Environmental Modelling & Software 24 (2009) 879–888

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

An integrated system for publishing environmental observations data

Jeffery S. Horsburgh [a,*], David G. Tarboton [a], Michael Piasecki [b], David R. Maidment [c], Ilya Zaslavsky [d], David Valentine [d], Thomas Whitenack [d]

[a] Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, USA
[b] Department of Civil, Architectural, and Environmental Engineering, Drexel University, Philadelphia, PA, USA
[c] Center for Research in Water Resources, University of Texas at Austin, Austin, TX, USA
[d] San Diego Supercomputer Center, University of California at San Diego, San Diego, CA, USA

ARTICLE INFO

Article history:
Received 1 August 2008
Received in revised form
2 January 2009
Accepted 5 January 2009
Available online 25 January 2009

Keywords:
Cyberinfrastructure
Data publication
Water resources data
Web services
Data model
Semantic heterogeneity
Controlled vocabulary
Data intensive science

ABSTRACT

Over the next decade, it is likely that science and engineering research will produce more scientific data than has been created over the whole of human history. The successful use of these data to achieve new scientific breakthroughs will depend on the ability to access, integrate, and analyze these large datasets. Robust data organization and publication methods are needed within the research community to enable data discovery and scientific analysis by researchers other than those that collected the data. We present a new method for publishing research datasets consisting of point observations that employs a standard observations data model populated using controlled vocabularies for environmental and water resources data along with web services for transmitting data to consumers. We describe how these components have reduced the syntactic and semantic heterogeneity in the data assembled within a national network of environmental observatory test beds and how this data publication system has been used to create a federated network of consistent research data out of a set of geographically decentralized and autonomous test bed databases.

1. Introduction

New technology and data resources are often instrumental in the emergence of new scientific discoveries. Because results from local research projects can be aggregated across sites and times, in many cases by investigators other than those who originally collected the data, the potential exists to advance science and research significantly through the publication of research data (Borgman et al., 2007; Research Information Network, 2008). There is a need, therefore, for standardized and robust methods to organize and publish environmental observations data as resources that can be discovered and used for scientific analysis.

Indeed, environmental research and education have recently become increasingly data-intensive as a result of the proliferation of digital technologies, instrumentation, and pervasive networks through which data are collected, generated, shared, and analyzed (National Science Foundation, 2007). Over the next decade, it is likely that science and engineering research will produce more scientific data than has been created over the whole of human history (Cox et al., 2006). Successfully using these data to achieve new scientific breakthroughs and increase understanding of the world around us, as well as in making sound and informed resource management decisions, will depend in large part on the ability to access, organize, integrate, and analyze these large datasets.

Comprehensive infrastructure that is being used to capitalize on dramatic advances in information technology has been termed "cyberinfrastructure" and integrates hardware for computing, data and networks, digitally enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools (National Science Foundation, 2007). This paper describes new cyberinfrastructure that enables the publication of point observations (i.e., measurements made at a point in space such as a weather station or water quality monitoring site). This cyberinfrastructure has been developed as part of a Hydrologic Information System (HIS), which is a distributed network of data sources and functions that are integrated using web services and that provide access to data, tools, and models that enable synthesis, visualization, and evaluation of hydrologic system behavior (http://his.cuahsi.org). Although the data publication system described in this paper has been developed primarily to advance the information science knowledge base and available data resources for water resources research, the general system

* Corresponding author. Tel.: +1 435 797 2946; fax: +1 435 797 3663.
E-mail address: jeff.horsburgh@usu.edu (J.S. Horsburgh).

1364-8152/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2009.01.002

architecture could be extended to many other types of point observations.

The HIS consists of four major components: data publication, data curation, data discovery, and data delivery. Publication is the process by which data are made available to users other than those that collected the data. Curation is the long term preservation of data to ensure that they persist indefinitely. Discovery involves tools that allow users to find published data, and delivery involves the transmittal of data to users in formats that they can use. In this paper, we focus mainly on the data publication component, although we include some discussion of the other components to place data publication in the context of the overall HIS.

Publication of research data involves persistent storage, management, and communication of data to potential users. Within and across research sites, multiple investigators and organizations are involved in both collecting and consuming data. To be effective, data publication systems must facilitate interoperation and mediation among data sources and their consumers. One challenge that arises in the design of data publication systems is heterogeneity within the formats and vocabularies that support the data (Colomb, 1997; Morocho et al., 2003; Sheth and Larson, 1990). Additionally, data consumers may not have intimate knowledge of the data collection process, requiring that the data be published with sufficient metadata to enable unambiguous interpretation (Gray et al., 2005). These metadata should include information about the location at which the observations were made, the variable that was observed or measured, the source of or organization that created the data, the procedures used to create the data, data qualifying comments, quality assurance and quality control information, time support, spacing, and extent, and other important attributes (Horsburgh et al., 2008).

In this paper, we describe a data publication system that overcomes the challenges in publishing research data through the use of a standard observations data model populated using controlled vocabularies for environmental and water resources data along with web services for transmitting data to consumers. Section 2 describes existing data publication efforts for environmental and water resources data. Section 3 describes new cyberinfrastructure efforts ongoing within several scientific domains. Section 4 describes syntactic and semantic heterogeneity and their implications for the publication, search for, and interpretation of existing environmental and water resources data. Section 5 describes how this heterogeneity can be overcome. Section 6 describes the design of a cyberinfrastructure that supports publication of environmental observations. Finally, Section 7 provides an implementation case study that describes how the components of the data publication system have been applied to create a federated network of consistent research data out of a set of geographically decentralized and autonomous databases from 11 environmental observatory test beds, effectively creating a publically available, community data resource from data that might otherwise have been confined to the private files of the individual investigators.

2. Existing data publication methods

Within the United States, many organizations and individuals measure hydrologic variables such as streamflow, water quality, groundwater levels, soil moisture, and precipitation. Several national data collection and publication networks operated by government agencies have arisen over the years. These include the USGS WATer Data STOrage and REtrieval System (WATSTORE), which has been replaced by the National Water Information System (NWIS) (http://waterdata.usgs.gov/nwis), the USEPA STOrage and RETrieval (STORET) System (http://www.epa.gov/storet/), the USDA SNOpack TELemetry (SNOTEL) System (http://www.wcc.nrcs.usda.gov/snow/) and Soil Climate Analysis Network (SCAN) (http://www.wcc.nrcs.usda.gov/scan/), the NOAA National Climatic Data Center (NCDC) (http://www.ncdc.noaa.gov/oa/ncdc.html), and a host of others. These national data repositories contain a wealth of data, but, in general, they have different data storage systems and formats, different data retrieval systems, and different data publication formats. Synthesizing data from these disparate sources into a single analysis can be difficult because each one presents users with the task of navigating through pages, menus, and files to access the data and metadata that they contain.

Recent times have also seen a push in the publication of data from existing experimental watersheds such as Reynolds Creek (Slaughter et al., 2001), the Little River (Bosch et al., 2007), and Walnut Gulch (Moran et al., 2008; Nichols and Anson, 2008). The technical details and much of the metadata for these datasets have been described in journal publications, and the data themselves have been made available as files that can be retrieved from public websites. Similarly, the Long Term Ecological Research (LTER) Network has made climatic and hydrologic data collected at LTER sites available through their ClimDB/HydroDB climate and hydrology database projects website (http://www.fsl.orst.edu/climhy/).

Although these efforts represent considerable progress, none of the data publication systems that have been developed have been embraced as a standard for the academic and scientific research communities. Because of this, data and metadata resulting from academic research in water resources continue to be published in peer-reviewed journals (Helly, 2006). Interpretations and figures based on data are widely published and archived in libraries, while most of the primary data are confined to the research files of the investigators, making verification of research results difficult. More recently, however, the idea of publishing observational data along with analysis results is gaining ground within the research community as the technology for doing so becomes more generally accessible (Research Information Network, 2008).

3. New cyberinfrastructure efforts

There are currently several large-scale cyberinfrastructure activities underway that are aimed at creating and sharing multidisciplinary datasets, facilitating collaborative and interdisciplinary research, and creating infrastructure to enable scientific discoveries. These activities include: the National Ecological Observatory Network (NEON), which is planning the deployment of networked sensors and cyberinfrastructure to gather data on compelling ecological challenges (http://www.neoninc.org); the Long Term Ecological Research Network (LTER), which is a network of research sites that promotes synthesis and comparative research across sites and ecosystems (http://www.lternet.edu/); the Geosciences Network (GEON), which has developed infrastructure for discovering, accessing and integrating earth sciences data and tools (http://www.geongrid.org/); EarthScope, which is an earth science program to explore the structure and evolution of the North American Continent and understand processes controlling earthquakes and volcanoes (http://www.earthscope.org/); and many others.

The data collected within each of these communities are diverse and are, in most cases, distributed across a number of research sites and study areas. To overcome these challenges, most of these cyberinfrastructure initiatives are developing web service based service-oriented architectures (SOA). Web services are applications that provide the ability to pass information between computers over the Internet, usually formatted using a platform independent markup language such as eXtensible Markup Language (XML) (Goodall et al., 2008). SOAs rely on a collection of loosely coupled, self-contained services that communicate with each other through

the Internet and that can be called from multiple clients (e.g., Excel, MATLAB, Visual Studio, etc.) in a standard fashion (Maidment, 2008). Web services can be distributed at many different locations, eliminating the need to consolidate data in a central location.

4. Syntactic and semantic heterogeneity in environmental observations data

Syntactic heterogeneity refers to a difference in how data and metadata are organized (e.g., rows vs. columns) and encoded (e.g., text files versus Excel spreadsheets), while semantic heterogeneity refers to the variety in language and terminology used to describe observations. Syntactic heterogeneity arises where there are methodological inconsistencies. For example, data downloaded from automated data loggers are generally encoded as delimited text files, whereas data generated as a result of chemical analysis of water samples in a laboratory may be entered by hand from a hard-copy laboratory report into an Excel spreadsheet. In addition to these methodological differences, different software applications have given rise to the proliferation of different file types and formats.

Semantic heterogeneity occurs when there is disagreement about the meaning, interpretation, or intended use of the same or related data (Sheth and Larson, 1990). Among observational data, this heterogeneity can be generalized into two types: 1) structural – i.e., the language used to describe the names of observation attributes; and 2) contextual – i.e., the language used to encode observation attribute values. Structural heterogeneity begs the questions – what are the common attributes of environmental observations, and what should those attributes be called? For example, should the location at which an observation was made be called a "monitoring site" or a "station?" Should the measured quantity be called a "variable" or a "parameter?" This type of semantic heterogeneity is structural because it determines the structure of any model that is used to represent the data.

Contextual heterogeneity lies in the attribute values themselves. For example, one attribute of scientific observations is the name of the variable that was measured. It is common for different investigators to use different names for the same variable (e.g., "discharge" versus "streamflow"), or the same name for different variables (e.g., using a single term "temperature" to represent both air temperature and water temperature). Many of the semantic differences that arise in research datasets are a result of investigator preference and inconsistencies among scientific domains. Table 1 provides examples of semantic heterogeneity in data from two popular water resources data sources and demonstrates both structural and contextual semantic heterogeneity.

The implications of syntactic and semantic heterogeneity in publishing environmental observations data are threefold – first in users finding the data, second in decoding and organizing the data, and third in interpreting them. Within water resources research, data are available from many different sources that use different nomenclature, storage technologies, user interfaces, and even languages, making data discovery a difficult and time consuming task (Beran and Piasecki, accepted for publication). Data discovery is an important aspect of the cyberinfrastructure required to support publication of research data because scientists' ability to find, decode, and interpret available datasets will determine how or if the data are used for scientific analyses. Performance of queries and search mechanisms for data discovery can be significantly improved when syntactic and semantic heterogeneity among datasets is overcome (Madin et al., 2007; Beran and Piasecki, accepted for publication). After data are discovered, much research time and effort (up to 50% or more) is spent decoding, manipulating, and organizing observational data into a format that is useful (Bandaragoda et al., 2005; Ruddell and Kumar, 2006; Ramachandran et al., 2005). This process is also error prone. Specialized knowledge and expensive software may be required to handle files in different formats from disparate sources.

Serious errors in data use and interpretation can result from semantic heterogeneity in data from different sources. This was spectacularly demonstrated when navigators of NASA's $125 million Mars Climate Orbiter sent the spacecraft off course to its eventual loss because they assumed that data used to compute the effects of thruster firings on the trajectory of the spacecraft were in metric units when they were in fact in English units (Mars Climate Orbiter Mishap Investigation Board, 1999). Madnick and Zhu (2006) use this example as well as many others to describe how many perceived data quality problems are actually data misinterpretation problems that result from semantic heterogeneity. It is critical, therefore, that data are published with sufficient metadata so that they can be unambiguously interpreted.

5. Overcoming heterogeneity

Reconciling heterogeneity in data from different sources, which may be required both within and across research sites, is a complex problem that has a long history in information science (Bergamaschi et al., 2001; Colomb, 1997; Cox et al., 2006). This challenge is fueling much of the movement toward using standardized markup languages as self-describing, common data formats that can be used by data producers and data consumers. Examples include Earth Science Markup Language (ESML) (Ramachandran et al.,

Table 1
Examples of semantic heterogeneity in two popular water resources datasets demonstrating both structural and contextual semantic heterogeneity.

General description of attribute	USGS NWIS[a]	EPA STORET[b]
Structural semantic heterogeneity		
Code for location at which data are collected	"site_no"	"Station ID"
Name of location at which data are collected	"Site" OR "Gage"	"Station Name"
Code for measured variable	"Parameter"	?[c]
Name of measured variable	"Description"	"Characteristic Name"
Time at which the observation was made	"datetime"	"Activity Start"
Code that identifies the agency that collected the data	"agency_cd"	"Org ID"
Contextual semantic heterogeneity		
Name of measured variable	"Discharge"	"Flow"
Units of measured variable	"cubic feet per second"	"cfs"
Time at which the observation was made	"2008-01-01"	"2006-04-04 00:00:00"
Latitude of location at which data are collected	"41°44'36"	"41.7188889"
Type of monitoring site	"Spring, Estuary, Lake, Surface Water"	"River/Stream"

[a] United States Geological Survey National Water Information System (http://waterdata.usgs.gov/nwis/).
[b] United States Environmental Protection Agency Storage and Retrieval System (http://www.epa.gov/storet/).
[c] An equivalent to the USGS parameter code does not exist in data retrieved from EPA STORET.

2005), Ecological Metadata Language (EML) (EML Project Members, 2008), Water Markup Language (WaterML) (Zaslavsky et al., 2007), and the Open Geospatial Consortium's (OGC) Observations and Measurements (O&M) (Cox, 2006). Other methods that have been used for this task include the use of standard data models, controlled vocabularies, and ontologies. In evaluating these methods, an important distinction must be made between technologies for data communication (i.e., the formats and mechanisms used to transmit data to consumers) and technologies for persistent data storage and management (i.e., the formats and mechanisms used by the data source for long term storage and management). Approaches for handling heterogeneity within these two distinct data publication tasks can be quite different, but both should be addressed in the publication of research datasets.

Existing published data sources such as NWIS, NCDC, and STORET provide a good example of the data publication problem. Data stored within these systems hold much value for scientific research, but each has its own autonomous methods for storing, managing, and communicating its data. Providing consistent access to the datasets from each of these federal data providers is important in leveraging these data for scientific research, but it requires mediating across the different data formats and vocabularies of each of these systems. Overcoming heterogeneity in these existing data repositories is mainly an issue of data communication (i.e., can the data from each of these systems be provided to users in a format that is syntactically and semantically similar regardless of their source?) because the data sources do not have the same underlying persistent storage or data communication mechanisms.

Standardized markup languages such as ESML, EML, WaterML, O&M, and others provide a structured syntax for communicating data from multiple sources as XML documents. These markup languages can be used to transmit data in a format that resolves syntactic heterogeneity, but they generally do not place semantic constraints on the meanings of the document contents. Recognizing this, scientists have begun to use ontologies in concert with these markup languages to overcome semantic heterogeneity in scientific data (Beran and Piasecki, accepted for publication; Lin and Ludäscher, 2003; Madin et al., 2007). A domain ontology defines the terms used to describe and represent an area of knowledge and that are used by people, databases, and applications that need to share domain information (Heflin, 2004). Ontologies can be implemented as structured, machine-interpretable vocabularies that include definitions of basic concepts in a domain and the relationships among them, thus capturing the semantics of the data that they represent.

Within a scientific domain, ontologies can provide a conceptual view of data stored within a variety of databases, and, because they can be formalized into machine-interpretable forms, they are powerful tools for virtually integrating disparate data sources without replicating the data or changing its persistent storage mechanism. For example, Beran and Piasecki (accepted for publication) describe an ontology-aided search engine called Hydroseek (http://www.hydroseek.org) that was specifically designed to mediate across the disparate formats and vocabularies of several national hydrologic data providers and provide users with a single interface to query and retrieve consistently formatted data from each of these data repositories. Hydroseek does not replicate or store the data from each of these repositories; it simply retrieves data from its source and communicates it to a user in a consistent format. Hydroseek's data discovery mechanism is based on an ontology that stores the vocabulary terms (e.g., variable names) from each of the data sources and the relationships between them so that a search using a single term such as "discharge" can return results from multiple data sources, even if some of those data sources use a different but equivalent term such as "streamflow" to describe their data. One significant barrier in using this approach, however, is that constructing the ontology that mediates across the vocabularies used by each data source is a difficult task that is prone to error because the mapping of terms from one source to another must be done by people who know how to interpret both vocabularies and there isn't always a one-to-one translation or mapping of terms.

Because the underlying data formats, vocabularies, and communication mechanisms of existing national data sources are different for each source, tools such as standardized markup languages and ontologies are needed to mediate across the sources and provide consistent access to the data. Unlike existing national data networks, however, most research datasets have not been formally published, they have not adopted standard methods for either persistent data storage or for data communication, and they have not settled on a specific vocabulary or format that define the syntax and semantics of the data. The opportunity exists, therefore, for the community of scientists collecting environmental and water resources data to build and adopt common data models and common vocabularies to describe the observations data for both storage and management and communication of data that are collected. A standardized data publication system can be used to resolve heterogeneity in existing datasets, both at the storage and communication levels, and to prevent heterogeneity in data to be collected in the future. Obviously, the easiest way to resolve heterogeneity is for it to never exist in the first place.

6. Design of a cyberinfrastructure for publishing environmental observations

The objective of a cyberinfrastructure for publishing environmental observations is to enable disparate users to publish data in a way that makes them available, interpretable, and interoperable. Cyberinfrastructure that supports publication of environmental observations data must answer the following four questions: 1) what are the characteristics of data that enable them to be interpreted and how will they be persistently stored; 2) what is the mechanism by which users will access the data and in what format will the data be communicated to them; 3) how will data from multiple sources be made interoperable; and 4) how will users discover the data? The answers to these questions guide the design of cyberinfrastructure for data publication and may differ across scientific domains. Within the Hydrologic Science community, a Hydrologic Information System (HIS) is under development by the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) in efforts to address these questions for point observations data.

The following sections describe the design of the CUAHSI HIS data publication system and how components of the HIS address each of the questions above. Fig. 1 shows the general architecture of this system and illustrates the process for publishing data. Data collected in the field using in-situ sensors or other sampling techniques are stored in a variety of differently formatted files. Data from these files are loaded into a database with special attention given to populating metadata using controlled vocabularies. Web services then make the data available over the Internet. Last, the address of the web services is registered with a central registry, announcing the availability of the data to the public and enabling data discovery tools like Hydroseek, which provide map and context based search capabilities, to consume the data.

6.1. Persistent storage and management

Within the HIS, persistent storage and management of observations data and their associated metadata are accomplished using the Observations Data Model (ODM). ODM is a relational model

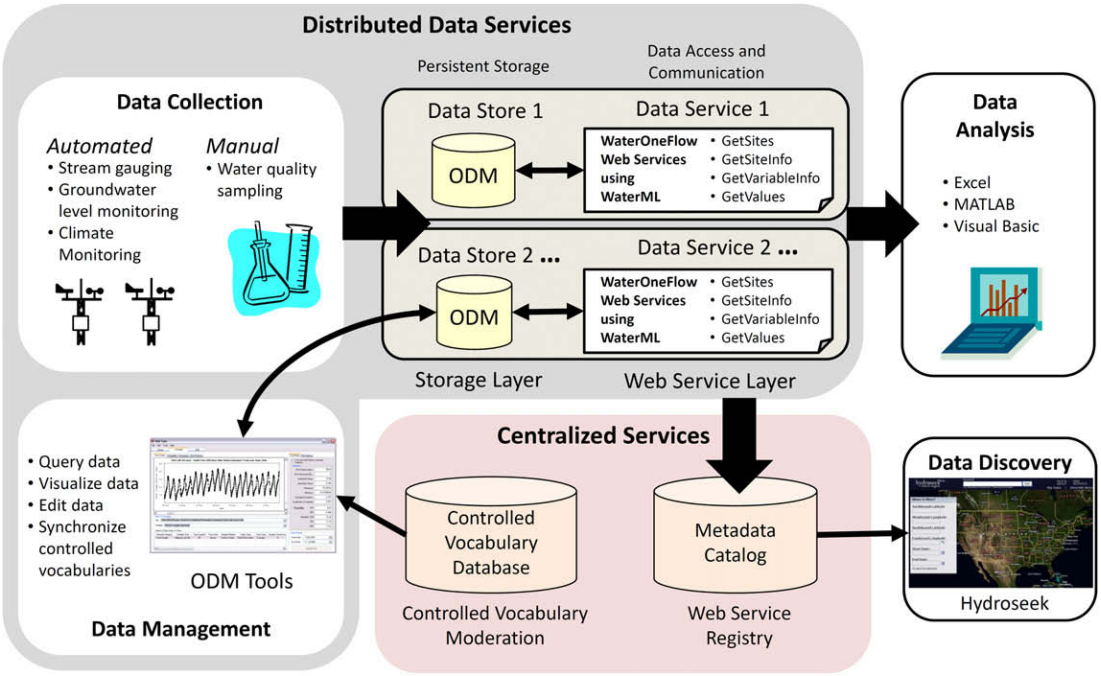

Fig. 1. General architecture of the CUAHSI HIS data publication system. Data are collected using field sensors and other observational procedures. Observational data with multiple formats are combined within a single ODM database where they are annotated with appropriate metadata using the ODM controlled vocabularies. Data are then published using ODM web services, which are registered with the central web services registry to enable integrated search and discovery.

that defines the persistent structure of data, including the set of attributes that accompany the data, their names, their data type, and their context (Horsburgh et al., 2008). 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.), which can greatly facilitate their use within common analyses. Not only can the data be standardized and appropriately qualified with metadata, which is required for them to be interpreted without ambiguity, but applications that interact with ODM can be harmonized, leading to greater cooperation, sharing (of both data and application code), and interoperability.

ODM was designed to be implemented within a Relational Database Management System (RDBMS). This provides several advantages. First, RDBMS are robust and technologically mature, having many different commercial implementations and having long ago demonstrated their reliability, scalability, and performance (Connolly and Begg, 2005). Additionally, RDBMS provide a standardized query language through which data can be manipulated, and, more recently, tools for advanced data analysis and manipulation such as online analytical processing (OLAP), data mining, and data warehousing.

6.2. Data access and communication

The main mechanism for communicating observational data stored in an ODM database to users is the WaterOneFlow web services. The WaterOneFlow web services respond to user queries and transmit data extracted from an ODM database encoded using WaterML (Zaslavsky et al., 2007). User queries are performed by calling methods that are exposed by the web services, such as *GetSites* for returning a list of sites within an ODM database along with the metadata for each site, *GetVariableInfo* for returning a list of variables within an ODM database along with the metadata for each variable, *GetSiteInfo* for returning a list of variables with data at a site, and *GetValues* for returning the time series of data for a site and variable combination. The web service methods can be called from many different programming languages and other software applications, including Microsoft Visual Basic, Microsoft Excel, MATLAB, and others from anywhere an Internet connection is available. Using web services, users can discover the data that they are interested in and then access it using the analysis software of their choice, rather than being forced to learn a new analysis system. The service-oriented architecture used by the HIS and represented by the WaterOneFlow web services serves to get the browser out of the way for data acquisition, thus enhancing environmental analysis and modeling capabilities through direct access to remote data sources from a wide range of software environments.

6.3. Data interoperability

The use of ODM as the persistent data storage mechanism and the WaterOneFlow web services as the data communication mechanism for the data publication system has several advantages in resolving data heterogeneity and promoting interoperability among datasets. First, ODM addresses the syntactic heterogeneity in the data (i.e., different file types, data formats, etc.) collected both within and across research sites. By loading data into an ODM database, data managers at each research site ensure that their data are syntactically similar to the data at all other sites that are using ODM. Second, because ODM defines the attributes that accompany the data and their context, loading data into ODM overcomes any structural semantic heterogeneity in the data.

Contextual semantic heterogeneity within and across research datasets is reduced through the use of controlled vocabularies 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). Fig. 2 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 (i.e., across research sites). 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.

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 in that it was developed by researchers working within the Hydrologic Science community. It is dynamic and growing; users can add new terms or edit existing terms by using the functionality available through the HIS website (http://his.cuahsi.org). If a data manager cannot find an appropriate term to describe data that is being added to an ODM database, he or she can navigate to the HIS website and use an online form to request addition of an appropriate term to the master controlled vocabulary. The ODM controlled vocabulary submission system (Fig. 3) 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 a software application called ODM Tools 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. The ODM Tools application was developed to provide data managers with a set of tools for managing data within an ODM database. Data managers can use functionality within ODM Tools 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 to facilitate the updating. Fig. 3 shows this interaction between the data manager, the ODM Tools application, and the ODM Controlled Vocabulary web services.

The WaterOneFlow web services preserve the semantic and syntactic homogeneity achieved by loading data into ODM because the data are transmitted over the Internet in a single format using a vocabulary that is consistent across research sites. They also promote the interoperability of the data through the use of standard web services protocols and XML formats that are platform and programming language independent. The WaterOneFlow web services are designed to be implemented on top of individual ODM databases so that the web services for each ODM database can be uniquely addressable. Each set of web services implements the same set of methods and returns data in the same format, but receives a unique URL for accessing the data in its underlying database. Because of this, users need only change the URL when

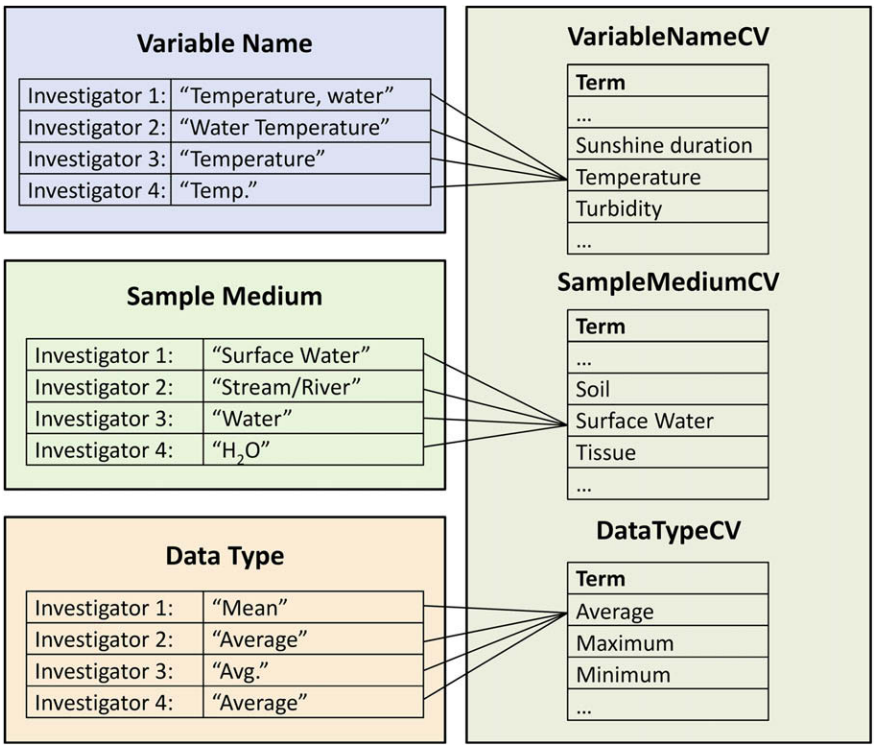

Fig. 2. Example of how contextual heterogeneity in the attributes of similar datasets from several different investigators can be reconciled through the use of the ODM controlled vocabularies.

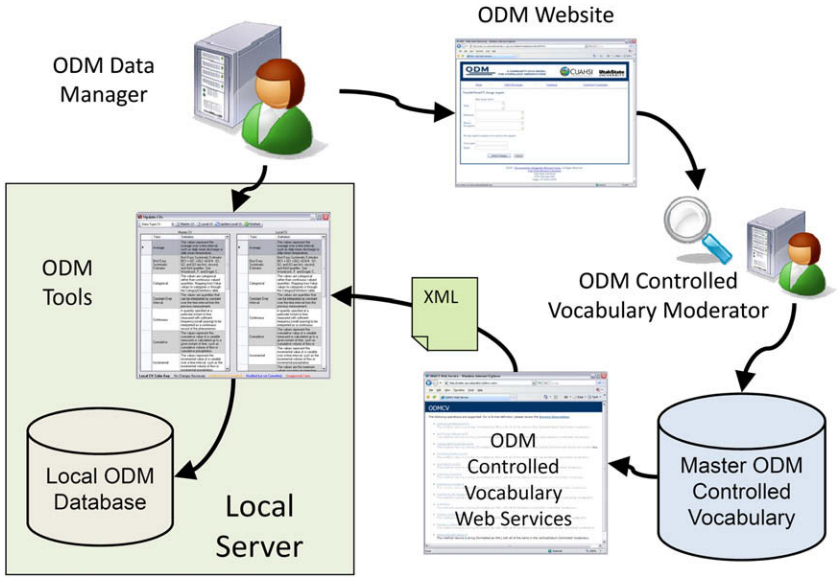

Fig. 3. The ODM controlled vocabulary system.

accessing data from multiple ODM databases via the WaterOneFlow web services. The WaterOneFlow web services for ODM are also consistent with WaterOneFlow web services that have been developed for the USGS NWIS, the USEPA STORET system, and other national hydrologic data providers (see http://his.cuahsi.org for a listing of all available web services). This means that data consumers can access data published using the HIS data publication system and data from national providers using a consistent set of methods, and data are returned in the same format from all of these sources.

6.4. Data discovery

Once data have been loaded into an ODM database and the WaterOneFlow web services have been implemented on top of that database, the data can be accessed over the Internet. However, making the data available on the Internet does not necessarily mean that they are easily discoverable. Because of this, the data publication process is not complete until the address of the web services has been registered with a central repository that stores links to each of the web services that make up the research data network and some metadata about each. The central web services registry is essentially a digital card catalog – it stores enough information about each of the databases and web services to know what they contain and how to access them, but it does not contain the published data. Users can navigate to the central web services registry from http://his.cuahsi.org and browse through the list of registered web services to determine which data are available. They can then query individual web services to get more detailed metadata and download the data.

Registering web services with the central registry also ensures that the data are available to centralized discovery, delivery, visualization, and analysis tools that have been developed as part of the HIS. For example, the Hydroseek application that was described previously has the capability to discover and deliver all of the data within databases and web services registered with the central registry. Simple keyword searches within Hydroseek return results from research sites that have implemented the HIS data publication system alongside data from other national data providers, and the data from all of these sources are delivered to users in a consistent and easy to use format.

7. A case study for publishing point observations data: creating a national research data network

Within the United States, a network of large-scale environmental observatories, which are integrated real-time observing systems that seek to improve understanding of the earth's water and biogeochemical cycles across multiple spatial and temporal scales, has been proposed under the premise that knowledge of the physical, chemical, and biological mechanisms controlling water quantity and quality is limited by lack of observations at the necessary spatial density and temporal frequency needed to infer the controlling processes (Montgomery et al., 2007). These observatories are intended to be community resources, and advanced cyberinfrastructure will be required as a central component in the planning and design of the observatory network to support collection, management, use, and publication of the datasets that are generated.

As part of the process of planning for the network of large-scale environmental observatories, 11 test bed projects, which are part of the Water and Environmental Research Systems (WATERS) Network (http://www.watersnet.org) and are located across the United States, have been established to develop techniques and technologies that will be used in the design of the network of large-scale observatories. The data being collected differs from one test bed to the next, but examples include: discharge and water quality variables such as water temperature, dissolved oxygen concentration, and turbidity; samples of water quality constituents such as nutrients and sediment; groundwater levels and quality; and meteorological variables such as precipitation, air temperature, and solar radiation. Because data collection is occurring at a variety of spatial and temporal scales, spanning different scientific investigators and domains, and across a variety of different locations and watersheds, heterogeneity has emerged within the datasets that have been collected, especially from one test bed to the next. More information about the test beds and the data being collected at each can be found at the following URL (http://www.watersnet.org/wtbs/index.html).

The CUAHSI HIS has been adopted by each of the test beds as a common data publication cyberinfrastructure, with goals of enabling cross-domain analysis within individual test beds as well as cross-test bed sharing and analysis of data. Each test bed

received a computer server and a suite of software tools for accomplishing the task. The server was pre-configured with the required operating system, RDBMS, and web server software. Also installed on the servers was the ODM Data Loader, which is a software application that assists users in loading data into an ODM database, and the ODM Tools application mentioned above.

Data managers at each of the test beds created one or more ODM databases into which they loaded their point observations data. Creation of an ODM database required the data managers to download a blank ODM database schema and attach it to their RDBMS. The blank ODM schema contains the full implementation of ODM, including all of the tables, fields, relationships, constraints, and controlled vocabularies. The data managers then used either the ODM Data Loader or the native data import tools of the RDBMS to populate their blank ODM databases with their observational data, using the controlled vocabularies to populate the metadata fields in the database. Finally, the data managers implemented WaterOneFlow web services for each of their ODM databases and registered the web services with the central web services registry. Implementation of the web services involved downloading the web service application files, copying them to the hard drive of the server, and then configuring the application using the instructions and configuration tools provided with the download. Registration with the central web services repository required data managers to navigate to the central website and fill out a form with information about their database and web services (e.g., the location of the web services and a brief description of the contents of the database). The resulting set of registered WaterOneFlow web services created by the test bed data managers represents a national network of syntactically and semantically similar scientific research data.

A snap-shot summary of the data published within the research data network, which now includes data from the test beds and other external data sources that have joined the network, is provided in Table 2 and Fig. 4. The statistics for the research data network were compiled using Visual Basic code that was written to call each of the published web services and compile an overall list of sites and variables, along with a summary of the observations for each site and variable combination. Table 2 lists statistics for the entire network of research sites, and Fig. 4 shows the number of monitoring sites, variables, and data values collected at each research site that has been added to the network. In Fig. 4, each dot on the map represents an ODM database with a corresponding set of WaterOneFlow web services. The dots are plotted at the location of the average latitude and longitude of all of the monitoring sites stored in the ODM database.

The numbers in Table 2 and Fig. 4 represent a snap-shot in time because new sites, variables, and data values are continually being added to the research data network. The following definitions apply for Table 2 and Fig. 4: a data source is the organization that collected the data; a monitoring site is a location at which data are collected and is identified by its latitude and longitude coordinates; a variable is characterized by the combination of its name (e.g., temperature), the medium in which it was sampled (e.g., surface water), how the measurement was obtained (e.g., field observation), the time support interval over which the observation was made (e.g., hourly), its data type (e.g., average), and the method used to make the measurements (e.g., the type of temperature sensor used); and a data value is a single observation of a single variable at a single site on a particular date and time (e.g., the dissolved oxygen concentration at site x was 8.3 mg L^{-1} on April 7, 2008 at 3:00 PM).

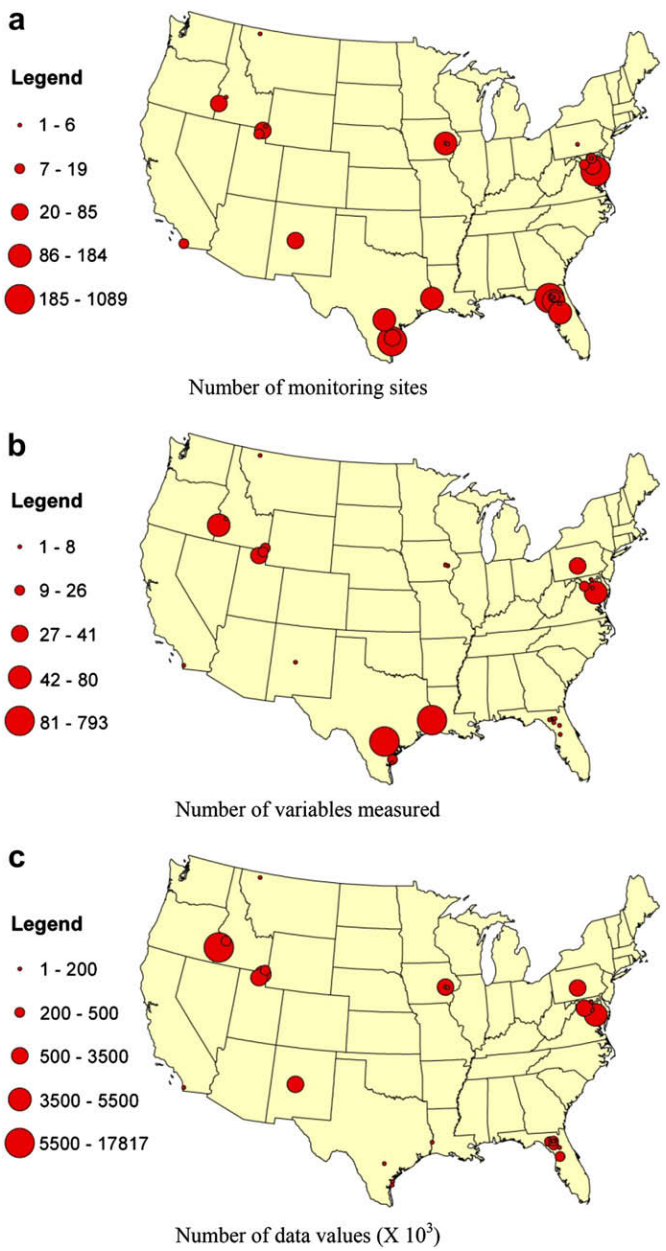

Fig. 4. Distribution of monitoring sites (a), variables (b), and data values (c) across the United States in the research data publication network as of June 17, 2008.

Table 2
Test bed data network summary as of June 17, 2008.

Item	Total number
ODM databases	31
Data sources	41
Monitoring sites	3767
Variables	793
Measurement methods	99
Data values	41,651,095

8. Discussion and conclusions

A standard method for publishing environmental and water resources point observations data has been presented. It provides a framework in which data of different types and from disparate sources can be integrated, while overcoming the syntactic and semantic heterogeneity in the data from each source. This has been the case at each site within a network of environmental

observatory test beds in the United States, where publishing observational data using this system has enabled a group of independent test bed investigators working on very different science problems to create a network of syntactically and semantically similar scientific data. The research data network, which is a full implementation of the data publication system as it was designed, now contains over 3700 data collection sites, nearly 800 measured variables, and nearly 42 million individual data values. The data publication system's flexibility in storing and enabling public access to similarly formatted data and metadata from multiple scientific domains and research sites has created a community data resource from data that might otherwise have been confined to the private files of the individual investigators.

Much of the success of the data publication system can be attributed to the federation of the individual databases. Each of the test beds maintains their own databases, and each is ultimately in charge of which data get published. Some have chosen to publish raw sensor data as it streams into their ODM database from field based sensors. Some have chosen to publish only data that have undergone quality control procedures. ODM stores data qualifying comments and information about the level of quality control data have been subjected to, and the WaterOneFlow web services transmit this information to ensure that users are aware of the quality and limitations of the data. Issues of data editing and cleansing, metadata population, data aggregation, and derived data generation are left to the data collectors who are most familiar with their datasets.

A significant challenge associated with this distributed data storage approach is that resources and expertise are required to implement the publication tools at each local research site. The data publication system requires a server on which an ODM database and a set of WaterOneFlow web services has been implemented. The server must be capable of hosting web applications, but does not have to be an expensive machine. Expertise with server administration, relational database management systems, and installing and configuring Internet applications is helpful for data managers; however, instructions for implementing ODM databases and the WaterOneFlow web services are contained in documentation available via the CUAHIS HIS website (http://his.cuahsi.org). Data managers with varying levels of expertise at the 11 test beds were able to successfully publish data using the system after having received a pre-configured server, although available documentation describes how to install and configure all of the components of the data publication system and a pre-configured server is not required. Once the ODM database and web services are set up, they require little maintenance apart from loading new data if and when it becomes available. Personnel (i.e., data manager) resources required to implement the system depend on the amount and complexity of the data to be published. The degree to which data acquisition is automated and the level of manual quality control to which the data are subjected are also drivers in the required personnel costs.

One advantage of this data publication system is that a standard, robust data model and controlled vocabularies ensure consistent and fully specified data and metadata, leading to higher quality analysis with less uncertainty and fewer data interpretation errors. Mapping datasets to the ODM schema and choosing appropriate controlled vocabulary terms can be challenging for data managers and can be error prone; however, the value of fully specified metadata cannot be overstated. Federation of individual databases (i.e., test bed or observatory databases) is also simplified because each of the databases has the same format and uses the same vocabulary. This simplifies the design of applications that facilitate data discovery across the entire network of published data. Additionally, because a consistent data model and vocabulary are used across sites, software application development can also be standardized and components reused at each site.

The ODM controlled vocabulary system provides a community resource for building a common vocabulary for environmental and water resources data and is a good example of how common systems can support a larger community. Other software tools include the WaterOneFlow web services, data loading and editing tools for ODM, and data visualization and retrieval tools that interact with the WaterOneFlow web services. Readers are referred to the CUAHSI HIS website for details of these software applications (http://his.cuahsi.org). The free availability of these software tools is a significant asset to investigators who cannot afford or do not have the expertise to develop sophisticated and interactive data publication websites on their own.

The data publication system described in this paper is not limited to test beds or environmental observatories, and, because of this, the network of available data is expected to grow. Data from several research sites outside of the original 11 test beds have already been published using this system. Investigators working outside of the environmental observatory community can adopt the methods and available software tools to publish their own data. By doing so, the network of observatories and other data sources that adopt the same infrastructure, although separated in space, will become an integrated network of consistent data like NWIS, STORET, and other national repositories. Sophisticated tools such as ontologies may still be needed to integrate research datasets with those from other national data providers, but one level of complexity (i.e., semantic and syntactic heterogeneity among the network of research datasets) can be avoided through the adoption of a common data publication system and common vocabulary.

Last, the conceptual framework of the data publication system presented in this paper (i.e., a common data model, a centralized controlled vocabulary system, web services for communicating data from federated data sources, and a central registry for web services) can be applied within any domain in which a community of diverse investigators is collecting data.

9. Software and data availability

The software components described in this paper, including ODM, the ODM Data Loader, ODM Tools, the ODM controlled vocabulary system, the WaterOneFlow web services, and the central web services registry can be accessed through the CUAHSI HIS website http://his.cuahsi.org, where they are distributed for free under the Open Source Berkeley Software Distribution (BSD) license. The test bed data described in this paper can be accessed through the individual web services for each test bed, which are listed in the central web services registry, also available through the HIS website.

Acknowledgments

The data publication system described in this paper has been 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 grants EAR 0413265 and 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

Bandaragoda, C.J., Tarboton, D.G., Maidment, D.R., 2005. "User Needs Assessment." In: Maidment, D.R. (Ed.), Hydrologic Information System Status Report, Version 1. pp. 48–87 (Chapter 4).<http://www.cuahsi.org/docs/HISStatusSept15.pdf>. (last accessed 06.04.08.).

Beran, B., Piasecki, M., Engineering new paths to water data. Computers and Geosciences, accepted for publication. doi:10.1016/j.cageo.2008.02.017.

Bergamaschi, S., Castano, S., Vincini, M., Beneventano, D., 2001. Semantic integration of heterogeneous information sources. Data & Knowledge Engineering 36 (3), 215–249, doi:10.1016/S0169-023X(00)00047-1.

Borgman, C.L., Wallis, J.C., Enyedy, N., 2007. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries 7 (1), 17–30, doi:10.1007/s00799-007-0022-9.

Bosch, D.D., Sheridan, J.M., Lowrance, R.R., Hubbard, R.K., Strickland, T.C., Feyereisen, G.W., Sullivan, D.G., 2007. Little river experimental watershed database. Water Resources Research 43 (W09470), doi:10.1029/2006WR005844.

Colomb, R.L., 1997. Impact of semantic heterogeneity on federating databases. The Computer Journal 40 (5), 235–244, doi:10.1093/comjnl/40.5.235.

Connolly, T., Begg, C., 2005. Database Systems: a Practical Approach to Design, Implementation, and Management, fourth ed. Addison-Wesley, Harlow, U.K.

Cox, S., Ed., 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 23.01.08.).

Cox, S., Jones, R., Lawrence, B., Milic-Frayling, N., Moreau, L., 2006. Interoperability Issues in Scientific Data Management (Version 1.0). Technical report, The Technical Computing Initiative, Microsoft Corporation. March 2006. <http://download.microsoft.com/download/f/b/3/fb3d02b8-2210-4d0d-a747-9519eafae6c1/ScientificDataManagement4.18.07.pdf>.

EML Project Members, 2008. Ecological Metadata Language (EML). <http://knb.ecoinformatics.org/software/eml> (last accessed 26.02.08.).

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. Environmental Modelling & Software 23 (4), 404–411.

Gray, J., Liu, D.T., Nieto-Santisteban, M., Szalay, A., DeWitt, D.J., Heber, G., 2005. Scientific data management in the coming decade. ACM SIGMOD Record 34 (4), 34–41, doi:10.1145/1107499.1107503.

Heflin, J. (Ed.), 2004. OWL Web Ontology Language Use Cases and Requirements W3C Recommendation 10 February 2004. <http://www.w3.org/TR/webont-req> (last accessed 26.02.08.).

Helly, J.J., 2006. Digital library technology for hydrology (Chapter 3). In: Kumar, P.K., Alameda, J., Bajcsy, P., Folk, M., Markus, M. (Eds.), Hydroinformatics Data Integrative Approaches in Computation, Analysis, and Modeling. CRC Press, Boca Raton, FL, pp. 21–37.

Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., Zaslavsky, I., 2008. A relational model for environmental and water resources data. Water Resources Research 44 (W05406), doi:10.1029/2007WR006392.

Lin, K., Ludäscher, B., 2003. A system for semantic integration of geologic maps via ontologies. In: Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data (SCISW). Sanibel Island, Florida. October 20, 2003. <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-83/sia_2.pdf>. (last accessed 21.02.08.).

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., Villa, F., 2007. An ontology for describing and synthesizing ecological observation data. Ecological Informatics 2 (3), 279–296, doi:10.1016/j.ecoinf.2007.05.004.

Madnick, S., Zhu, H., 2006. Improving data quality through effective use of data semantics. Data & Knowledge Engineering 59 (2), 460–475, doi:10.1016/j.datak.2005.10.001.

Maidment, D.R. (Ed.), 2008. CUAHSI Hydrologic Information System: Overview of Version 1.1. Consortium of Universities for the Advancement of Hydrologic Science Inc., Washington, D.C., p. 92. Available from: <http://his.cuahsi.org/documents/HISOverview.pdf>.

Mars Climate Orbiter Mishap Investigation Board, 1999. Mars Climate Orbiter Mishap Investigation Board Phase I Report. November 10, 1999. <ftp://ftp.hq.nasa.gov/pub/pao/reports/1999/MCO_report.pdf>. (last accessed 22.02.08.).

Montgomery, J.L., Harmon, T., Kaiser, W., Sanderson, A., Haas, C.N., Hooper, R., Minsker, B., Schnoor, J., Clesceri, N.L., Graham, W., Brezonik, P., 2007. The WATERS Network: an integrated environmental observatory network for water research. Environmental Science and Technology 41 (19), 6642–6647. <http://pubs.acs.org/subscribe/journals/esthag/41/i19/pdf/100107feature_waters.pdf> (last accessed 30.01.08.).

Moran, S.M., Emmerich, W.E., Goodrich, D.C., Heilman, P., Holifield Collins, C.D., Keefer, T.O., Nearing, M.A., Nichols, M.H., Renard, K.G., Scott, R.L., Smith, J.R., Stone, J.J., Unkrich, C.L., Wong, J., 2008. Preface to special section on Fifty Years of Research and Data Collection: U.S. Department of Agriculture Walnut Gulch Experimental Watershed. Water Resources Research 44 (W05S01), doi:10.1029/2007WR006083.

Morocho, V., Saltor, F., Perez-Vidal, L., 2003. Ontologies: solving semantic heterogeneity in federated spatial database system. In: Proceedings of 5th International Conference on Enterprise Information System, Angers, France. April, 2003. pp. 347–352. <http://citeseer.ist.psu.edu/morocho03ontologies.html>. (last accessed 21.02.08.).

National Science Foundation, 2007. Cyberinfrastructure Vision for 21st Century Discovery. NSF 07-28.<http://www.nsf.gov/pubs/2007/nsf0728/index.jsp>. (last accessed 06.04.08.).

Nichols, M.H., Anson, E., 2008. Southwest Watershed Research Center Data Access Project. Water Resources Research 44 (W05S03), doi:10.1029/2006WR005665.

Ramachandran, R., Christopher, S.A., Movva, S., Li, X., Conover, H.T., Keiser, K.R., Graves, S.J., McNider, R.T., 2005. Earth Science Markup Language: a solution to address data format heterogeneity problems in atmospheric sciences. Bulletin of the American Meteorological Society 86 (6), 791–794, doi:10.1175/BAMS-86-6-791.

Research Information Network, 2008. To Share or Not to Share: Publication and Quality Assurance of Research Data Outputs. Report commissioned by the Research Information Network (RIN). <http://www.rin.ac.uk/data-publication>. (last accessed 17.06.08.).

Ruddell, B.L., Kumar, P., 2006. Hydrologic Data Models (Chapter 5). In: Kumar, P.K., Alameda, J., Bajcsy, P., Folk, M., Markus, M. (Eds.), Hydroinformatics Data Integrative Approaches in Computation, Analysis, and Modeling. CRC Press, Boca Raton, FL, pp. 61–79.

Sheth, A.P., Larson, J.A., 1990. Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys 22 (3), doi:10.1145/96602.96604.

Slaughter, C.W., Marks, D., Flerchinger, G.N., Van Vactor, S.S., Burgess, M., 2001. Thirty-five years of research data collection at the Reynolds Creek Experimental Watershed, Idaho, United States. Water Resources Research 37 (11), doi:10.1029/2001WR000413.

Zaslavsky, I., Valentine, D., Whiteaker, T. (Eds.), 2007. CUAHSI WaterML. OGC Discussion Paper OGC 07-041r1. Version 0.3.0. <http://www.opengeospatial.org/standards/dp>. (last accessed 23.01.08.).