Environmental Modelling & Software 52 (2014) 62–73

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Managing a community shared vocabulary for hydrologic observations

Jeffery S. Horsburgh[a,*], David G. Tarboton[a], Richard P. Hooper[b], Ilya Zaslavsky[c]

[a] Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, USA
[b] Consortium of Universities for the Advancement of Hydrologic Science, Inc., Medford, MA, USA
[c] San Diego Supercomputing Center, University of California San Diego, San Diego, CA, USA

ARTICLE INFO

Article history:
Received 20 February 2013
Received in revised form
12 October 2013
Accepted 15 October 2013
Available online

Keywords:
Controlled vocabulary
Data publication
CUAHSI HIS
Hydrologic observations

ABSTRACT

The ability to discover and integrate data from multiple sources, projects, and research efforts is critical as scientists continue to investigate complex hydrologic processes at expanding spatial and temporal scales. Until recently, syntactic and semantic heterogeneity in data from different sources made data discovery and integration difficult. The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS) was developed to improve access to hydrologic data. A major semantic challenge related to data sharing and publication arose in development of the HIS. No accepted vocabulary existed within the hydrology research community for describing hydrologic observations, making it difficult to discover and synthesize data from multiple research groups even if access to the data was not a barrier. Additionally, the hydrology research community relies heavily on data collected or assembled by government agencies such as USGS and USEPA, each of which has its own semantics for describing observations. This semantic heterogeneity across data sources was a challenge in developing tools that support data discovery and access across multiple hydrologic data sources by time, geographic region, measured variable, data collection method, etc. This paper describes a community shared vocabulary and its supporting management tools that can be used by data publishers to populate metadata describing hydrologic observations to ensure that data from multiple sources published within the CUAHSI HIS are semantically consistent. We also describe how the CUAHSI HIS mediates across terms in the community shared vocabulary and terms used by government agencies to support discovery and integration of datasets published by both academic researchers and government agencies.

1. Introduction

The ability to discover and integrate data from multiple sources, projects, and research efforts is critical as scientists continue to investigate complex hydrologic processes at expanding spatial and temporal scales. Until recently, syntactic and semantic heterogeneity in data from different sources made data discovery and integration difficult (Bergamaschi et al., 2001; Lutz et al., 2009). Data from different sources were typically encapsulated within files or databases with unique structures and schemas (Pallickara et al., 2010). Semantic heterogeneity across sources made data discovery, integration, and synthesis difficult given that different investigators commonly use different terms to describe the same concepts, sometimes disagree about the meaning of terms, and rarely share observational data in such a way that they are annotated with sufficient attribute information, or metadata, to make their interpretation unambiguous by anyone other than the data originators (Beran et al., 2009; Haines et al., 2012).

The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS) was developed to improve access to hydrologic data and to address these needs within the hydrology research community. The vision of CUAHSI HIS is to bring together hydrologic observations from multiple sources across the United States into a uniform, standards-based, service-oriented environment where heterogeneous data can be seamlessly integrated for advanced computer-intensive analysis and modeling (Tarboton et al., 2009, 2011). CUAHSI HIS includes several software tools and standards that together enable publishing and accessing hydrologic data collected at point locations (e.g., time series of observations from stream gages, water quality monitoring sites, weather stations, etc.). Since the introduction of CUAHIS HIS, a national network of servers on which hydrologic data are published has emerged. These servers host data from numerous sources, including the United States Geological

* Corresponding author. Tel.: +1 435 797 2946.
E-mail address: jeff.horsburgh@usu.edu (J.S. Horsburgh).

1364-8152/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.10.012

Survey's (USGS) National Water Information System (NWIS), the U.S. Environmental Protection Agency's (USEPA) STORage and RETrieval (STORET) system, and from academic research groups collecting data within experimental watersheds across the United States.

Each hydrologic data server participating in the CUAHSI HIS hosts one or more water data web services that publish observational data on the Internet using a standard format, or syntax called Water Markup Language (WaterML) (Zaslavsky et al., 2007; Version 1.1 at the time of this writing). This federated system of observational data web services comprises perhaps the largest repository of syntactically homogenous hydrologic observations of its kind in the world. As of December 2012, there were 97 water data web services registered with the CUAHSI HIS, having data for approximately 2.8 million sites and observations of over 32,000 variables. The CUAHSI HIS provides access to nearly 34 million observational time series comprised of approximately 18 billion individual observations, all of which can be accessed in WaterML 1.1 format.

In developing CUAHSI HIS, semantic challenges related to data sharing and publication arose. As in other scientific communities (e.g., Lutz et al., 2009; Graybeal et al., 2012), there was no single, common vocabulary for describing hydrologic observations within the hydrology research community. Here, we define a vocabulary as a set of terms used to describe particular units of information. A controlled vocabulary is a set of consistent terms used within a specific knowledge domain (Ma et al., 2010), and a community shared vocabulary is a controlled vocabulary in which the terms have been selected and agreed upon by a community of people. Given that most hydrologic research has been done by individual investigators or small groups of investigators on project-based funding, data management within the hydrology research community has been done on a project or research group basis. Each group defined data management practices and semantics, selecting vocabulary terms to meet the needs of their project or group and making it difficult to synthesize data from multiple research groups even if access to the data was not a barrier. Additionally, the hydrology research community relies heavily on data collected or assembled by government agencies such as USGS and USEPA, each of which has its own vocabulary for describing observations driven by the agency's needs and culture. This semantic heterogeneity across data sources has been a challenge in developing tools that support data discovery and access across multiple hydrologic data sources by time, geographic region, measured variable, data collection method, etc.

While others have described the established practices and differences in how names and descriptions of observed hydrologic variables are published by different agencies and research groups (Zaslavsky et al., 2012), the technical approaches to management of semantics within the CUAHSI HIS have been only partially described (e.g., Piasecki and Beran, 2009). These previous efforts also focused primarily on the semantics surrounding the names of measured variables and have not addressed other important attributes used to describe hydrologic observations, each of which require differing levels of community consensus and enforcement to support data discovery and interpretation. For example, our experience in developing the CUAHSI HIS showed that community consensus and enforcement of semantics surrounding names of measured variables was more important than conventions for naming monitoring sites in supporting data discovery.

In this paper, we describe a major component of the CUAHSI HIS that aids in building community consensus around terms used to describe hydrologic observations – a centrally-managed, community shared vocabulary and associated management tools. The community shared vocabulary consists of a set of term lists that contain community vetted and accepted terms for describing metadata elements for hydrologic observations data. The intent of the shared vocabulary is to reduce semantic heterogeneity in data published by multiple sources by providing consistent terms that data publishers can use to describe their data and that data consumers can use to discover and interpret the data. We also describe how the CUAHSI HIS mediates across terms in the shared vocabulary and terms used by government agencies to support discovery and integration of datasets published by both academic researchers and government agencies.

Section 2 provides background, describing the problem of semantic heterogeneity in hydrologic observations. Section 3 describes the CUAHIS HIS. Section 4 illustrates specific semantic challenges in the design of HIS. Section 5 describes the cyberinfrastructure components that supported development and use of the shared vocabulary, and Section 6 describes the use of these components within the CUAHSI HIS to mediate across the vocabularies of data from academic sources and those from major agencies in support of data publication, cataloging of metadata, and enabling discovery and access.

2. Semantic heterogeneity in hydrologic observations

The problem of semantic heterogeneity in environmental datasets is well described (Beran and Piasecki, 2009; Lutz et al., 2009; Graybeal et al., 2012). Lutz et al. (2009) summarize the problem at three levels: 1) at the metadata level, semantic heterogeneity impedes the discovery of data; 2) at the schema level, semantic heterogeneity impedes the retrieval of data; and 3) at the content level, semantic heterogeneity impedes the interpretation, integration, and exchange of data.

One approach for providing data discovery services across multiple hydrologic data sources or repositories relies on cataloging metadata describing data available from each source. Piasecki and Beran (2009) report that overcoming semantic heterogeneities in the metadata across hydrologic repositories is one of the most difficult challenges in creating a large-scale metadata catalog because the meaning of the words in the vocabularies used by each data source and their intentions are subject to different interpretation. Indeed, in many cases it is difficult to assert whether terms used by different data sources are equivalent without extensive knowledge of the data and vocabularies used by each source.

At the schema level, the way data are formatted or encoded can constrain the ease with which they can be retrieved and used. For example, the use of discipline specific file formats or proprietary software and data formats can impede access to and reuse of data. There are, however, many notable efforts underway to develop both standard web service interfaces and standard data encoding schemas for exchanging data, including a host of standards being developed under the umbrella of the Open Geospatial Consortium (OGC). These include Web Map Services (WMS), Web Feature Services (WFS), and Web Coverage Services (WCS) as standard web service interfaces for geospatial datasets, a suite of standards under OGC's Sensor Web Enablement initiative for publishing observational data on the Internet, the development of the WaterML 1.1 (Zaslavsky et al., 2007) data encoding schema by the CUAHSI HIS project, and the recent development and acceptance of WaterML 2.0 as an OGC standard for encoding time series of hydrologic observations for transfer over the Internet (Taylor, 2012).

While data encoding schemas such as WaterML solve much of the syntactic heterogeneity among datasets, in many applications they are essentially containers for data. The semantics of the container are defined (i.e., the names of the elements within the schema and the structure in which they appear are set), but it is left to the user to determine which vocabulary terms to use for specifying content inside the elements of the schemas. There is not always agreement among data publishers within a community about the source of the terms to be used or whether/how the terms should be constrained, leading to gaps in community consensus around the use of terms and how they should be enforced. At the content level, then, there is also an opportunity to define common semantics for use by a community in populating the elements that describe data to ensure that terms used to describe a concept are consistent across all instances of that concept.

3. The CUAHSI HIS service oriented architecture

The design of the CUAHSI HIS follows an open, service-oriented architecture (SOA) model. SOA relies on a collection of loosely-coupled, self-contained services that communicate through the Internet and can be called from client applications in a standard fashion (Erl, 2005; Josuttis, 2007; Goodall et al., 2008). At the physical level, CUAHSI HIS infrastructure is comprised of a collection of computer servers that publish hydrologic observations data using a standardized web services protocol. Such servers can operate by means of the CUAHSI "HydroServer" software stack (Horsburgh et al., 2010) or by means of custom software that exposes the same web service interfaces. There is now a large and growing amount of hydrologic observations data available via web services published in this fashion (Horsburgh et al., 2009).

The core HIS SOA web services, called WaterOneFlow, provide a uniform protocol for accessing repositories of water observations data. Each HydroServer hosts one or more WaterOneFlow web services, each of which contains two types of web service methods: 1) a data delivery method called GetValues, which publishes the values of water observations; and 2) metadata delivery methods, including GetSites, GetSiteInfo, GetVariables, and GetVariableInfo, which identify and describe collections or series of data values associated with particular spatial locations. Results from WaterOneFlow web service calls are returned in WaterML 1.1 format, and because the WaterOneFlow web services and WaterML 1.1 standardize the data access protocol and data encoding language used to transmit data between servers and clients, data from every HydroServer are syntactically and, to some degree, semantically similar.

Even though the WaterML schema provides a consistent syntax, the semantic content of the WaterML elements returned from web service calls is dependent on the underlying data storage system. Data published using WaterOneFlow services may be stored on a HydroServer in a CUAHSI Observations Data Model (ODM) database (Horsburgh et al., 2008). In this case, both data delivery and metadata web service methods serve data directly from the ODM database. The information model that is shared by ODM and WaterML 1.1 resulted from a community consensus process aimed at identifying metadata elements required for unambiguous interpretation of hydrologic observations. Use of ODM for storing observations ensures that they are published and transmitted with metadata that conform to this community consensus process. The combination of ODM, WaterOneFlow, and WaterML 1.1 make it possible for disparate investigators and organizations to publish their hydrologic observations using a common protocol, in a common markup language with common syntax, thus alleviating much of the heterogeneity in hydrologic datasets from different sources (Horsburgh et al., 2009).

In other cases, for example with some government agency datasets, WaterOneFlow web services have been created by CUAHSI as an intermediary layer on top of an existing database structure, with metadata methods supplied by CUAHSI using copies of the data provider's metadata. These cases arise because of a particular agency's need to maintain existing internal data structures and due to the lack of a mechanism for publishing metadata in the format required by a WaterOneFlow web service. In these cases, CUAHSI serves as a mediator, providing a WaterOneFlow web service that translates the syntax and, in some instances, the semantics of the agency database to the standard formats and semantics of the CUAHSI HIS.

WaterOneFlow web services are registered with a central metadata cataloging service called HIS Central. Upon registration of a new web service (and regularly updated afterward), the central metadata catalog executes web service calls and harvests all of the metadata describing the datasets provided by the registered service. The harvested metadata are stored in a metadata catalog database, which then serves as the basis for a data discovery web service interface that can be accessed by client applications to search the contents of all of the web services registered with HIS Central.

4. Semantic challenges in the CUAHSI HIS

Given that the CUAHSI HIS exposes data from both academic sources that use ODM and agency sources that do not, two specific challenges emerged in development of the CUAHSI HIS. First, existing agency data systems (e.g., USGS NWIS and EPA STORET) had equivalent elements for some, but not all of the elements needed to satisfy the ODM/WaterML information model. Additionally, in some cases agency data systems also combined multiple concepts from the ODM/WaterML information model into single terms. For example, Fig. 1 shows a portion of the WaterML 1.1 encoding for streamflow data collected by the USGS and by Utah State University at two different stream gages in the Little Bear River of northern Utah, USA. The data were retrieved from separate WaterOneFlow web services, one for each data source. While the WaterML elements are the same, fewer of the WaterML elements are populated in the data from USGS, whereas the data from Utah State University, which are stored in an ODM database, expose a richer set of descriptive information about the measured variable. In the USGS data, the units element contains only an abbreviation for the units, but the variableName element contains both the name of the variable and unit information. In the Little Bear River data, the name of the variable and the units are provided in separate elements, along with the additional elements (e.g., valueType, dataType, generalCategory, sampleMedium, speciation) whose contents conform to controlled term lists and provide a more detailed description of the variable. The challenge, then, was one of structural semantic heterogeneity resulting from the fact that some existing data sources lacked values for some metadata elements and overloaded multiple concepts into other elements. We wanted to minimize this in new datasets added to the system.

Second, while ODM, WaterOneFlow, and WaterML 1.1 provide a common information model, web service protocol, and syntax for publishing hydrologic observations on the Internet, consistency in the set of terms used to populate the elements of the WaterML 1.1 schema could not be enforced across all data sources because we could not change the vocabularies used by major agency data sources. This led to contextual semantic heterogeneity among data sources. For example, the portion of the WaterML 1.1 schema that describes the variable that was observed includes an attribute called "VariableName." It is clear that this element should contain a name for the observed variable. However, because each data provider may use a different list of variables names, the WaterML 1.1 schema does not constrain how variables can be named or the list of terms that are acceptable as variable names. The same is true for many of the other elements in the WaterML 1.1 schema as demonstrated in Fig. 1, where the terms used to populate the WaterML elements for the two data sources are not the same.

```
- <variable>
    <variableCode network="NWISUV" default="true"
      vocabulary="NWISUV">00060</variableCode>
    <variableName>Discharge, cubic feet per second</variableName>
    <variableDescription>Streamflow, ft3/s</variableDescription>
    <valueType>Derived Value</valueType>
    <dataType>Unknown</dataType>
    <units unitsCode="cfs"/>
  - <options>
      <option name="Statistic" optionCode="00011"/>
    </options>
    <NoDataValue>-999999.0</NoDataValue>
  </variable>
```

(a)

```
- <variable>
    <variableCode variableID="44" default="true"
      vocabulary="LittleBearRiver">USU44</variableCode>
    <variableName>Discharge</variableName>
    <valueType>Derived Value</valueType>
    <dataType>Average</dataType>
    <generalCategory>Hydrology</generalCategory>
    <sampleMedium>Surface Water</sampleMedium>
  - <unit>
      <unitName>cubic feet per second</unitName>
      <unitType>Flow</unitType>
      <unitAbbreviation>cfs</unitAbbreviation>
      <unitCode>35</unitCode>
    </unit>
    <noDataValue>-9999</noDataValue>
  - <timeScale isRegular="true">
    - <unit>
        <unitName>minute</unitName>
        <unitType>Time</unitType>
        <unitAbbreviation>min</unitAbbreviation>
        <unitCode>102</unitCode>
      </unit>
      <timeSupport>30</timeSupport>
    </timeScale>
    <speciation>Not Applicable</speciation>
  </variable>
```

(b)

Fig. 1. A portion of the WaterML 1.1 encoding of variable information for stream discharge data from the USGS (a) and from Utah State University (b) for gages in the Little Bear River.

The difference in availability of information to populate WaterML elements across existing data sources and semantic differences in the terms used to describe the data were significant challenges that had to be overcome in developing the data discovery and integration capabilities of the CUAHSI HIS. These challenges also motivated development of a shared vocabulary for the hydrology research community. While we had no control over the vocabularies used by agencies and data sources outside of the academic community, the opportunity existed to develop and promote a shared vocabulary and associated tools for managing it to avoid incomplete specification of metadata and proliferation of many different vocabularies within the hydrology research community.

5. A shared vocabulary for the hydrology research community

In some scientific communities, multiple vocabularies have evolved as independent subsets of community members have worked together. In some of these communities, efforts are ongoing or have been made to develop tools capable of mediating across vocabularies developed by subgroups (e.g., Graybeal et al., 2012) given the difficulty in achieving community consensus on terminology. Bermudez et al. (2005) describe the approach used by the Marine Metadata Interoperability Project (MMI) to develop vocabularies and formulate relationships among them using the Web Ontology Language (OWL). They describe multiple steps in ontology development, including identification or creation of vocabularies by subject matter experts and expressing relationships between terms in these vocabularies. Instead of promoting a single vocabulary, their approach seeks to harmonize across multiple vocabularies using a mapping process. This approach has been used by the U.S. Integrated Ocean Observing System (IOOS) to establish a vocabulary for observed properties, their definitions, and units (Haines et al., 2012) across multiple IOOS Regional Associations that had independently developed separate vocabularies. It has also been explored by the British Oceanographic Data Centre's National Environmental Research Council (NERC) DataGrid for development

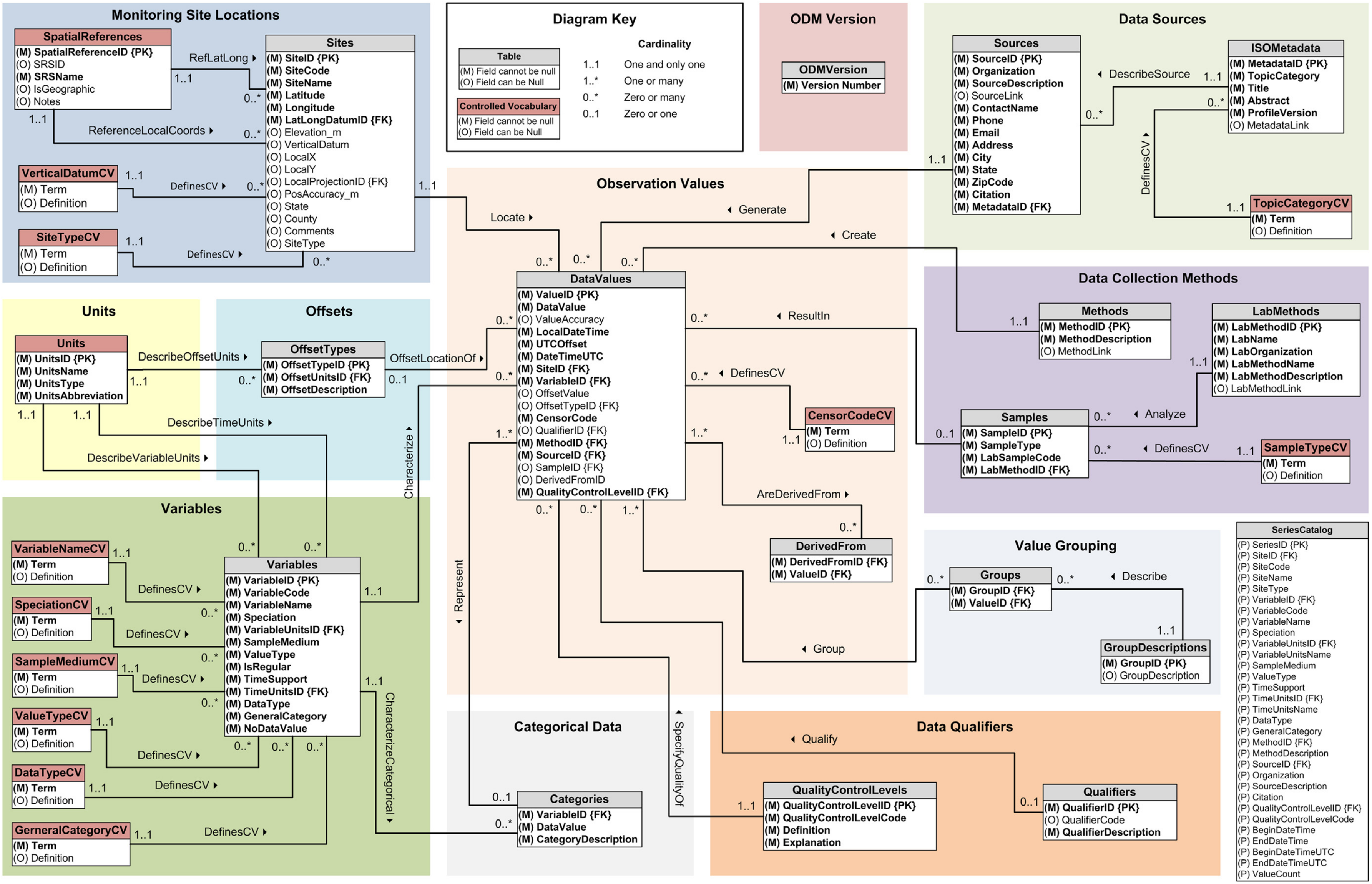

Fig. 2. CUAHSI HIS Observations Data Model Version 1.1.1. Shared vocabulary term lists are shown with red headers. Mandatory fields are shown in bold and are indicated by "(M)." Optional fields are indicated by "(O)." (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
CUAHSI HIS shared vocabulary term lists.

Controlled vocabulary name	Description	Examples
CensorCodeCV	Used to populate the CensorCode field in the ODM DataValues table. Describes whether an observation's value is unknown because is above or below a detection or quantitation limit	"lt," "gt," "nc"
DataTypeCV	Used to populate the DataType field in the ODM Variables table. Describes the type of the reported observation with regard to whether it is an instantaneous observation or a statistic calculated over a specific time support	"Minimum," "Maximum," "Mean"
GeneralCategoryCV	Used to populate the GeneralCategory field in the ODM Variables table. Assigns each measured variable to a coarse level searchable category of data	"Biota," "Climate," "Hydrology"
SampleMediumCV	Used to populate the SampleMedium field in the ODM Variables table. Describes the medium from which a sample was obtained or that an observed value represents	"Air," "Snow," "Soil," "Surface water"
SampleTypeCV	Used to populate the SampleType field in the ODM Samples table. Describes the type of sample that was collected	"Grab," "Core," "Trawl"
SiteTypeCV	Used to populate the SiteType field in the ODM Sites table. Describes the type of monitoring location at which an observation was made	"Canal," "Stream," "Well"
SpatialReferences	Used to specify the spatial reference of site coordinates in the ODM Sites table	"NAD27/UTM Zone 3N"
SpeciationCV	Used to populate the Speciation field in the ODM Variables table. Describes the chemical speciation of a reported observation	"P," "PO_4"
TopicCategoryCV	Used to populate the TopicCategory field in the ODM ISOMetadata table. Provides searchable topics at the dataset level and satisfies keyword requirements so that ISO compliant metadata can be extracted from an ODM database	"Biota," "Environment," "Inland Waters"
Units	Used to specify the units of measured variables as well as for time support in the ODM Variables table and to specify the units of offset values in the ODM OffsetTypes table	"Acre," "Meter," "Cubic feet per second"
ValueTypeCV	Used to populate the ValueType field in the ODM Variables table. Indicates how a reported observation resulted	"Field Observation," "Model Simulation Result"
VariableNameCV	Used to populate the VariableName field in the ODM Variables table. Specifies the name of a measured variable	"Temperature," "Discharge"
VerticalDatumCV	Used to specify the vertical datum of the elevation given for a location in the ODM Sites table	"MSL," "NDVD29"

of vocabularies to describe and enable discovery of measurements or model outputs (Lowry et al., 2005).

The geoscience community, on the other hand, has developed vocabularies for encoding digital geoscientific information using GeoSciML (CGI, 2013) through a Geoscience Terminology Working group that represents the committee. GeoSciML is an XML-based data transfer standard for the exchange of geoscientific information (e.g., the representation and description of the types of features typically found on geological maps). A number of attributes of geologic features and artifacts of geological investigations have been identified and controlled lists of terms have been created by the Geoscience Terminology Working Group. The term lists are then vetted by the geoscience community for defining the domain of values to be used for those attributes.

The larger hydrology community, including both hydrologic researchers and data collection agencies, is not unlike these communities. Differences in the vocabulary terms used by academic researchers and those used by government data collection agencies are evidence of this. In practice, the approach described in this paper incorporates elements of both of the approaches described above. The CUAHSI organization is somewhat unique among scientific communities, and, as an organization that represents members from the hydrology research community, CUAHSI has advanced, curated, and encouraged its members to use a common vocabulary rather than promoting multiple vocabularies and then mediating after the fact. Additionally, the CUAHSI HIS has developed methods for mediating across the vocabulary described in this paper and those used by agency data systems. The following sections describe in more detail development of the community vocabulary and tools that have been created to allow members of the hydrology research community to contribute to and use it.

5.1. Shared vocabulary within a common information model

A common way to resolve semantic differences across multiple data sources is to identify the informational elements that are common across all data sources and then map those elements to a common information model using a shared vocabulary that can represent the semantics of data from each source (Stock et al., 2011; Atkinson et al., 2012). In the case of CUAHSI HIS, this process began with a series of community surveys (e.g., Bandaragoda et al., 2006), evaluation of major agency data publication systems such as USGS NWIS and USEPA STORET, and culminated in the common information model used by CUAHSI HIS for organizing, describing, storing, and publishing observational time series collected at point locations. This information model has been formalized in ODM as a standard, relational data storage schema (Horsburgh et al., 2008), in WaterML as a data transfer schema (Zaslavsky et al., 2007), and is used by the CUAHSI HIS central metadata catalog and data discovery web services (Whitenack, 2010).

ODM, which is now in Version 1.1.1 (Fig. 2), was designed with a number of controlled term lists that define the terms that can be used to populate metadata attributes describing the entities of the information model within the ODM schema. The term lists are highlighted in Fig. 2 with names ending in "CV" and also include Units and SpatialReferences. The term list tables are linked to the tables containing the fields for which they define the allowable domain of values by formal relationships. Together with the ODM schema, these term lists, which are described in more detail in Table 1, make up a shared vocabulary for describing hydrologic observations. When observations stored in an ODM database that complies with the shared vocabulary are published using a WaterOneFlow web service, use of the ODM shared vocabulary minimizes contextual semantic heterogeneity in the data (i.e., heterogeneity resulting from different meanings of concepts and terms), while the use of ODM for data storage and the WaterML schema for data transfer minimizes structural and syntactic heterogeneity in the data.

The original terms in each of the shared vocabulary term lists were derived from common terms used by members of the hydrology research community to describe their data and from terms used to describe datasets that already existed in the CUAHSI HIS system. Additionally, some terms were derived from existing vocabularies that had hydrology-related terms. However, early experience with ODM quickly led to the conclusion that the original lists of shared vocabulary terms were inadequate to describe the diversity of observational data that users were loading into ODM databases. When users did not find the terms they needed to describe their data, they simply began inserting their own terms, which were, in some cases, inconsistent across ODM instances. It became clear that if the original shared vocabulary was to support large-scale implementation across the emerging network of HydroServers, successful implementation, widespread community adoption, and realization of the benefits associated with a shared vocabulary would depend on the ability of the vocabulary to respond to the needs of the community. A system by which the controlled vocabularies could be vetted and extended by the community while avoiding unnecessary duplication was needed.

5.2. Building, deploying, and managing the shared vocabulary

To meet the need for a shared vocabulary management system, we designed, built, and deployed a database and web application that serve as a registry for terms in the shared vocabulary (see http://his.cuahsi.org/mastercvreg.html). The database was implemented using Microsoft SQL Server and consists of a modified ODM database that contains two versions of each shared vocabulary term list, one with the list of approved terms, and another that lists all of the requests that have been made and documents the actions that have been taken on each request. The web application was developed using ASP.Net and is connected directly to the SQL Server database. ODM users can visit this web application, view existing terms in the shared vocabulary, and request the addition of new terms or modifications to existing terms. Although we have no quantitative data indicating how much time is required for users to learn how to use this system to propose changes, our experience was that once provided with the URL to the system, users were able to submit new terms and modifications to existing terms in all cases.

The system is human-moderated by a committee of community members to ensure that submitted terms are legitimate, relevant, and unique. Fig. 3 depicts the process by which changes to the shared vocabulary can be made. All requested changes to vocabulary terms are stored in the underlying database along with a record of the type of requested change (e.g., new term or change to existing term), the reason for the request, when the changes were submitted, the status of the change (e.g., approved or rejected), who submitted the change, and any notes that are added to the request by the system moderators, thus preserving the provenance of terms added to the shared vocabulary. Once changes are approved by the moderators, the new or modified terms become part of what is effectively a master shared vocabulary available for all ODM users.

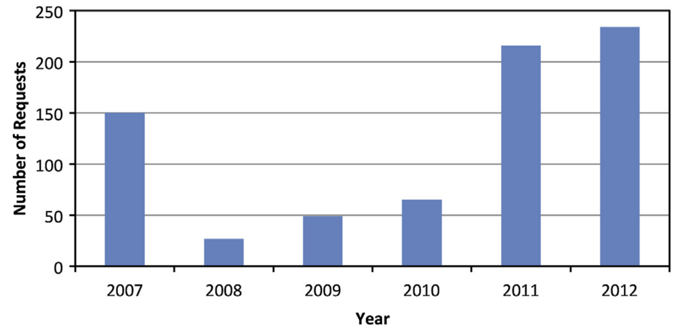

Fig. 4. Total number of change requests to the ODM shared vocabulary term lists described in Table 1 by year.

Perhaps the most important contribution of this system is that data publishers within the hydrology research community can actively modify the vocabulary to suit their needs, while making those modifications immediately available to the larger community. Fig. 4 shows the number of change requests that have been submitted to the shared vocabulary system by community members by year. The numbers shown are summed across all of the controlled term lists in the shared vocabulary. After an initial push when the shared vocabulary was first created in 2007, the number of requests has increased every year since. The vast majority of the submitted change requests have been related to the names of measured

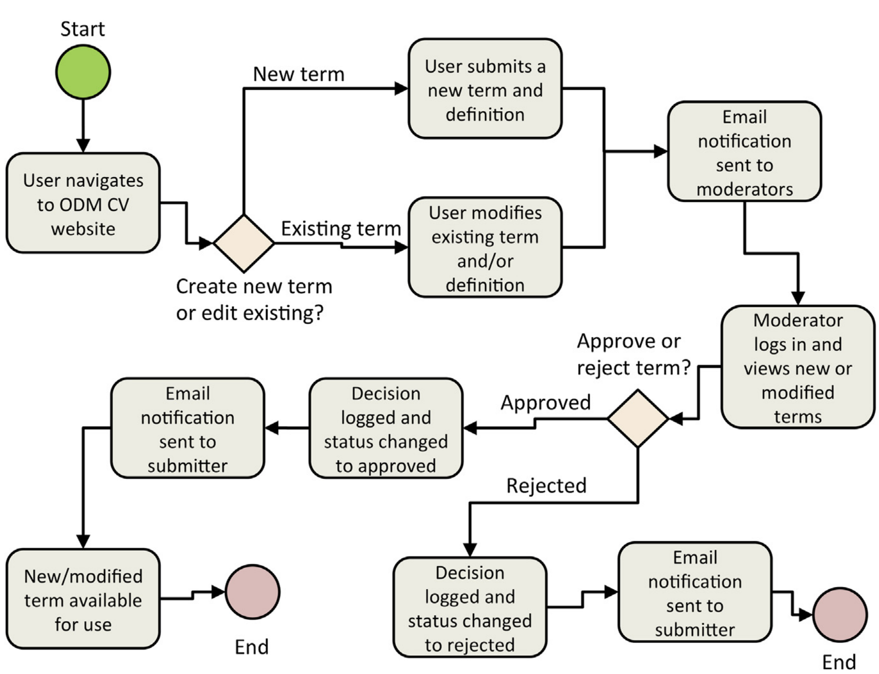

Fig. 3. Step-by-step process for requesting changes to the CUAHSI HIS shared vocabulary.

variables, units, and chemical speciation (approximately 93% of all requests), reflecting the growth in types of data published using the CUAHSI HIS. By far, the most common reason for requested changes was that submitted terms did not already exist in the shared vocabulary, with approximately 76% of all requests to the system being requests for new terms. Of the requests submitted to the system, approximately 85% were approved, with the most common reason for rejection being that an equivalent term already existed. The impact of each of these changes is that with each addition or modification, a richer and more consistent set of terms becomes available to describe data added to the CUAHSI HIS system. New data added to the system can be described using consistent terms, and, new or modified terms can be used by subsequent data publishers to share their data, rather than each publisher creating their own terms. Data consumers, who want to discover, access, and use the data, benefit as well because data within the system are described using consistent terms.

5.3. Publishing the shared vocabulary

Given the federated nature of ODM database instances within the CUAHSI HIS, managing the ODM shared vocabulary via a centralized website available to the community required a means for publishing the master terms so that distributed ODM users could easily synchronize their local database with the constantly-changing master terms. The master shared vocabulary terms were exposed via a SOAP-based web service interface that provides a single web service method for retrieving each vocabulary term list (e.g., GetVariableNameCV, GetDataTypeCV, etc.), with results encoded in a simple XML response. Fig. 5 shows a portion of the XML encoding of the return from one of the master shared vocabulary web service methods. This dynamic linkage between the master shared vocabulary managed via the website described above and the web services that publish the terms ensures that the most recent set of vocabulary terms is always available and can be accessed via any client application that can access SOAP-based web services.

5.4. Synchronizing local copies of the shared vocabulary

The ODM Tools software (Horsburgh et al., 2011) provides functionality for ODM database administrators to compare the vocabulary terms in their local ODM database side-by-side with the centrally-managed master shared vocabulary terms retrieved via the web services described above. Differences in terms are highlighted, and ODM Tools enables database administrators to add new terms to their local vocabulary, update existing terms that have changed, and re-map any of their own terms or terms that no longer exist in the master shared vocabulary to new terms from the master shared vocabulary. Fig. 6 shows the ODM Tools interface for managing the vocabulary terms in an ODM database. Panel (a) shows color (in the web version) highlighting that indicates new terms from the master vocabulary list that do not exist in the local vocabulary list (orange), a term that is in both the local and master lists but has been modified (blue), and a term that is in the local list, but not in the master list (red). Panel (b) shows how terms that are not in the master shared vocabulary list (old term) can be exchanged with approved terms from the master shared vocabulary (new term).

6. Mediating across hydrology domain vocabularies

While the shared vocabulary described above can be enforced at the database level for those data sources that use ODM, it is impossible to impose those vocabulary terms on agencies with existing database systems and vocabularies. CUAHSI has worked in partnership with agencies such as the USGS and USEPA to map their data and metadata elements to the CUAHSI HIS information model and to provide access to their data using standardized WaterOneFlow web services and WaterML encoding. This has been possible because the CUAHSI HIS information model was designed such that there are one-to-one mappings for most informational elements within these systems. However, since agency vocabularies differ from the CUAHSI HIS vocabulary, a solution was needed for mediating across the CUAHSI shared vocabulary and agency vocabularies so that consistent metadata cataloging and data discovery services could be created.

A variable name concept hierarchy was created for this purpose (Piasecki and Beran, 2009). This hierarchical concept tree begins with general concepts and then branches out to more specific concepts until it reaches a level that defines concepts slightly more general than the variable name terms used by the CUAHSI shared vocabulary or agency datasets. For example, a full path in the concept tree might look like "Hydrosphere/Physical/Temperature/


```
  <?xml version="1.0" encoding="UTF-8"?>
- <string xmlns="http://his.cuahsi.org/his/1.1/ws/">
  - <GetVariableNameCVResponse>
    - <Records count="368">
        ...
      - <Record>
          <Term>Discharge</Term>
          <Definition>Discharge</Definition>
        </Record>
        ...
      - <Record>
          <Term>Streamflow</Term>
          <Definition>The volume of water flowing past a fixed point. Equivalent to discharge.</Definition>
        </Record>
        ...
      - <Record>
          <Term>Temperature</Term>
          <Definition>Temperature</Definition>
        </Record>
      - <Record>
          <Term>Temperature, dew point</Term>
          <Definition>Dew point temperature</Definition>
        </Record>
        ...
      </Records>
    </GetVariableNameCVResponse>
  </string>
```


Fig. 5. A portion of the XML encoding of the CUAHSI HIS ODM shared vocabulary variable name list. The characters "…" indicate where portions of the list have been removed for brevity.

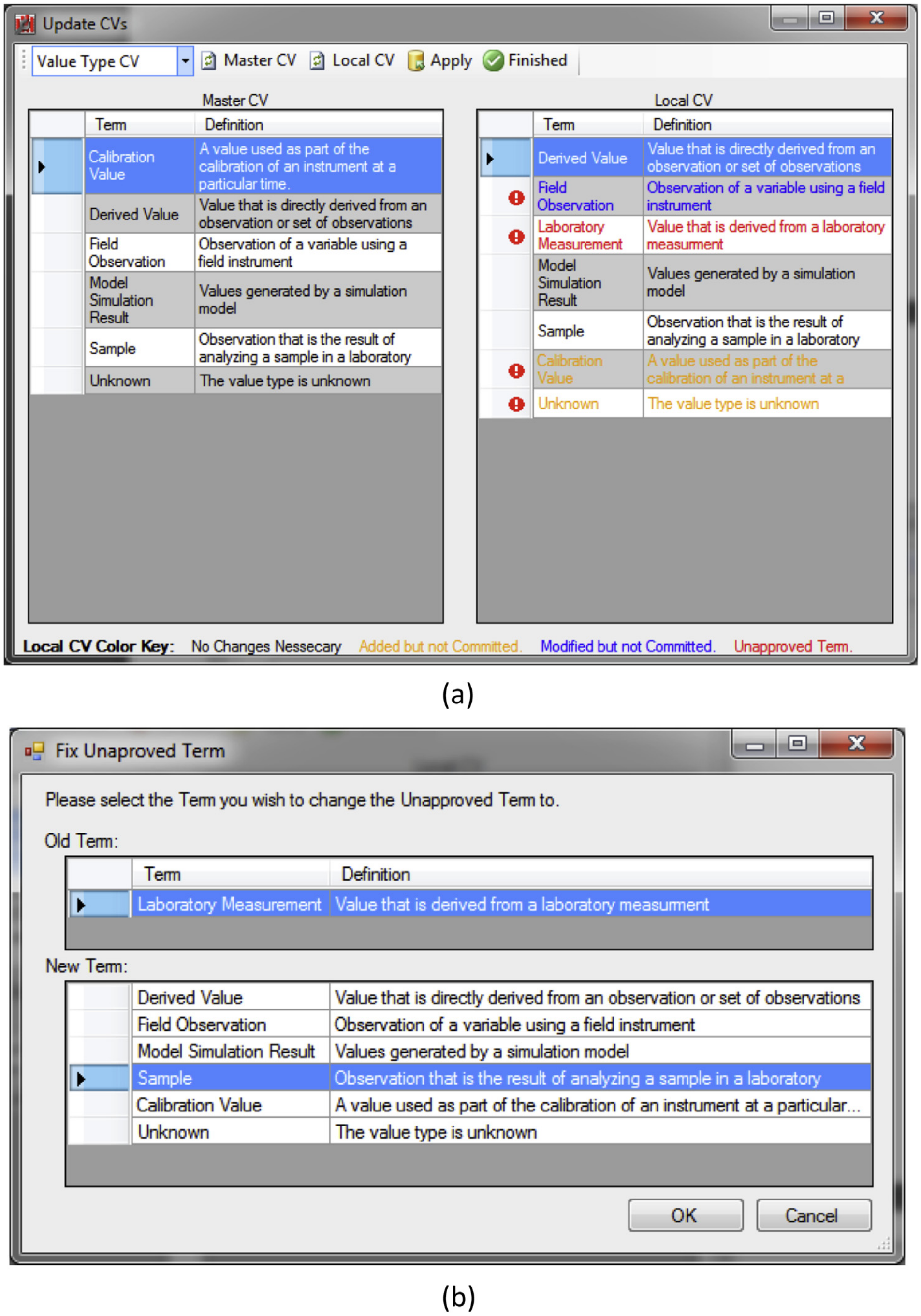

Fig. 6. The ODM Tools vocabulary editing interface.

Temperature, water," beginning with the most general term in the hierarchy ("Hydrosphere") and ending with a specific term that describes the name of an observed variable ("Temperature, water").

Once the metadata describing time series of hydrologic observations have been harvested from a WaterOneFlow web service into the central metadata catalog, each of the variable name terms from the service must be mapped to a corresponding concept within the variable name concept tree. For ODM-based data sources that conform to the CUAHSI shared vocabulary, mappings between variable name terms used in the ODM database and terms in the variable name concept tree are generated automatically when the metadata are harvested into the HIS Central metadata catalog database. For non-compliant services, these semantic mappings can be created using the semantic annotation tools described by Piasecki and Beran (2009).

The concept tree serves as the basis for variable name-based discovery of data within the CUAHSI HIS. Client application developers can enable users of their applications to navigate the concept tree, using terms at any level within the concept hierarchy as search terms. As an added feature, synonyms have been defined for many terms in the concept tree, enabling the expansion of terms used in data discovery queries. Fig. 7 shows an example of how a variable name-based data discovery query would be executed against the HIS Central metadata catalog and demonstrates how the concept tree aids in the discovery of data from multiple sources with differing semantics. A search term ("Streamflow") is passed in

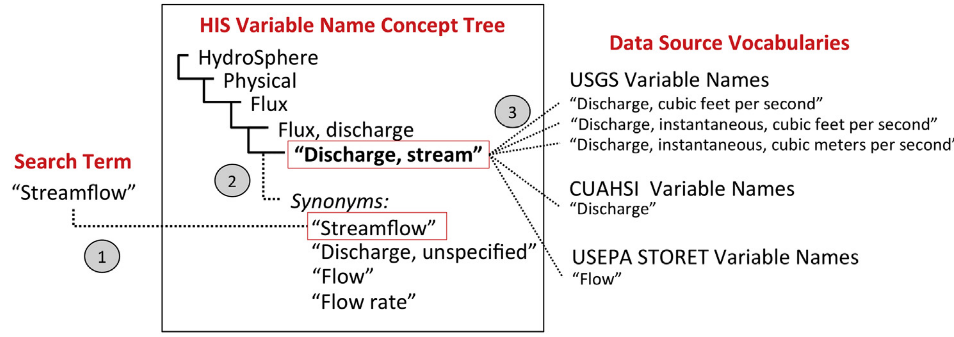

Fig. 7. Schematic of how data discovery queries are resolved at HIS Central, enabling discovery of equivalent data from sources using differing semantics.

code to the discovery web service (1) where it is evaluated against the list of searchable concepts in the variable name concept tree (2). Searchable concepts include both the terms in the concept tree and their synonyms. In the example, "Streamflow" matches a synonym for the primary term "Discharge, stream." Finally, the discovery service queries for variable names from cataloged data sources that have been mapped to the primary term "Discharge, stream" (3), in this case returning variable names describing data from the USGS NWIS system, the CUAHSI HIS, and the USEPA STORET database. Datasets with matching variable names can then be further narrowed using additional spatial and temporal search criteria.

As an extension to this example, a data discovery query was executed against the HIS Central data discovery service using the CUAHSI HIS client application HydroDesktop. "Streamflow" was used as the keyword, a time period from 1/1/2003–12/31/2013 was used as the time period for the search, and a polygon representing Cache County, Utah was used as the geospatial domain for the search. This query resulted in five time series from USGS's Daily Streamflow Values service, four time series from USGS's Unit Values service, seven time series from Utah State University's Little Bear River service, six time series from Utah State University's Logan River service, and 175 time series from USEPA's STORET Water Quality service. The spatial distribution of these search results is shown in Fig. 8.

7. Discussion and conclusions

The CUAHSI HIS makes data available via web services that transmit data using WaterML as a common data encoding syntax. This has enabled cataloging and indexing of the data from all service providers to provide data discovery services. This is analogous to the way standardization of the Internet on HTML enabled Internet search providers to catalog and index the contents of web sites to enable sophisticated searches. The subsequent addition of a standardized shared vocabulary for ODM-based data sources and the ability to mediate across additional agency vocabularies simplifies the use of variable names and other attributes for data publication, discovery, and interpretation.

The shared vocabulary and software tools described in this paper address the need within the hydrology research community for creating, managing, and achieving consensus around shared language that can be used to describe hydrologic observations data. These tools are being used as best practices within the hydrology research community to populate metadata and capture the semantics of hydrologic observations data in sharing them using the CUAHSI HIS. The shared vocabulary promotes greater semantic consistency in observations metadata published by multiple data sources. The open, moderated web application for modifying the shared vocabulary provides a simple mechanism for community users to examine existing terms and to request additions and changes. Moderation by community members ensures that submitted terms are relevant and unique. The shared vocabulary web services that publish the master vocabulary terms and client tools included in the ODM Tools software provide simple mechanisms through which distributed data publishers can update the vocabulary terms within their local ODM databases.

Since the work described in this paper began, USGS and USEPA have developed new, collaborative capabilities for the retrieval of water quality data from the USGS NWIS and USEPA STORET system using web services (Scott et al., 2008). The data are returned in a common format, and the two agencies have mapped the names of their measured variables to a vocabulary defined by the USEPA Substance Registry System (SRS – see http://www.epa.gov/srs/). Although the data returned by these new services are not in WaterML format, their availability, their common syntax, and their agreement on a single vocabulary for variable names will make it easier in the future to maintain the mapping of these data sources to the CUAHSI HIS concept hierarchy and provide data discovery services that enable discovery of USGS and USEPA data in the context of all of the other datasets hosted by the CUAHSI HIS. Convergence of USGS and USEPA on using the SRS for water quality variable names also presents an opportunity for potential wider adoption of the SRS vocabulary as an authoritative source for water quality variable names.

Our work has led to several observations about use of the system and potential areas for improvement. First, there are opportunities for additional vocabularies that currently do not exist in ODM. One illustrative example is that of the VariableNameCV. Because ODM 1.1.1 lacks a vocabulary to separately describe the fractionation of a sample, users have requested variable name terms such as "Phosphorus, total" and "Phosphorus, total dissolved" so that they can describe the difference between an observation that was made on an unfiltered water sample ("Phosphorus, total") and another made on a filtered sample ("Phosphorus, total dissolved"). While information about sample preparation steps can be encoded within the method information in ODM, users wanted to be explicit about this in their variable names. An approach that would reduce repetitive terms in the VariableNameCV would be to have a separate term list in the shared vocabulary for sample fraction.

Next, while this paper has focused on the shared vocabulary and related tools, the concept tree of searchable variable names was developed centrally, and no moderation system was developed to support it. A need that has emerged from this work is that of community vetting and expansion of the variable name concepts in the concept tree that supports data discovery. Because there is currently no automated mechanism for community members to suggest new searchable concepts for inclusion in the centralized system, the searchable concepts lag behind the terms in the shared

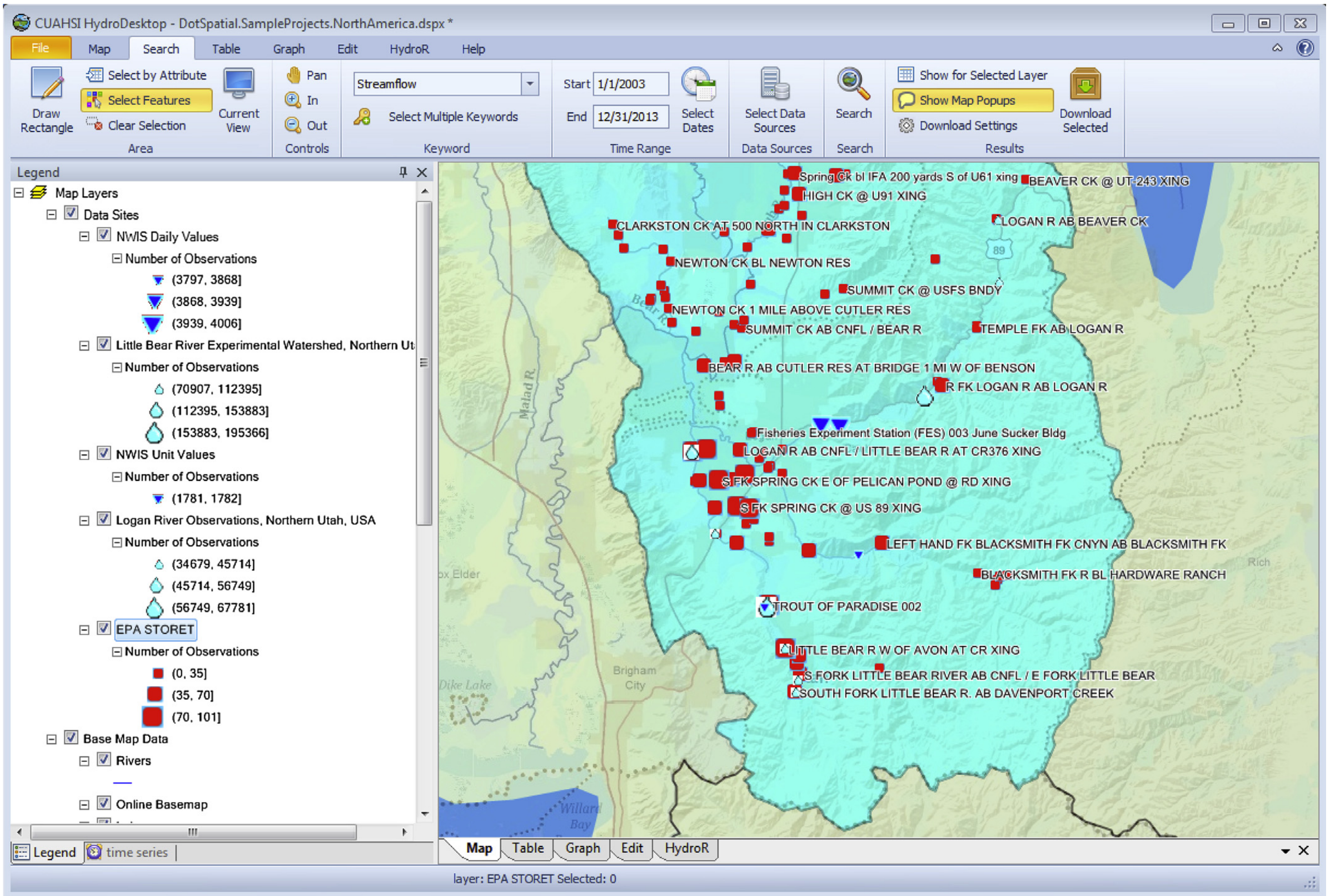

Fig. 8. Example data discovery query results in HydroDesktop.

vocabulary. This suggests that dynamic and community accessible tools are needed for updating and maintaining the concept tree to better support its expansion in concert with the shared vocabulary.

There is also an opportunity to develop new concept trees for additional ODM/WaterML information model attributes for which we have developed controlled term lists in the shared vocabulary. It would then be easier to create enhanced semantic search services that expose these additional metadata attributes as search facets for discovery. In a simple example, data consumers might then be able to specify a SampleMedium of “water” and discover observations having “surface water” or “groundwater” as their SampleMedium, by virtue of “surface water” and “groundwater” being child concepts of the “water” concept.

Last, there are potential interoperability benefits to exposing the shared vocabulary developed using the tools described in this paper using one or more of the existing vocabulary encoding standards that have emerged since development of the CUAHSI HIS system began. One candidate is the Simple Knowledge Organization System (SKOS) (Miles and Bechhofer, 2009), which is a standard data model for sharing and linking controlled vocabularies via the web. SKOS has a number of features, including identification of vocabulary terms using uniform resource identifiers (URIs), documentation of terms with various types of notes, and semantically relating terms to each other in hierarchies (Isaac and Summers, 2009). Representing the shared vocabulary described here using SKOS could enable it to be mapped to other existing hydrology domain vocabularies (e.g., the Substance Registry system used by USGS and USEPA for water quality data) as well as vocabularies from other domains, providing a mechanism for asserting the equivalence of terms among vocabularies. This could better enable its use outside of the hydrology research community. A number of new tools have emerged recently that may facilitate this. The TemaTres controlled vocabulary server (http://www.vocabularyserver.com/), for example, provides an open source vocabulary server and web application for managing and sharing vocabularies and supports the SKOS standard. TemaTres is being used by other scientific communities like the Long Term Ecological Research (LTER) Network for vocabulary management.

The architecture of the CUAHSI HIS shared vocabulary system has now been adopted by the integrated data management system of the Critical Zone Observatory (CZO) program. Scientists within the CZO community (and we anticipate many other communities) face many of the same challenges that the hydrology research community faces in describing observational data, and the shared vocabulary system described here provides them with a mechanism for creating consensus about terms used to describe observations made in the critical zone. The formal specification, encoding, and publication of these vocabularies represent key enablers for both human and computer interaction with published observational data.

Acknowledgments

The shared vocabulary, software, and systems described in this paper were developed as part of the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic

Information System (HIS) project whose goal is to advance information system technology for hydrologic science. This work was supported by the National Science Foundation under grant EAR 0622374 for the development of Hydrologic Information Systems. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Atkinson, R., Dornblut, I., Smith, D., 2012. An international standard conceptual model for sharing references to hydrologic features. J. Hydrol. 424-425, 24–36. http://dx.doi.org/10.1016/j.jhydrol.2011.12.002.

Bandaragoda, C., Tarboton, D., Maidment, D., 2006. Hydrology's efforts toward the cyberfrontier. EOS Trans. AGU 87 (1), 2. http://dx.doi.org/10.1029/2006EO010005.

Beran, B., Goodall, J., Valentine, D., Zaslavsky, I., Piasecki, M., 2009. Standardizing access to hydrologic data repositories through web services. In: Proceedings 2009 Conference on Advanced Geographic Information Systems & Web Services, Cancun, Mexico, 1–7 February. http://dx.doi.org/10.1109/GEOWS.2009.27.

Beran, B., Piasecki, M., 2009. Engineering new paths to water data. Comput. Geosci. 35, 753–760. http://dx.doi.org/10.1016/j.cageo.2008.02.017.

Bergamaschi, S., Castano, S., Vincini, M., Beneeventano, D., 2001. Semantic integration of heterogeneous information sources. Data Knowl. Eng. 36 (3), 215–249. http://dx.doi.org/10.1016/S0169-023X(00)00047-1.

Bermudez, L., Graybeal, J., Isenor, A.W., Lowry, R., Wright, D., 2005. Construction of marine vocabularies in the marine metadata interoperability project. In: Oceans, 2005. PRoceedings of MTS/IEEE, vol. 3, pp. 2578–2585. http://dx.doi.org/10.1109/OCEANS.2005.1640159.

Commission for the Management and Application of Geoscience Information (CGI), 2013. GeoSciML – Resource Repository, Version 3.2. http://www.geosciml.org/ (last accessed 01.11.13.).

Erl, T., 2005. Service-oriented Architecture: Concepts, Technology, and Design. Prentice Hall, New Jersey, p. 792.

Goodall, J.L., Horsburgh, J.S., Whiteaker, T.L., Maidment, D.R., Zaslavsky, I., 2008. A first approach to web services for the National Water Information System. Environ. Model. Softw. 23 (4), 404–411. http://dx.doi.org/10.1016/j.envsoft.2007.01.005.

Graybeal, J., Isenor, A.W., Rueda, C., 2012. Semantic mediation of vocabularies for ocean observing systems. Comput. Geosci. 40, 120–131. http://dx.doi.org/10.1016/j.cageo.2011.08.002.

Haines, S., Subramanian, V., Mayorga, E., Snowden, D., Ragsdale, R., Rueda, C., Howard, M., 2012. IOOS vocabulary and ontology strategy for observed properties. In: Proceedings MTS/IEEE Oceans'12, Hampton Road, VA, 14–19 October, pp. 1–9. http://dx.doi.org/10.1109/OCEANS.2012.6405083.

Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., Zaslavsky, I., 2008. A relational model for environmental and water resources data. Water Resour. Res. 44, W05406. http://dx.doi.org/10.1029/2007WR006392.

Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., Zaslavsky, I., 2011. Components of an integrated environmental observatory information system. Comput. Geosci. 37, 207–218. http://dx.doi.org/10.1016/j.cageo.2010.07.003.

Horsburgh, J.S., Tarboton, D.G., Piasecki, M., Maidment, D.R., Zaslavsky, I., Valentine, D., Whitenack, T., 2009. An integrated system for publishing environmental observations data. Environ. Model. Softw. 24 (8), 879–888. http://dx.doi.org/10.1016/j.envsoft.2009.01.002.

Horsburgh, J.S., Tarboton, D.G., Schreuders, K.A.T., Maidment, D.R., Zaslavsky, I., Valentine, D., 2010. Hydroserver: a platform for publishing space-time hydrologic datasets. In: Proceedings 2010 American Water Resources Association Spring Specialty Conference Geographic Information Systems (GIS) and Water Resources VI, Orlando, Florida. American Water Resources Association, Middleburg, Virginia, TPS, pp. 10–11.

Isaac, A., Summers, E., 2009. SKOS Simple Knowledge Organization System Primer. W3C Group Note 18 http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/.

Josuttis, N.M., 2007. SOA in Practice – the Art of Distributed System Design. O'Reilly Press, Sebastapol, CA, p. 324.

Lowry, R., Cramer, R., Gutierrez, M., Hughes, M., Kleese van Dam, K., Kondapalli, S., Latham, S., Lawrence, B., O'Neil, K., Wolf, A., 2005. Parameter vocabularies in the NERC DataGrid. In: Project Proceedings of the UK e-Science All Hands Meeting, 19th – 22nd September, Nottingham, UK. In: http://www.allhands.org.uk/2005/proceedings/papers/310.pdf (last accessed 01.11.13.).

Lutz, M., Sprado, J., Klien, E., Schubert, C., Christ, I., 2009. Overcoming semantic heterogeneity in spatial data infrastructures. Comput. Geosci. 35, 739–752. http://dx.doi.org/10.1016/j.cageo.2007.09.017.

Ma, X., Wu, C., Carranza, E.J.M., Schetselaar, E.M., van der Meer, F.D., Liu, G., Wang, X., Zhang, X., 2010. Development of a controlled vocabulary for semantic interoperability of mineral exploration geodata for mining projects. Comput. Geosci. 36, 1512–1522. http://dx.doi.org/10.1016/j.cageo.2010.05.014.

Miles, A., Bechhofer, S., 2009. SKOS Simple Knowledge Organization System Reference. W3C Recommendation http://www.w3.org/TR/2009/REC-skos-reference-20090818/.

Pallickara, S.L., Zupanski, M., Sullivan, S., 2010. Efficient metadata generation to enable interactive data discovery over large-scale scientific data collections. In: Proceedings 2nd IEEE International Conference on Cloud Computing Technology and Science, Indianapolis, IN, Nov. 30 – Dec. 3. http://dx.doi.org/10.1109/CloudCom.2010.99.

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

Scott, J.C., Gellenbeck, D.J., Young, D., Booth, N.L., 2008. U.S. Federal water quality web service collaboration. EOS Trans. Am. Geophys. Union 89 (52), 542–543. http://dx.doi.org/10.1029/2008EO520003.

Stock, K., Stojanovic, T., Reitsma, F., Ou, Y., Bishr, M., Ortmann, J., Robertson, A., 2011. To ontologize or not to ontologize: an information model for a geospatial knowledge infrastructure. Comput. Geosci.. http://dx.doi.org/10.1016/j.cageo.2011.10.021.

Taylor, P. (Ed.), 2012. OGC WaterML 2.0: Part 1 – Timeseries, OGC Candidate Standard OGC 10-126r2, Version 2.0.0. http://portal.opengeospatial.org/files/?artifact_id=48531 (last accessed 01.11.13.).

Tarboton, D.G., Horsburgh, J.S., Maidment, D.R., Whiteaker, T., Zaslavsky, I., Piasecki, M., Goodall, J., Valentine, D., Whitenack, T., 2009. Development of a community hydrologic information system. In: Anderssen, R.S., Braddock, R.D., Newham, L.T.H. (Eds.), 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, July 2009, ISBN 978-0-9758400-7-8, pp. 988–994.

Tarboton, D.G., Maidment, D., Zaslavsky, I., Ames, D., Goodall, J., Hooper, R.P., Horsburgh, J.S., Valentine, D., Whiteaker, T., Schreuders, K., 2011. Data interoperability in the hydrologic sciences: the cuahsi hydrologic information system. In: Proceedings of the Environmental Information Management Conference 2011, pp. 132–137. Santa Barbara, CA, 28-29 September http://eim.ecoinformatics.org/eim2011/eim-proceedings-2011/view.

Whitenack, T., 2010. CUAHIS HIS Central 1.2. Consortium of Universities for the Advancement of Hydrologic Science, Inc. http://his.cuahsi.org/documents/HIS_Central-1.2.pdf (last accessed 01.11.13.).

Zaslavsky, I., Valentine, D., Hooper, R., Piasecki, M., Couch, A., Bedig, A., 2012. Community practices for naming and managing hydrologic variables. In: Proceedings of the AWRA Spring Specialty Conference on GIS and Water Resources, New Orleans, LA, March, pp. 26–28.

Zaslavsky, I., Valentine, D., Whiteaker, T. (Eds.), 2007. CUAHSI WaterML. OGC Discussion Paper OGC 07-041r1. Version 0.3.0 http://portal.opengeospatial.org/files/?artifact_id=21743 (last accessed 01.11.13.).