution (BSD) license, go to our website: http://his.cuahsi.org.

Acknowledgment

Funding for this work, by the U.S. National Science Foundation under grant EAR 0622374 is greatly appreciated. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] D. G. Tarboton, D. Maidment, I. Zaslavsky, D. P. Ames, J. Goodall, and J. S. Horsburgh, "CUAHSI hydrologic information system 2010 status report," 2010. http://his.cuahsi.org/documents/CUAHSIHIS2010StatusReport.pdf.

[2] D. G. Tarboton, J. S. Horsburgh, D. R. Maidment, T. Whiteaker, I. Zaslavsky, M. Piasecki, J. Goodall, D. Valentine, and T. Whitenack, "Development of a community hydrologic information system," in *18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation*, 2009, pp. 988-994 http://www.mssanz.org.au/modsim09/C4/tarboton_C4.pdf.

[3] N. M. Josuttis, *SOA in practice - the art of distributed system design*. Sebastapol, CA: O'Reilly Press, 2007.

[4] R. Tomlinson, *Thinking about GIS*. Redlands CA: ESRI Press, 2003.

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

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

[7] J. S. Horsburgh, D. G. Tarboton, K. A. T. Schreuders, D. R. Maidment, I. Zaslavsky, and D. Valentine, "Hydroserver: A platform for publishing space-time hydrologic datasets," in *2010 AWRA Spring Specialty Conference Geographic Information Systems (GIS) and Water Resources VI*, Orlando Florida, 2010 http://www.awra.org/orlando2010/doc/abs/JefferyHorsburgh_7cb420e3_6602.pdf.