



CUAHSI HYDROLOGIC INFORMATION SYSTEM: OVERVIEW OF VERSION 1.1

**Edited by David R. Maidment
Center for Research in Water Resources
The University of Texas at Austin**

July 12, 2008

Distribution

Copyright © 2008, Consortium of Universities for the Advancement of Hydrologic Science, Inc.
All rights reserved.

Acknowledgement

Funding for this research was provided by the National Science Foundation under Grant No. EAR-0413265. In addition, much input and feedback has been received from the CUAHSI Hydrologic Information System development team, from the WATERS testbed network data managers, and from the CUAHSI Program Office. These contributions are gratefully acknowledged.

Table of Contents

Chapter 1.	Introduction	1
1.1	Motivation	1
1.2	CUAHSI Hydrologic Information System	2
1.3	Scope of This Report	4
1.4	References.....	5
Chapter 2.	Conceptual Framework.....	7
2.1	Information Model	8
2.2	System Model	10
2.3	Relational Data Persistence Model	12
2.4	Water Data Web Services	13
2.5	Metadata Catalog.....	18
2.6	Other Information Models	20
2.7	Concluding Remarks.....	24
2.8	References.....	24
Chapter 3.	Observations Data model	25
3.1	ODM Design	25
3.2	ODM Examples.....	29
3.3	Tools for working with ODM.....	32
3.4	ODM Controlled Vocabularies.....	37
3.5	Summary	39
3.6	References.....	40
Chapter 4.	Web Services and Water Markup Language.....	41
4.1	Service-oriented architecture	41
4.2	How web services work.....	42
4.3	What is WaterML?	44
4.4	Current status of CUAHSI HIS Web services.....	45
4.5	Web Services for National Data Repositories versus ODM Web Services	50
4.6	The evolution of water data web services	52
Chapter 5.	Semantic Mediation – Linking Data with Concepts	53
5.1	The need for a Global Data Search Environment.....	53
5.2	The Search Approach	54
5.3	The Search Ontology	55
5.4	HydroSeek: A global Search Engine.....	57
5.5	HydroTagger.....	58
5.6	References.....	59
Chapter 6.	Water Metadata Catalog and HIS Central.....	61
6.1	The Water Metadata Catalog.....	62
6.2	HIS Central - http://hiscentral.cuahsi.org	63
6.3	Registering and Testing a CUAHSI Water Data Service	64
6.4	Metadata Harvesting and Variable Tagging	66

6.5 HIS Central and the Hydrologic Community: The Path Forward	67
Chapter 7. Using Data in Analysis Applications	69
7.1 HydroObjects	69
7.2 HydroExcel - WaterOneFlow in Excel	69
7.3 HydroGET – Water Data Services in ArcMap	72
7.4 HydroGET and NHDPlus	74
7.5 Using Water Data Services in Your Application of Choice.....	78
Chapter 8. Using Data in Models.....	79
8.1 Introduction	79
8.2 What is HydroLink?	80
8.3 HydroLink Design	80
8.4 Tools for Building a WaterML File Cache	83
8.5 Implementation	84
8.6 Use Case Study	85
8.7 Alternative Approaches for Using HIS Data in Models.....	88
8.8 Summary Remarks and Future Work	90
Chapter 9. Conclusions	91
9.1 References.....	92

Chapter 1. INTRODUCTION

By David Maidment, The University of Texas at Austin

1.1 MOTIVATION

The advancement of hydrologic science is critically dependant on the assembly and synthesis of hydrologic data. Central to this task are observations data describing water conditions, such as streamflow, precipitation, water quality and groundwater levels, measured over time at gages and sampling points. To these must be added GIS data that describe the spatial context of the watersheds, aquifers, landscapes and stream networks through which water flows. Hydrology is driven by climate, so spatial grids of time-varying weather fields such as NEXRAD rainfall, air temperature, humidity, wind and solar radiation are important. Satellite remote sensing data provide time variations in spatial patterns of such quantities as the greenness of the landscape. Hydrologists want simple and efficient methods for discovering, accessing and acquiring high quality hydrologic data to describe the systems they are studying. The advent of the internet has provided a means of accessing vast quantities of such information.

The need for improved hydrologic data accessibility and management has been recognized for many years. The National Research Council (1991) report on “Opportunities in the Hydrologic Sciences” states (p.265) “Advances in hydrologic sciences depend on how well investigators can integrate reliable, large-scale, long-term data sets.” Of course, most of the available hydrologic data is collected by public agencies, such as the US Geological Survey, whose National Water Information System (NWIS) is the nation’s largest repository of water information. However, NWIS contains only data collected or approved by the USGS and large data holdings are accumulated by other federal agencies, such as the EPA’s STORET (STORage and RETrieval) water quality database, the climate archives maintained at the National Climatic Data Center, and the snow measurements in the western states collected by the US Department of Agriculture. For a study in any region of the nation, additional data collected by state and local agencies are important.

Moreover, hydrologic data are also collected by academic investigators, most often in project-based studies of limited duration designed to examine some particular hydrologic process or phenomenon. The National Research Council (1991) states further: “The data sets required to answer many of the open research questions in hydrology will be complex. Inevitably, many scientists from a variety of disciplines and backgrounds will be involved in data collection and analysis, over a significant period of time. How can diverse investigators and investigations produce compatible data sets, assure their quality, and confidently present them for larger, indeed public use and access? Creating effective data systems for assembling and distributing scientific data sets is not trivial and depends heavily on the personal efforts of active scientists. If the data systems are constructed within the scientific community by scientists themselves, rather than by independent data “experts”, there will be many scientific opportunities as well as technical and political challenges”.

The research described here addresses the data management challenges posed in the “Opportunities in the Hydrologic Sciences” report, and demonstrates how water observations data collected by many scientists in different projects with various science goals, can be stored and accessed through the internet in a consistent way. Our work relies on a synthesis of knowledge – from hydrologic science to assess what is needed, and from

computer science to create the means for meeting the needs. This research and development is carried out to build the CUAHSI Hydrologic Information System.

1.2 CUAHSI HYDROLOGIC INFORMATION SYSTEM

The Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) (<http://www.cuahsi.org>) is an organization representing 122 US universities (Figure 1-1), 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.

This report addresses mainly the first of these four goals. The report provides an overview of the activities of the HIS project which are focused on providing web data services, tools, standards and procedures that enhance access to more and better data for hydrologic analysis. We present three new water data capabilities developed as part of this research:

- **Data Storage** in an *Observations Data Model* (ODM), which is a standardized relational database structure for storing and describing hydrologic observations data measured at point locations;
- **Data Access** through internet-based *Water Data Services* that enable querying and accessing data stored at remote locations in ODM databases, and in the water databases of public agencies, and delivery of these data in a consistent data language, called WaterML;
- **Data Indexing** through a *National Water Metadata Catalog* that assembles in a consistent form the metadata that describe the water observation networks of the nation, and enables data searching across these networks.

The combination of these three capabilities creates a common window on water observations data for the United States unlike any that has existed before. More than eight million water data series measured at nearly two million locations are presently stored in the national water metadata catalog. Water data services have been established at eleven universities in the WATERS testbed network, whose data are all stored in the ODM and accessed in WaterML. The US Geological Survey is now publishing its NWIS daily values data in WaterML and the National Climatic Data Center is beginning to do the same for its Integrated Surface Data. Similar CUAHSI services have been developed for water data from EPA, USDA, and some regional and state water agencies.

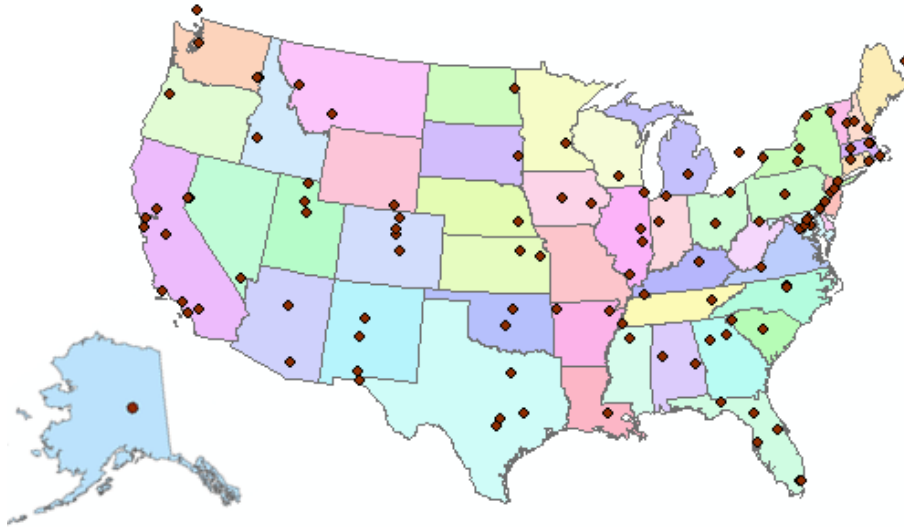


Figure 1-1 Locations of the 122 universities which are members of CUAHSI

The HIS project was one of several activities initiated by CUAHSI in 2000, by forming a Hydrologic Information Systems Committee with members drawn from nine CUAHSI universities. This Committee developed a conceptual plan for a Hydrologic Information System development program, and estimated the resources needed to implement that plan. (HIS Committee, 2002). In 2004, the National Science Foundation supported an initial investigation to better define HIS components, whose preliminary conclusions were summarized in an HIS Status Report (Maidment, 2005). NSF renewed its support for the HIS project with a 5-year grant which began in January 2007.

The HIS project team is balanced among hydrologic scientists at CUAHSI universities combined with computer scientists at the San Diego SuperComputer Center. Current project team members include:

- David Maidment (Project Leader), Tim Whiteaker, Ernest To, Kate Marney, Apurv Bhatia and Bryan Enslein at the Center for Research in Water Resources of the University of Texas at Austin;
- Ilya Zaslavsky, David Valentine and Tom Whitenack, at the San Diego Supercomputer Center;
- David Tarboton (Deputy Project Leader), Jeff Horsburgh, and Kim Schreuders at the Utah Water Research Laboratory of Utah State University;
- Michael Piasecki and Yoori Choi, at the Department of Civil Engineering of Drexel University;
- Jon Goodall and Tony Castronova at the Department of Civil Engineering of the University of South Carolina.

This report describes Version 1.1 of the CUAHSI Hydrologic Information System. Version 1.0 was produced in May 2007 and distributed for testing to partner institutions at eleven universities engaged in testbed projects to support the development of a possible network of hydrologic observatories called WATERS. As indicated in Figure 1-2, the information from the hydrologic information servers at each partner institution is catalogued at a facility called HIS Central operating at the San Diego Supercomputer Center (<http://hiscentral.cuahsi.org>).

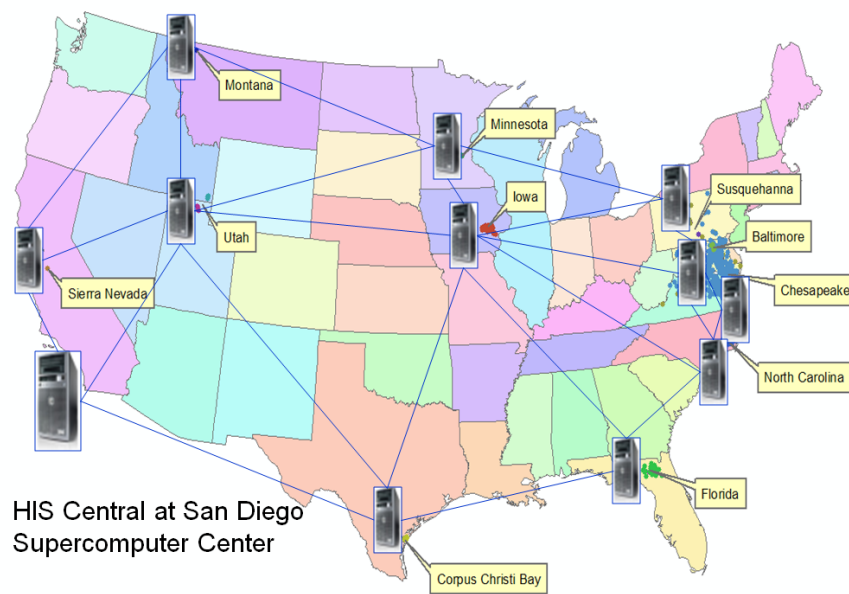


Figure 1-2 The network of hydrologic information servers deployed to support the WATERS testbed projects

Version 1.1 of CUAHSI HIS differs from Version 1.0 in the following ways:

- The Observations Data Model schema has been updated and a better data loader developed to put information into it;
- A National Water Metadata Catalog has been established at the San Diego Supercomputer Center (SDSC) that indexes water data collected at universities and public agencies in a common way;
- A procedure for publishing water data services has been formalized so that new services published at universities can be registered at the HIS Central, or the observations database itself can be sent to SDSC to be published at HIS Central.
- Many new CUAHSI water data services have now been published and existing services to government agencies have been reprogrammed to be more robust.
- Better tools for using water data services have been developed, in particular HydroExcel and HydroGet, which ingest water observations data into Excel and ArcGIS respectively.
- An improved HydroViewer is being developed to enable rapid map access to observations data.
- A connection has been made between WaterML data services and inputs for hydrologic modeling using the OpenMI model interface developed in Europe.

More detail about the tools and data services developed by the HIS project is provided at: <http://his.cuahsi.org>.

1.3 SCOPE OF THIS REPORT

This report contains nine chapters, of which this introduction is the first. Chapter 2 presents a conceptual framework for the components of the CUAHSI Hydrologic Information System and the water data services architecture that it is based on. Chapter 3 describes the structure of the Observations Data Model and the tools that have been created to store and edit data in it. Chapter 4 presents WaterML and the web services functions

that provide access to data in WaterML. Chapter 5 begins the synthesis of the data by means of a hydrologic ontology, or hierarchy of hydrologic concepts, to which the variables in the observation networks are linked. Chapter 6 outlines the mechanisms for registering a new water data service at the HIS Central facility developed at the San Diego SuperComputer Center. Chapter 7 describes some tools in Excel, Google Earth, and ArcGIS for ingesting and displaying data derived from CUAHSI water data services. Chapter 8 shows how hydrologic data in WaterML can be linked to hydrologic models through the OpenMI protocol or other methods. Chapter 9 presents some concluding thoughts on this research.

While the methods presented in this report do define a path-breaking approach for working with water observations data, it is important to recognize a number of significant limitations of the CUAHSI Hydrologic Information System as presently developed:

- It focuses on one class of information, water observations data measured at fixed point locations;
- While the observations data model can store “collections” of hydrologic data collected in field investigations because each data value is stored and indexed individually, our methods for indexing and accessing data are oriented around series of data measured through time about individual variables measured at particular locations, rather than collections of many variables measured simultaneously at one place and time;
- Research on other classes of hydrologic data, such as GIS, weather and climate grids, and remote sensing has been addressed to some degree but a true synthesis across these information classes has not yet been achieved;
- Our work on remote sensing is especially limited and we have not yet found a feasible way to extend our services-oriented architecture concept to remote sensing data;

Despite these limitations, water observations data measured at point locations is at the heart of many hydrologic research data collection studies, and the water monitoring activities of public agencies, and for this class of information we have new methods and technology to present in the following pages.

1.4 REFERENCES

- HIS Committee (2002), “CUAHSI Hydrologic Information Systems”, Technical Report #2, Consortium of Universities for the Advancement of Hydrologic Science, Inc, Washington, DC, 32p. http://www.cuahsi.org/publications/cuahsi_tech_rpt_2.pdf
- Maidment, D.R., (2005) (Ed.), “Hydrologic Information System Status Report”, Consortium of Universities for the Advancement of Hydrologic Science, Washington DC, 214p. <http://www.cuahsi.org/his/docs/HISStatusSept15.pdf>
- National Research Council (1991), “Opportunities in the Hydrologic Sciences”, National Academy Press, Washington, DC, 348p. http://www.nap.edu/catalog.php?record_id=1543

Chapter 2. CONCEPTUAL FRAMEWORK

By David Maidment, The University of Texas at Austin, David Tarboton, Utah State University, and Ilya Zaslavsky, San Diego Supercomputer Center

When contemplating the development of a hydrologic information system (HIS), it is appropriate to define the role that this system is intended to play. One useful approach has been established by the geographic information system (GIS) community. Tomlinson (2003, p.3) 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, so some broader context for accessing information is needed.

A services-oriented architecture is 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 (Josuttis, 2007). Indeed, when contemplating the plethora of web sites that present water observations data collected by government agencies and universities, a user is struck by the fact that no two web sites are the same, and there are no standards of consistency in how the information is provided – everyone uses their own data format and method of describing their data. It is almost like a “Tower of Babel” of data languages – everyone speaks their own.

The CUAHSI HIS project has as a goal the development of standards, systems, and software to overcome these inconsistencies and enhance the interoperability of the nation's water information. Ultimately we would like water data to be universally accessible and easy to use. The system that we envisage is built on interoperable components connected via the internet following the services-oriented architecture paradigm. Building such a system takes more than technology. HIS must also engage the community through partnerships of data providers, developers and collaborators. HIS must provide 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.

This chapter has seven sections. The first lays out the information model used for the representation of data in HIS. The second gives the system model, describing the conceptual framework for the components of the water data services architecture that underlies the CUAHSI HIS. This system requires that data be stored, transmitted and discovered. The third section describes the framework for persistent storage of water data in a relational model. The fourth section discusses the framework for data transmission using web services. These models for persistence and transmission of water information are core foundational concepts upon which the CUAHSI HIS described in following chapters, rests. The fifth section describes the use of a metadata catalog to support hybrid water data services where data is actually resident on a host agency server. The sixth section describes information models used in other scientific disciplines and some of the partnerships CUAHSI HIS has with these disciplines in terms of working towards interoperability among data services across these disciplines. The final section contains some concluding remarks.

2.1 INFORMATION MODEL

A model is a simplified view of the real world. A hydrologic simulation model is a simplified representation of hydrologic processes with equations that attempts to mimic the behavior of hydrologic phenomena. A statistical model of a time series of hydrologic observations is a summary using statistical parameters of an assembly of sampled data, sometimes associated with a mathematical probability model describing how the sampled phenomenon varies in nature. A geographic model of a landscape in a GIS is a simplified representation using themes or data layers of different classes of information that describe the characteristics of the landscape. All these concepts are relevant to a hydrologic information system – it needs a GIS component that can represent the physical landscape through which water flows; it needs an observational component that describes how the properties of water vary through time and space in the landscape; it needs a simulation component that can mimic and predict the functioning of hydrologic phenomena.

A common point of reference among GIS, observations and modeling for hydrology, is that all of these data deal with *variables*, that is, with quantities like the rainfall rate, streamflow discharge or slope of the land surface terrain, that vary through *space*, and may also vary through *time*. Indeed, water flow is one of the primary forces in shaping the landscape over millennia, so if we lengthen our time horizon to include geologic time scales, the landscape itself is also dynamic. So, in mathematical terms, suppose there are a set of n variables, denoted by V_1, V_2, \dots, V_n . Each variable may change in space, denoted by s , and in time denoted by t , and at a particular point of space and time, they have *values*, $v_1(s,t), v_2(s,t), \dots, v_n(s,t)$. The *region of space* over which these variables are defined may be defined by S , and the applicable *time horizon* by T , so, again in mathematical terms, $s \in S$ and $t \in T$. As shown in Figure 2-1, the value $v_i(s,t)$ denotes the value of the variable V_i at the spatial location s and time t , and one may think of i, s and t as the coordinates in a “data cube” which reference this particular value. A simple way to think about this is that it is a “what-where-when” model – “what” variable (V_i) is represented “where” (s), and “when”, (t).

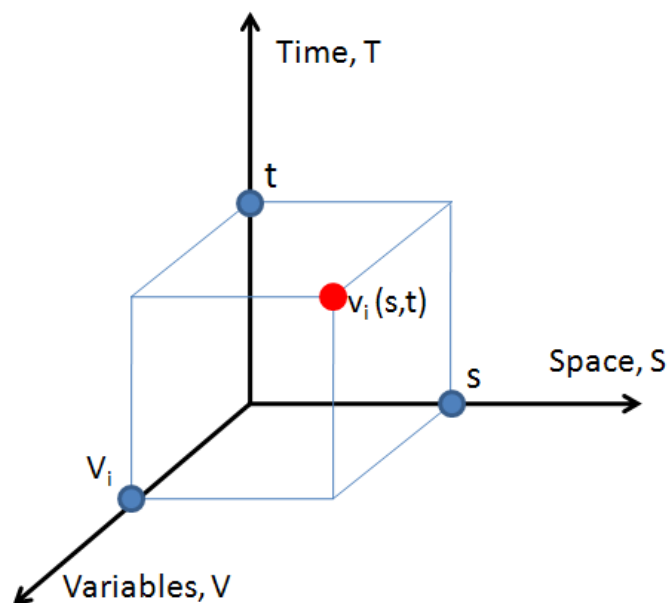


Figure 2-1 The “what-where-when” method of representing a value in a data cube

In a perfect world, all the variables would be defined precisely over the whole space-time domain, $V(S,T)$, and there would be no uncertainty about hydrologic conditions anywhere. In reality, however, we don't live in a perfect world. Observation of water properties can be done at gages and sampling sites at a few points in space and often irregularly in time, especially for water quality or biological sampling. Remote sensing creates continuous coverages across space at particular points in time, but the connection between remotely sensed images and hydrologic variables may be tenuous. There are many weather and climate grids available through time and across space for the United States and the world, but these are produced by computer models of variable accuracy. Even a quantity like land surface elevation, which seems so fixed and unchanging that it must be known without ambiguity, is in fact not so – the root mean square error in elevation of the National Elevation Dataset when compared to survey benchmarks is 2.34 m (National Research Council, 2007, p.5). So, rather than the perfect world of $V(S,T)$ without uncertainty, what actually exists is a much smaller data collection $\{v(s,t)\}$. Those data are of variable quality, and are produced by many organizations in many formats and descriptions.

Despite these limitations, the fundamental information model used in the CUAHSI Hydrologic Information System is simply this – a collection of variables defined sparsely and with some degree of uncertainty over a domain of space and time. It is important to note that this data model, while seemingly very general, has its limitations – it describes information measured in the natural environment, where location is measured in geographic coordinates and time is referenced to Coordinated Universal Time (UTC), which is the international standard for measurement of calendar time. We are not dealing here with describing data sets measuring water percolation through soil columns in a laboratory experiment, for example.

In contemplating how to implement this data model it is useful to distinguish between data values, data collections, data series, and data sets. A data *value* is a single entity that exists by itself and represents the value of one variable at one point of space and time. A data *collection* comprises the values of a group of variables all referenced to the same point of time and space. A data *series*, is a sequence of data values through time of a particular variable measured regularly or irregularly in time at a particular location in space. Data collections and data series are groups of values where the individual values may have no relation at all with one another. A data *set* is a collective entity where individual values have meaning only in relation to one another. For example, a set of results from a laboratory analysis of a water sample is a collection of data values. A multidimensional array of results from a climate simulation model defined over a regular grid in space and time is a dataset.

Another useful distinction is the concept of discrete and continuous space. A *discrete space* representation of a spatial region means that the region is represented by a collection of separate points, lines, areas or volumes, as in a GIS vector data set, such as a set of stream gage location points, river reach lines, watershed polygons or aquifer volumes. A continuous space representation of a region means the domain is covered with a regular mesh of cells or points such that the variables are defined continuously over the whole region. Some examples are digital elevation models of land surface terrain or grids of NEXRAD precipitation maps. In the computational world, discrete and continuous space representations are sometimes called *unstructured* and *structured* grids, respectively.

In a completely configured information model for a Hydrologic Information System, all these different concepts would be included. In what is being presented in this report, which focuses on water observations data measured repeatedly at gages and sampling sites, only data values and data series are being represented, and those only on a discrete space domain of a set of fixed point locations. In later versions of CUAHSI HIS, we intend to broaden our core information model to encompass a wider range of data types.

2.2 SYSTEM MODEL

The problem of sharing information in a consistent fashion is being addressed in the CUAHSI HIS through the services-oriented system model illustrated in Figure 2-2. The top three boxes in this figure depict different ways for a user to interact with and access data in the system. The bottom three boxes depict the servers that support the system. The red lines between these boxes depict data transfers using WaterML, the XML based markup format for data transmission, described below, that CUAHSI HIS web services use for data transfer. WaterML and the water data web services that serve to link the components are keys to the systems functioning across distributed servers and clients. Other modes of data transfer are depicted using thinner black lines. Within the HIS server boxes depicted as the leftmost two bottom boxes are ODM databases and a number of tools for working with ODM. ODM is the Observations Data Model that represents the standard for persistence of point observations in the CUAHSI HIS. WaterML and ODM are two key foundational concepts upon which the system is based.

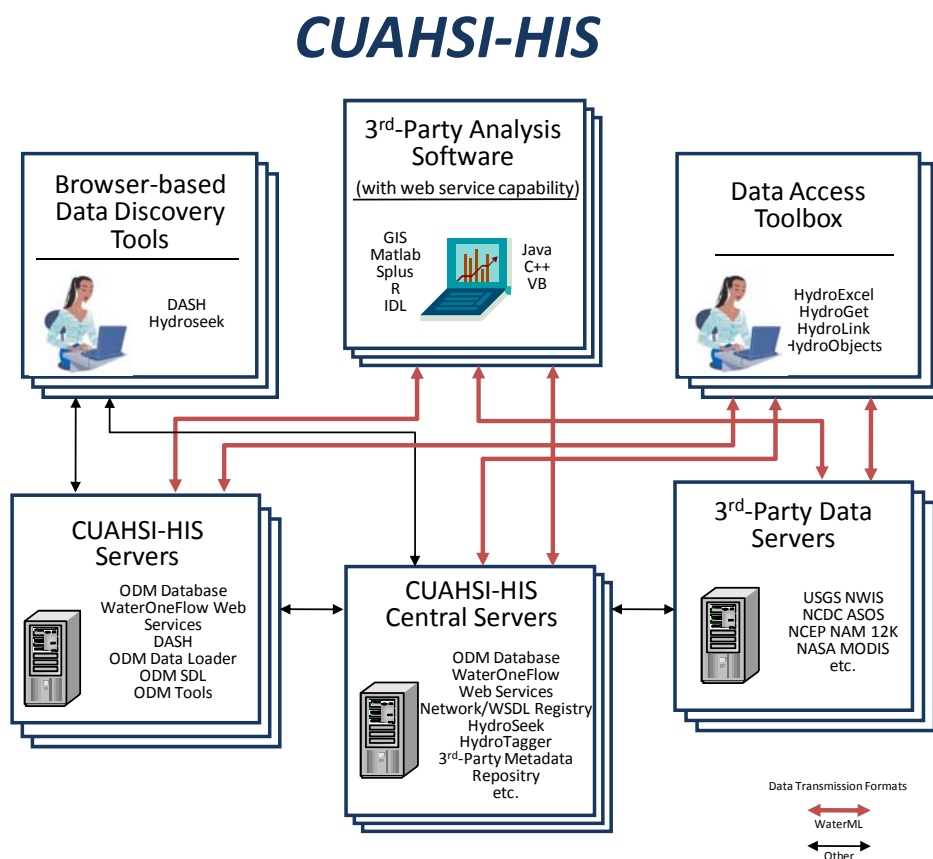


Figure 2-2 Services-oriented system model used by CUAHSI Hydrologic Information System

A strength of the system as conceptualized here is that it allows hydrologic data to be stored, found, accessed, interpreted and analyzed in a uniform way. The system is a distributed network of computers, linked together using the internet. The core of the system is a group of servers, referred to as HIS Central. In addition to this are numerous servers that are all based on a CUAHSI HIS design, called HIS Servers, which are operated by independent investigators or organizations that wish to easily publish hydrologic data. The third group of servers are large 3rd-party data sources such as USGS NWIS or the EPA STORET, that are of significant value to the hydrologic community, that the CUAHSI HIS project has integrated into the system, even though they are not based on the HIS Server design. To access the data on these various servers are a number of client tools, communication protocols and websites that are accessible to anyone with an internet connection. These various pieces function together as an integrated system, where investigators publish their data and everyone has easy access to both investigator data as well as data in a variety of large 3rd-party data repositories, in a standard way.

To get started using the system a user may use one of the browser based discovery tools. These interact directly with servers, either at HIS Central (middle bottom box) or distributed around the country, e.g., at observatories, (bottom left box) using standard browser http communications. The Data Access System for Hydrology (DASH) and Hydroseek components provide map and concept based data discovery capabilities. These are supported by metadata catalogs on the servers.

Once a user has discovered the data they want to use, then it can be accessed directly using water data web services (red lines). Web services are machine independent and available from a number of user application environments. This allows HIS users to use application analysis environments of their own choice to work with HIS data. This is depicted in the top middle box. HIS users accessing data this way do not need to learn a new system. Rather they need to learn new data access functions within the working environment of their choice. The system thus has a level of platform and application environment neutrality.

The top right box depicts components that CUAHSI HIS has developed to also provide data access. These include HydroExcel that allows users to directly access data from water data web services from within Microsoft Excel. HydroGet provides data access from within ArcGIS, while HydroLink provides capability to link web services to OpenMI compliant models, linking HIS and hydrologic modeling capabilities. OpenMI is the Open Model Interface (www.openmi.org) for linking hydrologic models.

Data available through the system includes academic data at CUAHSI HIS servers (bottom left and central boxes) as well as data that has become part of the system through the collaboration of third parties. Third party data servers are depicted in the bottom right box and comprise servers supported by the USGS, NCDC, EPA and others, where CUAHSI has established data access arrangements. The development and support of water data web services for these third party datasets is a key way that semantic heterogeneity is resolved. While many of these agencies have their own web pages and data sharing systems, each has historically been different, making access tedious and difficult. The WaterML water data services for these datasets provide a consistent access method across multiple agencies. Furthermore, collaboration with 3rd party agencies has resulted in a number of them moving towards supporting their own WaterML based web services. This data access mode is depicted by the red lines between the top boxes and 3rd party data servers.

The system also provides the capability to support the point observation needs of hydrologic observatories. Software for CUAHSI HIS servers (bottom left) is distributable and may be installed by individual investigators or research groups. Tools included provide capability to load, edit and quality control the data as well as publish it using CUAHSI water data services. Registration of published water data services with the Network/WSDL registry

(center bottom box) connects HIS server with the system allowing it to be discovered and accessed using the centralized discovery tools.

2.3 RELATIONAL DATA PERSISTENCE MODEL

While it is easy to define information models in the abstract, the difficulty comes when you try to implement them. A distinction is made in information technology between a logical data model and a physical data model. A *logical model* is a schema or structure for storing data that defines the various entities, their descriptions or attributes, and their relations with one another. The set of linked relational database tables defined in Chapter 3 for the CUAHSI Observations Data Model (ODM) is a logical model. This logical model can be implemented within any relational database system, such as Microsoft SQL/Server, which is the default environment for the CUAHSI HIS, Microsoft Access, the default database within Microsoft Office, or MySQL which is an open source relational database system. The physical data model refers to the actual positioning of the data elements within the data storage device – with relational databases, the user has no control over this and can interact with the data only through the functions that the makers of the relational database allow. In other words, depending on what particular relational database is used the actual physical location of the data elements stored in the computer's memory will be different, and is not accessible to the user. This rigid discipline makes relational databases very robust and indeed we are not aware of any data having been inadvertently lost so far from any implementation of the CUAHSI Observations Data Model. However, the same rigidity can make relational databases difficult to load with data and difficult to work with unless a good set of user tools is provided. We intend that the ODM loading and editing tools provided with CUAHSI HIS will help you overcome these difficulties.

Within the CUAHSI Observations Data Model, each data value is treated individually and all data values are stored in one large table, called the DataValues table, as the entity called "DataValue" listed second in the column of entities shown in Figure 2-3. This DataValue must be a number. ValueID {PK} is the *Primary Key* or the index number for the values in this table – ValueID is an integer that starts at 1 for the first record in the table and then continues incrementing by one for all subsequent records. The entities shown with the association {FK} are links called *Foreign Keys* to the similarly constructed primary keys in other tables of the database. For example, SiteID links the data value to the observation site at which it was measured, and VariableID links the value to the variable it describes. Time is actually represented by three entries – LocalDateTime for the time in the time zone where the measurements are made, DateTimeUTC for the corresponding Coordinated Universal Time (equivalent to Greenwich Mean Time), and UTCOffset to record the difference in hours between these two times. Other descriptors of the data value are signified by OffsetValue (if measurements are made at various depths within a water body, for example), Qualifier, Method, Source, Sample and Quality Control Level. And all these descriptors are applied to every single data value and could be different from one value to the next in the table.

The very great merit of this approach to storing observations data is its generality – physical, chemical and biological data describing conditions in water systems have all been stored in this structure in the many ODM implementations that have been made at our partner universities. We don't care whether data are recorded regularly or not since each data value has its own *time stamp* or value in calendar date and time when it applies. It is a relational database convention that all data that represent values over *intervals* of time are time stamped at the instant that the interval begins. This means that annual data, monthly data for January, and daily data for the first of January, are all time stamped at midnight on the beginning of the first day of January in that year, and time support metadata included in the Variables table of the ODM are needed to distinguish what kind of data series this is, if it is regularly recorded.

DataValues
ValueID {PK}
DataValue
ValueAccuracy
LocalDateTime
UTCOffset
DateTimeUTC
SiteID {FK}
VariableID {FK}
OffsetValue
OffsetTypeID {FK}
CensorCode
QualifierID {FK}
MethodID {FK}
SourceID {FK}
SampleID {FK}
DerivedFromID
QualityControlLevelID {FK}

Figure 2-3 The DataValues table from the CUAHSI Observations Data Model

In relational database design terms, this approach to data structuring is called a *star schema*, because all the actual data values are stored in a single table and other tables arranged around this one in a star configuration contain the metadata or descriptors of these values.

2.4 WATER DATA WEB SERVICES

The relational database design just described is a structure for internal storage of data in an HIS much like that envisaged by Tomlinson (2003) for a GIS. If one wishes simply to store observations data and work with them personally, then the Observations Data Model and its attendant ODM Tools function perfectly well as an independent entity for that purpose. Suppose, however, that the hydrologic scientist wishes also to publish these data publicly by providing automated external access to the data through the internet. We are now faced with an entirely different problem – how to communicate data rather than store them. One solution to this problem is to make a copy of the ODM database and put it on an ftp site so that others can download it. This simple, traditional approach has the merit that it completely conveys the entire content of the database. It is, however, greatly limited by the fact that any user of the data has to employ the same relational database system used in creating it – in other words, the physical data model now really matters when the data are moved to another computer.

A more flexible approach is to transform the database so that it becomes a *water data service*. This is accomplished by linking the database to a special set of functions constructed according to the conventions of the Web Services Description Language (WSDL) and a transmission protocol known as Simple Object Access Protocol (SOAP), a set of conventions defined by the World Wide Web Consortium that enable one computer to make legitimate requests of another computer at a remote location, and to receive appropriately constructed responses to those requests (<http://www.w3.org/TR/wSDL>, <http://www.w3.org/TR/soap/>). The CUAHSI HIS team has used these conventions to define a specialization of XML (<http://www.w3.org/TR/xml/>) called *WaterML* for

communicating water observations data through the internet, and a set of functions called *WaterOneFlow* web service methods that operate on the Observations Data Model and produce responses in WaterML to external requests.

Suppose, moreover, that analogous water data services could be constructed for existing national and regional water archives collected by government agencies even though the database system they are using internally for storing their data may be completely different than the CUAHSI Observations Data Model. In other words, when *WaterOneFlow* web service requests are made of these government databases, they also produce responses in WaterML. Now, the “Tower of Babel” of a plethora of water web sites and water data formats referred to earlier is vastly transformed – all the water data services have a unique *WSDL address* on the internet and provide data in WaterML to *WaterOneFlow* web service requests.

For example, http://his02.usu.edu/littlebearriver/cuahsi_1_0.asmx?WSDL defines a CUAHSI water data service provided by Utah State University, which presents data from observations they are making in the Little Bear River in Utah, and http://ccbay.tamucc.edu/CCBayODWS/cuahsi_1_0.asmx?WSDL defines a similar CUAHSI water data service for measurements being made in Corpus Christi Bay by researchers at Texas A&M University – Corpus Christi. Similarly, <http://river.sdsc.edu/wateroneflow/NWIS/DailyValues.asmx?WSDL> is the WSDL address for the CUAHSI water data service for all the Daily Values data from more than 30,000 sites in the USGS National Water Information System, and http://cbe.cae.drexel.edu/CIMS/cuahsi_1_0.asmx?WSDL is the WSDL address for water observations for Chesapeake Bay from the Chesapeake Information Management System linked by researchers at Drexel University in Philadelphia.

The point observations information model used to define *WaterOneFlow* web service requests is illustrated in Figure 2-4. It operates in hierarchical fashion using the following definitions:

- A **data source** is an organization or individual operating an observation network;
- A **network** is a set of observation sites;
- A **site** is a point location where one or more variables are measured, defined geospatially by latitude and longitude coordinates;
- A **variable** is a property describing the physical, chemical or biological conditions in the water;
- A **value** is an observation of a variable at a particular time;
- A **metadata** attribute provides additional information to qualify the value.

For example, in Figure 2-4, Utah State University operates an observation network in the Little Bear River Basin, one of whose sites is located on the Little Bear River at Mendon Rd, where a number of variables are measured, one of which is dissolved oxygen, which had a value of 9.78 mg/L at 5PM on 1 October 2007.

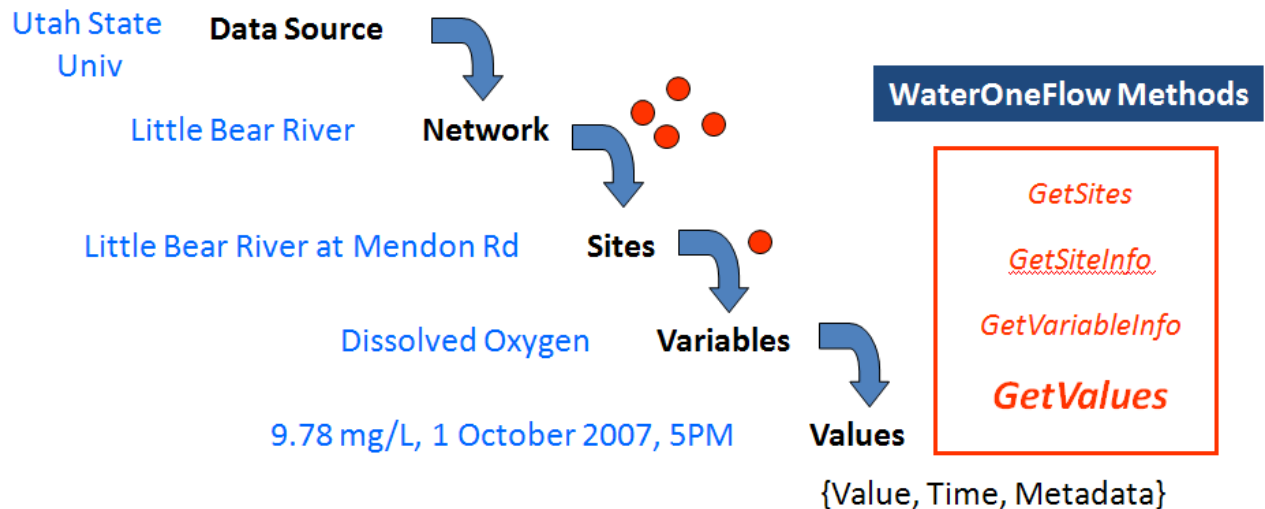


Figure 2-4 Point observations information model used in WaterOneFlow web services

There are currently four WaterOneFlow web service methods.

- **GetSites** – for a given observation network, return the list of observation sites in the network;
- **GetSiteInfo** – for one site in the network, return the list of variables measured there, the count of the number of values available, the begin date time and end date time of the first and last measurements, and metadata about the site and the variables;
- **GetVariableInfo** – for a given network, return the list of variables measured at any location on that network (not all variables need be measured at each site but all sites must use the same variable definitions and units), or – if a variable code is specified - return metadata about that variable such as its units and time support;
- **GetValues** – for one site and one variable, return all data values recorded between a begin date time and an end date time.

WaterML is one example of an eXtensible Markup Language (XML) specification, which is a generalization of the HyperText Markup Language (HTML) used to define normal web pages. In XML specifications, there are a hierarchy of *elements* indicated by `<element> ... </element>` tags, and descriptors or *attributes*. For example, a simple element for *siteName*, which has no attributes appears as:

```
<siteName>Little Bear River at Mendon Road near Mendon, Utah</siteName>
```

A more complicated element, such as that for a data *value* appears as:

```
<value censorCode="nc" dateTime="2007-10-01T17:00:00"
qualityControlLevel="Raw data" methodID="19" sourceID="1">9.78</value>
```

Which means that there is a `<value ...>9.78</value>` element whose attributes are its “nc” for not censored (i.e., it is not below detection limits), “2007-10-01T17:00:00” is the time of measurement (assumed to be local time zone in Utah by default); “Raw data” is the quality control level, “19” indexes the measurement method, and “1” indexes who is the source of this data value (these terms are explained more fully in the full metadata file given below).

The WaterML response to a GetValues call for dissolved oxygen data at Little Bear River at Mendon Rd on 1 October 2007, between 3PM and 8PM, shown below, is comprised of several parts.

First, there is a block of WaterML response describing the query itself – when it was made and for what site, variable, and time period. When no time is included in the date field, it is assumed that the time instant is at midnight at the beginning of the day indicated. The long time stamp value of 2008-07-10T23:31:13.125Z for creationTime (the time the query was made) is an International Standards Organization specification of the Year-Month-Day (T) Hour-Minute-Second.Millisecond (Z) where T indicates the transition between the date and time parts of the time stamp, and Z indicates that this is “Zulu” or Universal Coordinated Time.

```
- <TimeSeriesResponseType xmlns:xsi="http://www.w3.org/2001/XMLSchema
- <queryInfo xmlns="http://www.cuahsi.org/waterML/1.0/">
  <creationTime>2008-07-10T23:31:13.125Z</creationTime>
  - <criteria>
    <locationParam>LittleBearRiver:USU-LBR-Mendon</locationParam>
    <variableParam>LBR:USU32</variableParam>
    - <timeParam>
      <beginDateTime>2007-10-01T15:00:00</beginDateTime>
      <endDateTime>2007-10-01T20:00:00</endDateTime>
    </timeParam>
  </criteria>
  <note>OD Web Service</note>
</queryInfo>
```

Second, there is a block of WaterML response describing the measurement site: its site name, site code (unique identifier in the network); its location in latitude and longitude coordinates and in local UTM Easting and Northing coordinates, and some descriptive information about where the site is located in Utah.

```
- <timeSeries xmlns="http://www.cuahsi.org/waterML/1.0/">
- <sourceInfo xsi:type="SiteInfoType">
  <siteName>Little Bear River at Mendon Road near Mendon, Utah</siteName>
  <siteCode network="LittleBearRiver" siteID="1">USU-LBR-Mendon</siteCode>
  - <geoLocation>
    - <geogLocation xsi:type="LatLonPointType" srs="EPSG:4269">
      <latitude>41.718473</latitude>
      <longitude>-111.946402</longitude>
    </geogLocation>
    - <localSiteXY projectionInformation="NAD83 / UTM zone 12N">
      <X>421276.323</X>
      <Y>4618952.04</Y>
    </localSiteXY>
  </geoLocation>
  <verticalDatum>NGVD29</verticalDatum>
  <note title="County">Cache</note>
  <note title="State">Utah</note>
  <note title="Site Comments">Located below county road bridge at Mendon Road crossing</note>
</sourceInfo>
```

Third, there is a block of WaterML response describing the variable being measured – this is dissolved oxygen in mg/L whose recorded values have been averaged over a 30 minute time interval.

```

- <variable>
  <variableCode vocabulary="LBR" default="true" variableID="32">USU32</variableCode>
  <variableName>Oxygen, dissolved</variableName>
  <valueType>Field Observation</valueType>
  <dataType>Average</dataType>
  <generalCategory>Water Quality</generalCategory>
  <sampleMedium>Surface Water</sampleMedium>
  <units unitsAbbreviation="mg/L" unitsCode="199">milligrams per liter</units>
  <NoDataValue>-9999</NoDataValue>
- <timeSupport isRegular="true">
  - <unit UnitID="102">
    <UnitDescription>minute</UnitDescription>
    <UnitType>Time</UnitType>
    <UnitAbbreviation>min</UnitAbbreviation>
  </unit>
  <timeInterval>30</timeInterval>
  </timeSupport>
</variable>

```

Fourth, there is a block of WaterML response providing the data values themselves and some metadata about them.

```

- <values unitsAbbreviation="mg/L" unitsCode="199" count="11">
  <value censorCode="nc" dateTime="2007-10-01T15:00:00" qualityControlLevel="Raw data" methodID="19" sourceID="1">10.61833</value>
  <value censorCode="nc" dateTime="2007-10-01T15:30:00" qualityControlLevel="Raw data" methodID="19" sourceID="1">10.42</value>
  <value censorCode="nc" dateTime="2007-10-01T16:00:00" qualityControlLevel="Raw data" methodID="19" sourceID="1">10.245</value>
  <value censorCode="nc" dateTime="2007-10-01T16:30:00" qualityControlLevel="Raw data" methodID="19" sourceID="1">10.04333</value>
  <value censorCode="nc" dateTime="2007-10-01T17:00:00" qualityControlLevel="Raw data" methodID="19" sourceID="1">9.78</value>
  <value censorCode="nc" dateTime="2007-10-01T17:30:00" qualityControlLevel="Raw data" methodID="19" sourceID="1">9.568334</value>
  <value censorCode="nc" dateTime="2007-10-01T18:00:00" qualityControlLevel="Raw data" methodID="19" sourceID="1">9.388333</value>
  <value censorCode="nc" dateTime="2007-10-01T18:30:00" qualityControlLevel="Raw data" methodID="19" sourceID="1">9.2</value>
  <value censorCode="nc" dateTime="2007-10-01T19:00:00" qualityControlLevel="Raw data" methodID="19" sourceID="1">9.063334</value>
  <value censorCode="nc" dateTime="2007-10-01T19:30:00" qualityControlLevel="Raw data" methodID="19" sourceID="1">8.955001</value>
  <value censorCode="nc" dateTime="2007-10-01T20:00:00" qualityControlLevel="Raw data" methodID="19" sourceID="1">8.875</value>

```

Finally, there is a block of WaterML response providing information about the method used to collect the data and the researcher to be contacted in the event there are questions about it.

```

- <method methodID="19">
  <MethodDescription>Dissolved oxygen measured using a Hydrolab MS5 Water Quality Multiprobe.</MethodDescription>
</method>
- <source sourceID="1">
  <Organization>Utah State University Utah Water Research Laboratory</Organization>
  <SourceDescription>Continuous water quality monitoring by Utah State University as part of the USDA CEAP Grant</SourceDescription>
- <ContactInformation>
  <ContactName>Jeff Horsburgh</ContactName>
  <TypeOfContact>main</TypeOfContact>
  <Phone>1-435-797-2946</Phone>
  <Email>jeff.horsburgh@usu.edu</Email>
  <Address xsi:type="xsd:string">8200 Old Main Hill ,Logan, UT 84322-8200</Address>
</ContactInformation>
  <SourceLink>http://www.bearriverinfo.org</SourceLink>
</source>
</values>
</timeSeries>
</TimeSeriesResponseType>

```

It can be seen from examination of these blocks of information that a WaterML data file is a complete data package unto itself – it gives you the data values and it tells you where they were collected, by whom, by what method, and also what data quality you should expect from this information – in the example shown, the data are labeled as “raw data” because the measurements from the observation site are being streamed into the observations data model continuously at this site and are thus being published on the web as they are being measured. It is not required to provide such real-time access to the data and a hydrologic scientist may wish to

store and edit the data for quality control purposes before publishing them at a later time when their interpretation is complete.

2.5 METADATA CATALOG

A CUAHSI water data service can be built in two ways:

- A *unified* water data service, where all four WaterOneFlow methods are supported by observations and metadata in a single ODM database – this is the case in nearly all the water data services publishing data from academic data collections;
- A *hybrid* water data service, where the three metadata functions (GetSites, GetSiteInfo, GetVariableInfo) are supported by information in an ODM database, but the GetValues function which extracts the data goes to the web site or web service of the host water agency – this is the case for all the CUAHSI Water Data Services presently operating for large federal databases, such as those of the US Geological Survey.

In the case of the hybrid water data service, the usual path for building the GetValues function is to have a “web page scraper” that programmatically mimics the action of a human user going to an agency web site and downloading data from it. It turns out that about 50% of the hits on the USGS National Water Information System actually come from automated data mining systems built as web page scrapers. The advantage of a web page scraper is that it accepts the agency’s method of publishing its data and doesn’t disrupt its security procedures for protecting its data archive. The disadvantage is that any time the data provider “improves” the format of the web page, the web page scraper breaks, and has to be reprogrammed to operate again.

A more reliable means of supporting the GetValues function is for the data agency to provide a custom-programmed web services function with WaterML as its output. The US Geological Survey has done this for CUAHSI to provide access to its Daily Values database (<http://www.cuahsi.org/docs/usgs-cuahsi-webservices.pdf>), and is presently programming a similar web service to expose its Instantaneous Data Archive, or historical record of instantaneous measurements within a day.

Whether the GetValues function is supported by a web page scraper or by a custom-programmed web service, there is still the need to harvest the metadata needed to support site and variable metadata function. For CUAHSI’s water data services providing access to USGS data, this is done by the USGS doing a metadata dump from NWIS and providing the resulting files to CUAHSI, where they are reformatted to fit the ODM structure.

For a given data service, the metadata describing its measurements are summarized data series, as shown in Figure 2-5. Thus, for variable V_i and Site S_j , there is a count, C , of the number of measurements and a time horizon from a begin date time t_1 to an end date time t_2 between the first and last measurements. It does not matter whether the measurements were regularly recorded or not (that is described as part of the Variable metadata); all that matters is how many measurements there are over what time interval. Hence, one can summarize all such series within a data service as a collection of all such records for variables V_i and sites S_j , over all values of i and j , in other words for all sites and all variables.

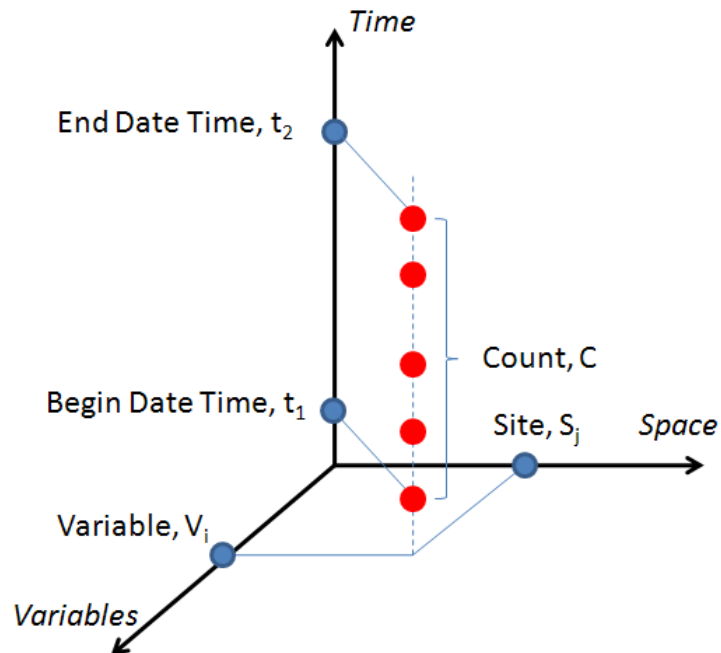


Figure 2-5 Metadata for a data series

It should be noted, however, that this is not the only way that metadata records could be built for observation values in a data cube – one could equally well assemble them as data collections where, for a given point in time and space, the number of variables having values at that point is noted. This might be a good way of indexing results from field sampling for water quality and in particular for aquatic biology species, where the number of variables is very large and a measurement site may be visited only once, so there is no such thing as a time series for any variable. One method of metadata assembly does not preclude others, and it may be that in future versions of CUAHSI HIS other methods than by data series will be used. But, for now, all metadata is summarized as data series.

Within the Observations Data Model, there is one table, the SeriesCatalog, which is not input by the user, but is constructed using a stored procedure by the database itself, once data loading is complete. The SeriesCatalog is shown in Figure 2-6. Its Primary Key is the SeriesID, which indexes how many series there are, and each series has a set of descriptors listed subsequently. The VariableID and SiteID are the indices i and j identifying variables V_i and sites S_j in Figure 2-5, and the BeginDateTime, EndDateTime and ValueCount are specified at the end of the table. The data source, measurement method and quality control level are also tabulated here for each series, so that searches can be made for data from particular organizations, or using particular methods, or with defined standards of quality control.

SeriesCatalog
SeriesID {PK}
SiteID {FK}
SiteCode
SiteName
VariableID {FK}
VariableCode
VariableName
Speciation
VariableUnitsID {FK}
VariableUnitsName
SampleMedium
ValueType
TimeSupport
TimeUnitsID {FK}
TimeUnitsName
DataType
GeneralCategory
MethodID {FK}
MethodDescription
SourceID {FK}
Organization
SourceDescription
Citation
QualityControlLevelID {FK}
QualityControlLevelCode
BeginDateTime
EndDateTime
BeginDateTimeUTC
EndDateTimeUTC
ValueCount

Figure 2-6 SeriesCatalog for a data series

Now, imagine that each of these data series is similarly cataloged over all measurement networks, and all that information is accumulated in a single database. This is what has been done at the HIS Central at the San Diego Supercomputer Center where more than eight million data series are presently cataloged from measurements at nearly two million locations in the nation. We call this the National Water Metadata Catalog and it is a valuable information resource in its own right – it does not store any data, but only metadata about who has measured what and where and when, but that is very valuable when you are searching for particular kinds of information in particular regions and periods of time.

One last major step still remains to render the nation’s water information scientifically searchable – the variables have to be tagged using a set of scientific concepts so that searches by concept can operate over all the observation networks without regard to the fact that the same variable may be described differently by different organizations. This process is called semantic mediation, and its conceptual framework is presented in Chapter 5 of this report, so will not be discussed further here.

2.6 OTHER INFORMATION MODELS

The previous sections have described in some detail the conceptual framework that underlies the three key elements of the current CUAHSI HIS that are presented in Chapter 1, namely data storage, data access and data indexing. By these means we have succeeded to a significant degree in creating a single window on water observations data measured and stored by different organizations at many geographic locations. There are many

limitations on our work, however, and this section briefly describes some of the emerging elements of what we see will become parts of CUAHSI HIS in the future.

2.6.1 INTEROPERABILITY WITH OTHER SCIENTIFIC DISCIPLINES

A key requirement for successful cyberinfrastructure is that it be interoperable among scientific disciplines. Water is important as a physical environment which defines its physical and chemical properties; water is important as a living environment for fish and aquatic organisms, and to support plant life; water is important to the human environment – water accounts for more than half of the weight of a human body. Although water interacts with many disciplines, active cyberinfrastructure efforts in four disciplines are particularly relevant for hydrologic science – those in atmospheric science, geological science, ocean science and ecological science. The first three of these are the environments above, below, and surrounding the domain of hydrology of the land surface and near subsurface. The fourth, ecological science, connects hydrology with living things. A set of emerging NSF Critical Zone Observatories combine data from several of these disciplines.

Atmospheric Science

Unidata is the equivalent organization to CUAHSI HIS within the domain of the Universities Corporation for Atmospheric Research (<http://www.unidata.ucar.edu/>). Located in Boulder, Colorado, Unidata has been operational since 1983 and is the means of supplying real-time atmospheric data to more than 100 universities in the United States. Technology developed by Unidata is also being used within the National Weather Service and related agencies to access and transmit atmospheric data. Two elements of Unidata's function are critical for CUAHSI HIS – netCDF and THREDDS. NetCDF is a file format for storing multidimensional arrays of data, where the index dimensions, such as those for time and space are called *coordinate dimensions*, and dimensions that store actual data values like temperature or relative humidity are called *variable dimensions*. NetCDF is a very good format for storing datasets which are documented and interpreted as a whole, rather than individual data values as we are doing with our Observations Data Model and WaterML web services.

THREDDS (Thematic Real-Time Environmental Data Distribution Services) (<http://www.unidata.ucar.edu/projects/THREDDS/>) is a means like WaterOneFlow web services of publishing multidimensional array information on the internet without regard to the underlying form of the gridded data. CUAHSI HIS has adopted netCDF as its default file format for multidimensional array data and intends to include publication of such arrays using THREDDS in a future version of its HIS Server. One existing CUAHSI water data service, that for the North American Forecast Model (NAM), uses the THREDDS server as a data source, and returns a time series in WaterML of forecast weather conditions for a single geographic point selected anywhere within the spatial domain of the model. CUAHSI water data services to other data sources published in THREDDS could similarly be developed, such as for NEXRAD precipitation data published in THREDDS by the National Climatic Data Center, in Asheville, NC.

Geological Science

The principal NSF cyberinfrastructure project in geological sciences is GEON (<http://www.geongrid.org/>), a collaboration between geological scientists and computer scientists, whose computer science base, like that of CUAHSI HIS, is also at the San Diego Supercomputer Center. One GEON activity of particular interest to CUAHSI HIS is its capacity to process large point clouds of LIDAR information for land surface terrain collected in NSF projects. We are working with GEON to add processing functions to their terrain processing system that will yield

information products of interest to hydrologists, such as watershed and stream network delineation. LIDAR data is much more precise than conventional National Elevation Dataset terrain data but is so voluminous that processing it is beyond the range of desktop computers for significant regional coverage.

Ocean Science

The International Ocean Observing System, IOOS, operates an observations registry (<http://obsregistry.org/index.php>) which catalogs the measurement of water conditions along the coasts of the United States in a similar way the SeriesCatalog approach that we have developed. One of these networks is the Texas Coastal Ocean Observing Network (TCOON) (<http://lighthouse.tamucc.edu/TCOON/>). As part of building a set of Texas Water Data Services (<http://data.cwr.utexas.edu>), using separate funding from our NSF project, we succeeded in building a hybrid water data service for TCOON where the metadata was read from the IOOS XML metadata specification and the GetValues function was made operational using a web scraper on the TCOON web site, so that coastal water observations data are published as a water data service just the same as water data services constructed for inland waters. It may later be possible to accomplish that goal for other networks in the IOOS Observations Registry.

Ecological Science

The ecological science community and the National Center for Ecological Analysis and Synthesis (NCEAS) have created an Ecological Metadata Language (EML) (<http://www.nceas.ucsb.edu/ecoinfo/approach>) whose purpose is to describe ecological datasets. These datasets can use any physical format and are not restricted to the rigid structure we have used in our CUAHSI Observations Data Model. The focus of EML is on the digital description of data so that it can be subsequently searched for and discovered, rather than homogeneity of format and meaning of terms, as we are doing with CUAHSI water data services. EML is the standard metadata language used at all the Long Term Ecological Research (LTER) sites, and some discussion has taken place between CUAHSI HIS and the LTER Network Office in Albuquerque about greater interoperability of cyberinfrastructure. The LTER Network has a TRENDS project that documents long-term trends in annual data measured at LTER sites. This would be an excellent candidate to be published as a CUAHSI water data service.

Critical Zone Observatories

The NSF Earth Sciences Division has established three Critical Zone Observatories at Penn State University, the University of Colorado, and at the University of California-Merced, respectively. The Critical Zone extends “from the outer limits of vegetation down to and including the zone of groundwater. This zone sustains most terrestrial life on the planet.” (<http://www.czen.org/>) Through interaction with the principal investigators, a data inventory was made of information that is being or will be collected for each observatory. These data fall into four categories:

1. Time series data measured at a fixed location at regular intervals;
2. Analyses of manually collected water samples at fixed location at regular or irregular intervals;
3. Analyses of one time samples – rock and soil samples – from known position and depth;
4. Other data – data types not otherwise classified, such as LIDAR, geophysics, land surveys, tree surveys, stream channel surveys.

The CUAHSI Hydrologic Information System is well suited for storing the first two classes of data (which will make up the bulk of the observational data at the Critical Zone Observatories) but is less effective for the third type of data because it indexes the data by series rather than collections. Another means, such as Digital Libraries, could be used to store and publish the fourth class of data being measured at these observatories.

2.6.2 INTEROPERABILITY WITH OTHER TECHNOLOGIES

Geographic Information Systems

Although observations data have been our principal focus to date, CUAHSI HIS has made a considerable effort to involve GIS in our work. In the first version of the HIS server used at the WATERS testbed sites, ArcGIS Server version 9.2 was used to enable interactive map querying of the data in an application called DASH (Data Access System for Hydrologists). This application proved to operate rather slowly, and after some feedback about this from project scientists, we are now working with the Environmental Systems Research Institute¹ on a new approach, called HydroViewer, which is a much lighter and faster mapping application for viewing observations data. HydroViewer uses preprocessed images of the NHDPlus dataset and a range of spatial scales to facilitate rapid panning from continental to local scale. We anticipate in the future a greater degree of reliance on the NHDPlus dataset, which is the geospatial integration of the National Hydrography Dataset, National Elevation Dataset and National Land Cover Dataset. Each of the 2.9 million river and water body reaches in the nation (average length 2 km) has an associated catchment defined (average area 3 square km) whose upstream and local land cover characteristics are defined. Linking our National Water Metadata Catalog to NHDPlus would provide a powerful combination of geographic and observational water data.

Remote Sensing

We have one CUAHSI water data service, that to MODIS, which accesses NASA remote sensing information and supplies a time series of atmospheric water properties averaged over a latitude/longitude box in space and a month in time, and converted into a time series in WaterML. This data service was built by adapting the NASA MODIS Online Visualization and Analysis System (MOVAS) in an analogous manner to the web page scraping described earlier. NASA does use web services to internally convey remote sensing information from one office to another, but these services are not publicly accessible. Another factor complicating access to NASA remote sensing data is that access to water data is spread over several Distributed Active Archive Centers (DAACs), which each operate differently with respect to providing data access tools. It is clear that the hydrologic science community would like better access to remotely sensed data, but the way to adapt our services-oriented architecture for water data to include remotely sensed products, is not apparent at present.

Digital Libraries

One method of storing and accessing water data products is to store them in a digital library. The library systems of the nation's universities are gradually developing *institutional repositories*, or digital data holdings representing the intellectual output of the university, much like the digital equivalent of the traditional binding a dissertation as a book, and archiving it in the university library. Various technologies are used for this purpose, such as Dspace

¹ The CUAHSI HIS team wishes to record its appreciation for the significant assistance and collaboration by ESRI, which has been provided entirely free of charge to the project.

(<http://www.dspace.org/>), and the critical thing is not the particular technology that is used but rather that all data entities in a digital library have a unique URL address, which means that they can be indexed by search engines such as Google and permanently accessed. Some examples are the Texas Digital Library <http://www.tdl.org/> and the Iowa Digital Library <http://digital.lib.uiowa.edu/>. As part of the development of Texas Water Data Services at the University of Texas, we have linked documents in the Texas Digital Library with water data services.

2.6.3 COLLABORATIONS

Open Geospatial Consortium

The Open Geospatial Consortium (<http://www.opengeospatial.org/>) is an international organization with 370 member institutions which has built up over the last 20 years, a set of standard methods for conveying geospatial information through the internet. In particular, they have developed the Geographic Markup Language, GML, for defining the properties of geographic objects. We would like to work towards adapting WaterML to make it more GML compliant so that it could convey water data series describing arbitrarily shaped areas like watersheds and aquifers, not just points and boxes as now. The OGC is considering forming a hydrology working group, which would work towards standards harmonization for hydrologic data transfer.

International Collaborations

The CUAHSI HIS project has two active international collaborations, one with the European OpenMI consortium for interoperability of data and hydrologic simulation models, described more fully in Chapter 8, and the other with the Australian Water Resources Information Project of the Bureau of Meteorology, who are particularly interested in our work on water data services. We hope that as these collaborations mature they will enable us to leverage water information system developments created elsewhere for the benefit of the hydrologic science.

2.7 CONCLUDING REMARKS

The concept of water data services for point observations is well understood and CUAHSI HIS has built and proven in operation an effective services-oriented architecture for this purpose. It is considered, among the federal water agencies that focus on data collection and archiving, that the CUAHSI approach to this task has “broken through the stovepipes” and thus there is no real need to develop an alternative water web services data system, but rather, as individual agencies, to work with CUAHSI to provide access to that agency’s data as CUAHSI water data services. The need of hydrologic scientists to be able to publish and access academically collected data is quite a different task, because the observations database to load and store the data has to be provided, and its web services implemented, sometimes in remote locations with not very much experience in the intricacies of such computer systems. Despite these difficulties, a good foundation has been established.

2.8 REFERENCES

- Josuttis, N.M., (2007), “SOA in practice – the art of distributed system design”, O’Reilly Press, Sebastopol, CA, 324p.
- National Research Council (2007), “Elevation data for floodplain mapping”, Report of the Committee on Floodplain Mapping Technologies, National Academy Press, Washington DC, 151p.
- Tomlinson, R., (2003), “Thinking about GIS”, ESRI Press, Redlands CA, 283p.

Chapter 3. OBSERVATIONS DATA MODEL

By David Tarboton and Jeff Horsburgh, Utah State University

Observations are fundamental to hydrology and water resources, and the way these data are organized and manipulated either enables or inhibits the analyses that can be performed. The Observations Data Model (ODM) presented here serves as the foundation for the systematic organization, storage and retrieval of point observations within the CUAHSI HIS. At a conceptual level, the 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 collected by multiple investigators. Within the 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 is based upon a relational database model that 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. The conceptual organization of point observations provided by the ODM serves as a foundation for the representation of hydrologic observations throughout much of the HIS, including WaterML, the WaterOneFlow web services, and the various metadata catalogs that support HIS functionality.

The ODM represents a new, systematic way to organize and share data that overcomes many of the syntactic and semantic differences between heterogeneous datasets, thereby facilitating a more integrated understanding of water resources based on fully specified information. This chapter reviews the design principles and features of ODM, focusing on how it can be used to enhance the organization, publication, and analysis of point observations data while retaining a simple relational format. We then describe tools and HIS system components that work with the ODM. These include tools to load data into the ODM and view and edit ODM data. Also described is the centralized web based system for editing ODM controlled vocabularies (CV) that allows users to submit new terms for consideration as part of CUAHSI's standard vocabulary. The ODM CV system supports web services that allow users to synchronize their local ODM controlled vocabularies with the standard set, thereby limiting semantic heterogeneities. This chapter is written at a fairly high level, summarizing the key concepts and ideas. For details, the reader is referred to the ODM Design specifications (<http://his.cuahsi.org/documents/ODM1.1DesignSpecifications.pdf>), paper in Water Resources Research (Horsburgh et al., 2008) and ODM loader and tools manuals.

3.1 ODM DESIGN

The ODM was designed to store environmental observations along with sufficient metadata to provide traceable heritage from raw measurements to usable information, allowing observations stored in ODM to be unambiguously interpreted and used. An observation is an event that results in a value describing some phenomenon (Open Geospatial Consortium Inc., 2006). Observation values are not self describing, and, because of this, interpretation of a particular set of observations requires contextual information, or metadata. Metadata is the descriptive information about data that explains the measurement attributes, their names, units, precision, accuracy, and data layout, as well as the data lineage describing how the data was measured, acquired, or

computed (Gray et al., 2005). The importance of recording fundamental metadata to help others discover and access data products is well recognized (Gray et al., 2005; Bose, 2002; Michener et al., 1997).

Table 3-1 ODM attributes associated with an observation

Attribute	Definition
Value	The observation value itself
Accuracy	Quantification of the measurement accuracy associated with the observation value
Date and Time	The date and time of the observation (including time zone offset relative to UTC and daylight savings time factor)
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. latitude and longitude)
Units	The units (e.g. m or m ³ /s) and unit type (e.g. length or volume/time) associated with the variable
Interval	The interval over which each observation was collected or implicitly averaged by the measurement method and whether the observations are regularly recorded on that interval
Offset	Distance from a reference point to the location at which the observation was made (e.g. 5 meters below water surface)
Offset Type/ Reference Point	The reference point from which the offset to the measurement location was measured (e.g. water surface, stream bank, snow surface)
Data Type	A 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
Data Qualifying Comments	Comments accompanying the data that can affect the way the data is used or interpreted (e.g. holding time exceeded, sample contaminated, provisional data subject to change, etc.)
Analysis Procedure	An indication of what method was used to collect the observation (e.g. dissolved oxygen by field probe or dissolved oxygen by Winkler Titration)
Source	Information on the original source of the observation (e.g. from a specific instrument or investigator 3 rd party database)
Sample Medium	The medium in which the sample was collected (e.g. water, air, sediment, 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

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 type of variable 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 to make observations, qualifying comments about the observation, and information about the organization that made the observation.

Table 3-1 presents general attributes that are 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 3-1, except for the value of the observation itself, can be considered metadata.

The ODM logical data model (Figure 3-1) has been designed to store observation values and their supporting metadata in a structured way.

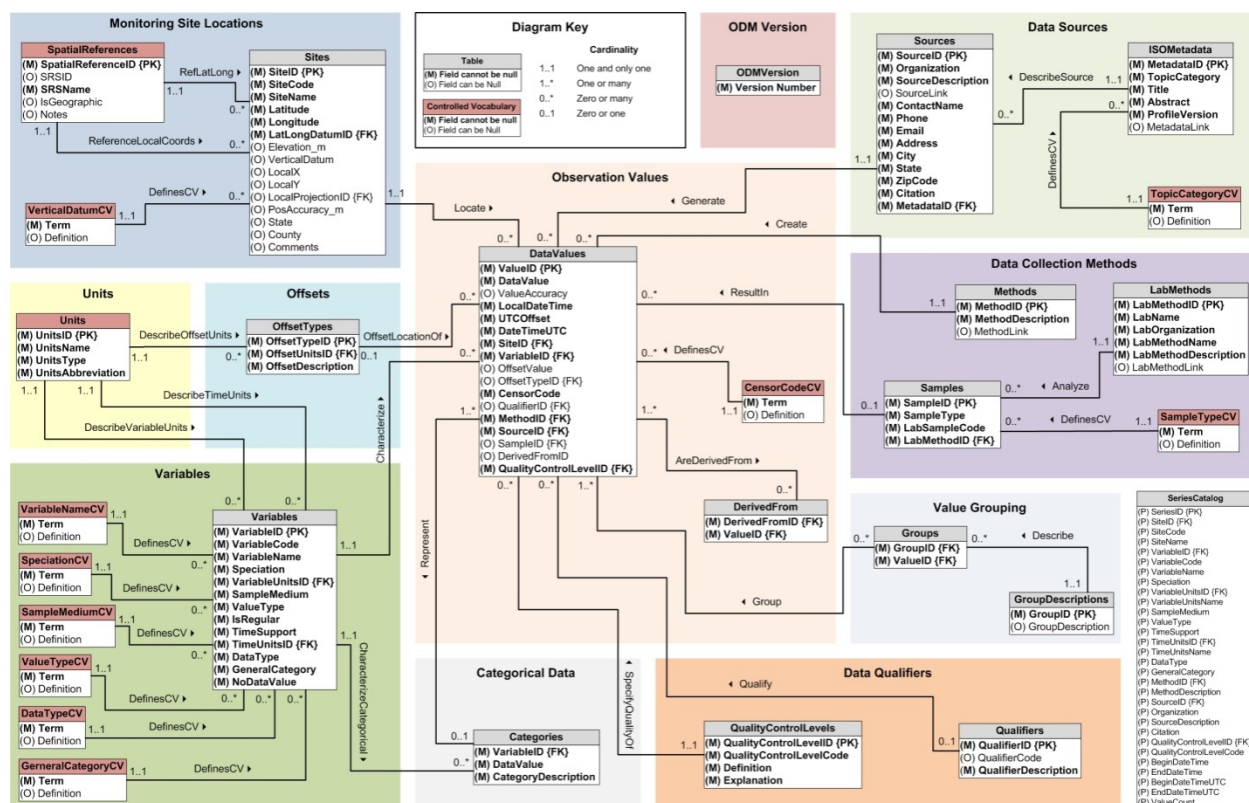


Figure 3-1 ODM logical data model. The primary key field for each table is designated with a {PK} label. Foreign keys are designated with a {FK} label. The lines between tables show relationships with cardinality indicated by numbers and labeled with the name and directionality of the relationship. Required (Mandatory) data fields are indicated with an M and are in bold, while optional data fields are indicated with an O.

The DataValues table at the center of Figure 3-1 stores the numeric values for observations and links (foreign keys) to all of the data value level attributes. Most of the attribute details are stored in the tables surrounding the DataValues table to avoid redundancy. The relationships between tables are shown, along with all of the required primary and foreign keys. Each of these relationships has a name, which is indicated by a text label, and a

directionality that is indicated by an arrow. For example, the relationship between the Sources table and the DataValues table is named “Generate” and has directionality that points from the Sources table to the DataValues table. This indicates that data sources *generate* data values. Additionally, the cardinality, or numeric relationship between entities in each of the tables, is shown at either end of each of the relationship lines. For example, the relationship line between the Variables and DataValues tables has “1..1” at the Variables end, and “0..*” at the DataValues end, indicating that there is one and only one variable associated with 0 or many DataValues (i.e., there is a one-to-many relationship between variables and data values) and that variables *characterize* data values.

The ODM evolved from an initial design presented at a CUAHSI workshop held in Austin during March, 2005 (Maidment, 2005) that was then widely reviewed with comments being received from 22 individuals (Tarboton, 2005). These reviews served as the basis for a redesign that was presented at a CUAHSI workshop at Duke University during July, 2005 and presented as part of the CUAHSI HIS status report (Horsburgh et al., 2005). Following this presentation of the design, the data model was reviewed and commented on by a number of other individuals and groups, including the CLEANER (Collaborative Large-scale Engineering Analysis Network for Environmental Research) cyberinfrastructure committee. Further versions of the Observations Data Model were circulated in April, June and October 2006, during which time tables were added to give spatial reference information, metadata information, and to define controlled vocabularies. Version 1.0 of ODM, which was the first public version of ODM, was released in May 2007. At the release of this version many of the fields and tables were changed to reconcile the nomenclature with that independently developed for our web services. Following the release of ODM 1.0, it has been implemented and tested within the WATERS Network test bed sites and was documented in Water Resources Research (Horsburgh et al., 2008). Based on the use in the WATERS Network test bed sites we identified the need to formally encode constraints and preferred database practices, as well as add additional information. The current version, ODM Version 1.1, has just been released (May 2008).

The following changes were made going from Version 1.0 to 1.1:

- A Citation field was added to the Sources table to provide a place for a formal citation for data in the database.
- A Speciation field was added to the Variables table. This provides a place to store information about the speciation of chemistry observations and allows a more comprehensive description of variables that are defined in terms of a chemical species. A SpeciationCV controlled vocabulary table was added to define this field.
- The controlled vocabulary was relaxed on the QualityControlLevels table to allow more detailed versioning of data series. A QualityControlLevelCode was also added to this table to facilitate this. Now rather than QualityControlLevels being limited to 0, 1, ..., 5, quality control levels such as 1.1 can be specified to allow data managers greater control where they record data as it goes through quality control processing.
- All integer IDs serving as the primary key for tables in ODM have been changed to auto number/identity fields to be consistent with preferred database practice.
- Text field lengths have been relaxed in some cases and have been standardized according to the following scheme: codes = 50 characters, terms = 255 characters, links = 500 characters, definitions/explanations = unlimited. This allows fields that may contain lengthy descriptions, such as abstracts and method descriptions to be better accommodated.
- Check constraints have been defined for the Latitude and Longitude fields in the Sites table.
- Check constraints have been added to many of the fields in ODM to constrain the characters that are valid for those fields (for example special characters that could cause problems with the web services are no longer allowed as part of an integer code).
- Relationships have been added between controlled vocabulary tables and the tables that contain the fields that they define. This was done to more rigorously enforce the ODM controlled vocabularies.

- Unique constraints were placed on both SiteCode in the Sites table and VariableCode in the Variables table.
- An ODMVersion table was added to store the version number of the database.
- The SeriesCatalog table was updated based on the fields added above.

Through all these changes, the fundamental design has not changed since the status report presentation of the model (Horsburgh et al., 2005),.

The design specifications (<http://his.cuahsi.org/documents/ODM1.1DesignSpecifications.pdf>) and ODM paper (Horsburgh et al., 2008) describe in detail features of how monitoring site geography, variables, units, sources, quality control, value accuracy, data qualifying comments and other important aspects of annotating point observations are handled in the ODM design, so these details will not be repeated here. The ODM has proven to be a simple but powerful construct for the representation of point observations data. The ODM has been implemented at WATERS Network test bed locations around the country, as well as adopted by some other experimental watersheds (e.g. Reynolds Creek, http://idahowaters.uidaho.edu/RCEW_ODWS/) and even the Australian Water Resource Observation Network (WRON). In these applications it has proven its capability to store many sets of observational data, including physical, chemical and biological variables.

3.2 ODM EXAMPLES

The following examples illustrate the capability of the ODM data model to store different types of point observations. The examples present selected fields and tables chosen to illustrate key capabilities of the data model. These examples are presented using table names and field names shown in Figure 3-1.

3.2.1 STREAMFLOW - GAGE HEIGHT AND DISCHARGE

Figure 3-2 illustrates how both stream gage height measurements and the associated discharge estimates derived from the gage height measurements can be stored in the ODM. Note that gage height in feet and discharge in cubic feet per second are both in the same data table but with different VariableIDs that reference the Variables table, which specifies the variable name, units, and other quantities associated with these data values. The link between VariableID in the DataValues table and Variables table is shown. In this example, discharge measurements are derived from gage height (stage) measurements through a rating curve. The MethodID associated with each discharge record references into the Methods table that describes this and provides a URL that contains metadata details for this method. The DerivedFromID in the DataValues table references into the DerivedFrom table that references back to the corresponding gage height in the DataValues table from which the discharge was derived.

ValueID	DataValue	ValueAccuracy	LocalDateTime	UTCOffset	SiteID	VariableID	MethodID	DerivedFromID
1	4.18		05/01/2006 0:00:00.000	-7	1	1	1	
97	748		05/01/2006 0:00:00.000	-7	1	2	1	
193	722	22.89831642	05/01/2006 0:00:00.000	-7	1	3	100	
2	4.18		05/01/2006 0:15:00.000	-7	1	1	1	
98	748		05/01/2006 0:15:00.000	-7	1	2	2	
3	4.17		05/01/2006 0:30:00.000	-7	1	1	1	
99	742		05/01/2006 0:30:00.000	-7	1	2	3	
4	4.17		05/01/2006 0:45:00.000	-7	1	1	1	
100	742		05/01/2006 0:45:00.000	-7	1	2	4	
5	4.17		05/01/2006 1:00:00.000	-7	1	1	1	
101	742		05/01/2006 1:00:00.000	-7	1	2	5	
6	4.17		05/01/2006 1:15:00.000	-7	1	1	1	
102	742		05/01/2006 1:15:00.000	-7	1	2	6	

DerivedFromID	ValueID
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13

VariableID	VariableCode	VariableName	VariableUnitsID	ValueType	IsRegular	TimeSupport	TimeUnitsID	DataType	GeneralCategory
1	NWIS:00065	Gage height		1 Field Observation	<input checked="" type="checkbox"/>	15	5 Continuous	Hydrologic	
2	NWIS:00060	Discharge		2 Derived Value	<input checked="" type="checkbox"/>	15	5 Continuous	Hydrologic	
3	NWIS:00060	Discharge, daily average		2 Derived Value	<input checked="" type="checkbox"/>	24	6 Average	Hydrologic	
4	NWIS:00300	Dissolved oxygen concentration		3 Field Observation	<input type="checkbox"/>	0	5 Sporadic	Water Quality	

UnitsID	UnitsName	UnitsType	UnitsAbbreviation
1	Feet	Length	ft
2	Cubic feet per second	Flow	ft ³ /s
3	Milligrams per liter	Concentration	mg/L
4	Meters	Length	m
5	Minutes	Time	min
6	Hours	Time	hr

MethodID	MethodDescription
1	Gage height measured with continuous data logger
2	Discharge derived from water stage using site specific rating curve
3	Daily average discharge derived from 15 minute continuous discharge values
4	Dissolved oxygen measured with a Hydrolab multiprobe field instrument
7	Precipitation from a tipping bucket gage. 0 values not logged.
8	Daily precipitation from a tipping bucket gage. 0 values not logged.

Figure 3-2 Excerpts from tables illustrating the population of ODM with streamflow gage height (stage) and discharge data

3.2.2 STREAMFLOW - DAILY AVERAGE DISCHARGE

Figure 3-3 shows excerpts from tables illustrating the population of ODM with both continuous discharge values and derived daily averages. Daily average streamflow is reported as an average of continuous 15 minute interval data values. The record giving the single daily average discharge with a value of $722 \text{ ft}^3 \text{ s}^{-1}$ in the DataValues table has a DerivedFromID of 100. This refers to multiple records in the DerivedFrom table, with associated ValueIDs 97, 98, 99, ... 113 shown. These refer to the specific 15 minute discharge values in the DataValues table used to derive the average daily discharge. VariableID in the DataValues table identifies the appropriate record in the Variables table specifying that this is a daily average discharge with units of $\text{ft}^3 \text{ s}^{-1}$ from UnitsID referencing in to the Units table. MethodID in the DataValues table identifies the appropriate record in the Methods table specifying that the method used to obtain this data value was daily averaging.

ValueID	DataValue	ValueAccuracy	LocalDateTime	UTCOffset	SiteID	VariableID	MethodID	DerivedFromID
1	4.18		05/01/2006 0:00:00.000	-7	1	1	1	1
97	748		05/01/2006 0:00:00.000	-7	1	1	2	1
193	722	22.89831642	05/01/2006 0:00:00.000	-7	1	3	3	100
2	4.18		05/01/2006 0:15:00.000	-7	1	1	1	1
98	748		05/01/2006 0:15:00.000	-7	1	1	2	2
3	4.17		05/01/2006 0:30:00.000	-7	1	1	1	1
99	742		05/01/2006 0:30:00.000	-7	1	1	2	3
4	4.17		05/01/2006 0:45:00.000	-7	1	1	1	1
100	742		05/01/2006 0:45:00.000	-7	1	1	2	4
5	4.17		05/01/2006 1:00:00.000	-7	1	1	1	1
101	742		05/01/2006 1:00:00.000	-7	1	1	2	5
6	4.17		05/01/2006 1:15:00.000	-7	1	1	1	1
102	742		05/01/2006 1:15:00.000	-7	1	1	2	6

VariableID	VariableCode	VariableName	VariableUnitsID	ValueType	isRegular	TimeSupport	TimeUnitsID	DataType	GeneralCategory
1	NWIS:00065	Gage height		1 Field Observation	<input checked="" type="checkbox"/>	15	5 Continuous	Hydrologic	
2	NWIS:00060	Discharge		2 Derived Value	<input checked="" type="checkbox"/>	15	5 Continuous	Hydrologic	
3	NWIS:00060	Discharge, daily average		2 Derived Value	<input checked="" type="checkbox"/>	24	6 Average	Hydrologic	
4	NWIS:00300	Dissolved oxygen concentration		3 Field Observation	<input type="checkbox"/>	0	5 Sporadic	Water Quality	

UnitsID	UnitsName	UnitsType	UnitsAbbreviation
1	Feet	Length	ft
2	Cubic feet per second	Flow	ft^3/s
3	Milligrams per liter	Concentration	mg/L
4	Meters	Length	m
5	Minutes	Time	min
6	Hours	Time	hr

MethodID	MethodDescription
1	Gage height measured with continuous data logger
2	Discharge derived from water stage using site specific rating curve
3	Daily average discharge derived from 15 minute continuous discharge values
4	Dissolved oxygen measured with a Hydrolab multiprobe field instrument
7	Precipitation from a tipping bucket gage. 0 values not logged.
8	Daily precipitation from a tipping bucket gage. 0 values not logged.

Figure 3-3 Excerpts from tables illustrating the population of ODM with daily average discharge derived from 15 minute discharge values

3.2.3 WATER CHEMISTRY FROM A PROFILE IN A LAKE

Reservoir profile measurements provide an example of the logical grouping of data values and data values that have an offset in relationship to the location of the monitoring site. These measurements may be made simultaneously (by multiple instruments in the water column) or over a short time period (one instrument that is lowered from top to bottom). Figure 3-4 shows an example of how these data would be stored in ODM. The OffsetTypes table and OffsetValue attribute are used to quantify the depth offset associated with each measurement. Each of the data values shown has an OffsetTypeID that references into the OffsetTypes table. The OffsetTypes table indicates that for this OffsetType the offset is “Depth below water surface.” The OffsetTypes table references into the Units table indicating that the OffsetUnits are meters, so OffsetValue in the DataValues table is in units of meters depth below the water surface.

Each of the data values shown has a VariableID that in the Variables table indicates that the variable measured was dissolved oxygen concentration in units of mg liter^{-1} . Each of the data values shown also has a MethodID that in the Methods table indicates that dissolved oxygen was measured with a Hydrolab multiprobe. The combination of the variable name, units, and method are sufficiently general to describe what has been measured. Within the ODM controlled vocabularies, the convention is that the units remain generic, whereas the variable names are more specific. For example, “dissolved phosphorus as P” is a different variable name than “dissolved phosphorus as PO_4 ,” but the units of both are mg liter^{-1} .

Additionally, the data values shown are part of a logical group of data values representing the water chemistry profile in a lake. This is represented using the Groups table and GroupDescriptions table. The Groups table

associates GroupID 1 with each of the ValueIDs of the data values belonging to the group. A description of this group is given in the GroupDescriptions table.

The figure displays six database tables with the following data excerpts and relationships:

- DataValues Table:** Contains records for ValueID (194-201), DataValue, LocalDateTime, UTCOffset, SiteID, VariableID (4), OffsetValue (0.2-7), OffsetTypeID (1), and MethodID (4). A blue circle highlights VariableID 4, a red circle highlights OffsetTypeID 1, and a green circle highlights MethodID 4.
- Variables Table:** Contains records for VariableID (4), VariableCode (0300), VariableName (Dissolved oxygen concentration), VariableUnitsID (3), ValueType (Field Observation), IsRegular, TimeSupport, TimeUnitsID (5), DataType (Sporadic), and NoDataValue (-9999). A blue circle highlights VariableID 4.
- GroupDescriptions Table:** Contains records for GroupID (1) and GroupDescription (Echo Reservoir Profile 9/4/2003). A purple circle highlights GroupID 1.
- Units Table:** Contains records for UnitsID (3, 4), UnitsName (Milligrams per liter, Meters), UnitsType (concentration, length), and UnitsAbbreviation (mg/L, m). A red circle highlights UnitsID 3 and a green circle highlights UnitsID 4.
- OffsetTypes Table:** Contains records for OffsetTypeID (1), OffsetUnitsID (4), and OffsetDescription (Depth below water surface). A red circle highlights OffsetTypeID 1 and a green circle highlights OffsetUnitsID 4.
- Methods Table:** Contains records for MethodID (4), MethodDescription (Dissolved oxygen measured with a Hydrolab multiprobe field instrument), and MethodLink (http://www.hydrolab.com). A green circle highlights MethodID 4.

Relationships indicated by lines:

- Blue line: VariableID 4 in DataValues to VariableID 4 in Variables.
- Red line: OffsetTypeID 1 in DataValues to OffsetTypeID 1 in OffsetTypes.
- Green line: MethodID 4 in DataValues to MethodID 4 in Methods.
- Purple line: GroupID 1 in GroupDescriptions to GroupID 1 in Groups.
- Red line: UnitsID 3 in Units to VariableUnitsID 3 in Variables.
- Green line: UnitsID 4 in Units to OffsetUnitsID 4 in OffsetTypes.

Figure 3-4 Excerpts from tables illustrating the population of ODM with water chemistry data from a profile in a lake

3.3 TOOLS FOR WORKING WITH ODM

3.3.1 DATA LOADING APPLICATIONS

Within CUAHSI HIS Servers, the ODM has been implemented as a Microsoft SQL Server 2005 database. The following applications are available for loading data into an ODM database.

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

The ODM Data Loader and Streaming Data Loader functionality will be described in what follows. For data loading with SSIS users are referred to the SSIS manuals. A presentation and example script that illustrates the use of SSIS for loading data into ODM is available at <http://www.cuahsi.org/meetings/wtb-presentations.html>.

ODM DATA LOADER

The ODM Data Loader (ODMDL) application was created to allow administrators of local instances of the ODM to load data into an instance of the ODM. The development of the ODMDL application has several advantages. First, ODMDL protects the security and consistency of an ODM database because it provides users with a set of tools for validating and loading their data into ODM. This minimizes the potential for human caused errors in loading these data into an ODM database. The ODMDL input file formats are similar to the table structures in ODM, but they do provide users with some flexibility in specifying the required metadata. Users do not need to perform any specialized programming to parse and load the data, and ODMDL ensures that the data are fully qualified with valid metadata when they are loaded.

The general concept behind ODMDL is that it should accept as input, data in table format (Microsoft Excel or comma separated values .csv) that is sufficient that it can be loaded into ODM without violating any ODM constraints. Tables have a one row header that uses ODM field names in the header, followed by the data in subsequent rows. The ODMDL has the following functionality:

- Bulk data loading – Bulk data loading provides the capability to load data from a flat file containing in each row a data value and all the metadata necessary to annotate the data value. Bulk data loading does not require a user to organize the data into the table format of ODM; rather this is done by the loader. The user does however need to ensure that, at a minimum, all required metadata is present, or the loader will give an error.
- Loading of individual ODM data tables – Individual ODM table loading provides the capability to independently load data into each ODM table from separate input files.
- Sequential data loading – Data values cannot be loaded in to ODM without the required metadata being present. If data and metadata files are being loaded separately it is therefore important to load information in the correct order so that dependencies are not violated. The ODMDL includes a Wizard that guides a user through the steps of loading ODM tables in the correct order.
- Command line Interface and batch data loading – It is often useful to be able to script data loading operations. ODMDL supports command line execution so that loads of multiple files can be scripted and so that data loading tasks can be automated.

For more details about ODMDL, see <http://his.cuahsi.org/odmdataloader.html> where you can download the software, detailed software manual and functional specifications.

ODM STREAMING DATA LOADER

The Streaming Data Loader (ODM SDL) was designed for streaming continuously measured sensor data generated by a monitoring and telemetry system into an ODM database. Similar to the regular Data Loader, the ODM SDL is a file based data loader, but takes as input datalogger files that have a single date column and potentially multiple columns of data. The ODM SDL provides simple visual tools for mapping streaming data files to the ODM schema and for specifying all of the required metadata, which means that users do not need to perform any specialized programming to parse and load the data and that the data are fully qualified with valid metadata when they are loaded.

The ODM SDL is implemented as two separate executable programs. The first is the ODM SDL Configuration Wizard, which allows users to create and save the mapping of their sensor data file and all associated metadata to the ODM schema. The second executable is the ODM SDL Data Loader. It has no user interface and was designed to be run automatically as a Windows scheduled task. It reads the configuration file generated by the

Configuration Wizard, parses the streaming data file, and loads the data into the ODM database according to the settings in the configuration file. The ODM SDL Data Loader executable can be scheduled as a Windows task to run automatically on any user defined interval, or it can be run manually through the Configuration Wizard. This means that loading of sensor data with multiple reporting frequencies can be run automatically and optimized according to a user defined schedule. The ODM SDL supports table based, comma- or tab-delimited text files, where the date and time of each observation are stored in one column and the observed values are stored in subsequent columns (one column for each variable) delimited by commas.

For more details about ODM SDL, see <http://his.cuahsi.org/odmsdl.html> where you can download the software, detailed software manual and functional specifications.

3.3.2 DATA EDITING AND QUALITY CONTROL WITH ODM TOOLS

Data editing and quality control capability is presently provided through ODM Tools, an application that was created to allow administrators of local instances of the ODM to visualize, manage, manipulate, edit, and export data that have been imported to their local instance of the ODM. The development of the ODM Tools application has several advantages. First, ODM Tools protects the security and consistency of a work group HIS ODM database because it provides users with a set of automated tools for performing many of the most common database transactions. Second, ODM Tools allows users to export data from their ODM instance with an accompanying metadata file. This allows users to work with local copies of data series exported from their ODM database while preserving the provenance of the data via the metadata file. ODM Tools also provides a mechanism by which users can interact with the ODM database without having to learn the complexities of its relational structure. Finally, for more advanced users, the source code of the ODM Tools application provides an example of how applications can be built on top of the CUAHIS HIS ODM.

The main objective of the ODM Tools application is to provide managers and users of work group instances of the ODM with a set of value added tools that they can use to better manage their data. These tools are organized into three general areas: 1) query and export; 2) visualize; and 3) edit. The Query and export functionality allows users to find the data that they are interested in and export it to a simple format that can be used with a variety of analysis software. The Visualize functionality allows users to quickly plot and summarize data using a variety of plot types and descriptive statistics. The Edit capability of ODM Tools was designed to provide users with a simple set of tools that they can use to edit existing data series and to create new data series from existing data series.

QUERYING AND EXPORTING DATA SERIES

The CUAHIS HIS ODM has within it the concept of a “data series.” Each data series in the ODM represents a unique Site, Variable, Method, QualityControlLevel, and Source combination, and the SeriesCatalog table in the ODM provides a listing of all of the distinct series of data values stored in the ODM. ODM Tools provides the ability to query an instance of the ODM for specific data series based on information contained in one or more fields in the SeriesCatalog table. Once specific data series are identified, users can then export them to a delimited text file in the CUAHIS HIS MyDB format. Figure 3-5 shows the Query tab of the ODM Tools application that provides an interface for identifying and querying series within an ODM. Once a set of data series have been identified using the query options, they can be exported to a delimited text file that can easily be loaded into many data visualization and analysis software programs such as Microsoft Excel. ODM Tools also includes functionality to view and export the metadata associated with one or more selected data series. The exported metadata file contains a snapshot of all of the metadata stored in ODM for the data series that are being exported.

ODM Tools

File Edit Tools Help

Query Visualize Edit

☐ Query by Site

☐ Choose Sites from a list

NWIS:10010400 - EAST FK BEAR RIVER NR EVANSTON, WYOMING
 NWIS:10010500 - HILLIARD-E FK CANAL NR ST LINE NR EVANSTON,
 NWIS:10011200 - WEST FORK BEAR RIVER AT WHITNEY DAM, NR O.
 NWIS:10011400 - WEST FK BEAR RIVER BL DEER CR NR EVANSTON,
 NWIS:10011500 - BEAR RIVER NEAR UTAH-WYOMING STATE LINE

☐ Query by Site Name

☐ Query by Site Code

Multiple Entries (:)
☒ AND
☐ OR

☐ Query by Variable

☐ Choose Variables from a list

NWIS:00010 - Temperature, water
 NWIS:00020 - Temperature, air
 NWIS:00028 - Agency analyzing sample, code
 NWIS:00060 - Discharge
 NWIS:00060 - Discharge, daily average

☐ Query by Variable Name

☐ Query by Variable Code

Multiple Entries (:)
☒ AND
☐ OR

☐ Query by Source

☐ Organization (:)

☐ Source Description (:)

Multiple Entries (:)
☒ AND ☐ OR

Other Query Options

☐ General Category
 Hydrology
 Water Quality

☐ Value Type
 Derived Value
 Field Observation
 Sample

☐ Sample Medium
 Surface Water

☐ Data Type
 Average
 Instantaneous

☐ Quality Control Level
 0 - Raw data
 1 - Quality controlled d
 2 - Derived products
 3 - Interpreted product
 4 - Knowledge product

☐ # of Observations
☐ > 1
☐ <=

☐ Time Period
 from: 3/ 6/2007
 to: 3/ 6/2007

☐ Method (:)

Site	Variable	Variable Units	General Category	Value Type	Sample Medium	Data Type	Quality Control Level	Method Description	# of Obser

Export Checked Metadata Export Checked Data Query

Figure 3-5 ODM Tools Query tab

VISUALIZING AND SUMMARIZING DATA SERIES

ODM Tools provides users with the capability to visualize data series using a variety of plot types and to generate simple descriptive statistics for data series. The data series visualization and statistical summary tools are contained within the “Visualize” tab of the ODM Tools application (Figure 3-6). The following plot types are available in ODM tools.

- Time Series Plot
- Probability Plot
- Histogram
- Box and Whisker Plot

The plots generated by ODM Tools can be exported for use in documents, presentations, etc.

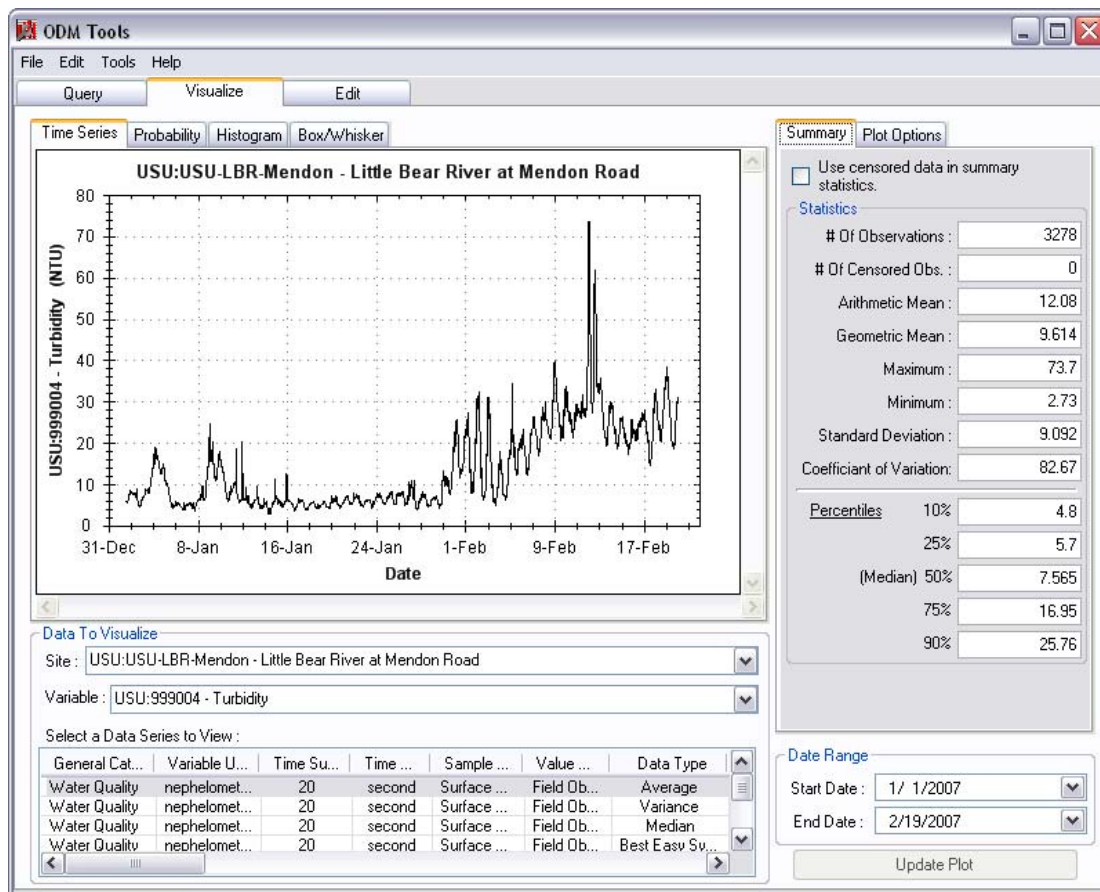


Figure 3-6 ODM Tools Visualize tab

EDITING EXISTING AND DERIVING NEW DATA SERIES

ODM Tools includes functionality to edit the data values and some of the attributes of the data values within data series stored within an instance of the ODM. This is useful, for example, in performing manual quality assurance and quality control of data series, where some data values may need to be deleted, adjusted, or interpolated. In addition, ODM Tools provides functionality to derive new data series from existing data series. For example, daily average data values can be derived from more frequent observations using ODM Tools' aggregate functions. All of the data series editing and creation tools are on the Edit tab of the ODM Tools application (Figure 3-7). ODM Tools does not allow raw data series to be edited. In order to use all of the data editing functionality of ODM Tools to perform quality assurance and quality control for a raw data series, a copy of the data series must first be created. All data editing is then performed on the copy. Within ODM, raw data series are specified with a Quality Control Level of 0, and quality controlled data series are generally specified with a Quality Control Level of 1.

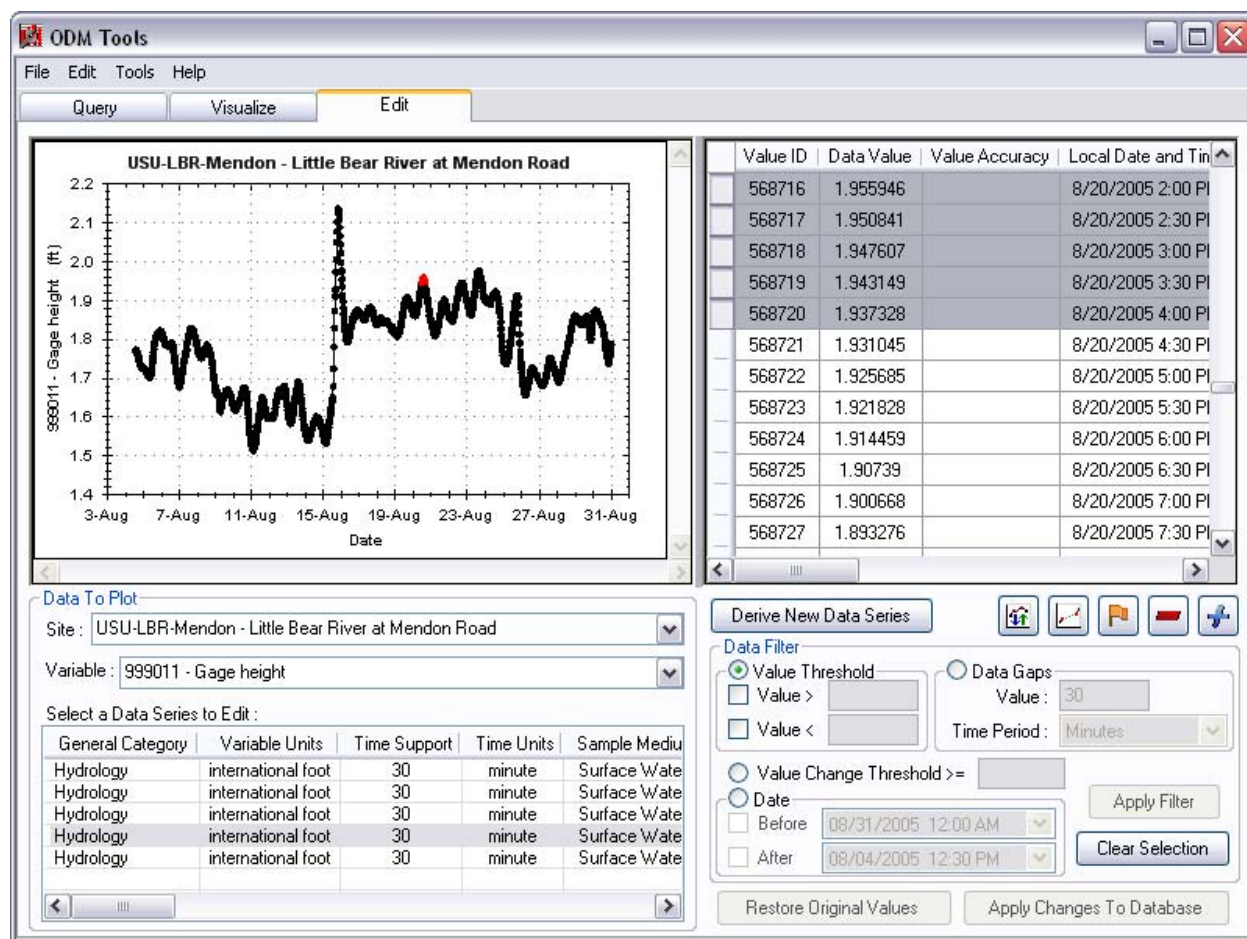


Figure 3-7 ODM Tools Edit tab

For more details about ODM Tools, see <http://his.cuahsi.org/odmtools.html> where you can download the software, detailed software manual and functional specifications.

3.4 ODM CONTROLLED VOCABULARIES

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. Multiple datasets added to an ODM database are reconciled through the use of appropriate and consistent controlled vocabulary terms to describe the data. Since the controlled vocabularies within ODM list the terms that are acceptable for use 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 the data managers to select the appropriate controlled vocabulary terms, the advantage is that the terms in the ODM controlled vocabularies are unique and devoid of ambiguity (i.e., only a single term exists in a controlled vocabulary for each concept described). Figure 3-8 provides an example of how contextual heterogeneity in attributes of datasets from multiple investigators is reconciled through the use of the ODM controlled vocabularies.

Resolving the contextual heterogeneity in datasets using the ODM controlled vocabularies ensures that datasets are consistently described within each ODM database. In addition, it assures that datasets are consistently

described across ODM databases. The controlled vocabularies form the basis of the metadata within ODM and provide specific language to describe characteristics of the data to aid in its identification, discovery, assessment, and management.

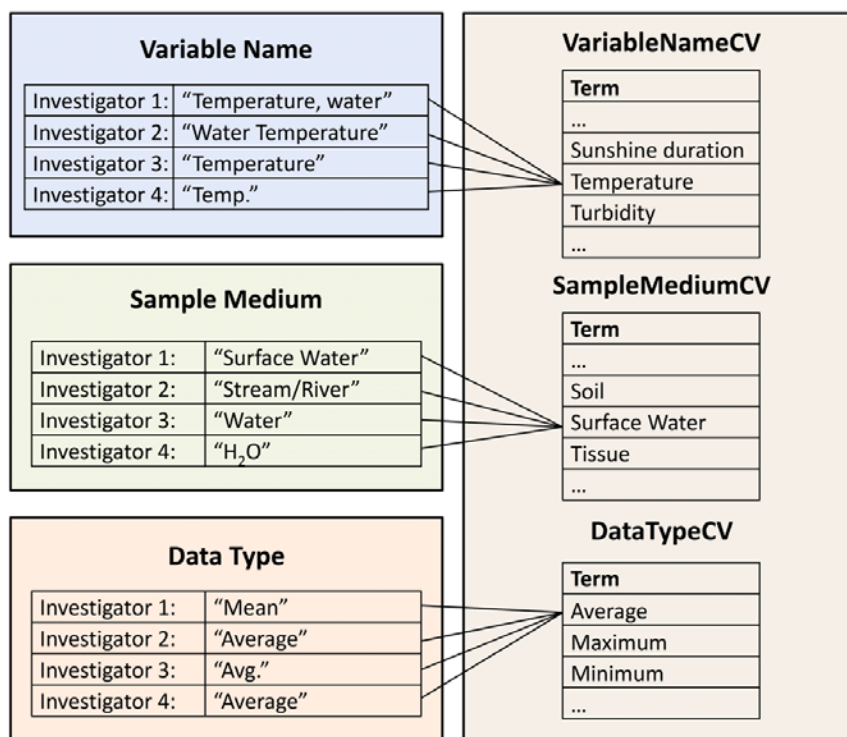


Figure 3-8 Example of how contextual heterogeneity in the attributes of similar datasets from several different investigators can be reconciled through the use of the ODM controlled vocabularies

Centralized Controlled Vocabulary System

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://water.usu.edu/cuahsi/odm/>). 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 ODM website and use an online form to request addition of an appropriate term to the master controlled vocabulary repository. The ODM controlled vocabulary submission system (Figure 3-9) is moderated to ensure that submitted terms are appropriate, unique, and unambiguous. Once a new term is accepted, 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 of this, and 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 through the ODM Tools software application and the ODM Controlled Vocabulary web services.

The ODM Controlled Vocabulary web services are implemented on top of the master controlled vocabulary repository database and broadcast the terms within the master repository in XML format. Data managers can use functionality within the ODM Tools application to compare their local controlled vocabulary with the master repository and download any updated or added terms. ODM Tools gets the controlled vocabulary terms from the local database, accesses the ODM Controlled Vocabulary web services and automatically parses the XML messages that are returned, and then presents a tabular, side-by-side comparison of local and master terms. Users can then compare the terms in their local database with those in the master list and add any new or updated terms to their local database. Figure 3-9 shows this interaction between the data manager, the ODM Tools application, and the ODM Controlled Vocabulary web services.

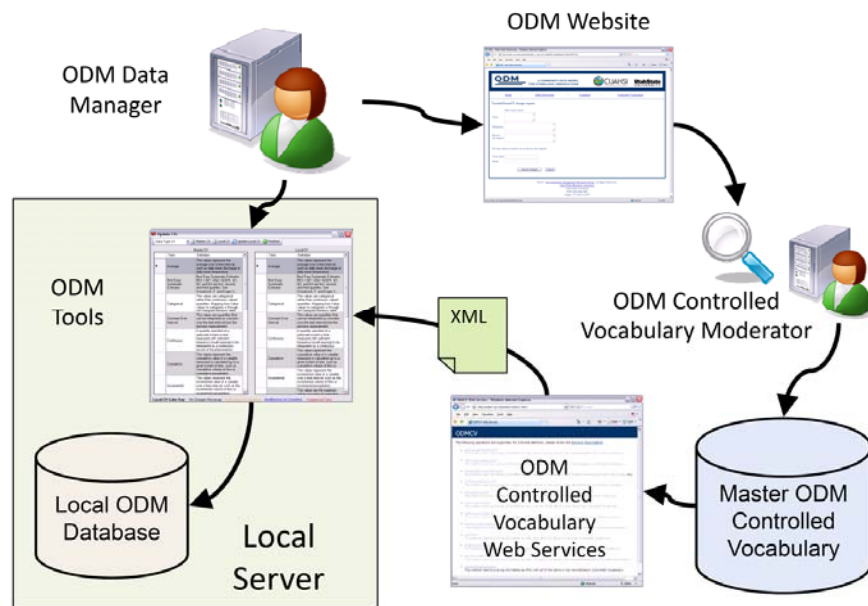


Figure 3-9 The ODM controlled vocabulary system

3.5 SUMMARY

A data model for storing and managing environmental observations has been presented. The importance of metadata in describing environmental observations data cannot be overstated. It is critical that the data be carefully documented and annotated with metadata so that it can be unambiguously interpreted and used by investigators other than those that collected the data. The co-location of observational data and their associated metadata within a single, integrated ODM database enables easy and automated access.

The reliance of ODM on relational database technology provides several advantages. First, implementation of ODM within a relational database management system enables users to take advantage of the mature technology and advanced tools available in relational database systems. These include data import and export tools, a standardized, high level query language, and tools for advanced data analysis and manipulation.

Next, ODM provides a framework in which data of different types and from disparate sources can be integrated. For example, data from multiple scientific disciplines can be assembled within a single ODM instance (e.g.,

hydrologic variables, water quality variables, climate variables, etc.). This has been the case at each site within the WATERS Network Test Beds where publishing observational data using ODM and the WaterOneFlow web services has enabled both multi-disciplinary and cross-Test Bed access to a national network of consistent data.

The number of characteristics used to describe observations can potentially be large and different across data sources. One significant advantage of ODM is that, along with the observation values, it provides a place to store a standard set of the most commonly used attributes of environmental observations. As with any other model, this representation has some limitations. However, once assembled within ODM, observations can be presented in a consistent way – negating the need for users to learn the diverse data formats of multiple scientific communities. This can be useful when data from multiple disciplines need to be combined into a single analysis or simulation model.

Last, a consistent data model enables the standardization of software application development. These software tools include the WaterOneFlow web services, data loading and editing tools, and data visualization and retrieval tools. Readers are referred to the CUAHSI HIS website for details of these software applications (<http://his.cuahsi.org>).

3.6 REFERENCES

- Bose, R., (2002), "A conceptual framework for composing and managing scientific data lineage," Proceedings of the 14th International Conference on Scientific and Statistical Database Management, Edinburgh, Scotland, July 24-26.
- Gray, J., D. T. Liu, M. Nieto-Santisteban, A. Szalay, D. J. DeWitt and G. Heber, (2005), "Scientific data management in the coming decade," ACM SIGMOD Record, 34(4): 34-41.
- Horsburgh, J. S., D. G. Tarboton and D. R. Maidment, (2005), "A Community Data Model for Hydrologic Observations, Chapter 6," in Hydrologic Information System Status Report, Version 1, Edited by D. R. Maidment, p.102-135, <http://www.cuahsi.org/his/docs/HISStatusSept15.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.
- Maidment, D. R., (2005), "A Data Model for Hydrologic Observations." Paper prepared for presentation at the CUAHSI Hydrologic Information Systems Symposium, University of Texas at Austin. March 7, 2005. <http://www.crrw.utexas.edu/cuahsi/symposium05/HODatabase/Documents/HydroObsDataModel.doc>
- Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner and S. G. Stafford, (1997), "Nongeospatial metadata for the ecological sciences," Ecological Applications, 7(1): 330-342.
- Open Geospatial Consortium Inc., (2006), "Observations and Measurements," OGC Best Practices Document, OGC 05-087r4 Version 0.14.7, Simon Cox Editor, <http://www.opengeospatial.org/standards/bp> Last accessed January 23, 2008.
- 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>.

Chapter 4. WEB SERVICES AND WATER MARKUP LANGUAGE

By Ilya Zaslavsky and David Valentine, San Diego Supercomputer Center

4.1 SERVICE-ORIENTED ARCHITECTURE

The CUAHSI Hydrologic Information System design follows an open service-oriented architecture model. A service-oriented architecture relies on a collection of loosely coupled self-contained services that communicate with each other through the internet and can be called from multiple clients in a standard fashion. Common benefits associated with a service-oriented architecture include: scalability, security, easier monitoring and auditing; standards-reliance; interoperability across a range of resources; plug-and-play interfaces. Internal service complexity is hidden from service clients, and backend processing is decoupled from client applications. In other words, different types of clients, including web browsers and such desktop applications as MATLAB, ArcGIS and Excel (which are identified as the primary desktop client environments by the CUAHSI user needs assessment), are able to access the same service functionality, leading to a more transparent and better managed system.

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, which can be remote or locally-stored in ODM instances, or in public agency databases. These services use SOAP (Simple Object Access Protocol), which is a standard set of protocols established by the World Wide Web Consortium for enabling one computer to appropriately request services of another. These services, and the markup language they implement, WaterML (Water Markup Language), are the lingua franca of CUAHSI HIS: communications between servers and clients in this service-oriented architecture follow this standard protocol.

At the physical level, the infrastructure represents a collection of HIS Servers, which support publishing hydrologic observations data. There are two types of regional and testbed HIS servers: HIS Servers that rely on commercially available components (SQL Server, ArcGIS Server) and can both serve data and display them on maps and charts, and HIS Server Lite whose software stack is composed of freely available components only (including SQL Server Express). HIS Server Lite installations can only store observations data and publish them as WaterOneFlow web services, but do not provide online mapping functionality (Data Access System for Hydrology, or DASH, which relies on ESRI's ArcGIS Server.) HIS servers are installed at various universities and public agencies, and serve local observations data. Besides the databases, web service templates and the customizable online mapping application (DASH), the servers include a software suite for publishing and curating observations data. A Central HIS Server is maintained at SDSC, which provides access to the central metadata catalog that integrates metadata across multiple observation networks, and includes a system for registering additional observation networks. Once observations data are published as web services, they can be discovered and retrieved from multiple online (DASH, Hydroseek, Google Earth – based) and desktop (ArcGIS, MapWindow, Excel, MATLAB, various programming languages) clients. A high-level view of this organization is shown in Figure 4-1.

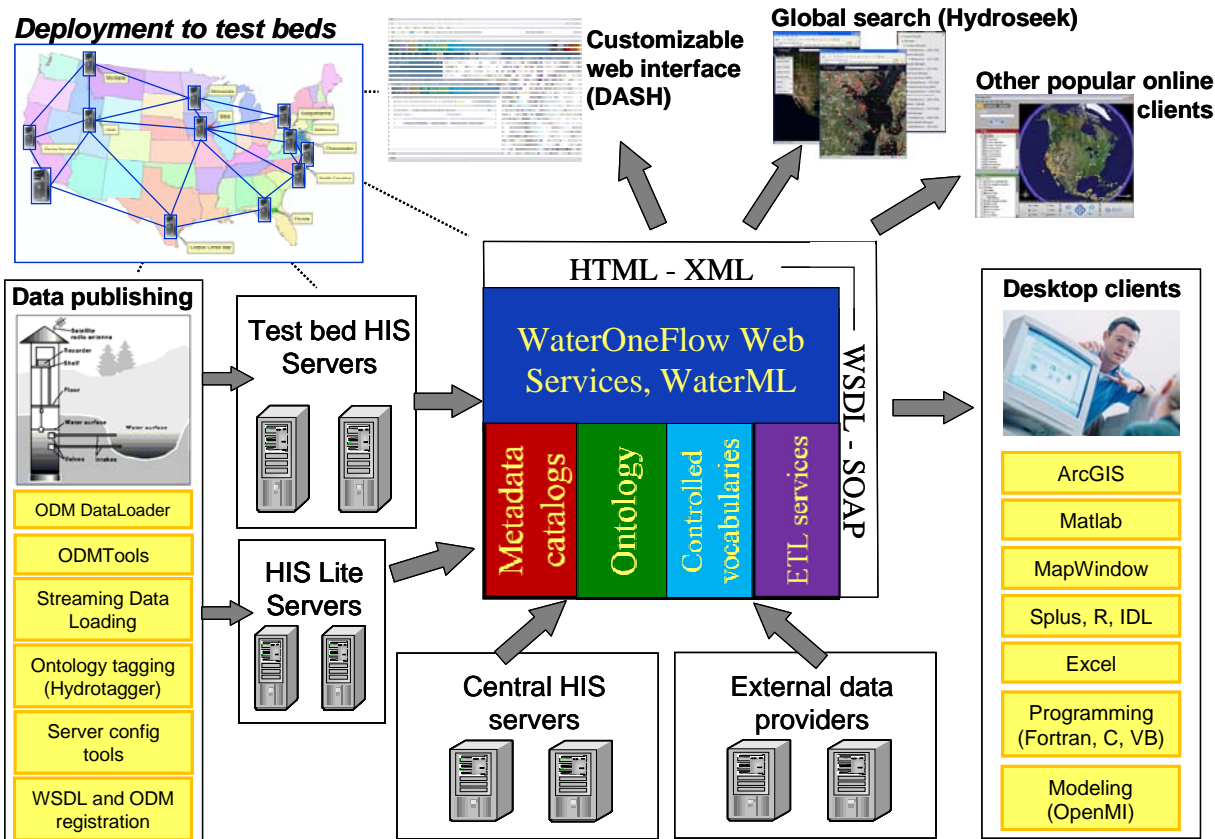


Figure 4-1 CUAHSI HIS services-oriented architecture

4.2 HOW WEB SERVICES WORK

CUAHSI HIS web services are called Water Data Services, and in version 1 are also referred to as WaterOneFlow services. On standard user requests, they return hydrologic metadata and data in a standard format called WaterML. In version 1.0, WaterOneFlow services include the following four methods:

- **GetSites:** returns a list of measurement sites in a particular observation network;
- **GetSiteInfo:** 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:** returns metadata describing each variable such as its units of measurement; A **GetVariableInfo** method call with no parameters returns the list of all variables measured on this network (only some variables may be measured at any given site);
- **GetValues:** 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 methods are descriptive or metadata methods that return information about the sites and variables. The final method, GetValues, is the one that actually provides the observations data.

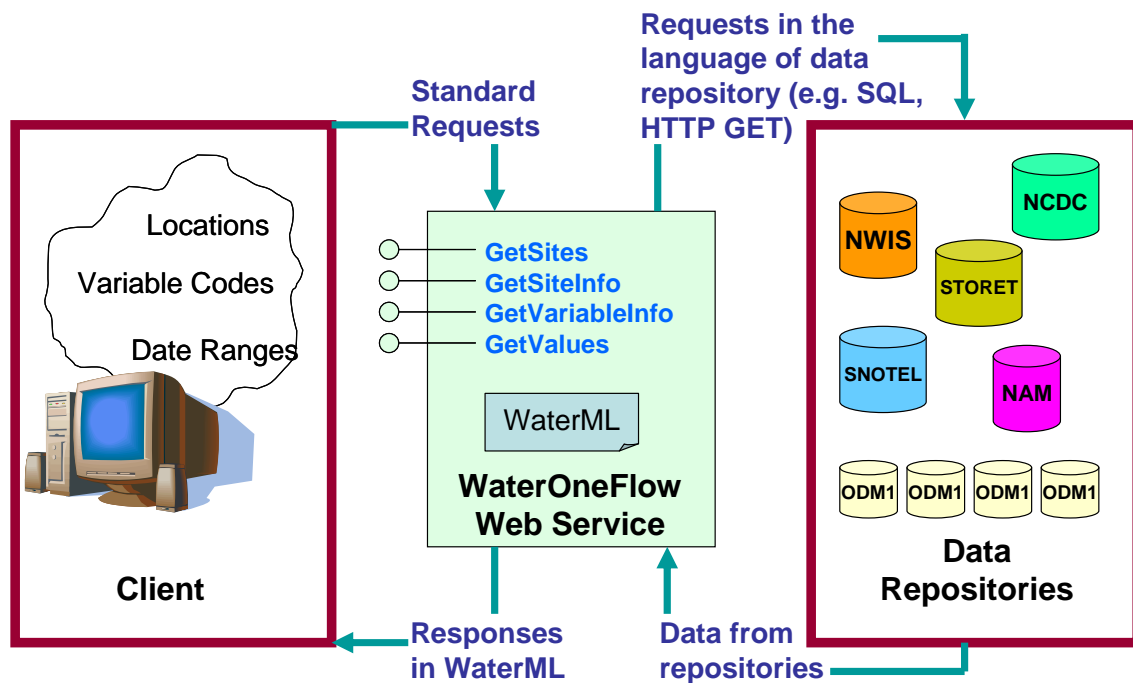


Figure 4-2 Data transformations that take place inside CUAHSI HIS web services

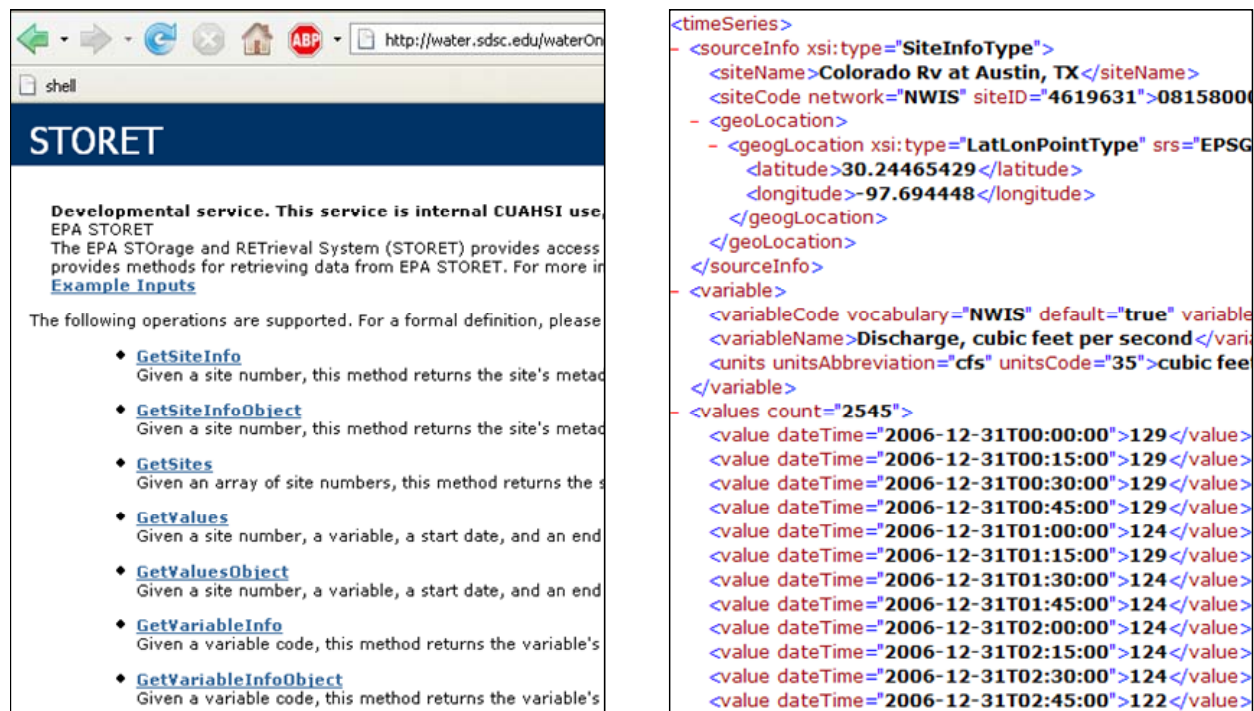
In practice, we have discovered that WaterOneFlow services methods need to be duplicated, and return both a standard XML schema-based response, and a serialized XML schema response. The standard XML schema-based responses are better suited for advanced client application environments than methods returning strings. Each of the methods returning XML schema objects is postfixed with “object” (GetSiteInfoObject, GetVariableInfoObject, GetValuesObject) or “XML” (GetSitesXML). The definition of an XML schema also makes it possible to represent WaterML as a REST-based service, and enables storing a result on disk for later use.

When a client makes a data or metadata request using a CUAHSI HIS web service, these requests are made in standard manner, following the CUAHSI HIS web service signatures – regardless of how the underlying data source may be organized. Also, regardless of the format in which the data are returned by the source, the web services respond to requests by returning the data in a standard format of WaterML. As long as the underlying data are mapped to the point observations information model – for example, loaded into an instance of the Observations Data Model – it doesn’t matter for the client application that these data repositories reside at different physical locations, have different sets of tables, different access mechanisms, etc. Because of this, CUAHSI HIS can expand to accommodate a large number of observations data sources without dramatic changes percolating through the entire system. CUAHSI HIS web services can also be custom programmed to operate from different database structures than the CUAHSI ODM. It is also possible to operate the four metadata methods using information stored in a CUAHSI ODM and then access the actual data using the GetValues method directly from a public agency database using a “web page scraper” or a custom programmed function that mimics the action of a human user of that agency’s web pages. In this way, specially published and existing water data can be harmonized and delivered through the internet using WaterML.

The uniformity and scalability is accomplished in the web services through two sets of transformations. First, the standard requests (GetSites, GetSiteInfo, GetValues, etc.) are transformed into specific requests against the data repositories. For ODM sources, they are transformed into SQL statements that are executed against ODM-compliant databases. For web-accessible repositories (such as USGS NWIS, EPA STORET), the requests are transformed into HTTP GET or HTTP POST requests, or even other SOAP or REST calls if the repository provides a web service interface (REST, or Representational State Transfer, and SOAP, Simple Object Access Protocol, are two common web service standards). The second transformation happens in web services when the underlying source returns data or metadata. This information is transformed into a common XML format, called Water Markup Language, or WaterML. The data transformations that take place within the web services as they access various data sites are shown in Figure 4-2.

4.3 WHAT IS WATERML?

CUAHSI WaterML is a standard output schema for CUAHSI HIS WaterOneFlow web services. Its formal specification is available as an OGC discussion paper at www.opengeospatial.org/standards/dp/. The goal of WaterML design has been to capture semantics of hydrologic observations discovery and retrieval and express the point observations information model as an XML schema. To a large extent, it follows the representation of the information model as adopted by the ODM relational design. Another driver of WaterML design is specifications and metadata adopted by USGS NWIS, EPA STORET, and other federal agencies, as it seeks to provide a common foundation for exchanging both agency data and data collected in multiple academic projects. Another WaterML design principle was to create, in version 1 of HIS in particular, a fairly rigid and simple XML schema which is easy to generate and parse, thus creating the least barrier for adoption by hydrologists. An example of WaterOneFlow web services, and the output of a GetValues request in WaterML format, are shown in Figure 4-3.



The figure consists of two side-by-side panels. The left panel is a screenshot of the EPA STORET web service interface. The browser address bar shows 'http://water.sdsc.edu/waterOneFlow/'. The page title is 'STORET'. Below the title, it says 'Developmental service. This service is internal CUAHSI use.' and 'The EPA STORage and RETrieval System (STORET) provides access to EPA STORET data. For more information, see the Example Inputs.' Below this, it lists several operations supported: GetSiteInfo, GetSiteInfoObject, GetSites, GetValues, GetValuesObject, GetVariableInfo, and GetVariableInfoObject. Each operation is followed by a brief description of its parameters and return value.

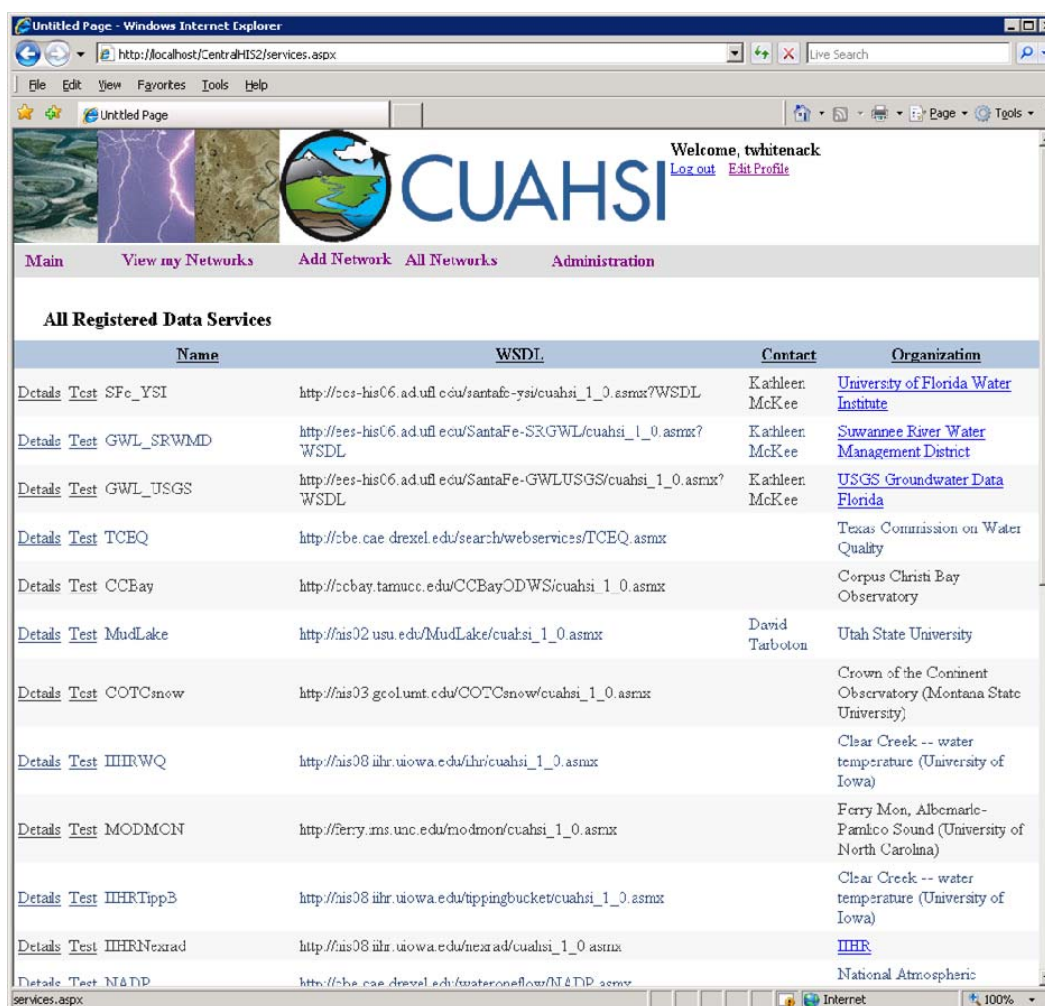
The right panel shows a sample WaterML output for a GetValues request. The XML is as follows:

```
<timeSeries>
  <sourceInfo xsi:type="SiteInfoType">
    <siteName>Colorado Rv at Austin, TX</siteName>
    <siteCode network="NWIS" siteID="4619631">0815800</siteCode>
  </sourceInfo>
  <geoLocation>
    <geogLocation xsi:type="LatLonPointType" srs="EPSG:4326">
      <latitude>30.24465429</latitude>
      <longitude>-97.694448</longitude>
    </geogLocation>
  </geoLocation>
  <variable>
    <variableCode vocabulary="NWIS" default="true" variableName="Discharge, cubic feet per second" units="cfs" unitsCode="35">cubic feet</variableCode>
  </variable>
  <values count="2545">
    <value dateTime="2006-12-31T00:00:00">129</value>
    <value dateTime="2006-12-31T00:15:00">129</value>
    <value dateTime="2006-12-31T00:30:00">129</value>
    <value dateTime="2006-12-31T00:45:00">129</value>
    <value dateTime="2006-12-31T01:00:00">124</value>
    <value dateTime="2006-12-31T01:15:00">129</value>
    <value dateTime="2006-12-31T01:30:00">124</value>
    <value dateTime="2006-12-31T01:45:00">124</value>
    <value dateTime="2006-12-31T02:00:00">124</value>
    <value dateTime="2006-12-31T02:15:00">124</value>
    <value dateTime="2006-12-31T02:30:00">124</value>
    <value dateTime="2006-12-31T02:45:00">122</value>
  </values>
</timeSeries>
```

Figure 4-3 List of standard web service methods in CUAHSI HIS service, and a sample WaterML output

4.4 CURRENT STATUS OF CUAHSI HIS WEB SERVICES

As of early June 2008, there are more than 50 observation networks from around the country that can be accessed via CUAHSI HIS WaterOneFlow services, with a total of over 1.7 million observation stations. These networks are listed in the HIS Central online application (Figure 4-4 – also see Chapter 6 of this document). The bulk of hydrologic observation values come from federal agencies: USGS NWIS, EPA STORET, USDA SNOTEL, NCDC (Table 4-1). In addition, several state agencies - in Texas, Florida, Idaho - collaborate with CUAHSI HIS universities to jointly serve observations data in WaterML. There are also many academic projects, primarily those supported by NSF through the WATERS Hydrologic Observatory Testbed and CEO:P (Cyberinfrastructure for Environmental Observatories: Prototype) initiatives, that are either publishing their observations data using CUAHSI HIS Server, or otherwise adopting ODM and WaterML specifications. WaterML-compliant services are being developed by KISTERS, an international water resources data management company based in Germany. Various online and desktop clients of WaterML services have been developed by MapWindow open source GIS team, by ESRI, and by the Australian Water Resource Observation Network (WRON) project.



The screenshot shows a web browser window displaying the CUAHSI HIS Central online application. The page has a header with the CUAHSI logo and a navigation bar with links: Main, View my Networks, Add Network, All Networks, and Administration. Below the navigation bar is a table titled "All Registered Data Services". The table has four columns: Name, WSDL, Contact, and Organization. The table lists 12 data services, each with a "Details" and "Test" link. The services include SFe_YSI, GWL_SRWMD, GWL_USGS, TCEQ, CCEBay, MudLake, COTCsnow, IIHRWQ, MODMCN, IIHRTippB, IIHRNexrad, and NADTD.

Name	WSDL	Contact	Organization
Details Test SFe_YSI	http://ces-his06.ad.ufl.edu/santaFe-ysi/cuahsi_1_0.asmx?WSDL	Kathleen McKee	University of Florida Water Institute
Details Test GWL_SRWMD	http://ces-his06.ad.ufl.edu/SantaFe-SRGWL/cuahsi_1_0.asmx?WSDL	Kathleen McKee	Suwannee River Water Management District
Details Test GWL_USGS	http://ces-his06.ad.ufl.edu/SantaFe-GWLUSGS/cuahsi_1_0.asmx?WSDL	Kathleen McKee	USGS Groundwater Data Florida
Details Test TCEQ	http://tce.cae.drexel.edu/search/webservices/TCEQ.asmx		Texas Commission on Water Quality
Details Test CCEBay	http://ccbay.tamucc.edu/CCBayODWS/cuahsi_1_0.asmx		Corpus Christi Bay Observatory
Details Test MudLake	http://his02.usu.edu/MudLake/cuahsi_1_0.asmx	David Tarboton	Utah State University
Details Test COTCsnow	http://his03.geol.umt.edu/COTCsnow/cuahsi_1_0.asmx		Crown of the Continent Observatory (Montana State University)
Details Test IIHRWQ	http://his08.ihr.uiowa.edu/ihr/cuahsi_1_0.asmx		Clear Creek -- water temperature (University of Iowa)
Details Test MODMCN	http://ferry.ms.unc.edu/modmon/cuahsi_1_0.asmx		Ferry Mon, Albemarle-Pamlico Sound (University of North Carolina)
Details Test IIHRTippB	http://his08.ihr.uiowa.edu/tippingbucket/cuahsi_1_0.asmx		Clear Creek -- water temperature (University of Iowa)
Details Test IIHRNexrad	http://his08.ihr.uiowa.edu/nexrad/cuahsi_1_0.asmx		IIHR
Details Test NADTD	http://tce.cae.drexel.edu/wateroneflow/NADTD.asmx		National Atmospheric

Figure 4-4 The Central HIS catalog of registered data services

Even more importantly, the WaterML specification is being adopted by federal agencies. The experimental USGS NWIS Daily Values web service (<http://waterservices.usgs.gov/NWISQuery/GetWSService?wsdl>) returns WaterML-compliant TimeSeriesResponse. The National Climatic Data Center is also prototyping WaterML for data delivery, and has developed a REST-based service that generates WaterML-compliant output for the NCDC ASOS network. Such agency-supported web services coming online provide a much more efficient way to deliver agency data compared to the web site scraper services that the CUAHSI HIS project has developed initially.

Table 4-1 Federal and other national and international data sources

Organization	Dataset	Description
US Geological Survey (USGS)	National Water Information System (NWIS)	The USGS National Water Information System (NWIS) provides access to more than one million sites measuring streamflow, groundwater levels, and water quality. This web service provides methods for retrieving daily values data, such as discharge and water levels, from NWIS. For more information about NWIS, see the NWIS home page at http://waterdata.usgs.gov/nwis .
Environmental Protection Agency (EPA)	STORET (STORage and RETrieval)	STORET is a repository for water quality, biological, and physical data and is used by state environmental agencies, EPA and other federal agencies, universities, private citizens, and many others. For information see: http://www.epa.gov/storet/ .
NOAA National Climate Data Center	Automated Surface Observing System	The ASOS program is a joint effort of the National Weather Service (NWS), the Federal Aviation Administration (FAA), and the Department of Defense (DOD). The ASOS system serves as the nation's primary surface weather observing network. ASOS is designed to support weather forecast activities and aviation operations and, at the same time, support the needs of the meteorological, hydrological, and climatological research communities. For more information see: http://www.nws.noaa.gov/asos/ .

Organization	Dataset	Description
NASA	Moderate Resolution Imaging Spectroradiometer (MODIS)	MODIS is a key instrument aboard the Terra (EOS AM) and Aqua (EOS PM) satellites. Terra's orbit is timed so that it passes from north to south across the equator in the morning, while Aqua passes south to north over the equator in the afternoon. Terra MODIS and Aqua MODIS are viewing the entire Earth's surface every 1 to 2 days, acquiring data in 36 spectral bands, or groups of wavelengths (see MODIS Technical Specifications). More information can be obtained at: http://modis.gsfc.nasa.gov/about/ .
U.S. Department of Agriculture (USDA) Natural Resources Conservation Service(NRCS) National Water and Climate Center	SNOWpack TElemetry (SNOTEL)	Snowpack and related climatic data in the Western United States
University Corporation for Atmospheric Research (UCAR)	NCEP North American Mesoscale (NAM) Weather Research and Forecasting (WRF) mode	Data from the NCEP North American Mesoscale (NAM) Weather Research and Forecasting (WRF) model. More information can be obtained at: http://www.meted.ucar.edu/ .
National Center for Atmospheric Research (NCAR)	Daily Meteorological and Climatological Summaries (DAYMET)	Daymet is a 1km grid of daily temperature, precipitation, humidity, wind speed and radiation interpolated from gage data for the continental US

Table 4-2 Academic data sources catalogued in HIS Central

Organization	Dataset	Description
Chesapeake Bay Information Management System		Chesapeake Bay physical and chemical observations.
Corpus Christi Bay Observatory	Hypoxia in Corpus Christi Bay	Corpus Christi Bay physical and chemical observations.
Idaho Waters	Reynolds Creek Experimental Watershed	Historical monitoring of the watershed has included climate, precipitation, snow accumulation and redistribution, snowmelt, frozen soils and frost depth, soil water and temperature, streamflow and sediment yield, and vegetation. The watershed is instrumented with three meteorological stations and seven stream gaging stations. Multiple sub-basin sites are instrumented for ongoing investigations into geochemistry, groundwater recharge, infiltration, basin precipitation processing, soil water distribution, streamflow generation, and runoff over multiple scales.
	Dry Creek Experimental Watershed	
Montana State University	Crown of the Continent Observatory	Meteorological, snow accumulation and stream gauge height.
National Atmospheric Deposition Program	The National Atmospheric Deposition Program/National Trends Network (NADP/NTN).	The precipitation at each station is collected weekly according to strict clean-handling procedures. It is then sent to the Central Analytical Laboratory where it is analyzed for hydrogen (acidity as pH), sulfate, nitrate, ammonium, chloride, and base cations (such as calcium, magnesium, potassium and sodium).
San Diego River Park Foundation		Meteorology and hydrology observations
Susquehanna River Basin Hydrologic Observatory System		Meteorology, air and hydrology observations

Organization	Dataset	Description
Suwannee River Water Management District	Ground Water Levels	Groundwater level (GWL) records are available for 1,089 wells. The majority of these wells are only measured during record high or record low periods and 396 wells are inactive. One hundred and eighty-one (181) wells are measured monthly in the GWL network; of these, 77 have continuous recorders. Groundwater levels are stored in the District's GWL database. In feet above NGVD29
Texas Commission for Environmental Quality	TRACS	TRACS (TCEQ Regulatory Activities and Compliance System) water quality data
Texas Instream Flow Program	Lower Sabine	Biological fish species and aquatic environment data from the lower Sabine River, Texas
Texas Parks and Wildlife Department	Coastal water surveys	Water Level, turbidity, salinity, temperature, and dissolved oxygen levels.
University of Florida Water Institute	Waters Testbed Project	Data from nitrate sensor: YSI 9600 Nitrate Monitor allows a user to continuously record nitrate levels at a variable sample interval and providing improved detection of nitrate dynamics vis-a-vis grab sampling.
University of Iowa	Clear Creek	Precipitation
University of North Carolina	Ferry Mon, Albemarle-Pamlico Sound	Water quality observations
Utah State University	Little Bear River WATERS Test Bed	Continuous water quality monitoring of the Little Bear River to investigate the use of surrogate measures such as turbidity in creating high frequency load estimates for constituents that cannot be measured continuously.
	Mud Lake	Continuous water quality monitoring to investigate the sediment and nutrient budget of Mud Lake within the Bear Lake National Wildlife Refuge, Idaho.

Table 4-3 Statistics by organization of data available from web services cataloged in HIS Central. Information is available for 342 million data values distributed over 1.75 million sites.

Organization	Site Count	Data Value Count
USGS	1449881	281287171
EPA	273069	41626359
TCEQ	8407	7591675
CIMS	894	5445889
SRBHOS	5	2091755
SEV	41	1642555
IIHRNexrad	142	1196776
LittleBear	10	993018
MudLake	5	246476
NADP	19	213801
IIHRTippB	3	135524
CCBay	47	85880
SantaFeGWL	21	59564
BaltOD	5	34238
COTCsnow	1	18867
SantaFeISUS	1	6888
SantaFeYSI	1	3096
SantaFeFLstoret	17	1024
NCDC	22405	Unknown
TOTAL	1,754,974	342,680,556

4.5 WEB SERVICES FOR NATIONAL DATA REPOSITORIES VERSUS ODM WEB SERVICES

Two types of web services are being developed in CUAHSI HIS: services that provide access to independently maintained and updated online data repositories (which is typical for large national and state data repositories), and services that provide access to ODM instances. While both types of services return WaterML-compliant documents, internally they have different organization, and have different objectives.

The development of the CUAHSI web services has evolved over the last several years, and we can trace several phases corresponding to the evolution in methods of data access. Originally, the CUAHSI web services accessed the online remote data repositories, just like a normal user did. Typically, the user discovers and retrieves observations data from a remote repository, such as NWISWeb or EPA's STORET, by going through a sequence of online forms that help her narrow down the request: for example, selecting a state and county, retrieving a list of stations for the county, selecting a station, retrieving a list of parameters measured at the station, formulating a data retrieval request against one or several parameters, and retrieving an HTML page with the results table or a chart. In the absence of other entry points into the remote repository, to provide programmatic access to the repository a software application is written, which simulates the actions of the web user. The software, usually

called a web site wrapper, supplies query parameters to the repository, retrieves the query result as an HTML page, and converts the data content of the HTML page into a form that computers can understand (this procedure is sometimes referred to as web page scraping, or screen scraping). In case of CUAHSI data services, the software has been setup as a web service, and the retrieved HTML page with results have been converted into a WaterML-compliant XML document. Within this model, a change in the user workflow or parameter list, as well as changes in web page layout, often broke access to the data, and thus required constant monitoring to ensure that the service is operational and reliable.

While the initial services standardized the access to data resource of the large federal data sources, they still required users to know what information (e.g. specific site and variable codes) they needed to formulate valid data requests. Because for such metadata information the services had to query the remote repositories, they lacked fast discovery of listing of all the information in a repository. In order to improve discovery, and speed access, it was determined that a metadata catalog database was needed. Metadata catalogs for remote repositories were originally constructed based on a list of site codes, using the `GetSiteInfo` and `GetVariableInfo` methods of the original web services, and included information about sites, variables, periods of record, and other repository metadata expected by the point observations information model. Presently, the CUAHSI HIS team works in close collaboration with many agency data providers, and often receives database updates directly from the source, eliminating the need for catalog harvesting. Once a local metadata catalog is established the web services can be recoded to take advantage of faster metadata browsing, retrieval, and data summarization : if the information is available locally there is no need to connect to the remote repository except for retrieving the actual data via the `GetValues` method.

Recently, some large remote repositories of observations data started supporting other ways to access their content. For example, EPA started to provide access to the EPA STORET data warehouse via WQX services. In such a case, a CUAHSI data service would call the agency web services internally, and translate their output into WaterML documents, for use by CUAHSI clients. Now, as federal or state agencies turn to publishing their own web services, we turn to this second approach, because it is usually more efficient and reliable than screen scraping. It is especially effective if the service output complies with the WaterML standard: then wrapping such a service is either not needed, or trivial.

While following the same service signature and output schema (WaterML) conventions, ODM web services differ from the above model in several important respects. These services work in conjunction with local instances of the Observations Data Model, as part of HIS Server and HIS Server Lite software stacks. ODM web services connect to an ODM instance, translating standard CUAHSI HIS web service calls (`GetSites`, `GetSiteInfo`, `GetVariableInfo`, `GetValues`, and their variants that return objects) into SQL `SELECT` statements against respective ODM tables, and converting the results into WaterML documents. They are used mostly for academic datasets, and for datasets that are relatively small (compared to large federal or state data sources). They are typically used to serve “frozen” data, or data for which ODM updates are managed by CUAHSI-developed tools (e.g. ODM Streaming Data Loader that streams observations values from LoggerNet directly into ODM’s `DataValues` table). This is in contrast to large federal and state repositories that typically serve observations data values after elaborate internal QA/QC effort and often use proprietary database schemas which are not ODM-compliant and typically not accessible from outside the organization.

Information about setting up CUAHSI ODM web services is provided in the CUAHSI HIS Server manuals. The procedure, however, is very simple. There is a standard ODM web service application template supplied as part of the HIS Server. After your data are loaded into an ODM instance, with the help of ODM data loading mechanisms,

you copy the ODM web service template into a new folder, and edit the web.config file for the web service application. You will need to change several lines in this file, specifying the name of your ODM instance, the connection string, and several additional metadata fields. When you are done, you have an HIS Server that serves observations data in WaterML compliant form and can be included in the DASH application, registered at HIS Central (Chapter 6) or called from any client application that can connect with CUAHSI data services. Please refer to <http://hiscentral.cuahsi.org> for the current list of data services, both for national and state repositories and for ODM instances, which are registered at the HIS Central.

4.6 THE EVOLUTION OF WATER DATA WEB SERVICES

The CUAHSI water data web services will continue to serve as the main communication mechanism within CUAHSI HIS, connecting a variety of data sources with a growing set of web service clients being developed in both academia and the commercial sector. The driving forces for the development of web services continue to be:

- Application experience and needs of the growing number of CUAHSI HIS users, who experiment with additional data types, analysis modes, data browsing and searching strategies, and provide feedback to WaterML developers;
- Evolution of the point observations information model, which is also expressed in the ODM relational schema;
- Data description requirements posed by various federal and state agencies;
- Harmonization with standards being adopted or developed in neighboring communities, in particular the relevant standards being explored within the Open Geospatial Consortium.

WaterML version 1.1 described in this Chapter enhances the original WaterML 1.0 schema in several respects. It adds elements to fully expose all ODM 1.1 information, removes enumerations that constrained the dynamic controlled vocabulary management, and allows value arrays inside the value elements. To maintain a working CUAHSI HIS, WaterML 1.0 and WaterML 1.1 services will co-exist for the time needed for transitioning to the new version, while newly developed clients will be encouraged to follow the 1.1 specification. A similar transition period, where several versions of the web services will co-exist, is expected when WaterML 2.0 (compliant with the GML Simple Features Schema) becomes available.

Chapter 5. SEMANTIC MEDIATION – LINKING DATA WITH CONCEPTS

By Michael Piasecki, Drexel University

5.1 THE NEED FOR A GLOBAL DATA SEARCH ENVIRONMENT

When the CUAHSI HIS team conducted an initial survey to learn more about what the community would want from an IT effort of this scale, among the first responses was the statement “...give me better access to data ...”. This is matched by statements of many who have lamented the fact that much time is consumed in finding, retrieving and then reformatting data before it can actually be used “to do the real science”. Hence, it is one of the main objectives of the HIS team to build a one-stop data search environment in which it would not be necessary to know all the details about the way data providers describe their data and a specific search could be spawned simultaneously across many data repositories.

Large nationwide mission agencies like EPA (STORET) or USGS (NWIS), and regional water systems like the Chesapeake Information Management System (CIMS), have their own data access portals which allow querying their data holdings in order to search for and retrieve desired data sets. If you were to routinely use a dozen or so data providers then this requires you to learn how to operate and use a dozen different access mechanisms that return data in a dozen different formats with a dozen different descriptions requiring a dozen different ways to reformat and a dozen different ways to read and understand what these data were all about. In other words, heterogeneity of water data is ubiquitous.

Water data heterogeneity manifests itself in two ways – *syntactic heterogeneity*, or variation in data *format*, and *semantic heterogeneity*, or variation in data *description*. Syntactic heterogeneity has been addressed by the HIS team through the creation of WaterML and WaterOneFlow web services. Semantic mediation among disparate data sources is the topic of this chapter. There are three types of semantic heterogeneity: synonymy, hyponymy, and polysemy. In linguistics, a *synonym* refers to the situation where two words have identical or at least similar meanings. Consider, as an example, one agency calling its surface water level measurements “gauge height”, while a second calls this data “stage height”, a third refers to them as just “stage”, and fourth uses the term “water level”. While there may be subtle differences among these four agencies (in vertical reference datum for example) the four data providers are using different names to refer to the same water data concept. A *hyponym* is a word or phrase whose semantic range is included within that of another word, for example “Groundwater Level”, “Stage Height”, and “Reservoir Level” are all hyponyms of “Water Level”. A *polyseme* is a word that can have multiple meanings, such as “stage”.

The above issues suggest what needs to be done: if you were to search for water level data in a certain geographic region, you would want to find all data that is available. In addition, if you want to use the keyword “gauge height” as a search criterion then the semantic mediator should be able to find all water level related data regardless of what it is called within a certain data repository.

5.2 THE SEARCH APPROACH

When developing a search approach, additional aspects need to be taken into consideration that go beyond semantic mediation. These concern the user friendliness of the search engine, graphical display support, simplicity in using the interface, but also issues of search response volume and the management of the returned results (Beran and Piasecki, 2008). There is also the need to integrate it with the other CyberInfrastructure appliances developed by the HIS team, such as the use of the Observations Data Model (ODM) database implementations that are being deployed at the test bed sites as well as by a growing number of other users. We won't address all them here but some warrant closer inspection.

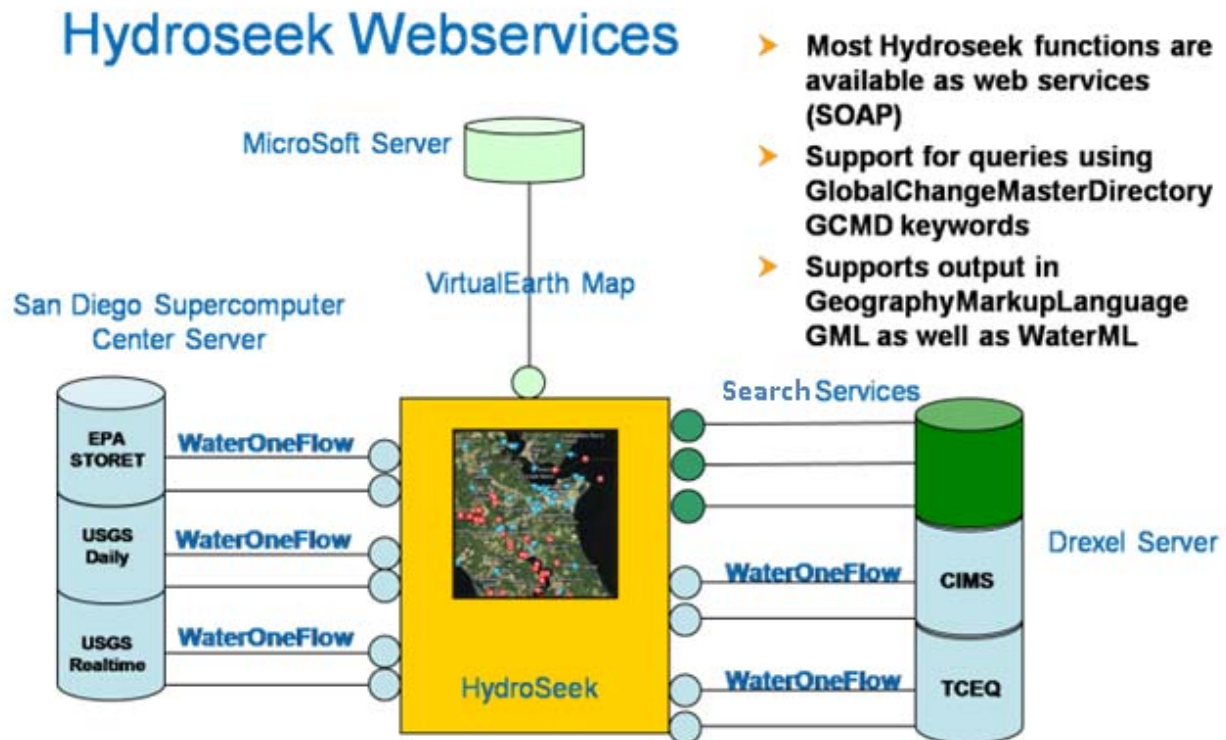


Figure 5-1 Basic Schema of the Web-services that the search engine "HydroSeek" uses and where they are located

One of the key issues that needs addressing is the "high precision -> low recall" and conversely "low precision -> high recall" problem. This is a problem that many experience when using Google and either getting no return (the keywords suggested were too specific or high precision) or getting inundated with hundreds of thousands of returns (the keyword was too generic or low precision) neither of which is very helpful. Hence, when using a specific keyword selection, the return of results should be manageable; not too many and not nothing. To this extent we have developed a hierarchical concept or keyword ensemble, called an *ontology* that permits the use of a range of search keywords in terms of generality versus specificity.

We have also tried to make use of already existing CyberInfrastructure components that would permit re-use thus providing savings in development time, and follow the design guidelines of the CUAHSI HIS appliances and the technologies deployed there, such as the use of web-services. Figure 5-1 presents a simplified schematic of how the HydroSeek search engine is constructed and what technologies and services it makes use of that are actually

provided elsewhere. The graphical user interface is Microsoft's Virtual Earth which can be invoked from a Microsoft web server. Besides the standard set of WaterOneFlow web-services, there are some additional web services (here called "search services") that support and execute the search functions. In other words, once new data sources get registered with HIS Central, the search engine will automatically add those data to its search range.

5.3 THE SEARCH ONTOLOGY

The concept system consists of an ontology that contains various layers that move from general concepts at the top to more granular or finer concepts at the bottom, like a knowledge tree with a trunk and branches. The concept tree branches end in *leaf concepts* that are quite specific, and to which the variable names in the data sources are mapped. Figure 5-2 shows an example in which the specific forms of nitrogen are at the bottom and they are progressively assembled into higher level concept classes.

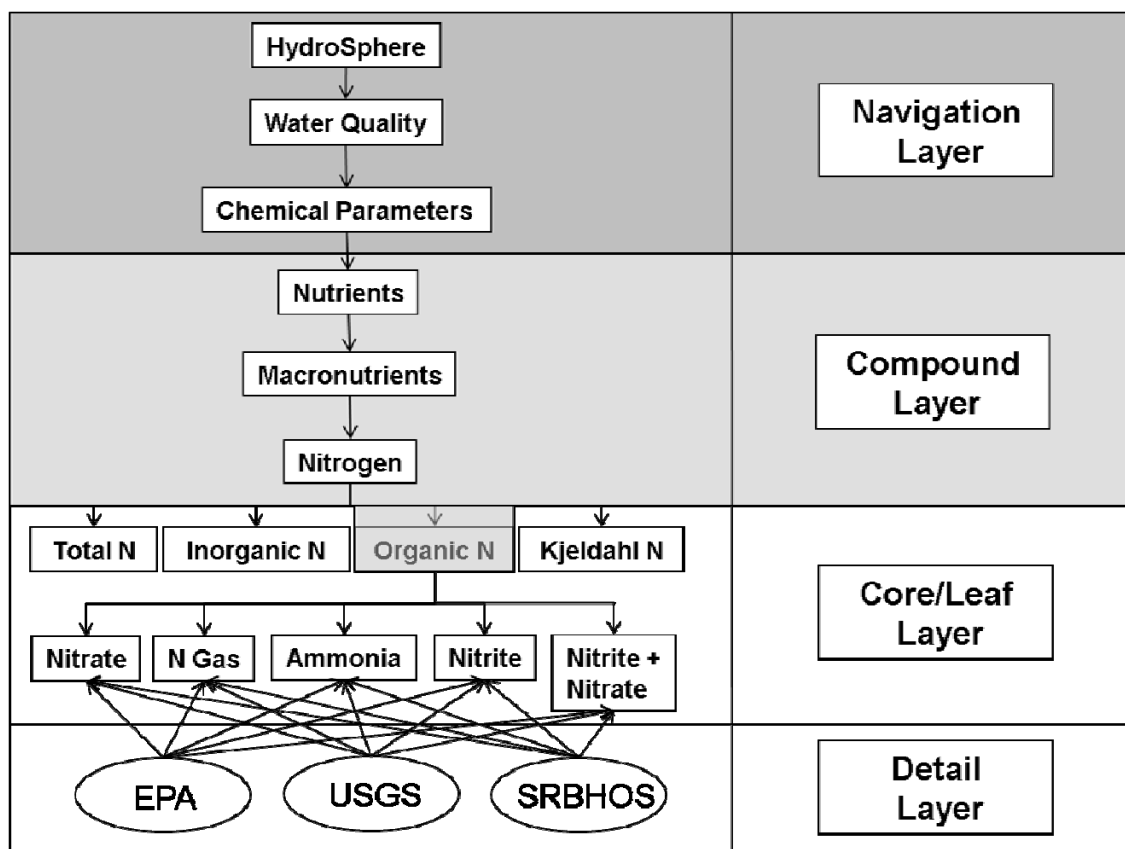


Figure 5-2 Ontology layer structure shown with the example of nitrogen

The term Nitrogen or Ammonia for instance is hardly ever used by any of the data sources to name a measured parameter because it often depends where it is measured, whether it is filtered or not and so on. Hence, parameter names either tend to be very specific or to have synonymous names like "NitConc" instead of "Nitrite Concentration", or "NO3ConcBot" for nitrate concentrations measured at the bottom of a water body. Consequently, "Nitrogen or Ammonia measurements" are common and well understood terms that lend themselves to be used as search keywords rather than as specific nitrogen variable names. The all important

search link is thus established by creating a look up table that pairs a source specific variable, like USGS NWIS:00060 (Discharge, cfs) with the corresponding concept in the ontology: Discharge Stream. The concept can then be used as a keyword for the search. It should be mentioned that one could also use synonyms like “Streamflow” or “Riverflow” to carry out the search as these synonyms are also stored.

We used NASA’s Global Change Master Directory as a start point for our keyword ontology. Our ontology currently holds about 250 concept leaves to which approximately 750 parameter names are mapped, as shown in Figure 5-3.

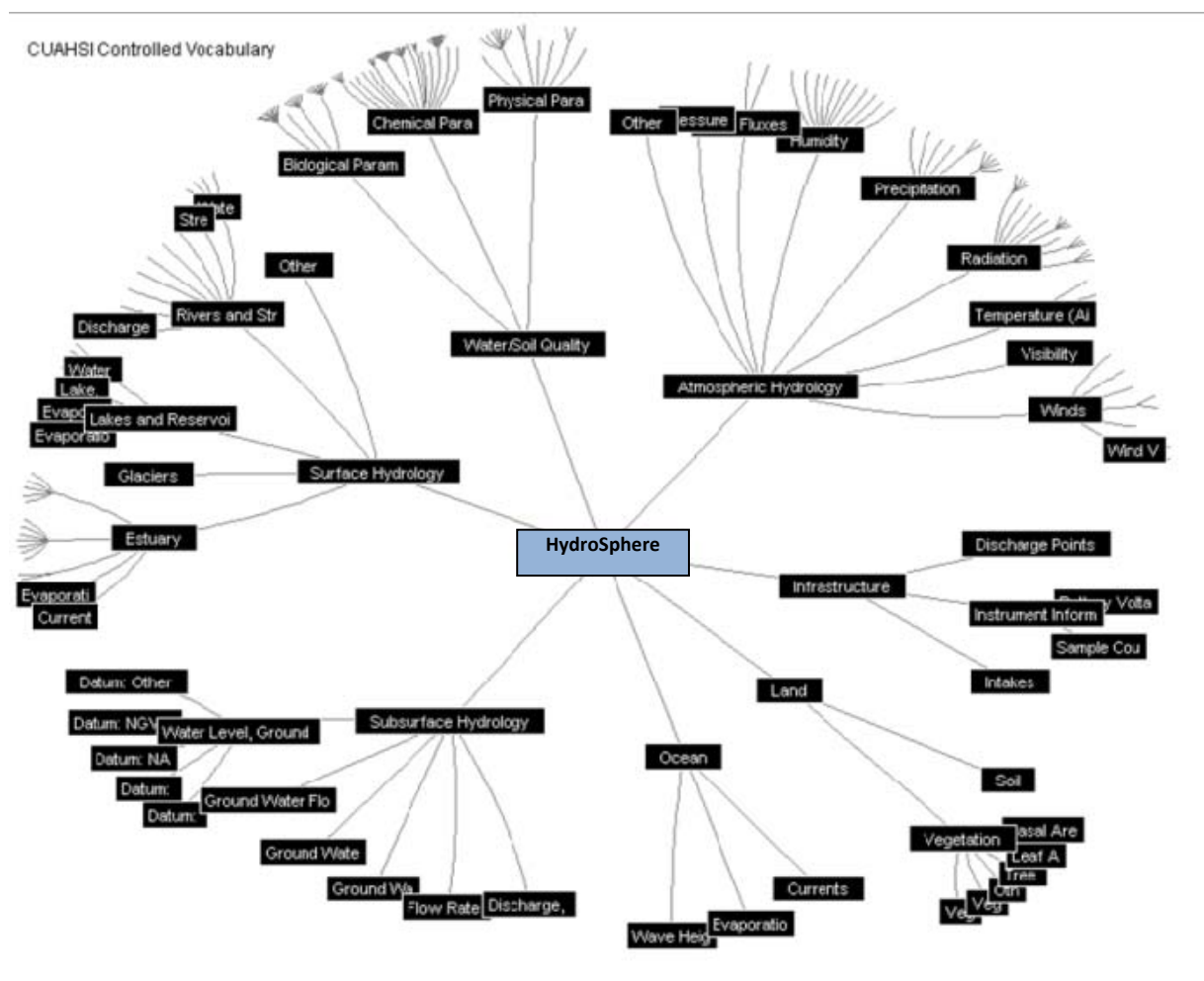


Figure 5-3 Screenshot of the central part of the hydrologic ontology using a hyperbolic StarTree viewer

The hydrologic ontology begins with an origin concept of “HydroSphere” and then branches into “Water Quality”, “SubSurface Hydrology”, “Surface Hydrology”, “Atmospheric Hydrology”, “Infrastructure”, “Ocean”, and “Land”. When traversing the ontology, a total of four different types of layers appear in the ontology, as shown in Figure 5-2: a Navigation Layer, a Core Layer, a Compound Layer and then the Detail Layer. The top level, called the *navigation layer*, provides structure to the ontology but class names in this layer cannot be used as keywords in the search process because they are too broad causing too many parameters to be returned. Beneath this is the *compound layer*, whose class names represent the first pool of permissible keyword entries for the search, such as “nutrients”. Below that is the *core layer*, like “nitrogen”, whose class names are more detailed but still broad

enough to be used as a search keyword. Finally, there is the *detail layer* which actually links to the variable names in each of the data sources. Variable names of individual data sources are not permitted in the search either because this would require the knowledge of an overwhelming number of names and codes, the avoidance of which is the primary motivation for building this search engine in the first place. Current limitations of the HydroSeek system are mostly associated with the number of available classes against which variables can be mapped; EPA's STORET and USGS' National Water Information System, NWIS, alone holds about 19,000 different variable codes. While not all of these are relevant to hydrologists, more need to be made accessible through the ontology.

5.4 HYDROSEEK: A GLOBAL SEARCH ENGINE

A result of these efforts is a search engine called “HydroSeek” (Beran, 2007). It's primary mission is to discover data by displaying observation sites that have data for specified geospatial and temporal bounds provided as well as a selected keyword concept. The geospatial bounds can either be a bounding box or a watershed that is classified by USGS HUC code system (HydroSeek's data base supports up to the 8-digit HUC). Once the data has been identified it can also be put into a data cart and downloaded for further processing. HydroSeek is not designed to provide an analysis environment or other data tools that a user might want to use. It is a very specialized race horse designed for one purpose but it is not good for plowing fields!

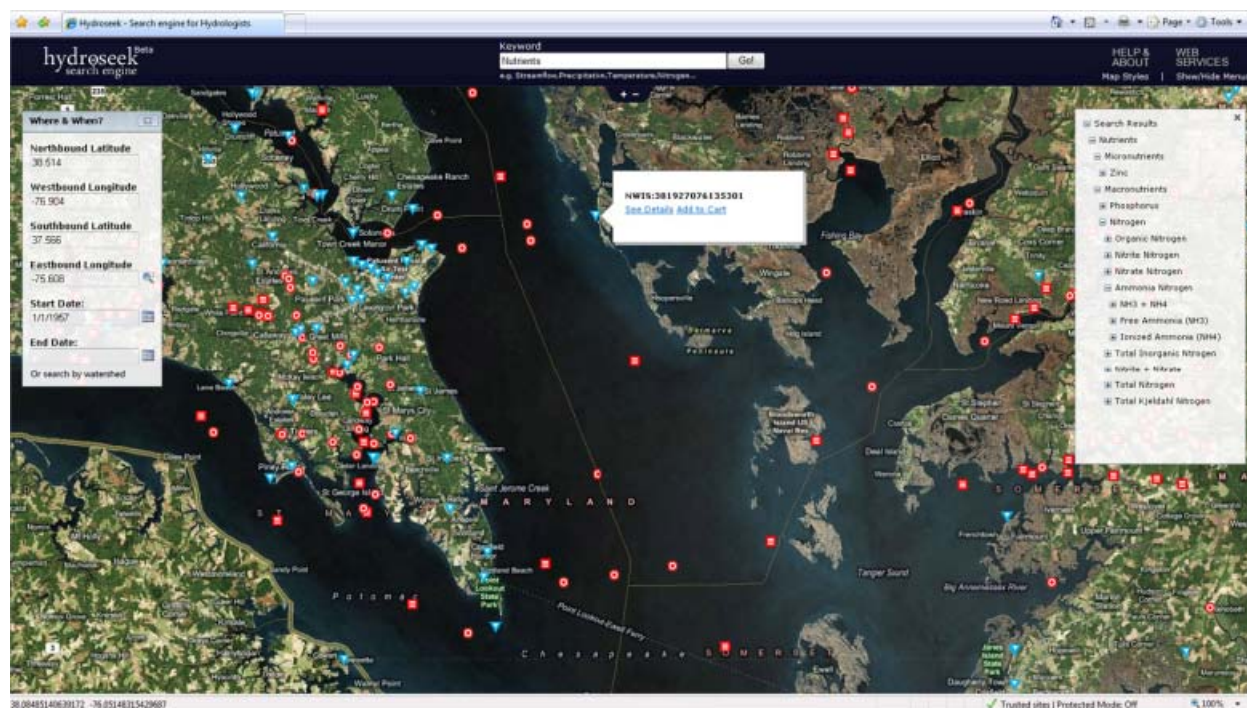


Figure 5-4 Screen shot of the HydroSeek application showing a number of sites for "nutrient" search in the Chesapeake Bay

Figure 5-4 shows a screen shot of a search result for “Nutrients” in the mid–section of Chesapeake Bay. Notice the different icons appearing in the image each type representing a different data source, i.e. blue triangles are USGS

NWIS, red circles are EPA STORET, and the red squares are the regional Chesapeake Bay Information Management System, CIMS. Also, notice the right in-set panel which shows the search results.

One very important feature of HydroSeek is its ability to return classified results instead of a long monotone list of hits. The classification is based in the ontology tree that follows the selected keyword (in this case “Nutrients”) thus providing structure to an otherwise tedious alphabetical listing. The various branches can be expanded and collapsed and the search result can be navigated by clicking on finer concepts which then automatically updates the display, i.e. the sites that do not have data for this finer concept disappear from the display. The HydroSeek application is available at www.hydroseek.net/search which is currently hosted at the SDSC. There is also a fail-over mirror site at the University of Texas at Austin which takes over the data search services in case the site at SDSC goes down thus providing back up services.

5.5 HYDROTAGGER

In order to support the HydroSeek search capabilities by establishing the necessary concept and variable linkages, the semantic mediation framework contains another application named HydroTagger. HydroTagger provides a graphical interface in which one can search for the appropriate concept to tag a variable against. HydroTagger operates off the same database tables as HydroSeek and has the main task of managing tagged and non-tagged variables. Non-tagged variables are “discovered” by a crawler that, currently once a week, trawls through all registered test bed services (those of the large nationwide and regional data sources are updated less frequently because they are very work intensive) to find out what has been added in the last week. If new variables have been found it will place those in a table that holds untagged variables and then offer them up in the HydroTagger interface whenever the data manager responsible for the network where it was found logs onto the system.

Once the tagging has been successfully carried out, the HydroTagger places it into the tagged table thus making the newly discovered variable available for HydroSeek. The system also permits the addition of new concepts in case a tagging cannot be carried out because the ontology does not feature an appropriate concept. This is an important aspect to recognize: the hydrologic ontology will remain work in progress for some time to come as it needs to grow as demand increases and the scope of the CUAHSI data services expands. It will be exposed to a larger audience so the conceptual underpinnings of the keyword search structure are agreed upon by the entire community. Finally, unlike HydroSeek, the HydroTagger application is not available for the general public and is restricted for usage by approved data managers and the CUAHSI HIS development team.

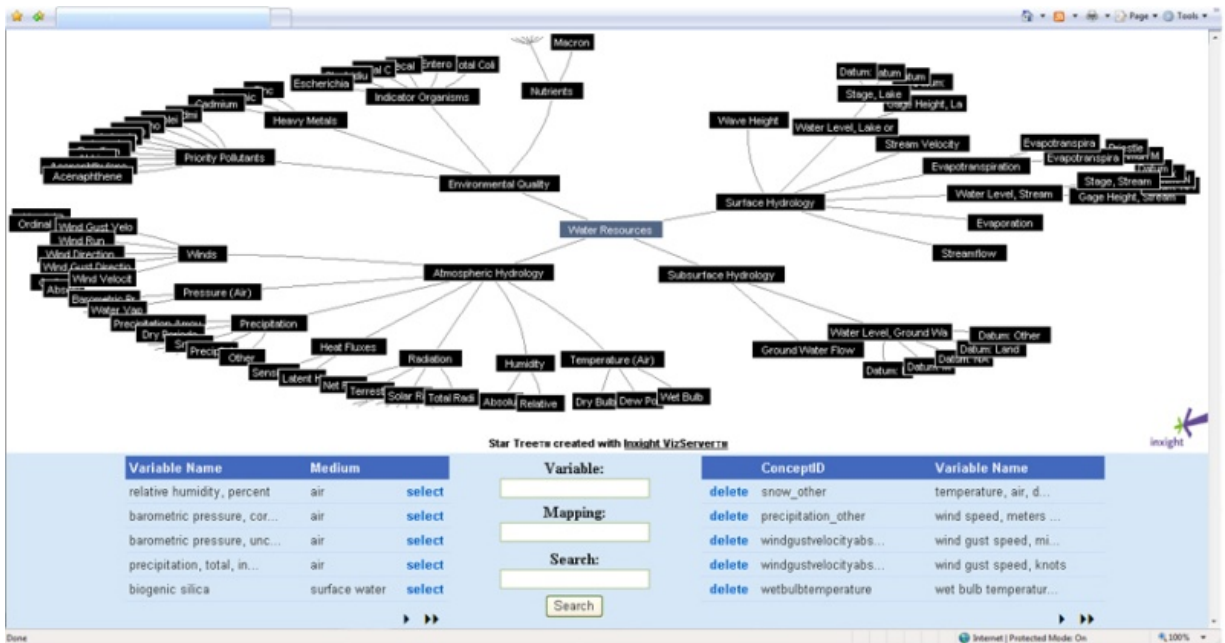


Figure 5-5 Screen shot of the HydroTagger application. Untagged variables are listed on the left while the tagging takes place in the middle section after which the variable/concept pair is placed on the right side.

5.6 REFERENCES

- Beran B., Piasecki M. (2008), "Engineering New Paths to Hydrologic Data", submitted to *Computers and GeoSciences*, Elsevier, Accepted for publication, February 2008.
- Beran B. (2007), HYDROSEEK: An Ontology-Aided Data Discovery System for Hydrologic Sciences, Ph.D. Thesis. 155 pp., Drexel University, Philadelphia, 5 September, 2007. <http://idea.library.drexel.edu/items-by-author?author=Beran%2C+Bora>

Chapter 6. WATER METADATA CATALOG AND HIS CENTRAL

By Tom Whitenack, David Valentine and Ilya Zaslavsky, San Diego Supercomputer Center

The CUAHSI HIS project creates a national inventory of hydrologic metadata and services. As described in Chapter 4, the project has established a standard protocol for exchanging hydrologic data and metadata, and developed a system for publishing hydrologic datasets. The data publication workflow implemented within HIS Server or HIS Server Lite includes several steps: loading the data into an instance of CUAHSI Observations Data Model, configuring CUAHSI Water Data Services to provide web service access to the database, and cataloging the services and metadata in a national water metadatabase. Once a new Water Data Service is published through this system, its metadata (stations, variables, etc.) and the data values become available for browsing and querying alongside data from many federal data repositories (USGS NWIS, EPA STORET, NCDC ASOS, USDA SNOTEL, etc.), state repositories, and multiple academic projects. Together they create a comprehensive portrait of the history and geography of hydrologic observations in the country. Data from multiple sources, once they are structurally and semantically reconciled with the each other, can be jointly used in hydrologic analysis and modeling. The more quality data sources become available through the system in a standard manner, the more valuable the system becomes in supporting hydrologic research and education. Hence the project's focus on encouraging hydrologic data sharing, and making publication, discovery and access of distributed hydrologic data straightforward to the users. The integrated Water Metadata Catalog, and the HIS Central application, are the core components of this vision.

As the HIS project continues to move forward, the number of CUAHSI Water Data Services is steadily increasing. Currently, data from more than 1.7 million observation sites is accessible through these services. The purpose of the water metadata catalog is to have a centralized repository of descriptive information about these sites, the variables measured at them, and the organizations that make these measurements, so that the body of water observations information can be accessed and queried as a whole across all data services. This capability is critical for enabling rapid access to distributed hydrologic data from mapping and analysis interfaces.

The HIS Central web application is used as a portal to the water metadata catalog where registered users can add and modify their CUAHSI water data services, and determine how their data will appear in the Hydroseek application. As illustrated in Figure 6.1, the HIS Central portal permits registration of new data services, harvesting of their metadata into the catalog, and then access to this metadata through the HydroSeek search engine. Registering data services at the HIS Central site is the last step in the hydrologic data publication workflow developed within the project. Other steps of the workflow (loading data into ODM, and setting up web services) are described in previous chapters. The semantic tagging of variables in the registered datasets, using HISCentral's HydroTagger application, is described in the chapter 5 focused on semantic mediation.

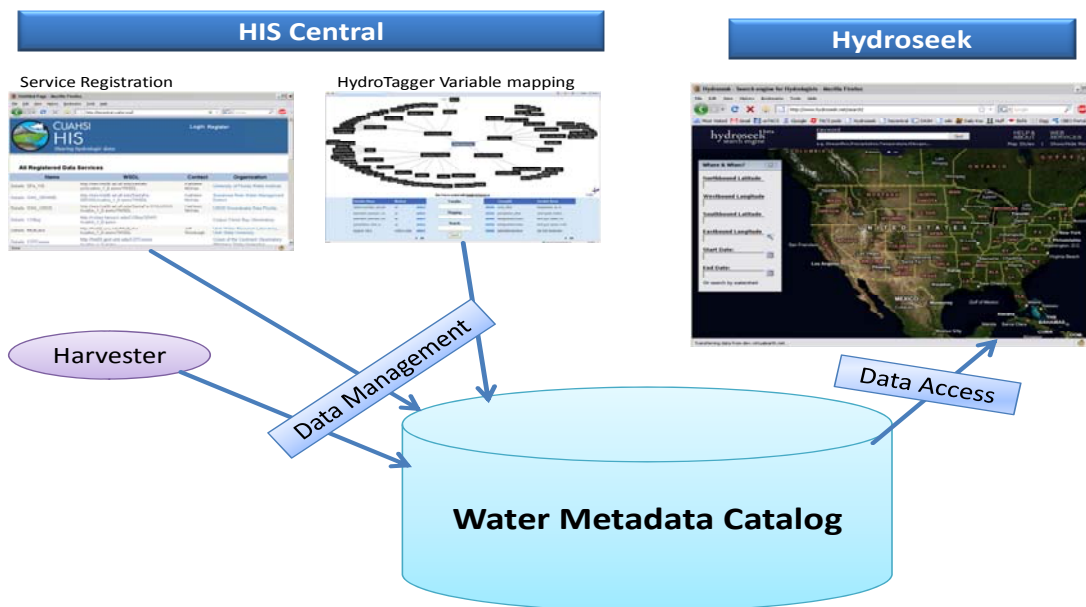


Figure 6-1 Diagram illustrating relationships between Water metadata catalog, HIS Central and Hydroseek

6.1 THE WATER METADATA CATALOG

The water metadata catalog is an integrated database that contains information about each of the registered CUAHSI Water Data Services:

- 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

To ensure that site and variable codes are unique, each site code and variable code is prefixed with the code of the observation network (e.g. NWIS:00060). With this information, client applications such as Hydroseek can effectively query data from multiple web services simultaneously. Figure 6.2 shows the key tables in the metadatabase and how they are linked. The key table is the SeriesCatalog, which contains a record for each individual data series. This table is linked to the Variables table via a VariableID which describes the variable and provides linkage to the corresponding concept. The SeriesCatalog is also linked to the Sites Table, via a SiteID which describes the location of the site. The Sites Table is linked to a table of data sources, which describe more about the organization or individual who provides the data.

By design, this table structure derives from the ODM relational database schema, so the meaning and description of each field can be obtained from ODM documentation. The four tables provide a powerful and at the same time lightweight foundation for the metadata catalog, supporting hydrologic metadata discovery from various CUAHSI

HIS client applications. The current (June 2008) content of the metadata catalog is described in the web services chapter (Chapter 4).

The primary application for managing the metadata catalog is the HIS Central described below.

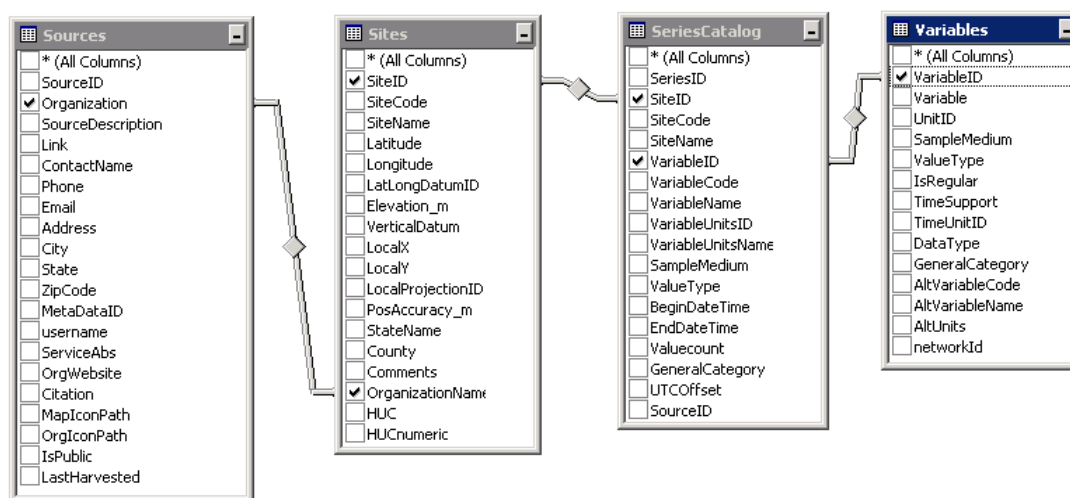


Figure 6-2 Primary metadata database tables and their relation to one another

6.2 HIS CENTRAL - [HTTP://HISCENTRAL.CUAHSI.ORG](http://hiscentral.cuahsi.org)

To announce the availability of their water data services to larger hydrologic research community, and make them available for indexing, querying and mapping from the growing number of online and desktop web service clients (DASH, Hydroseek, Google Earth-based client, HydroExcel, HydroGet, Matlab, MapWindow, etc.), data contributors may register the services at a hydrologic community portal. HIS Central is a web application which provides a web interface for adding services to the metadata catalog, and editing service registrations. It also displays the list of registered services available. To register a new water data service, users must create an account and login to the HIS Central. An overview of the HIS Central process is as follows:

1. Create a login account at <http://hiscentral.cuahsi.org>
2. Add a new Data Service (see section 6.3 for details)
3. Test the data service using the test page to make sure it works as you expect
4. Change the status of your service to public, to have it appear on the public version of Hydroseek
5. The HIS Central Administrator will trigger a harvest of your service.
6. You will receive an email notification of the results of the harvest and will be prompted to return to the HIS Central site to tag the harvested variables.
7. Once your variables are tagged, you can test the concept search functionality in the Hydroseek application, using a test version (<http://test.hydroseek.net/search>)
8. Upon approval of the new services, the updated metadata catalog will be linked to the production version of Hydroseek at SDSC (www.hydroseek.net) and replicated off-site at the University of Texas.

Figure 6.3 shows some of the currently registered CUAHSI Water Data Services.

Name	WSDL	Contact	Organization
Details SFe_YSI	http://ees-his06.ad.ufl.edu/santafe-ysi/cuahsi_1_0.asmx?WSDL	Kathleen McKee	University of Florida Water Institute
Details GWL_SRWMD	http://ees-his06.ad.ufl.edu/SantaFe-SRGWL/cuahsi_1_0.asmx?WSDL	Kathleen McKee	Suwannee River Water Management District
Details GWL_USGS	http://ees-his06.ad.ufl.edu/SantaFe-GWLUSGS/cuahsi_1_0.asmx?WSDL	Kathleen McKee	USGS Groundwater Data Florida
Details CCBay	http://ccbay.tamucc.edu/CCBayODWS/cuahsi_1_0.asmx		Corpus Christi Bay Observatory
Details MudLake	http://his02.usu.edu/MudLake/cuahsi_1_0.asmx?WSDL	Jeff Horsburgh	Utah Water Research Laboratory, Utah State University
Details COTCsnow	http://his03.geol.umt.edu/COTCsnow/cuahsi_1_0.asmx		Crown of the Continent Observatory (Montana State University)

Figure 6-3 Partial listing of CUAHSI Water Data Services

6.3 REGISTERING AND TESTING A CUAHSI WATER DATA SERVICE

At this step, the data service contributor is logged into HIS Central. Adding a new data service to the metadata catalog requires you to fill out a standard web based form.

- **Service Name:** this should correspond to your “Network” name, the value you provide when configuring a WaterOneFlow web service against an ODM data base. The term “ODM” is not allowed.
- **Service WSDL.** The web access point to your data server. Be sure to provide a URL that is accessible from the internet.
- **Source Info:** What is the name of the organization that is responsible for the data service? Do they have a web site?
- **Contact Info:** Who should be contacted with questions regarding this service? How?
- **Citation:** What should the citation text read in the files downloaded from your service?
- **Abstract:** A general description of service data. Where it came from, why was the study done, etc.
- **Is the service public?** This is unchecked, or private by default, this provides you with the ability to take your service offline from our system. Services which are tagged private, aren’t visible to other users and aren’t included in Hydroseek queries.

Figure 6.4 shows the interface used to enter this information.

CUAHSI HIS Central - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://hiscentral.cuahsi.org/

CUAHSI HIS
Sharing hydrologic data

Welcome, twhitenack
Log out Edit Profile

View my Data Services Add Data Service All Data Services Administration

Register Data Service

Service:
Name:
Service WSDL:

Source Info:
Organization:
URL:

Contact Info:
Name:
Email:
Phone:
☐ Is service public?

Citation:

Abstract:

Done

Figure 6-4 HIS Central Data Service registration Form

Once the service is registered, you will be able to upload an organizational logo and map icon for use in the Hydroseek client application. Hydroseek will use the map icon you supplied to symbolize stations in your newly registered network. Additionally, you will be able to edit the information you just entered, and once the service has been harvested, you'll be able to list your Sites and Variables, and you'll be able to Tag your variables. Figure 6.5 shows the result of having entered this information for a particular data service.

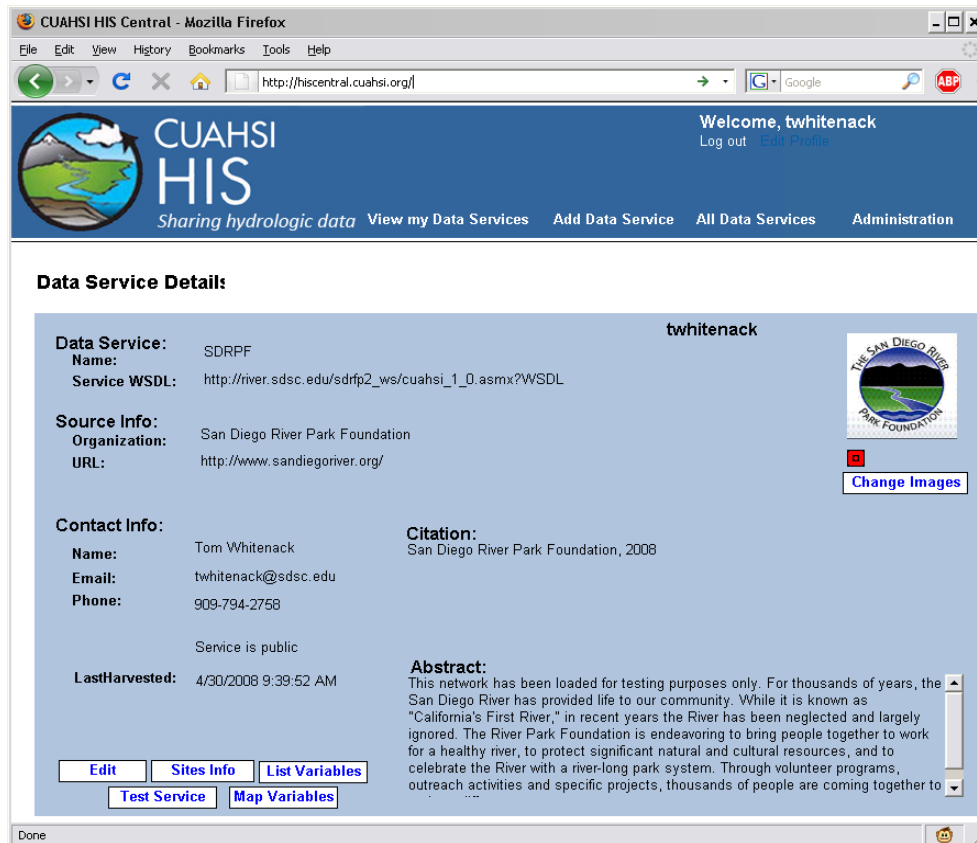


Figure 6-5 HIS Central Data Service details page

6.4 METADATA HARVESTING AND VARIABLE TAGGING

Once a data service has been registered, its metadata needs to be harvested in order to populate all of the tables of the Water Metadata Catalog. Harvesting is initialized by the HIS Central administrator. The harvesting process utilizes the methods available to all clients of WaterOneFlow web services. It starts with a call to the `getSites()` method, which returns a list of every site within the service. Then the harvester process iterates through each site, running the `getSiteInfo()` method to obtain the list of variables and data series collected at each site. Metadata harvesting is CPU intensive process for both the remote server as well as the metadata base server. Once the initial harvest is completed, the data service is re-harvested on a weekly basis to see if you have added more sites or variables, or extended the time period of available data.

Because the HydroSeek client application has ontological search capabilities it is necessary to tag, or associate the specific hydrologic concept with each variable. For that reason, the HydroTagger application described in Chapter 5 has been incorporated into the HIS Central to provide a graphical interface in which one can search for the appropriate concept to tag a variable against. Figure 6.6 shows the data tagging application within HIS Central.

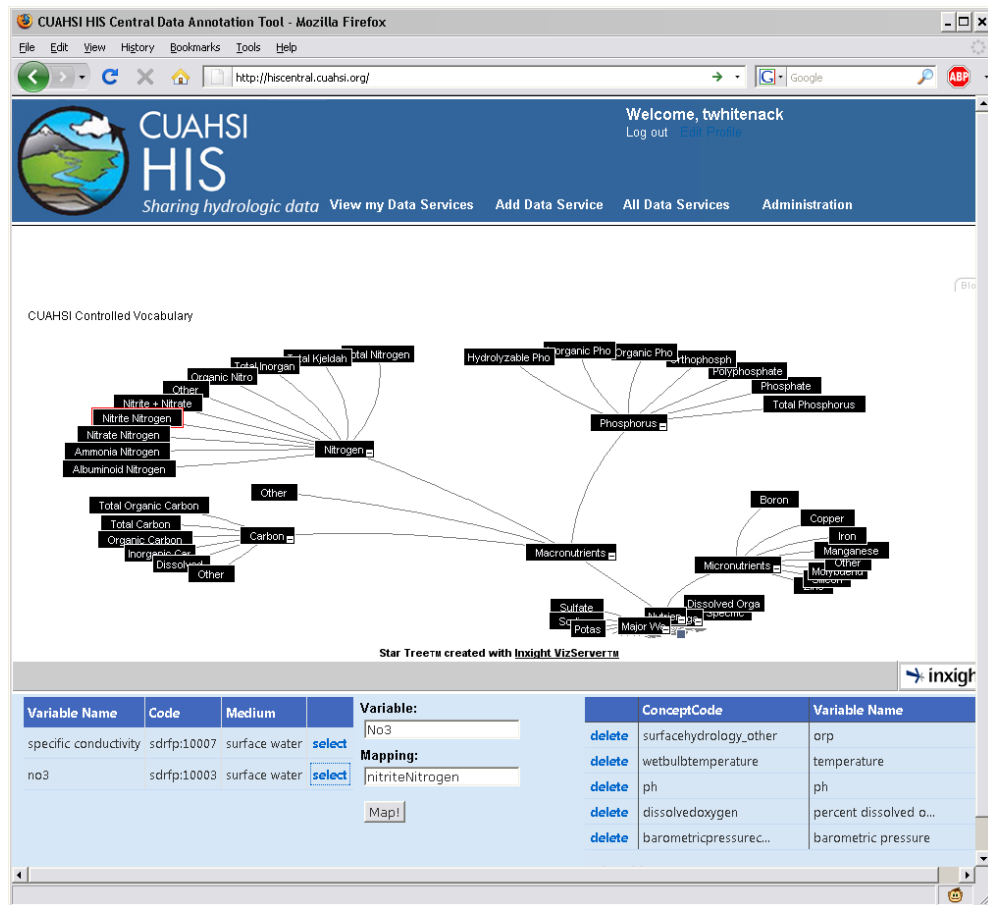


Figure 6-6 HydroTagger Application as part of the HIS Central Data Service publication process

6.5 HIS CENTRAL AND THE HYDROLOGIC COMMUNITY: THE PATH FORWARD

HIS Central is developed to be the focal point of data collection, data sharing and data discovery efforts within the CUAHSI HIS and the hydrologic community. Its mission is to integrate disparate observations data repositories, from government, academic and other sources, and maintain a comprehensive portrait of the geography and history of hydrologic observations in the US. What is described above is the current functionality of the HIS Central application: registration and management of geographically distributed Water Data Services, and support of global discovery of hydrologic observations data with the harvested hydrologic metadata catalog. We envision that, to fulfill its mission, HIS Central will evolve to have these additional features:

- A mapping interface, showing multiple geographically distributed data services;
- A web service interface providing programmatic access to the hydrologic metadata catalog;
- Ability to submit ODM instances to be published through the HIS Central (i.e. beyond the currently supported submission of Data Services);
- A comprehensive visualization and analysis system for the metadata catalog, based on the OLAP (Online Analytical Processing) data cube technology. This system would allow users to rapidly explore the growing

metadata catalog, answering queries such as “in which watersheds a given pair of variables was measured, having at least a 100 concurrent measurements”;

- An annotation/Wiki system, letting the community annotate data services in a managed fashion, and then discover such annotations using semantic and spatial queries;
- Registration of additional data types (beyond observations data) as they become incorporated in CUAHSI HIS (e.g. raster coverages, remote sensing data);
- Ability to discover data in other earth science related portals (in environmental science, geology, atmospheric sciences, oceanography) in a standard manner, to integrate them with hydrologic observations;
- Integration with online collaboration and communication tools.

HIS Central is a new application developed as part of CUAHSI HIS 1.1, and feedback from the community on its features, and the additional features we envision, would be appreciated.

Chapter 7. USING DATA IN ANALYSIS APPLICATIONS

By Tim Whiteaker, Ernest To, David Maidment, and Kate Marney, The University of Texas at Austin

While CUAHSI water data services support standardized, automated queries for hydrologic data, they may be difficult to use for those who are uninitiated into the world of web services. Therefore, CUAHSI HIS includes extensions to applications commonly used in hydrologic science in order to connect those applications with water data services. Two applications are presented here: HydroExcel (for Microsoft Excel) and HydroGET (for ESRI's ArcMap GIS). Both applications make use of an object library called HydroObjects in order to make dynamic connections to web services.

7.1 HYDROOBJECTS

HydroObjects is a package that underlies applications running in Microsoft Windows, such as Excel and ArcGIS, which allows these applications to access CUAHSI Water Data Services. More information about HydroObjects is available at: <http://his.cuahsi.org/hydroobjects.html>. HydroObjects has to be installed on your computer before you can run HydroExcel or HydroGet. HydroObjects is a .NET DLL with COM classes that supports hydrology applications. The key class in the library is WebServiceWrapper, which provides a method for calling Web Services from a COM (e.g., Visual Basic for Applications (VBA)) environment. This class can be used to call CUAHSI Water Data Services for downloading hydrologic time series.

7.2 HYDROEXCEL - WATERONEFLOW IN EXCEL

<http://his.cuahsi.org/hydroexcel.html>

HydroExcel is a Microsoft Excel spreadsheet that uses macros and HydroObjects to download hydrologic observations data from CUAHSI Water Data Services. This means that you can query for observation sites, variables, and time series data from online resources directly within Excel. As long as a web service follows WaterOneFlow specifications, HydroExcel will be able to communicate with it. Thus, HydroExcel provides a window into the nation's water data from within one of the most widely used applications within the hydrologic science community.

HydroExcel consists of nine worksheets, six of which provide access to WaterOneFlow web services. Each worksheet accesses a specific kind of data. For example, Figure 1-1 shows a screenshot of the Time Series worksheet, which is used for downloading a time series of values for a given variable at a given location.

GetValues	<input checked="" type="checkbox"/> Ignore NoData Value
Site Code/Location	NWIS:08158000
Variable Code	NWIS:00060
Start Date	5/1/2008 0:00
End Date	6/30/2008 0:00
Get Values	
DateTime	Value
5/1/2008 0:00	786
5/2/2008 0:00	820
5/3/2008 0:00	1170
5/4/2008 0:00	797
5/5/2008 0:00	975
5/6/2008 0:00	952
5/7/2008 0:00	969
5/8/2008 0:00	1100

Figure 7-1 Daily streamflow values (cfs) downloaded in HydroExcel

To use HydroExcel, you indicate the web service that you want to work with, and then click buttons in the spreadsheet to download information from the web service. Links to some existing WaterOneFlow web services are provided in the spreadsheet to get you started, as well as informative text, as shown in Figure 7-2.

The screenshot shows the 'Data Source' worksheet in HydroExcel. The layout includes a 'Data Source' section on the left with instructions and buttons. The main area contains a 'Specify the web service that will be used in all worksheets' section with a 'WSDL Location' input field and buttons for 'Get Capabilities', 'Open Service Web Page', 'Get Sites', 'Get Variables', and 'Get Site Catalog'. Below this is a table of 'Web Services for National Data Sources' and another table for 'Web Services for Academic Investigator Data'. Callouts highlight the 'Active Web Service' (WSDL Location), 'Informative Text' (left sidebar), 'Learning about the Service' (Open Service Web Page), 'Worksheet Shortcuts' (Get Sites, Get Variables, Get Site Catalog), and 'Example Web Services' (the tables of web services).

Data Source	WSDL Location	Description
United States Geological Survey	http://river.sdsc.edu/wateroneflow/NWIS/DailyValues.aspx?WSDL	NWIS daily values
United States Geological Survey	http://river.sdsc.edu/wateroneflow/NWIS/Groundwater.aspx?WSDL	NWIS groundwater
United States Geological Survey	http://river.sdsc.edu/wateroneflow/NWIS/UnitValues.aspx?WSDL	NWIS real time
United States Geological Survey	http://river.sdsc.edu/wateroneflow/NWIS/Data.aspx?WSDL	NWIS instantaneous
Oak Ridge National Laboratory	http://river.sdsc.edu/wateroneflow/DAYMET/Service.aspx?WSDL	Daymet Meteorology
National Centers for Environmental Prediction	http://river.sdsc.edu/wateroneflow/NAEM12k/Service.aspx?WSDL	North American Monsoon
Environmental Protection Agency	http://river.sdsc.edu/wateroneflow/EPA/cuahsi_1_0.aspx?WSDL	STORET water quality
NASA	http://river.sdsc.edu/wateroneflow/MODIS/Service.aspx?WSDL	Atmospheric Moisture

University	WSDL Location	Description
Utah State University	http://his02.usu.edu/littlebear/cuahsi_1_0.aspx?WSDL	Utah State University
Utah State University	http://his02.usu.edu/mudlake/cuahsi_1_0.aspx?WSDL	Utah State University
University of Iowa	http://his08.ihr.uiowa.edu/nexrad/cuahsi_1_0.aspx?WSDL	University of Iowa
University of Iowa	http://his08.ihr.uiowa.edu/water_quality/cuahsi_1_0.aspx?WSDL	University of Iowa
University of Iowa	http://his08.ihr.uiowa.edu/lincoln/cuahsi_1_0.aspx?WSDL	University of Iowa

Figure 7-2 Layout for Data Source worksheet

The worksheets and their functions are:

- **Introduction** – Introduce the worksheet and provide license information.
- **Data Source** – Set the web service that will be accessed in the spreadsheet.

- **Sites** – Download a list of sites available from the web service (Figure 7-3).
- **Variables** – Download a list of variables available from the web service.
- **Site Info** – Download information about a specific site, including a list of variables measured at the site.
- **Site Catalog** – Download site info for several sites at once.
- **Site Summary** – Use a pivot table to summarize the site catalog, typically by the number of values of a given variable measured at each site.
- **Time Series** – Download a time series of values for a given variable at a given location for a given time period.
- **Statistics and Charts** – Use a pivot table and chart to summarize time series data (Figure 7-4).

<div> <div>Get Sites Options</div> <div>Show sites in Google Earth after download</div> <div>FALSE</div> </div>			<div>About the Data You're Viewing</div> <div>Data Source</div> <div>http://his02.usu.edu/littlebearriver</div>		
<div> <div>Get Sites</div> <div>Create Site Catalog for These</div> </div>			<div>Obtained</div> <div>6/6/2008 11:47</div>		
<div>Site list</div>					
Site Code	Site Name	State	County	Latitude	Longitude
LittleBearRiver.USU-LBR-Mendon	Little Bear River at Mendon Road near Mendon, Ut	Utah	Cache	41.718473	-111.9464
LittleBearRiver.USU-LBR-Paradise	Little Bear River at McJurdy Hollow near Paradise	Utah	Cache	41.575552	-111.85522
LittleBearRiver.USU-LBR-ExpFarm	Utah State University Experimental Farm near We	Utah	Cache	41.666993	-111.89057
LittleBearRiver.USU-LBR-SFLower	South Fork Little Bear River below Davenport	Utah	Cache	41.506518	-111.81508
LittleBearRiver.USU-LBR-EFLower	East Fork Little Bear River at Paradise Canal	Utah	Cache	41.529212	-111.79932
LittleBearRiver.USU-LBR-EFWeather	Little Bear River Upper Weather Station near Avon	Utah	Cache	41.535543	-111.80595
LittleBearRiver.USU-LBR-SFUpper	South Fork Little Bear River above Davenport	Utah	Cache	41.495409	-111.81799

Figure 7-3 Sites worksheet layout

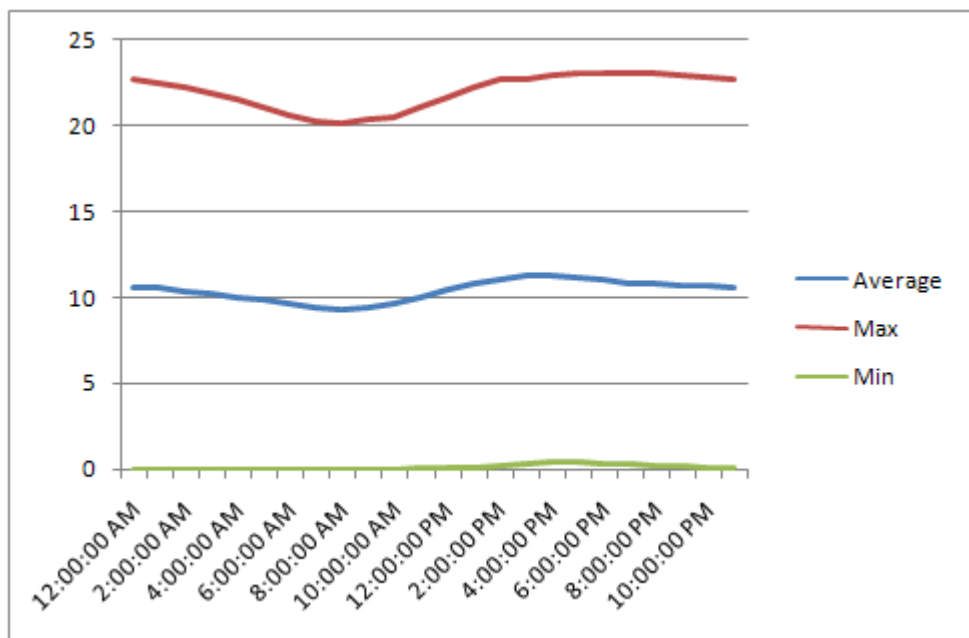


Figure 7-4 Diurnal variation of temperature (degrees C) at the Little Bear River at Mendon Road

When downloading site information, HydroExcel can build a KML file to show the sites in Google Earth. This provides a spatial component that complements the tabular nature of the Excel spreadsheet (Figure 7-5).

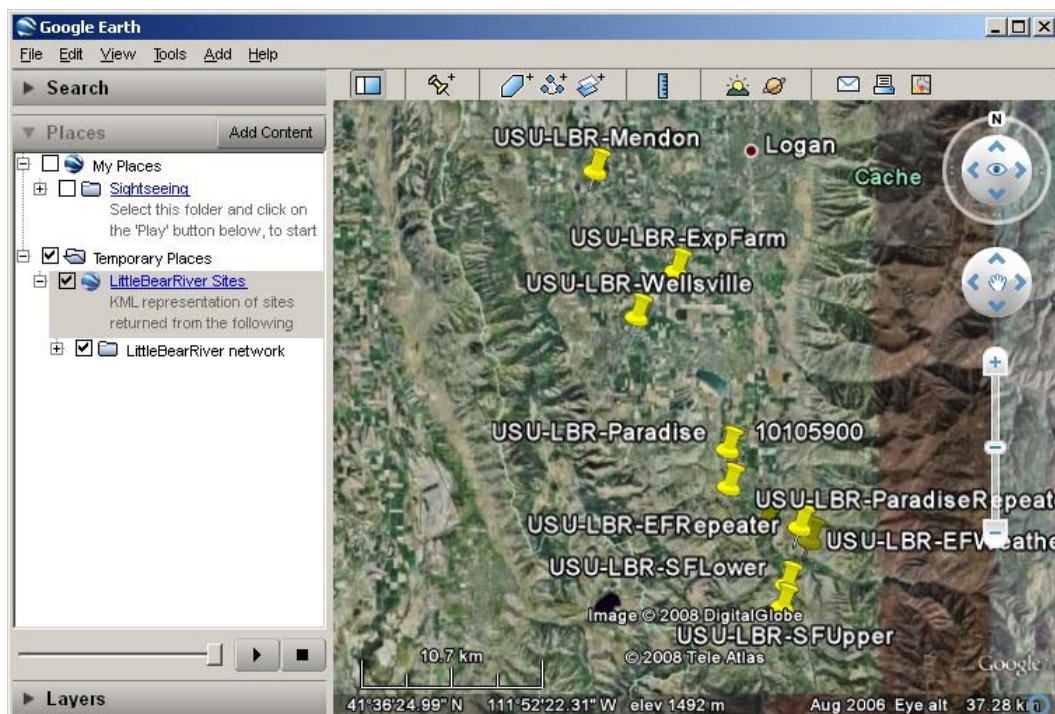


Figure 7-5 Little Bear River network sites in Google Earth

In summary, HydroExcel provides access to WaterOneFlow web services from the Excel application environment using HydroObjects and VBA macros.

7.3 HYDROGET – WATER DATA SERVICES IN ARCMAP

<http://his.cuahsi.org/hydroget.html>

CUAHSI's *HydroGET* (**Hydro**logic **GIS** **E**xtraction **T**ool) is a versatile tool that gives ArcGIS users the ability to ingest water observations data into ArcGIS. HydroGET stores the downloaded data using the time series format in the well-known data model, Arc Hydro, and its upcoming successor, Arc Hydro II. HydroGET can work with any web service as long as the web service complies with the WaterOneFlow protocol.

Figure 7-6 shows the program interface of HydroGET. It contains five tabs for the user to specify the kind of environmental data to be downloaded. These tabs are listed as follows:

- Atmospheric
- Surface
- Subsurface
- Custom (single point); and,
- Custom (multiple points).

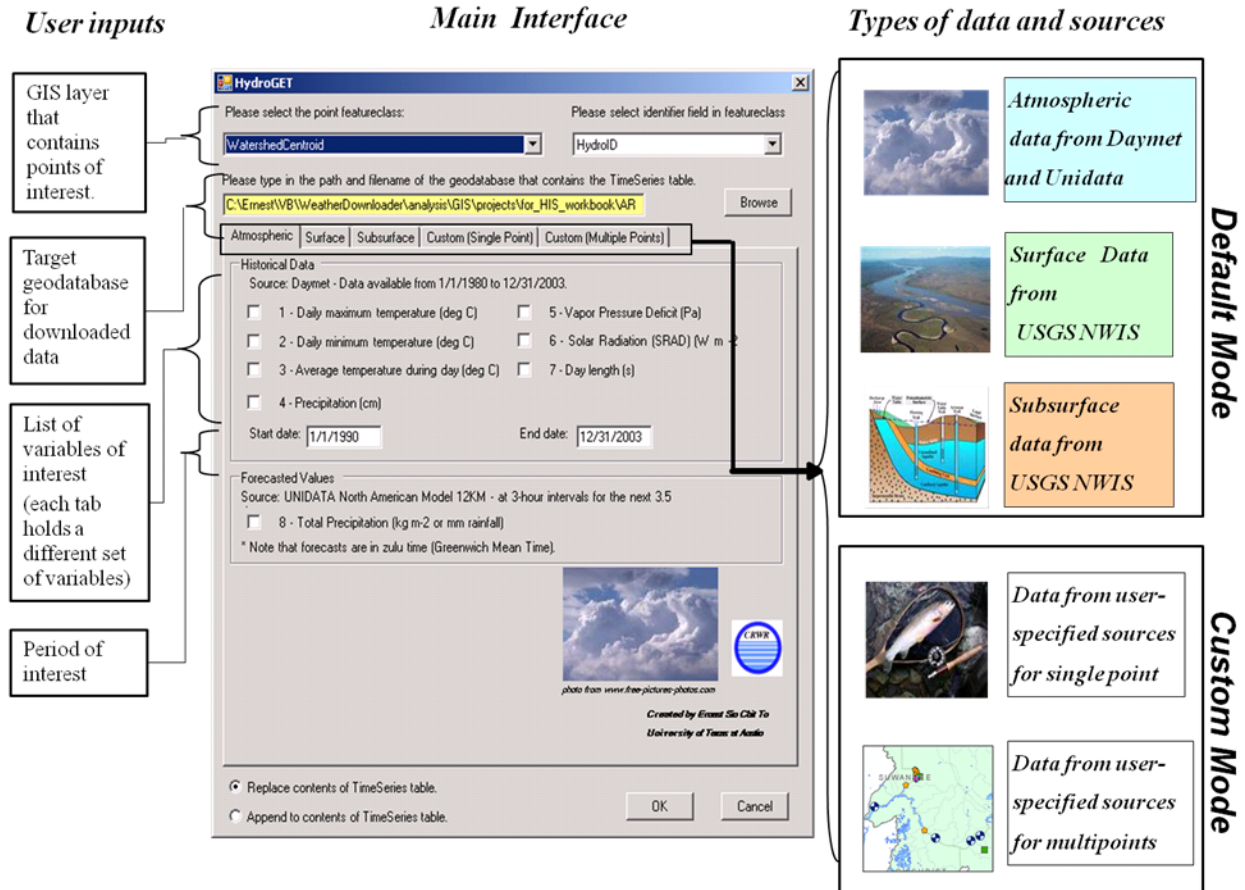


Figure 7-6 The HydroGET user interface

The **Atmospheric**, **Surface**, **Subsurface** tabs offer the user the ability to download data from preset web services to describe the different components of the hydrologic cycle. These web services access Daymet and Unidata for atmospheric data and USGS NWIS for surface water and groundwater data. Together, they enable the user to characterize the hydrologic characteristics of a geographic region of interest.

When the user utilizes only these three tabs to download data, HydroGET is operating in **Default Mode**. This means that HydroGET only calls the web-services and variables that have already been hard-wired into its code. When HydroGET is operating in **Default Mode**, it is very user-friendly and does not require the user to have any background knowledge of web services. However, its capabilities in this mode are limited as it cannot handle data sources other than Daymet, Unidata and NWIS.

The **Custom (single point)** and **Custom (multiple points)** tabs allow the user to retrieve data from any user-defined web services that comply with WaterML format.

When the user utilizes these two tabs to download data, HydroGET is operating in **Custom Mode**. **Custom Mode** requires the user to have slightly more knowledge about web services and is recommended for the intermediate user. In this mode HydroGET becomes truly a powerful tool. Not only does it have the ability to access a wide range of web services, it also has the ability to batch process multiple requests to different web services. In this mode it essentially becomes a harvester for data.

To use HydroGET to download data, the user supplies the following inputs via the program interface:

- 1) a point feature class in ArcGIS that denotes the locations of interest;
- 2) the variables of interest (e.g. precipitation, temperature), and;
- 3) the time periods of interest.

When executed, HydroGET cycles through each location in the point feature class, downloads the desired data through web services and writes them to the TimeSeries table of an Arc Hydro geodatabase. Figure 7-7 shows an illustration of these steps.

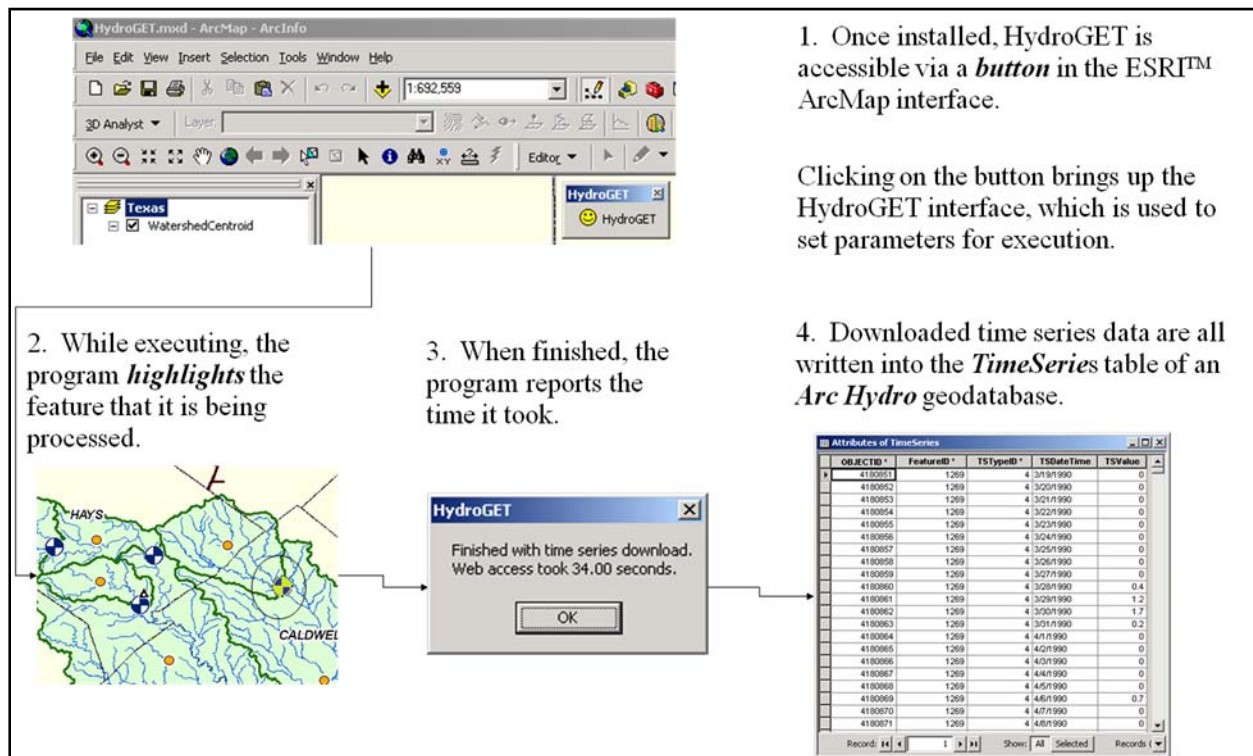


Figure 7-7 Steps to run HydroGET

In addition to updating the TimeSeries table, HydroGET also updates the TSType table with information of any new variables that have been downloaded. This information includes TSTypeID, variable name and units.

7.4 HYDROGET AND NHDPLUS

CUAHSI's HydroGET can also be used in conjunction with the NHDPlus dataset. The National Hydrography Dataset Plus (NHDPlus) is an integrated suite of application-ready geospatial data sets that contain information about surface water features such as lakes, ponds, streams, rivers, springs, and wells. NHDPlus data is available for download on the NHDPlus website: <http://www.horizon-systems.com/nhdplus/data.php>, as shown in Figure 7-8. By applying the HydroGET tool to an NHDPlus dataset, obtaining time series of hydrologic data via web services for stream gages, water bodies, and watersheds becomes a simple task.

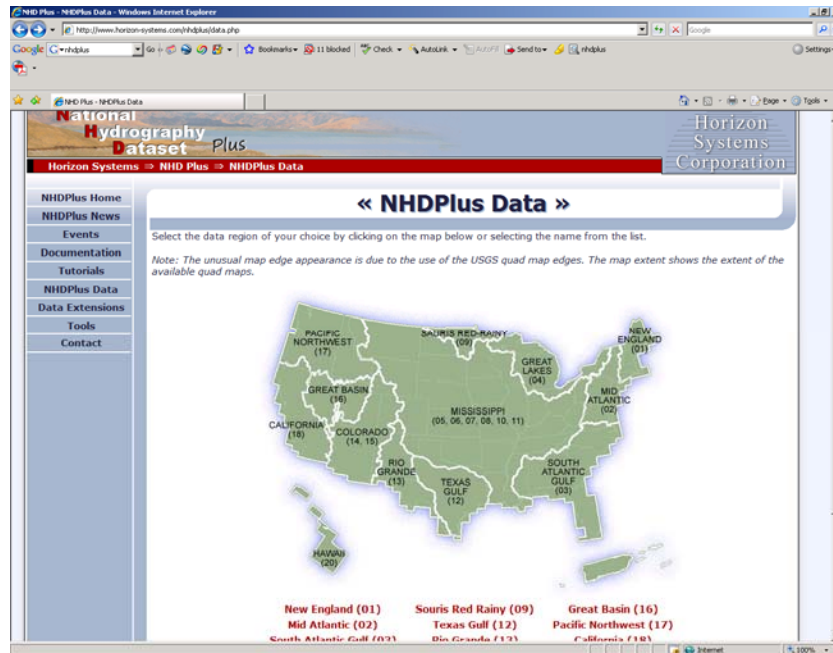


Figure 7-8 Website for downloading NHDPlus data

NHDPlus data is organized by drainage basin and contains many different shapefiles pertaining to hydrology. This includes gages, water bodies, flowlines, basins, subbasins, watersheds, catchments, etc. The Gages layer contains all of the USGS gaging sites in the United States, the NHDWaterbody layer contains all water bodies, and the Subbasin layer includes all of the 8-digit HUC subbasins.

One of the most fundamental of all hydrologic data types is daily streamflow. HydroGET is easily utilized to access and download USGS NWIS streamflow data using selected gages within the NHDPlus dataset. The attribute table of the Gages layer contains two fields, named DAY1 and DAYN, indicating the dates of record for each gage, as shown in Figure 7-9. This information can be used to select a period of interest.

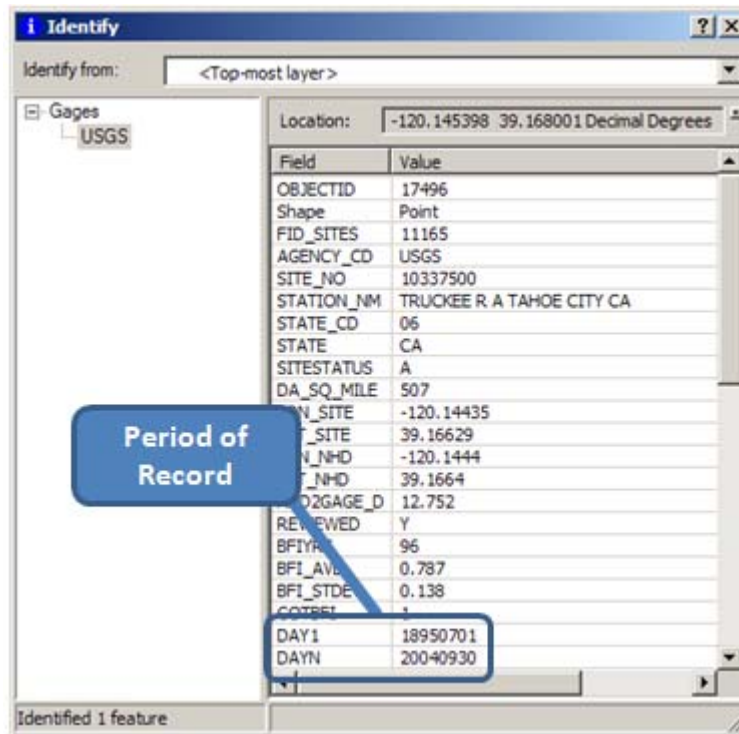


Figure 7-9 NHDPlus Gage selection and identification

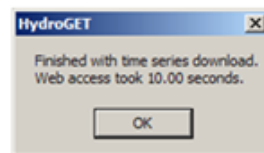
In HydroGet, the user selects the point feature class to be used (Gages), the identifier field (OBJECTID), the location of the TimeSeries table, streamflow data to be downloaded (daily streamflow data, indicated by a checked box), USGS gage number field (SITE_NO), and the start and end date, as shown in Figure 7-10.

1. User inputs criteria into HydroGET

The HydroGET window includes the following fields and options:

- Please select the point featureclass:** A dropdown menu with "Gages" selected.
- Please select identifier field in featureclass:** A dropdown menu with "OBJECTID" selected.
- Please type in the path and filename of the geodatabase that contains the TimeSeries table:** A text field containing "C:\arcfile\csw\Projects\EnvFlow\ERIS\CUAKSI\HydroGET\HydroGET.mdb" and a "Browse" button.
- Streamflow Data:**
 - Source: USGS National Water Information System (NWIS)
 - ☒ 9 - Daily Streamflow Data (cfs)
 - Please select field that contains USGS gage number: A dropdown menu with "SITE_NO" selected.
 - Start date: 1/1/2000
 - End date: 12/31/2003
- Buttons:** "Atmospheric", "Surface", "Subsurface", "Custom (Single Point)", "Custom (Multiple Points)".
- Radio Buttons:**
 - ☒ Replace contents of TimeSeries table.
 - ☐ Append to contents of TimeSeries table.

2. When finished, the program reports the time it took.



3. Downloaded time series data are all written into the *TimeSeries* table of an *ArcHydro* geodatabase.

OBJECTID *	FeatureID *	TSTypeID *	TSTime	TSValue
4599873	17496	9	1/1/2000	110
4599874	17496	9	1/2/2000	110
4599875	17496	9	1/3/2000	163
4599876	17496	9	1/4/2000	245
4599877	17496	9	1/5/2000	267
4599878	17496	9	1/6/2000	267
4599879	17496	9	1/7/2000	267
4599880	17496	9	1/8/2000	267
4599881	17496	9	1/9/2000	267
4599882	17496	9	1/10/2000	265
4599883	17496	9	1/11/2000	263
4599884	17496	9	1/12/2000	264
4599885	17496	9	1/13/2000	264
4599886	17496	9	1/14/2000	261
4599887	17496	9	1/15/2000	262
4599888	17496	9	1/16/2000	216
4599889	17496	9	1/17/2000	183
4599890	17496	9	1/18/2000	157
4599891	17496	9	1/19/2000	126

Figure 7-10 Using HydroGet to obtain streamflow with NHDPlus gage sites

The TSTypeID is a number that identifies the variable reported. The TSType table can be opened to verify which variable is represented by a TSTypeID, as shown in Figure 7-11.

OBJECTID *	TSTypeID	Variable	Units	IsRegular	TSTypeID	DataType	Origin
1	1	Daily maximum temperature	deg C	True	1Day	Maximum	Recorded
2	2	Daily minimum temperature	deg C	True	1Day	Maximum	Recorded
3	3	Average daily temperature	deg C	True	1Day	Average	Recorded
4	4	Precipitation	cm	True	1Day	Average	Recorded
5	5	Vapor Pressure Deficit	Pa	True	1Day	Maximum	Recorded
6	6	Solar Radiation (SRAD)	W m -2	True	1Day	Average	Recorded
7	7	Day length	s	True	1Day	Average	Recorded
8	8	Forecast. Tot Prcip (3 hr int)	kg m-2 or mm	True	3Hour	Cumulative	Generated
9	9	Streamflow	cfs	True	1Day	Average	Recorded
10	10	Groundwater level	feet below ground surface	False	Other	Instantaneous	Recorded

Figure 7-11 TSType table

Because HydroGET is designed to retrieve data for point features, it readily works with the USGS gages from NHDPlus. By calculating the centroid for polygon features using standard ArcGIS tools, you can enable HydroGET to work with additional NHDPlus features such as water bodies or subbasins. To generate a centroid for a polygon, use the ArcToolbox | Data Management Tools | Features | Feature to Point tool, as shown in Figure 7-12.

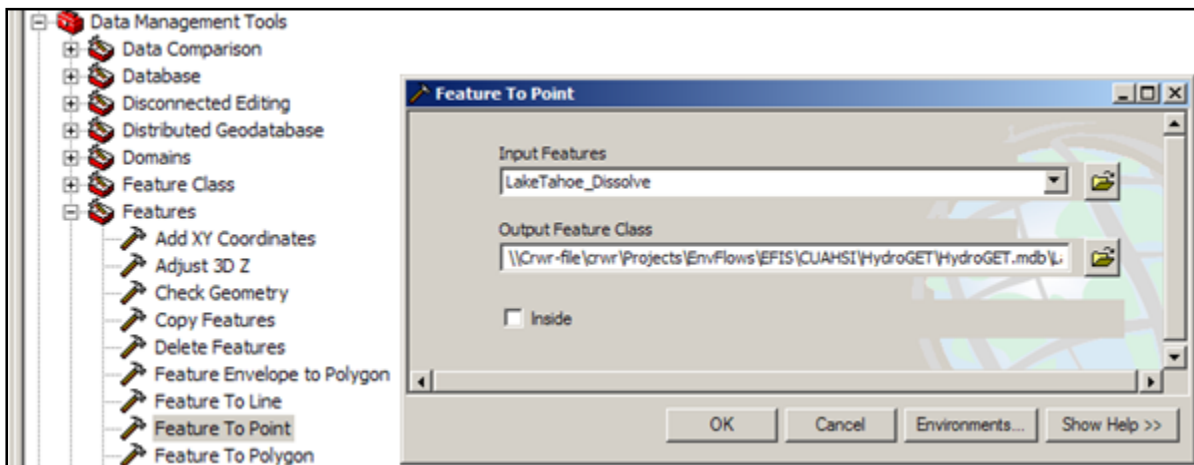


Figure 7-12 Creating a polygon centroid

By creating this new point feature class to represent a water body area or land area, HydroGET can be used to download time series data pertaining to a surface. This is useful for analyzing variables such as temperature over a water body's surface or precipitation over a catchment or subbasin.

7.5 USING WATER DATA SERVICES IN YOUR APPLICATION OF CHOICE

CUAHSI's Water Data Services can be used in many applications beyond what we've illustrated here. To learn more about WaterOneFlow web services, please visit <http://his.cuahsi.org/wofws.html>. That website describes WaterOneFlow and lists several services already in production. You can also find documentation about how to use the services in additional environments such as MATLAB, Visual Studio, and Java NetBeans IDE, and how to install WaterOneFlow web services if you want to publish your own data.

Chapter 8. USING DATA IN MODELS

By Jon Goodall and Tony Castronova, University of South Carolina

8.1 INTRODUCTION

Hydrologic science often requires both observations and models to test scientific hypothesis. Thus, it is important to consider approaches for using HIS data in models. While there are many approaches one could take to accomplish this, the primary goal of this chapter is to present a tool named *HydroLink* that provides a standardized way to connect the HIS with models through the Open Modeling Interface (OpenMI). While models are free to use data from the WaterOneFlow servers, ODM databases, or WaterML files directly, the OpenMI provides a translation layer between data and models so that it is easier to plug-and-play models and data within a component-based modeling system.

The OpenMI is a protocol for how models exchange data during a simulation run. It defines the interfaces for models so that two models can be coupled in terms of shared boundary conditions (Figure 8-1). The OpenMI was designed and developed through a project co-funded by the European Water Directive with participation from various research and modeling software companies in Europe. The Center for Ecology and Hydrology, UK, DHI, Wallingford Software, and Delft Hydraulics (now Deltares) are among the participants. Nearly €10 million has been invested in the standard and technical implementation of the standard, which is freely available and open source. The first version of the OpenMI was released in 2005. The standard is controlled by the OpenMI Association and is not in the control of any single company or organization. OpenMI has been or is in the process of being implemented for many different hydrology models (e.g. MIKE-SHE, Delft-3D, HEC-RAS, Modflow, and SWAT), but it could be used to wrap any hydrology process or set of processes into a linkable modeling component. More information about the OpenMI standard is available at www.openmi.org.

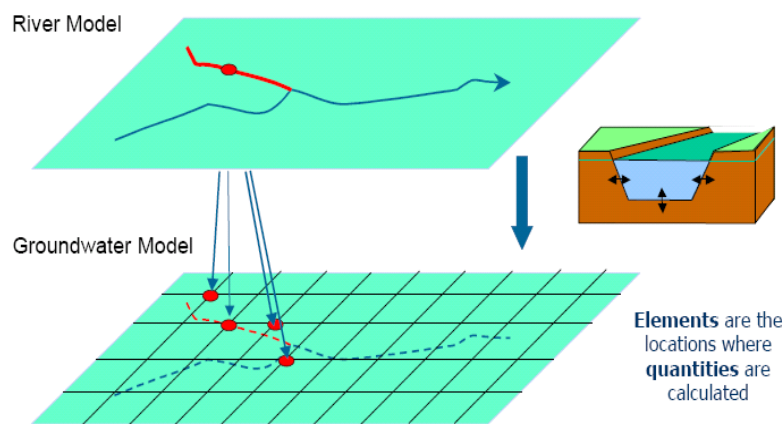


Figure 8-1 OpenMI defines a standard interface so that models can exchange values during a simulation run. For example, a groundwater model and river hydraulics model could be coupled through the exchange of groundwater heads and river seepage rates.

The OpenMI concept of a linkable component can be used to wrap models, databases, web services, file directories, or any other resource used in a modeling system that needs to share data with other resources within the system. Each component is a modular part of the system and can evolve separately from the other parts of the system, as long as it maintains its interface specification. The goal of a modeler when using component-based modeling systems is to define a configuration that specifies how components within the system are linked together in terms of data flow. For example, a forcing variable like precipitation might be available through HydroLink and delivered to a rainfall/runoff model component as a boundary condition. By standardizing the interface for each component within the system, OpenMI achieves modularity and expedites the process of transferring data between subcomponents of large, complex modeling and information systems.

OpenMI is a solution to the problem of integrating hydrologic models just as the HIS is a solution to the problem of integrating hydrologic data. The OpenMI and HIS, therefore, share similar ideas. For example, just as the WaterOneFlow web service defines a standard interface for describing and accessing data repositories, the OpenMI defines a standard interface for describing and executing models. Likewise, just as the OpenMI includes an object model for communicating data between models, the HIS defines the WaterML schema as an object model to communicate water observations between clients and servers. The two technologies leveraged together create a scalable hydrologic information system that includes both data and model interoperability.

8.2 WHAT IS HYDROLINK?

HydroLink is an OpenMI-compliant client application to the CUAHSI HIS data delivery services. From a user's perspective, HydroLink is an OpenMI linkable component that can be coupled to other OpenMI components to provide input data (e.g. boundary conditions, forcing data) to those components. From a programmers' perspective, HydroLink provides an Application Programming Interface (API) for reading HIS data and delivering this data to models. From either perspective, the goal is to define a component within a component-based modeling system that can serve HIS data to other components within that system.

8.3 HYDROLINK DESIGN

8.3.1 FILE CACHING

An important design decision in creating HydroLink was to incorporate an intermediate data cache between data delivery, accomplished through HIS web services, and the model input, accomplished by HydroLink. In other words, HydroLink does not make web service calls to CUAHSI Water Data Services during model execution, but instead works from a repository of WaterML files harvested before the model run begins. These files can be assembled from different CUAHSI Water Data Services or created by the modeler from other data unavailable from data services. However, although WaterML files can be created directly, it is highly recommended that the data used for modeling first be loaded into a HIS Server ODM database and then requested using WaterOneFlow web services. This is so that others can easily reproduce the modeling input files using the HIS.

WaterML files are the output from WaterOneFlow servers and provide an XML-based description of concepts such as an observational time series. While these WaterML files are typically transferred directly from a WaterOneFlow service call into a client program's memory, HydroLink requires that these files are cached into a particular directory structure that groups time series of the same quantity (or variable) into what OpenMI calls exchange items. HydroLink, therefore, references the path to this directory of cached WaterML files instead of

WaterOneFlow servers. This process of preparing WaterML files and using the HydroLink component are demonstrated in the *Use Case Study* section. We provide tools to aid in the process of creating WaterML caches including FetchWaterML and a WaterML class compiled into a .Net Assembly. These tools are discussed subsequently in the *Tools for Building a WaterML File Cache* section.

There are several reasons for choosing to implement this intermediate caching level within HydroLink design. First, the cache improves performance by eliminating the need for network service calls. Some web services, in particular those that expose large, federal databases, may take a few seconds to respond to a data request. By caching the data on the client-side, the latency issue is avoided. Second, the directory of WaterML files provides a clear documentation of the files used for a model run. The modeler can easily view these files using other applications and can edit the files to fill data gaps or incorrect values. Lastly, a modeler can archive the WaterML files in order to reproduce a model run. Presently, the CUAHSI Water Data Services are intended for data delivery and not data preservation. Therefore there is no guarantee that the same data request will always return the same information. This is a concern for modeling and is avoided by using a data cache. If there are changes to the HIS data, modelers can view the new data in relation to the cached data and make a decision about updating the data cache.

8.3.2 TRANSLATION FROM WATERML TO OPENMI

The OpenMI is designed specifically for modeling, thus it provides organization and access to data designed to accommodate modeling activities. A useful way to view HydroLink is that it provides access to HIS observational data in terms of OpenMI concepts instead of WaterML and WaterOneFlow concepts. The primary task implemented within HydroLink, therefore, is translating WaterML concepts into OpenMI concepts. The mapping between the WaterML and OpenMI object models is presented in Figure 8-2.

Some OpenMI concepts are not expressed in WaterML, for example unit conversions and dimensions. For the most part, however, the information expressed in WaterML can be used to populate the OpenMI objects. Future work will be aimed adding to the WaterML specification some elements that express unit conversion factors and dimensions.

8.3.3 STRUCTURE OF THE WATERML FILE CACHE

HydroLink requires a directory of WaterML files structured in a particular way. This directory is referred to as a cache because it is a representation of data that can be easily reproduced by a server (or servers). The WaterML file cache includes a grouping of individual WaterML files into folders where each folder represents an OpenMI Exchange Item. An OpenMI Exchange Item, as expressed in Figure 8-2, is composed of a quantity and an element set where values are measured or modeled over some period of time. WaterML files, on the other hand, contain data for one time series (one location, one variable, over time). Thus, by grouping WaterML files into folders, it is possible to represent OpenMI exchange items if all of the WaterML files within a folder have the same quantity and same time horizon. For example, to represent a collection of rainfall gaging stations as an OpenMI exchange item, each gage's precipitation time series is stored in a single WaterML file and the WaterML files for all of the gages are put into a folder within the WaterML cache directory (Figure 8-3). This design was followed because it was a simple and transparent way of grouping WaterML time series into OpenMI exchange items.

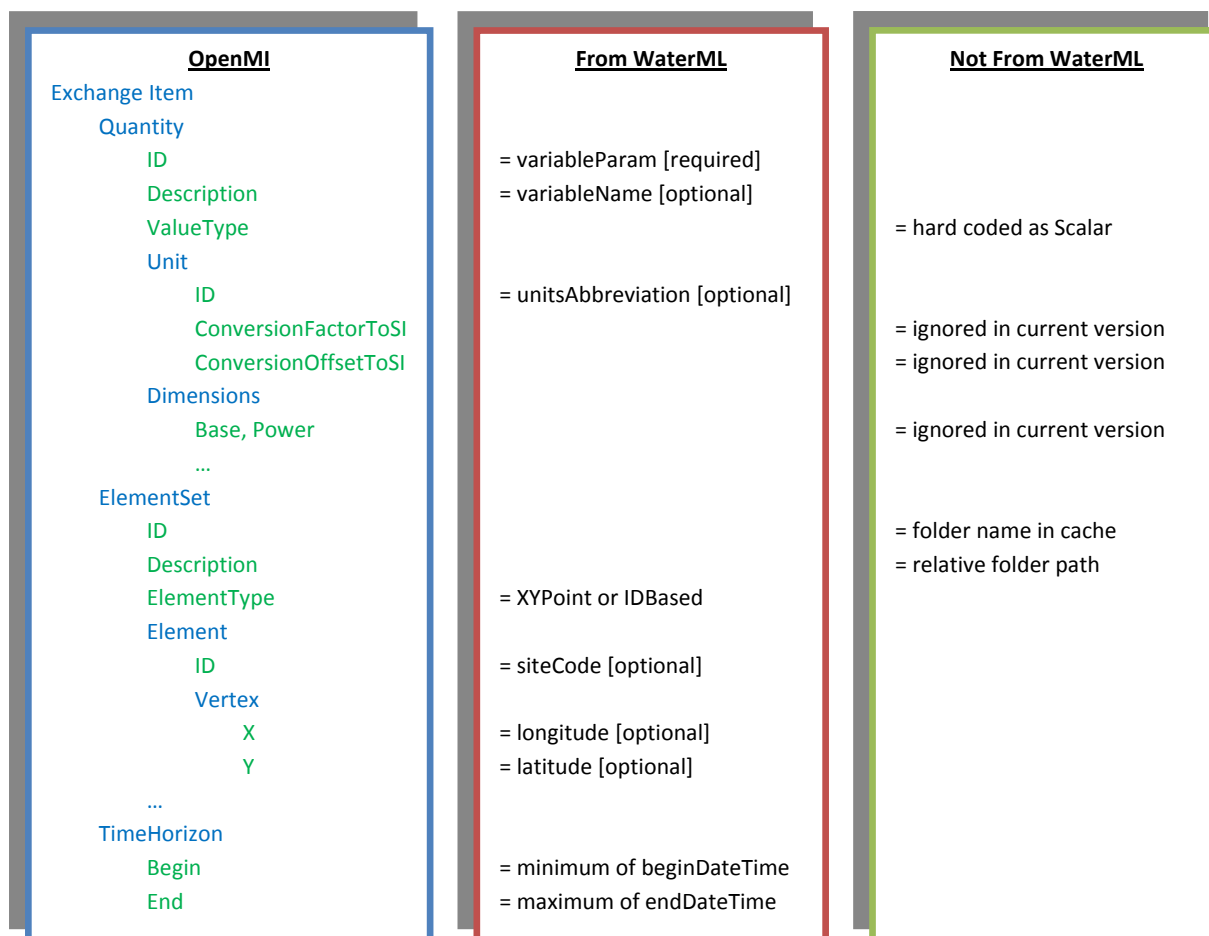


Figure 8-2 Mapping of concepts between WaterML and OpenMI. Blue text indicates objects while green text represents attributes.

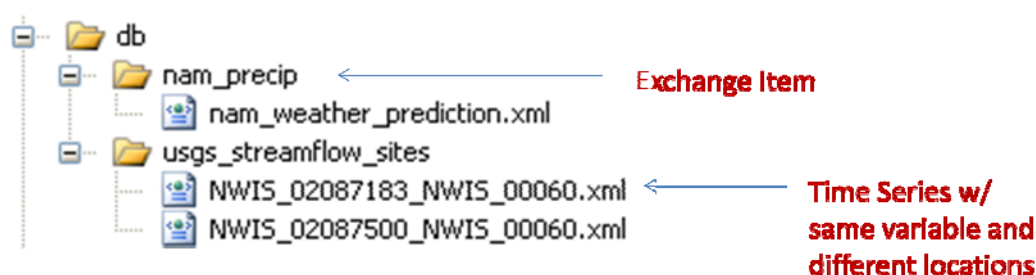


Figure 8-3 Example structure of a WaterML cache

It should be noted that the current design is meant to maximize ease of use and may have performance limitations for some types of modeling activities due to the verbose nature of XML. If the speed of reading the contents of the cached WaterML files becomes unacceptable, it is possible to add other file formats for storing the cached WaterML information beside XML. For example, a modeling file format like HDF or netCDF, or a relational database could supply the backend data storage for HydroLink without disrupting models that rely on HydroLink.

Feedback from users will determine whether such additions to the component are necessary, but it is important to realize that there is a clear path forward to improve data reading performance if it becomes a bottle neck.

8.4 TOOLS FOR BUILDING A WATERML FILE CACHE

8.4.1 FETCHWATERML

Building a WaterML file cache is likely the most time consuming process in using HydroLink. While other HIS tools aid in the process of finding relevant data stored in the HIS (e.g. HydroSeek), a simple downloading tool called FetchWaterML is provided to batch download a set of WaterML files from different WaterOneFlow servers. The tool takes as input a simple XML file that specifies the parameters for a set of GetValuesAsObject WaterOneFlow requests (Figure 8-4). It processes this input file and creates a WaterML file for each time series.

```
<?xml version="1.0" encoding="utf-8" ?>
<timeseries_catalog>
  <timeseries>
    <location>NWIS:02169506</location>
    <variable>NWIS:00065</variable>
    <begin>2008-06-25</begin>
    <end>2008-06-26</end>
    <wsdl>http://water.sdsc.edu/WaterOneFlow/NWIS/UnitValues.asmx</wsdl>
  </timeseries>
</timeseries_catalog>
```

Figure 8-4 Input XML file for the FetchWaterML tool. The tool will download a WaterML file for each time series element in this file.

```
// -- create an object to proxy the WaterOneFlow server (requires the
WaterOneFlow.dll assembly)
WaterOneFlow NWIS = new
WaterOneFlow(http://water.sdsc.edu/WaterOneFlow/NWIS/DailyValues.asmx?wsdl);

// -- request a time series from the server
TimeSeriesResponseType tsr = NWIS.GetValuesObject("NWIS:02087500",
"NWIS:02087500", "2000-01-01", "2000-02-01", "");

// -- serialize the data to an xml file.
XmlSerializer serializer = new XmlSerializer(typeof(TimeSeriesResponseType));
StreamWriter sr = File.CreateText("my_new_waterML_file.xml");
serializer.Serialize(sr, tsr);
```

Figure 8-5 A C# .Net code snippet that requests a time series from a WaterOneFlow server and serializes the time series response object file to disk in a WaterML file. The source code for FetchWaterML provides a more complete example. It is also possible deserialize an xml file into a programming object using the "Deserialize" method of the XmlSerializer class.

A C# .Net code snippet (Figure 8-5) shows how the tool uses a class to proxy a WaterOneFlow server, and then how the time series response object returned from that server can be written to an WaterML document using a process called serialization (serialization is described in more detail in the following section). Once the time series object is

serialized to the file, the modeler is free to examine, edit, and arrange these files into folders that represent OpenMI exchange items.

Because WaterML files are simply XML files, any text editor can be used to view and edit the contents of WaterML files. Applications like Excel provide enhanced viewing of WaterML files for those who are not accustomed to working with XML documents. Excel 2007 is able to load XML files and display their contents as tables. XML is a widely used file format, so there are many applications ranging from commercial to open source on likely every operating system available to view and edit WaterML files. The FetchWaterML tool is further described in the *Use Case Study* section.

8.4.2 CREATING NEW WATERML FILES

If one wishes to add data to HydroLink and the data is not available from a WaterOneFlow server, then a possible work around is to create a new WaterML file to store the data. One way to create new WaterML files is by leveraging XML serialization tools available through many programming languages. Serialization is the process of transferring objects between the program's memory and a persistent on disk file (usually XML-based). Understanding the tools available for serializing and de-serializing objects into and out of XML makes reading and writing WaterML files trivial. An alternative for creating WaterML files through serialization is to treat the XML as a standards ASCII text file. However, the best solution for creating new WaterML files is to load the data into an ODM database on a HIS Server and then to request WaterML files using WaterOneFlow web services. This not only simplifies the process of creating WaterML files, but it also allows others to reproduce modeling input datasets by simply knowing the parameters for a GetValues request.

8.5 IMPLEMENTATION

8.5.1 WINDOWS OPERATING SYSTEM

HydroLink is programmed in C# .Net and is freely available and open source (<http://his.cuahsi.org/openmi.html>). The .Net Framework is provided with the Windows Operating system. Version 2.0 of the Framework was used to create HydroLink. HydroLink is provided with a Windows installation file that includes an example WaterML cache and the necessary OpenMI assemblies to run the component.

8.5.2 OTHER OPERATING SYSTEMS

For those that prefer not to use a Windows operating system, there is a free, open source implementation of the .Net Framework called Mono (<http://www.mono-project.com>). Mono mimics the Microsoft .Net Framework and allows one to run .Net assemblies on various operating systems including Windows, Mac OS X, Linux, and Unix. HydroLink has undergone limited testing on the OpenSUSE 10.3 Linux operating system using Mono.

8.5.3 SCRIPTING USING PYTHON OR F# .NET

For those that prefer a scripting language, a possible approach for using HydroLink is through the Python .Net implementation called Iron Python (<http://www.codeplex.com/Wiki/View.aspx?ProjectName=IronPython>). Limited tests have been performed for running HydroLink using Iron Python. Alternatively, a language being designed by Microsoft Research called F# (<http://research.microsoft.com/fsharp/fsharp.aspx>) provides Python-like

syntax and also targets the .Net Framework. HydroLink has not been tested with F#.Net, but should work properly this F# is designed specifically to target the Microsoft .Net Framework.

8.6 USE CASE STUDY

As an example use case study for HydroLink to demonstrate how it would be used with an OpenMI-compliant model, consider the task of estimating water velocity in a river channel using Manning's Equation given the stage of the river, cross section at the gaging station, and a Manning's roughness coefficient. The tasks needed to complete this simple example are (1) download stage data from the USGS National Water Information System in WaterML format using FetchWaterML, (2) create a WaterML cache directory to store this file as an OpenMI exchange item, (3) link HydroLink with an OpenMI-compliant Manning's equation component to estimate velocity from stage given data about the channel cross section, slope, and roughness. This chapter will not go into detail in how the Manning's component was created, however readers are encouraged to contact the authors for further information on creating OpenMI-compliant models.

8.6.1 STEP 1: DOWNLOAD THE WATERML FILE

This example uses the Rocky Branch USGS station in Columbia, SC that measures gage height in real-time. The station identifier is 02169506 and the variable code for gage height is 00065. The FetchWaterML tool requires these values to be stored in an XML file named sites.xml that has a very simple structure (Figure 8-6).

FetchWaterML could be used to batch download a set of time series, even if those time series are stored in different WaterOneFlow servers, by repeating the time series element within the sites.xml file.

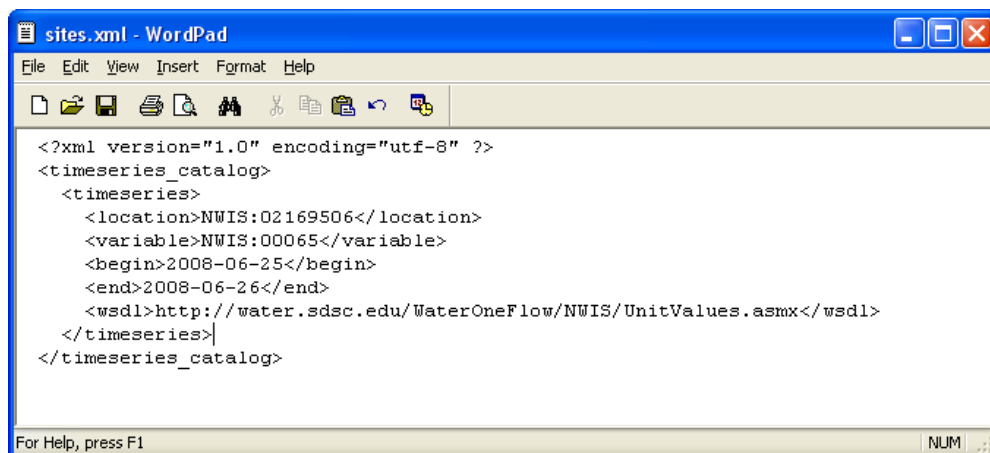


Figure 8-6 Input file for FetchWaterML tool. Each time series element is written to a WaterML file. [Note, the time period might need adjustments because only the previous 30 days are available from the server]

The resulting WaterML file is displayed as raw XML (Figure 8-7) and as a table in Excel 2007 (Figure 8-8). Other applications capable of working with XML files could be used to view and edit this WaterML file as necessary prior to running the model.

```

<?xml version="1.0" encoding="UTF-8" ?>
- <timeSeriesResponse xmlns:gml="http://www.opengis.net/gml" xmlns:xlink="http://www.w3.org/1999/xlink"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:wtr="http://www.cuahsi.org/waterML/" xmlns="http://www.cuahsi.org/waterML/1.0/">
- <queryInfo>
  <creationTime>2008-06-26T13:24:29.25-07:00</creationTime>
- <criteria>
  <locationParam>NWIS:02169506</locationParam>
  <variableParam>NWIS:00065</variableParam>
- <timeParam>
  <beginDateTime>2008-06-25</beginDateTime>
  <endDateTime>2008-06-26</endDateTime>
  </timeParam>
</criteria>
<note href="http://waterdata.usgs.gov/nwis/help/?provisional" title="USGS Data Provisional">All data are
  provisional, and subject to revision</note>
<note title="USGS URL">http://nwis.waterdata.usgs.gov/nwis/uv?format=rdb&date_format=YYYY-MM-
  DD&&site_no=02169506&parameter_cd=00065&date_format=YYYY-MM-DD&begin_date=2008-06-
  25&end_date=2008-06-26</note>
</queryInfo>
- <timeSeries>
- <sourceInfo xsi:type="SiteInfoType">
  <siteCode />
</sourceInfo>
- <variable oid="12582">
  <variableCode vocabulary="NWIS" default="true">00065</variableCode>
  <variableName>Gage height, feet</variableName>
  <dataTvne>Instantaneous</dataTvne>

```

Figure 8-7 A WaterML file of Gage Height for Rocky Branch, South Carolina viewed as raw XML (this is an incomplete file)

	ns1:UnitType	ns1:UnitAbbreviation	ns1:timeInterval	count	ns1:value	qualifiers	dateTime	ns1:q
11	Time	min		15	155	1.1 P	6/25/2008 1:45	Provisional data su
12	Time	min		15	155	1.1 P	6/25/2008 2:00	Provisional data su
13	Time	min		15	155	1.1 P	6/25/2008 2:15	Provisional data su
14	Time	min		15	155	1.1 P	6/25/2008 2:30	Provisional data su
15	Time	min		15	155	1.1 P	6/25/2008 2:45	Provisional data su
16	Time	min		15	155	1.1 P	6/25/2008 3:00	Provisional data su
17	Time	min		15	155	1.1 P	6/25/2008 3:15	Provisional data su
18	Time	min		15	155	1.1 P	6/25/2008 3:30	Provisional data su
19	Time	min		15	155	1.1 P	6/25/2008 3:45	Provisional data su
20	Time	min		15	155	1.11 P	6/25/2008 4:00	Provisional data su
21	Time	min		15	155	1.1 P	6/25/2008 4:15	Provisional data su
22	Time	min		15	155	1.1 P	6/25/2008 4:30	Provisional data su
23	Time	min		15	155	1.1 P	6/25/2008 4:45	Provisional data su
24	Time	min		15	155	1.1 P	6/25/2008 5:00	Provisional data su
25	Time	min		15	155	1.1 P	6/25/2008 5:15	Provisional data su
26	Time	min		15	155	1.1 P	6/25/2008 5:30	Provisional data su
27	Time	min		15	155	1.1 P	6/25/2008 5:45	Provisional data su
28	Time	min		15	155	1.1 P	6/25/2008 6:00	Provisional data su
29	Time	min		15	155	1.1 P	6/25/2008 6:15	Provisional data su
30	Time	min		15	155	1.1 P	6/25/2008 6:30	Provisional data su
31	Time	min		15	155	1.1 P	6/25/2008 6:45	Provisional data su

Figure 8-8 The same XML file viewed as a table in Excel 2007

8.6.2 STEP 2: BUILD THE WATERML CACHE DIRECTORY

The WaterML cache directory is simply a directory with folders at the root level that define exchange items and WaterML files within each of these root level folders that define the time series in an exchange item. In this simple example, there is only one exchange item represented by the folder stage_rocky_branch and one WaterML file within the exchange item stored in the WaterML file 02169506_00065.xml (Figure 8-9). HydroLink will look into this waterML file to determine the metadata about the time series (units, variable, site location, etc.)



Figure 8-9 An example WaterML cache directory

8.6.3 STEP 3: USE A MANNING'S COMPONENT TO ESTIMATE VELOCITY

OpenMI can be used to wrap any model, from a simple calculation like Manning's to a complex numerical code like Modflow. In each case, the model is wrapped so that it can take in and produce boundary conditions that are used to link that component with other components. Here we created a simple OpenMI component to perform Manning's Equation to estimate a velocity given the river stage and properties about the river cross section and roughness. HydroLink is linked to the Manning's component so that the stage data stored in WaterML can be utilized as input to the Manning's Equation component (Figure 8-10). The Manning's Equation component has a file that describes the river cross section and roughness, and it will produce a simple text file with the stream velocity time series.

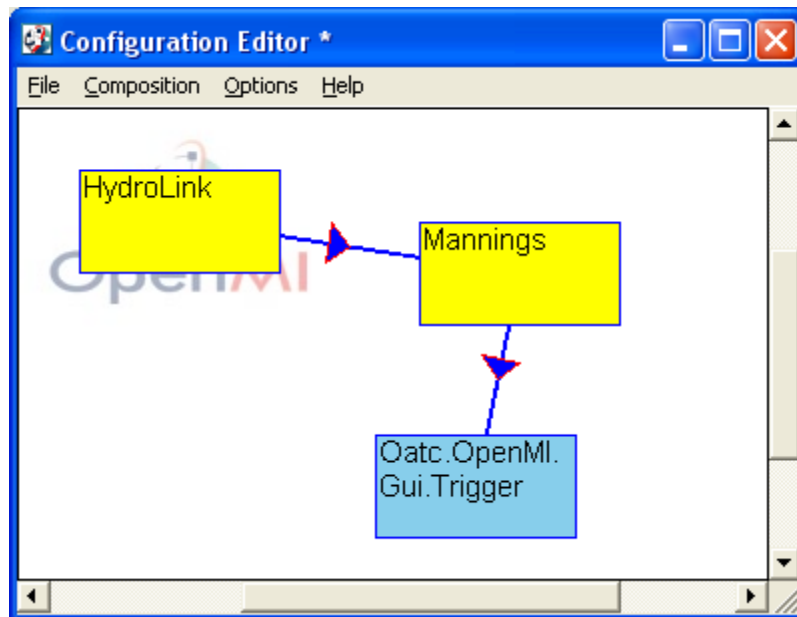


Figure 8-10 Linking HydroLink with other Models through the OpenMI Configuration Editor GUI application

The Manning's model writes an output file for velocity on each time step by making use of the stage values provided by the HydroLink component. These values are presented in Figure 8-11.

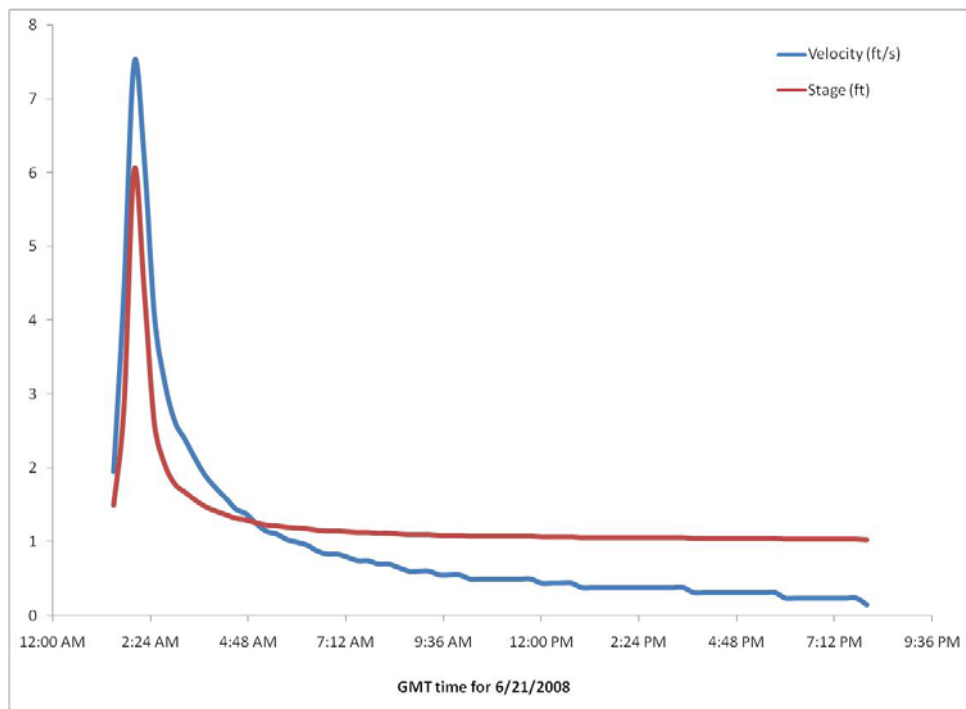


Figure 8-11 Stage and predicted velocity from the model

8.6.4 EXTENDING THE USE CASE STUDY

This simple example can be extended in a number of ways. Other components could be added to perform additional calculations, for example an equation to estimate discharge from velocity, or the Manning's component could be replaced with a rating curve calculation. It is also important to note that this configuration could be constructed programmatically as well as through a graphical user interface (GUI). One could create scripts that automate the process of adding components and configuring links between components, or one could edit the XML file that results from this Configuration Editor application that is provided by the OpenMI Association. These examples show how component-based modeling provides a scalable solution for representing complex configurations with many data resources and models interaction together.

8.7 ALTERNATIVE APPROACHES FOR USING HIS DATA IN MODELS

While OpenMI provides long term benefits of interoperability and component standardization, short term adoption of HIS data within models could be achieved by removing the OpenMI layer and directly connecting the HIS with a specific model. Three different approaches for using HIS data within models are presented in this section. When considering these approaches, it is important to realize that no single best approach exists for using HIS data within models. The best approach will always depend on the needs of the modeling application.

8.7.1 APPROACH 1: EXTRACT, TRANSFORM AND LOAD (ETL) FROM WATERONEFLOW TO A PREDETERMINED MODEL INPUT FILE FORMAT

Extract, Transform, and Load (ETL) is a common phrase used in data management to describe the process of moving data from one data model to another data model. For the case of the HIS, the source data model will most often be WaterML since it is the common language between different data providers. However, there may be cases where researchers wish to directly connect to their own ODM Relational Database Management System (RDMS) in which case the ODM schema is the source data model. The destination data model will describe the structure of the information required by the model. A script could be written to automate the ETL process of moving data from the HIS to the model input files.

This approach has the benefit that it requires no changes to the model input data structure and format. Because the HIS is web accessible and provide access to federal databases, scripts can be reused to more quickly construct a model for a particular place, to check for new data present in the HIS, or to allow colleagues to recreate the model input files directly from the HIS. The modeler would be required to load their data into the HIS, but the benefit would be that the data is preserved yet accessible using the HIS protocols. The primary disadvantage of this approach is that each model will require its own code to automate the ETL process.

8.7.2 APPROACH 2: EMBED HIS WEB SERVICE CALLS WITHIN A MODEL

In the first approach, the HIS and models are kept as separate entities. A transformation routine is used to extract, transform, and load HIS data into a model input file, but the model itself has no ability to directly request data from the HIS. Some models may benefit from a more direct access to HIS data by embedding the code for performing WaterOneFlow service calls within the model. The model could reference a CUAHSI Water Data Service WSDL address and dynamically pull data from this server from within the model. Because these services follow industry standards for web services (SOAP, WSDL), there are a number of software libraries available for accessing the CUAHSI HIS and connecting it to models. The disadvantage of this approach is that it relies on a remote server that may have significant latency in responding to data requests, or may have downtimes in which the model would be unable to run. These disadvantages could be minimized by using caching and multithread processing, however the approach of relying on remote servers is only recommended if there is a strong need for access to data that is frequently updated.

8.7.3 APPROACH 3: EMBED WATERML SERIALIZATION WITHIN A MODEL

The third approach is similar to the second approach except that instead of making web service calls, the model is written to make use of WaterML files as a data input file. The WaterML files would be stored in a local directory to insure data is available for model runs. The WaterML files could be automatically updated or refreshed from WaterOneFlow servers. Many programming languages provide tools for serializing XML documents into objects that could be used to easily read WaterML files within a model. This is the process used in HydroLink when reading WaterML files and remapping WaterML objects into OpenMI objects. It could be used within a model to remap WaterML objects directly into modeling objects without going through the OpenMI API.

8.8 SUMMARY REMARKS AND FUTURE WORK

The purpose of this chapter is to present approaches for using HIS data in models. A recommended approach is to use the HydroLink tool that uses the Open Modeling Interface (OpenMI) as an intermediate layer between the data and models. OpenMI provides a common communication language between components within a modeling system which will allow for interoperability between models and data sources from different organizations. It is important to note, however, that using HIS data in models does not require the adoption of the OpenMI. If modelers prefer not to use the Open Modeling Interface, they are free to use the HIS data directly in their models without going through an OpenMI layer. Three possible approaches are presented as alternatives for using HIS data directly within models.

Future work will be directed toward improving HydroLink, specifically in how the component is able to perform "on-the-fly" data transformations when the data required by a model does not exactly match the data available in WaterML files. The OpenMI follows the mantra "ask for what you need, get what you ask for", which means that the HydroLink component should include the logic to rescale WaterML data in space and time, and perform unit conversions to meet the needs of a requesting model. For example, if WaterML files have time series on a 15 minute time step, and a model requires as input a time series on a 1 hour time step, the HydroLink component should be able to perform this data transformation, saving the modeler from creating a new WaterML file with 1 values on a 1 hour time step. OpenMI defines the concept of a data operation for encapsulating such algorithms, and the OpenMI Software Development Kit (SDK) provides some basic data transformations that can be reused within HydroLink (i.e. linear interpolation, weighted average, etc.). Implementing data transformations within HydroLink and documenting how to correctly apply these data operations to transform WaterML data for modeling will make HydroLink an even more powerful tool for supplying HIS data to models.

Chapter 9. CONCLUSIONS

By David Maidment, The University of Texas at Austin

This report summarizes the current status of a large NSF-supported research and development effort that has taken place over a period of several years. The goal of NSF research is to be “transformative”, which is described by the National Science Board (2007) as “research that has the capacity to revolutionize existing fields, create new subfields, cause paradigm shifts, support discovery, and lead to radically new technologies”. Although it may be a little presumptive to claim this, the authors of this report believe that their research is transformative in this sense.

We have achieved a genuine synthesis of knowledge between hydrologic science and computer science, hydrologic science bringing the assessment of what is needed, and computer science bringing the capacity to meet those needs with a radically new approach, namely a service-oriented architecture for water resources data. This provides hydrologic scientists with a capacity to publish their water data, and to seamlessly integrate those data with corresponding water data derived from observations by public agencies at all levels of government. We have found a way to build and implement a common water data sharing language called WaterML in a geographically and institutionally distributed information environment that is otherwise a “Tower of Babel” of incompatible websites, data formats and descriptions. This accomplishment is unprecedented in the water resources field in the United States or elsewhere. Within the large water data collection agencies of the federal government, the perception is that the water data services approach developed by CUAHSI HIS has “broken through the stovepipes” and is a principal means by which water data for the nation will be provided using web data services in the future.

The CUAHSI HIS team has worked with eleven partner universities in the WATERS testbed network and implemented with them the Hydrologic Information System described here so that they are each producing consistent water data services from their measured information even though the science goals of the individual projects are very different from one another. The CUAHSI Observations Data Model and its accompanying WaterOneFlow web services have proven to be an extremely robust and reliable means of storing and providing internet access to these data. By this means, water data collected by individual hydrologic scientists can be published and made permanently informative in the same manner as is the data collected by the US Geological Survey, the nation’s principal water science agency. The USGS and the National Climatic Data Center are also beginning to publish their data using the WaterML language.

While much has been accomplished, it is also important to recognize that the gap between what is available and what is needed is still large. In particular, our CUAHSI Hydrologic Information System has the following limitations:

- It focuses on one class of information, water observations data measured at fixed point locations;
- While the observations data model can store “collections” of hydrologic data collected in field investigations because each data value is stored and indexed individually, our methods for indexing and accessing data are oriented around time series of data, rather than collections of many variables measured simultaneously at one place and time;
- Research on other classes of hydrologic data, such as GIS, weather and climate grids, and remote sensing has been addressed to some degree but a true synthesis across these information classes has not yet been achieved;
- Our work on remote sensing is especially limited and we have not yet found a feasible way to extend our services-oriented architecture concept to remote sensing data.

Looking back over the past several years, reveals an experience of growth and discovery for our HIS team, both in defining what needs to be done and how it might be done, but also in learning how to transition from a group that does interesting research to a group that produces and supports robust and useful technology. Probably, to a

certain degree, our team is still on this learning curve. This report is a benchmark that helps us assess where we are and what we have accomplished to date. It will help us plan our work into the future in a more rational way. We hope that the information presented here is informative to you and that you will find the technology described here helpful in your work. For more information, please go to our web site: <http://his.cuahsi.org>.

9.1 REFERENCES

National Science Board (2007), Enhancing support of transformative research at the National Science Foundation, Document NSB-07-32, National Science Board, Washington DC, 29p., http://www.nsf.gov/nsb/documents/2007/tr_report.pdf