18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Development of a Community Hydrologic Information System

Tarboton, D.G. [1], **J.S. Horsburgh** [1], **D.R. Maidment** [2], **T. Whiteaker** [2], **I. Zaslavsky** [3], **M. Piasecki** [4], **J Goodall** [5], **D. Valentine** [3], **T. Whitenack** [3]

[1] *Utah Water Research Laboratory, Utah State University, Logan, Utah, USA*
Email: david.tarboton@usu.edu
[2] *Center for Research in Water Resources, University of Texas at Austin, Austin, Texas, USA*
[3] *San Diego Supercomputer Center, University of California at San Diego, San Diego, California, USA*
[4] *Department of Civil, Architectural, and Environmental Engineering, Drexel University, Philadelphia, Pennsylvania, USA*
[5] *Department of Civil and Environmental Engineering, University of South Carolina*

Abstract: Over the next decade, it is likely that science and engineering research will produce more scientific data than has been created over the whole of human history. The successful use of these data to achieve new scientific breakthroughs will depend on the ability to access, integrate, and analyze these large datasets. The way these data are organized and manipulated either enables or inhibits the analyses that can be performed. The Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) is developing information technology infrastructure to support advanced hydrologic analysis and education. The CUAHSI Hydrologic Information System (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 HIS is founded upon an information model for observations at stationary points that supports its data services. The CUAHSI Observations Data Model (ODM) provides community defined semantics needed to allow sharing of hydrologic information. Following this uniform semantic model, CUAHSI HIS web services provide access to multiple disparate data sources: from national repositories such as the USGS National Water Information System (NWIS) and USEPA Storage and Retrieval System (STORET), to distributed databases published by academic researchers and community groups on their own servers in a standard format. These web services are registered to a central HIS website, where they become searchable and accessible through centralized discovery and data access tools. HIS utilizes both an XML and relational database schema for transmission and storage of data respectively. WaterML (Water Markup Language) is the XML schema used for data transmission that underlies machine to machine communications, while ODM is implemented as a relational database model for persistent data storage. Web services support access to hydrologic data stored in ODM or other repositories from application software such as Excel, MATLAB and ArcGIS that have Simple Object Access Protocol (SOAP) capability. HIS Desktop provides a local data repository and set of tools that also facilitates the integration and analysis processes. A significant value of both the web services and HIS desktop derives from the capability to use them from within a user's preferred analysis environment, using community defined semantics, rather than requiring a user to learn new software.

This paper describes the technology and tools developed as part of the CUAHSI HIS that provide: (1) **Data Storage** in a relational data model (ODM); (2) **Data Access** through internet-based Water Data Services using a consistent data language, called WaterML; (3) **Data Indexing** through a National Water Metadata Catalog; and (4) **Data Discovery** through a federated map and thematic keyword search system. The combination of these capabilities creates a common window on water observations data for the United States unlike any that has existed before, and is also extensible worldwide.

For more information about the CUAHSI HIS, or to obtain the software, freely distributed under the Berkeley Software Distribution (BSD) license, go to our website: http://his.cuahsi.org.

Keywords: *Hydrologic Information System, Web Services, Data Model*

1. 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 128 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 (Maidment, 2008) has as a goal the development of standards, systems, and software to enhance access to and interoperability among water data from multiple sources. The system that we envisage is built on interoperable components connected via the internet following a services-oriented architecture paradigm (Josuttis, 2007), which is a common design strategy for adopted by large information system projects (e.g. GEON http://www.geongrid.org, LEAD https://portal.leadproject.org). Reliance on common data exchange standards developed through the World Wide Web Consortium (e.g. XML, WSDL, SOAP), and an Open Geospatial Consortium (OGC) initiative to establish a framework of open standards for sensors and sensor systems (Botts et al., 2007), are a common component of achieving interoperability across disparate physically distributed data sources. However, achieving hydrologic data interoperability takes more than following established standards. HIS has also focused on scientific, organizational, and infrastructure aspects of hydrologic data integration. These included, respectively: establishing and formalizing the semantics of data discovery and retrieval in hydrology based on identified research use cases; creating an organizational framework for data sharing by engaging the community through partnerships of data providers, developers and collaborators; and developing an operational services infrastructure that currently provides access to the largest collection of hydrologic observations in the United States. HIS is providing leadership in laying the key foundations that others can participate in and build on. In this context of partnering, the mission of HIS is to build an access and sharing system for academic and public water observation data and data about the water environment, and to enable the linking of data and models to understand how water systems function. It is key to the success of this effort that US federal agencies such as the USGS and NCDC have started to use CUAHSI's water markup language, (WaterML, Zaslavsky et al., 2007) to publish their data. It is also key that HIS use existing and emerging standards. There is ongoing work to advance the (WaterML, Zaslavsky et al., 2007) within the OGC, in particular harmonizing it with OGC Sensor Web Enablement (SWE) suite of specifications (Botts et al., 2007), in the framework of the recently established OGC Hydrology Domain Working Group. In addition, we are incorporating existing OGC and UNIDATA geospatial data formats (e.g. NetCDF and OGC Web Feature Services and Web Coverage Services) in HIS servers.

The CUAHSI HIS is comprised of a set of components (Figure 1) that collectively represent a system using service oriented architecture to publish, catalog and share hydrologic data to support discovery, access, modeling and analysis. Central to this system is WaterML and the WaterOneFlow web services described in section 2. Persistent data storage uses the Observations Data Model (ODM, Horsburgh et al., 2008) relational database schema implemented within HIS Server described in section 3. HIS Server software and WaterOneFlow web services are available for anyone to establish a server and publish their data using WaterML. The CUAHSI Water Data Publication System is described in section 4. HIS Central is described in section 5 and includes a centralized metadata catalog and registry of WaterOneFlow web services; Hydroseek (Beran and Piasecki, 2009), a web based search engine for data discovery based on keywords from a concept ontology; and web service links to a number of 3rd party repositories, such as the

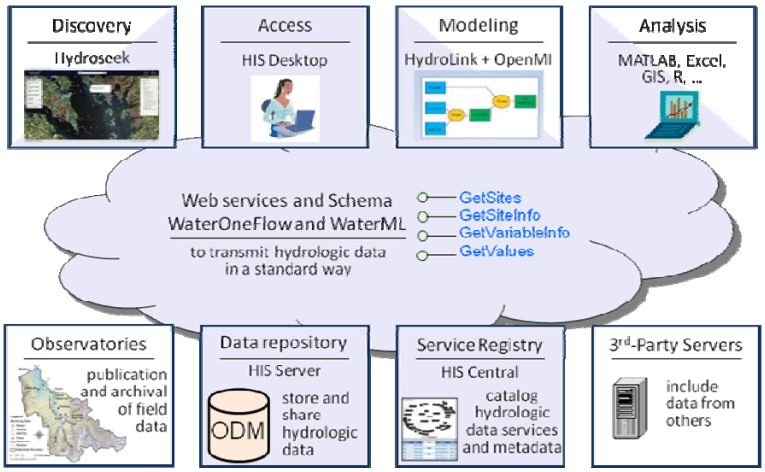

Figure 1. Services oriented conceptual system model used by the CUAHSI Hydrologic Information System Water Data Services. Shaded components are components of the HIS, while open boxes are components the system interacts with.

United States Geological Survey (USGS) National Water Information System (NWIS) and the United States Environmental Protection Agency (USEPA) Storage and Retrieval (STORET) database.

The top row of boxes in Figure 1 illustrates user interactions with HIS. Data discovery is through Hydroseek, which provides geographic and keyword based search capability across all registered web services, and data download. Analytical access to data is through HIS desktop, which is currently under development. HIS desktop will be based on the open source MapWindow GIS (http://www.mapwindow.org/) and will provide the capability to download and organize HIS data for analysis and modeling. Modeling using HIS data is also being enabled through the development of an OpenMI interface for both HIS Desktop and WaterOneFlow web services. OpenMI (http://www.openmi.org/) is a data exchange standard that facilitates the modeling of process interactions by enabling models to exchange data with each other and other modeling tools each time step. By using OpenMI, HIS supports interoperability among models that adopt the OpenMI standard and makes HIS data available to these models. The analysis box on the top right of Figure 1 depicts the use of WaterOneFlow web services directly from within 3rd-party analysis environments such as Excel, MATLAB and ArcGIS. Connection to web services from Excel requires an object library, referred to as HydroObjects, to translate WaterML into Excel content. An Excel spreadsheet document, named HydroExcel, distributed on the HIS website (http://his.cuahsi.org/hydroexcel.html), uses Excel macros to provide access to WaterOneFlow web services from within Excel. MATLAB has functionality to directly read web services using the CreateClassFromWsdl MATLAB function (see http://his.cuahsi.org/documents/workshop-cuahsi08/MatlabTutorial.pdf). Access to data from within ArcGIS is facilitated using a plugin called HydroGet (http://his.cuahsi.org/hydroget.html). Connection to web services using this functionality has the advantage of allowing users to access WaterOneFlow web services directly from within their preferred analysis environment, effectively getting the browser out of the way and limiting the need for users to learn new software. It also provides opportunities to script the download of data from multiple services.

2. WATERONEFLOW WEB SERVICES

The core of the HIS service-oriented architecture is a collection of WaterOneFlow web services, which provide uniform access to multiple repositories of observation data. Data can be remote or locally-stored in ODM instances, or in public agency databases. WaterOneFlow services use SOAP (Simple Object Access Protocol), which is a standard protocol established by the World Wide Web Consortium for enabling one computer to request services of another. WaterOneFlow services, and the markup language they implement, WaterML, are the lingua franca of CUAHSI HIS: communications between servers and clients in this service-oriented architecture follow this standard protocol.

WaterOneFlow services include the following four pairs of methods:

- **GetSitesXML/GetSites:** returns a list of measurement sites in a particular observation network;
- **GetSiteInfo/GetSiteInfoObject:** returns detailed site metadata, the set of variables actually measured at the site, with the period of record and count of available values for each variable;
- **GetVariableInfo/GetVariableInfoObject:** returns metadata describing each variable such as its name, data type and units of measurement;
- **GetValues/GetValuesObject:** returns a series of values of a variable measured at a given site between a given start date and time, and end date and time.

The first three of these method pairs are descriptive, or metadata, methods that return information about the sites and variables. The final method pair, GetValues/GetValuesObject, is the one that actually provides observations data.

The methods listed above are in pairs because it is useful to have methods that return standard XML schema based responses (GetSites, GetSiteInfoObject, GetVariableInfo-Object, GetValuesObject) and serialized responses that return XML strings (GetSitesXML, GetSiteInfo, GetVariableInfo, GetValues). The standard XML schema-based responses are suited to client application environments that can automatically parse the output into objects. The serialized responses are useful for clients to access the XML text for cases where they cannot properly parse attributes into the client object format.

There are two groups of WaterOneFlow web services. The first is for publishing data from ODM databases and the second is for publishing data from 3rd-party data sources. ODM WaterOneFlow web services are used in conjunction with an ODM database to publish observation data. The second group of web services are custom programmed to support an individual 3rd-party data source such as those from NWIS and STORET. The model for HIS providing access to large national databases for which it is unrealistic to expect the agency to convert their systems to ODM/HIS Server is to either develop translation services that map from

the agency format onto WaterML, or to encourage the agency to develop and support its own WaterML web services. Sometimes the former is a path to the latter. Both the National Climate Data Center and USGS have WaterML based web services at some level of development and testing (see http://waterdata.usgs.gov/nwis/?DailyValues_Service_Instructions).

3. HIS SERVER

An HIS Server is a computer with a collection of databases, web services, tools and front-end applications that allow an investigator and their data manager to store, publish and analyze observation data. The server is designed to permit local control of the data, while still being part of a distributed system allowing universal access to the data. ODM is the core of an HIS server and is used for the systematic organization, storage and retrieval of point observations. At a conceptual level, ODM is a schema for the representation of point observations using tables linked by associations or relationships between key fields. The ODM schema has been implemented in a relational database designed to facilitate integrated analysis of large datasets. This section provides only a very cursory overview of ODM. For details see Horsburgh et al. (2008) and the design specifications (http://his.cuahsi.org/documents/ODM1.1DesignSpecifications.pdf).

Within ODM, observations are stored with sufficient ancillary information (metadata) about the observations to allow them to be unambiguously interpreted and to provide traceable heritage from raw measurements to useable information. The design exposes each single observation as a record, taking advantage of the capability in relational database systems for querying based upon data values and enabling cross dimension data retrieval and analysis. Environmental observations are identified by the following fundamental characteristics: (1) the location at which the observations were made (space); (2) the date and time at which the observations were made (time); and (3) the quantity that was observed, such as streamflow, water quality concentration, etc. (variable). In addition to these fundamental characteristics, there are many other attributes that provide additional information necessary for interpretation of observational data. These include the methods used, qualifying comments, and information about the organization that made the observation.

Table 1 presents general attributes important in interpreting and establishing the provenance of an observation. This list of attributes was compiled from comments received from a community review of a preliminary version of ODM (Tarboton, 2005). All of the information contained in Table 1, except for the value of the observation itself, can be considered metadata.

Table 1. ODM attributes associated with an observation.

Attribute	Definition
Value	The observation value itself
Accuracy	The measurement accuracy associated with the observation value
Date & Time	The date and time of the observation (including time zone info)
Variable Name	The name of the physical, chemical, or biological quantity that the value represents (e.g. streamflow, precipitation, water quality)
Location	The location at which the observation was made (e.g. lat. and long.)
Units	The units (e.g. m or m^3/s) and unit type (e.g. length or volume/time)
Interval	The interval over which each observation was collected or implicitly averaged by the measurement method
Offset	Distance from a reference point to the location at which the observation was made (e.g. 5 meters below water surface)
Offset Ref. Point	The reference point from which the offset to the measurement location was measured (e.g. water surface, stream bank)
Data Type	Descriptor of the measured quantity (e.g. an instantaneous or cumulative measurement)
Organization	The organization or entity providing the measurement
Censoring	An indication of whether the observation is censored or not
Qualifying Comments	Comments to indicate problems or exceptions (e.g. holding time exceeded, sample contaminated, estimated)
Analysis Procedure	An indication of what method was used to collect the observation
Source	Information on the original source of the observation
Medium	The medium in which the sample was collected (e.g. water, air, etc.)
Quality Control Level	An indication of the level of quality control the data has been subjected to (e.g., raw data, checked data, derived data)
Value Category	An indication of whether the value represents an actual measurement, a calculated value, or is the result of a model simulation

The ODM logical data model (Figure 2) has been designed to store observation values and their supporting metadata in a structured way. The DataValues table at the center of Figure 2 stores the numeric values for observations, some data value level attributes, and links (foreign keys) to surrounding tables with non value level attribute details to avoid redundancy.

Within CUAHSI HIS Servers, ODM has been implemented as a Microsoft SQL Server database. The following applications are available for loading data into an ODM database.

- ODM Data Loader. This is an interactive software application that loads data into ODM from spreadsheets and comma separated tables in simple format.
- ODM Streaming Data Loader. This is a pair of software applications (the configuration wizard and loader) that facilitates the loading of data from data logger files on a prescribed schedule.
- SQL Server Integration Services. This is a Microsoft application accompanying SQL Server useful for programming complex loading or data management functions.

In HIS Server, data editing and quality control capability is provided through ODM Tools, an application that allows data managers to visualize, manage, manipulate, edit and export data from their instance of the ODM.

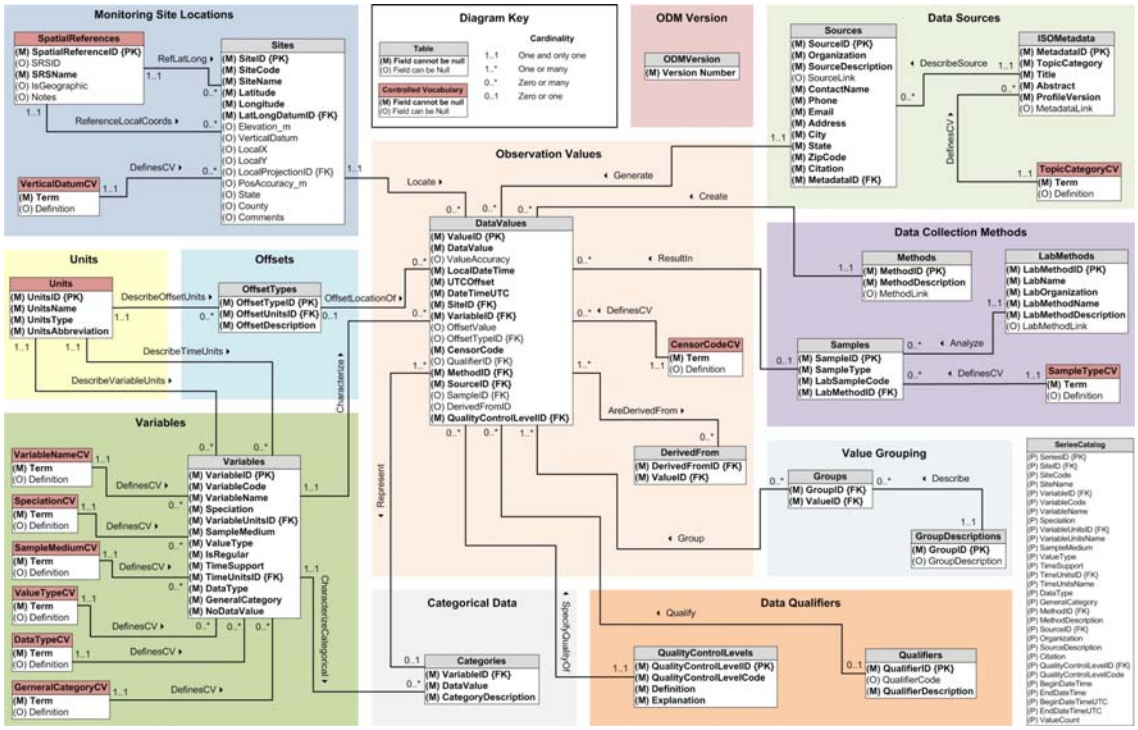

Figure 2. ODM logical data model. The primary key field for each table is designated with a {PK} label. Foreign keys are designated with a {FK} label. The lines between tables show relationships. Required (Mandatory) data fields are bold and indicated with an M, while optional data fields are indicated with an O.

In an effort to reduce contextual semantic heterogeneity within and across ODM databases, controlled vocabularies have been specified for many of the attributes within ODM. Since the controlled vocabularies list the terms that are acceptable within many fields in the database, data managers choose from the list of acceptable terms when loading data into the database rather than using their own, potentially inconsistent terms. While this places a burden on data managers to select appropriate controlled vocabulary terms, the advantage is that this helps promote consistent terminology.

A master list of approved controlled vocabulary terms is maintained within a central database. This central repository represents a community vocabulary for describing environmental and water resources data. It is dynamic and growing; users can add new terms or edit existing terms by using the functionality on the ODM website (http://his.cuahsi. org/mastercvreg.html). If a data manager cannot find an appropriate term to describe data that is being added to an ODM database, he or she can use the ODM website to request addition of a term to the master controlled vocabulary repository. Once the moderator accepts a new term, it becomes part of the master database.

The ODM controlled vocabularies are duplicated within each ODM database to maintain the integrity of data and to ensure that data loaded into local databases are connected with the required metadata. Because new terms are continually being added to the master list, local databases must be synchronized periodically with the master repository to ensure the availability of the controlled vocabulary terms within each local database. This is accomplished using ODM Tools and the ODM Controlled Vocabulary web services, which broadcast the terms within the master repository in XML format.

4. CUAHSI WATER DATA PUBLICATION SYSTEM

WaterOneFlow web services and HIS Server support the sharing of data through the CUAHSI Water Data Publication System (Figure 3). This section provides only a very cursory review of the water data publication system (for details see Horsburgh et al., 2009). The steps involved in data publication are:

1. Establish a HIS Server with ODM and WaterOneFlow web services.
2. Load data. Figure 3 illustrates loading of data from Excel or text files as well as from datalogger files received from an observing system telemetry network. The web services on top of ODM expose its content immediately in near real time. Figure 3 also illustrates data editing using ODM tools to publish raw data that has been quality controlled at a higher quality control level.
3. Register with HIS Central. A registry of data services maintains metadata on each registered data service.

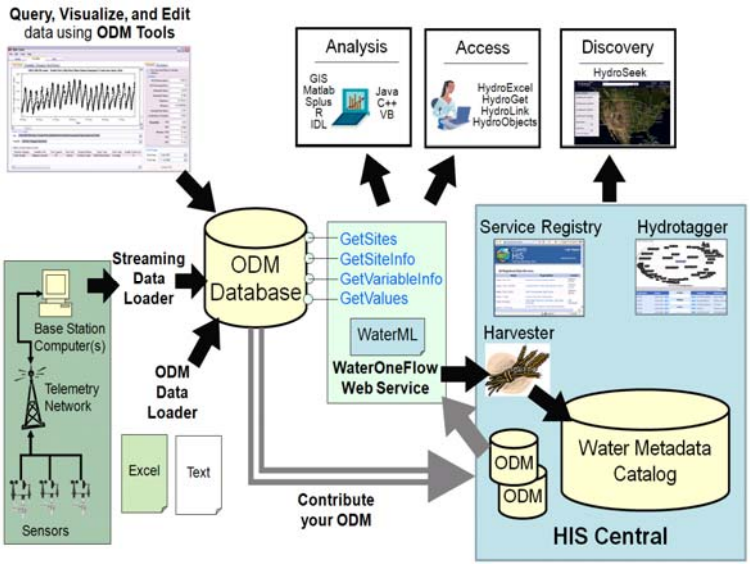

Figure 3. Conceptual architecture of CUAHSI Water Data Publication System.

4. Tag variables. HIS Central also maintains the CUAHSI ontology of concept keywords. Once variables within a WaterOneFlow web service have been catalogued, they need to be associated with an appropriate leaf concept in the ontology. This is done using the Hydrotagger, and once complete the data become discoverable from concept keyword searches.

5. HIS CENTRAL

HIS Central maintains a central registry of WaterOneFlow web services, and a metadata catalog containing metadata for each registered service. Users of HIS Central can browse a list of public services, and register, manage and test their own services, including registering variables to a common ontology and defining how hydrologic measurement points will appear on the national map. HIS Central indexes both ODM WaterOneFlow services that provide access to data served from an ODM database on an HIS Server, and WaterML-compliant service wrappers that provide access to 3rd-party data repositories such as NWIS and STORET..

The HIS Central water metadata catalog is an integrated database that contains information about each registered WaterOneFlow data service:

- For each service, there is a listing of the sites in its observation network.
- Each site has one or more data series.
- Each data series has details on what variable is being monitored, when it began and ended, and how many data values there are.
- Each variable is tagged with a concept from the ontology

The metadata catalog is intended to provide enough information to help users assess the contents of data served by a data service and to formulate requests to services to retrieve the data. It is maintained current by a harvester application, which periodically calls each registered WaterOneFlow service to retrieve information on the sites, variables, periods of record, etc., available from the respective data source, and updates the metadata catalog as needed. The HIS Central web application is used as a portal to the catalog where registered users can register and modify the description of their WaterOneFlow web services, and determine how their data will appear in the Hydroseek application.

Hydroseek (Beran and Piasecki, 2009), is a web based search engine for data discovery based on keywords from a concept ontology that are connected to variables within the registered data services. HIS Central

maintains these connections through a web application called Hydrotagger which data managers use to relate the variables that have been harvested to leaf concepts in the standard ontology.

6. CONCLUSION

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); (2) **Data Access** through internet-based Water Data Services using a consistent data language, called WaterML; (3) **Data Indexing** through a National Water Metadata Catalog; and (4) **Data Discovery** through a federated map and thematic keyword search system. 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 combination of these 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 academic data that might otherwise have been confined to the private files of the individual investigators and serves as a prototype for the infrastructure that will be required to support a network of large scale environmental observatories as well as other observatory efforts and 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.

ACKNOWLEDGMENTS

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

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.

Botts, M., G. Percivall, C. Reed and J. Davidson, (2007), "OGC Sensor Web Enablement: Overview And High Level Architecture," Open Geospatial Consortium White Paper OGC 07-165, http://www.opengeospatial.org/projects/groups/sensorweb.

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.

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

Maidment, D. R., ed. (2008), *CUAHSI Hydrologic Information System: Overview of Version 1.1*, Consortium of Universities for the Advancement of Hydrologic Science, Inc, Washington, DC, 96 p, http://his.cuahsi.org/documents/HISOverview.pdf.

Tarboton, D. G., (2005), "Review of Proposed CUAHSI Hydrologic Information System Hydrologic Observations Data Model." Utah State University. May 5, 2005. http://www.engineering.usu.edu/dtarb/HydroObsDataModelReview.pdf.

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.