Advancing Solutions for an EarthCube Design. What can be learned from the CUAHSI HIS experience?

An EarthCube Design Approaches Paper

Prepared by: David Tarboton[1], David Maidment[2], Ilya Zaslavsky[3], Jon Goodall[4], Dan Ames[5], Richard P Hooper[6], Jeffery Horsburgh[1], Kim Schreuders[1], Ray Idaszak[7], Alva Couch[8]

October 17, 2011

Introduction

The EarthCube goal of creating an infrastructure for managing and integrating knowledge across all geosciences in an open, transparent and inclusive manner aligns well with the overarching goals for the work that has been done in CUAHSI on Hydrologic Information Systems. CUAHSI has taken some steps in this direction, but a lot remains to be done. This paper describes these steps and makes some suggestions for the path ahead.

CUAHSI is the Consortium of Universities for the Advancement of Hydrologic Science, Inc., established in 2001 to promote research infrastructure that advances Hydrologic Science. Membership now stands at over 100 Universities and research organizations. Hydrologic Information Systems (HIS) have been part of this infrastructure right from the conception of CUAHSI. The goals of CUAHSI HIS are:

- Data Access – providing better access to a large volume of high quality hydrologic data;
- Hydrologic Observatories – storing and synthesizing hydrologic data for a region;
- Hydrologic Science – providing a stronger hydrologic information infrastructure;
- Hydrologic Education – bringing more hydrologic data into the classroom.

This paper reviews HIS capability developed to address these goals by promoting data sharing and interoperability in the Hydrologic Sciences with the ultimate goal of enabling hydrologic analyses that integrate data from multiple sources. We then abstract and distill the key lessons learned and conclusions drawn from this work regarding the general architectural framework and cyber-infrastructure functional components that could apply more broadly across the geosciences domain pertinent to EarthCube.

This paper is one of a group of EarthCube papers related to CUAHSI. The paper "The Open Geospatial Consortium and EarthCube" being submitted by David Maidment describes how an Open Geospatial

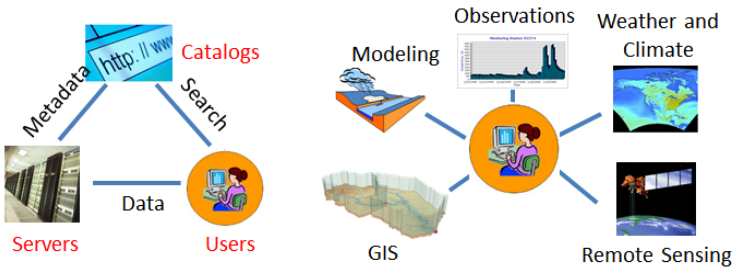
[1] Utah Water Research Laboratory, Utah State University
[2] Center for Research in Water Resources, University of Texas at Austin
[3] San Diego Supercomputer Center, University of California, San Diego
[4] University of South Carolina
[5] Idaho State University
[6] Consortium of Universities for the Advancement of Hydrologic Science, Inc.
[7] RENCI, University of North Carolina
[8] Tufts University

Consortium process could be useful for broad sharing of information in the geosciences. The paper "Improving the Interoperability of Earth Observations" being submitted by Jeff Horsburgh describes ideas for enhancing a common observations information model by linking data to the geo-environment and capturing the knowledge content of data. The paper "The chain of trust: making shared data plausible" being submitted by Alva Couch addresses issues related to trust and uncertainty in aggregated data sets from a Bayesian perspective. The paper "CZO as an EarthCube prototype" by Ilya Zaslavsky presents ideas drawn from work on the Critical Zone Observatory integrated data system. The paper "Use Cases to Test OGC O&M Profile" by Richard Hooper examines how the Open Geospatial Consortium's Observations and Measurement (O&M) profile can be adapted to capture broader contextual information about data in a series of use cases presented.

The CUAHSI HIS

The CUAHSI HIS is an internet based system to support the sharing of hydrologic data (see http://his.cuahsi.org for details beyond the scope of this paper). It is comprised of hydrologic databases and servers connected through web services as well as software for data publication, discovery and access. The system that has been developed provides new opportunities for the water research community to approach the management, publication, and analysis of their data systematically. The system's flexibility in storing and enabling public access to similarly formatted data and metadata has created a community data resource from public and academic data that might otherwise have been confined to the private files of agencies or individual investigators. HIS provides an analysis environment for the integration of data from multiple sources and serves as a prototype for the infrastructure to support a network of large scale environmental observatories or research watersheds. As presently implemented HIS focuses on a relatively narrow class of data, namely in situ point time series of observations with some capability for including ex-situ samples and simple geospatial information. Nevertheless many of the concepts developed translate to the much broader class of data that EarthCube will need to include.

Two concepts, (1) the services oriented architecture; and (2) the desktop hydrologic information system underlie the architecture of the CUAHSI HIS (Figure 1).



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

Figure 1. Hydrologic Information System Overarching Vision

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

The concept of HIS desktop application software is somewhat analogous to Geographic Information System (GIS) desktop software that supports storage and analysis of logically linked data (Tomlinson, 2003). Our implementation, "HydroDesktop" provides an analysis environment within which data from multiple sources can be discovered, accessed and integrated.

We have developed prototype functionality for all three components of the services oriented architecture and data transmission formats for the data exchanges between them. In terms of the desktop hydrologic information system, we have developed a prototype desktop application that combines the analysis of GIS, modeling and observations. It downloads, stores and operates on the information on a local desktop computer. Our present implementation is still under active development and has not yet developed the capability to integrate weather, climate and remote sensing data illustrated in Figure 1, but does synthesize GIS, point observations and time series and modeling.

The HIS services-oriented architecture is comprised of three classes of functionality: 1) data publication (HydroServer), 2) data cataloging (HydroCatalog), and 3) data discovery, access and analysis (HydroDesktop) (Figure 2). This functionality follows the general paradigm of the Internet. HydroServer publishes data similar to the way Internet web servers publish content. HydroDesktop consumes data published from HydroServer, similar to the way web browsers consume Internet content. HydroCatalog supports data discovery based on indexed metadata similar to the way search engines support the discovery of Internet content. Syntactic (file types and formats) and semantic consistency has been a focus of HIS with an ontology and community controlled vocabulary used to harmonize the terminology used and support thematic key word based data discovery.

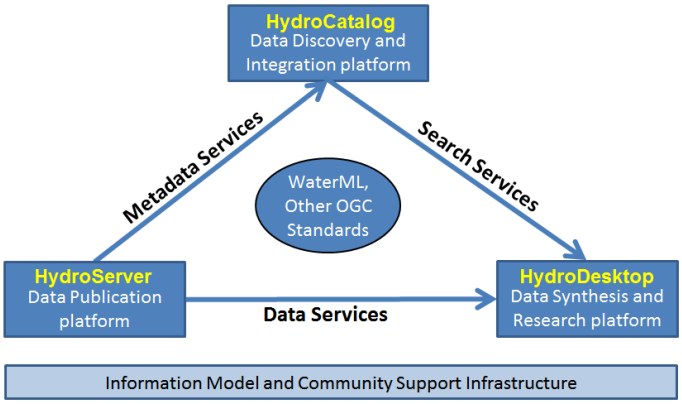


Figure 2. Components of CUAHSI HIS Services Oriented Architecture.

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

- Data Services – which convey the actual data.
- Metadata Services – which convey metadata about specific collections or series of data.
- Search Services – which enable search, discovery, and selection of data and convey metadata required for accessing data using data services.

The formats for transmission of information between these systems and the interfaces that enable the communication between them (the connecting arrows in Fig. 2) are critical to the functioning of the system. CUAHSI HIS has developed WaterML, an XML based language for transmitting observation data via web services (Zaslavsky et al., 2007). The web services are referred to as WaterOneFlow web services. CUAHSI HIS also relies on other established standards such as World Wide Web Consortium Simple Object Access Protocol (SOAP) and Open Geospatial Consortium (OGC) Geographic Markup Language (GML) for transmission of information between the three primary components.

At the base of Figure 2 is depicted the information model and community support infrastructure upon which the system is founded. The information model is the conceptual model used to organize and define sufficient metadata about hydrologic observations for them to be unambiguously interpreted and used. Within HydroServer, it is encoded using the Observations Data Model (ODM) (Horsburgh et al., 2008) relational database and the HydroServer Capabilities Database to ensure that data and metadata are stored together. The information model also serves as the conceptual basis for WaterML to ensure that data and associated metadata are transmitted with fidelity when data are downloaded. HydroDesktop implements the information model within its data repository database ensuring that local copies of data retrieved from a server maintain their original context. ODM includes a number of controlled vocabularies for metadata such as units, variable names, sample media etc., where semantic consistency in describing observations is important. The information model also includes a defined ontology used to represent a hierarchy of concepts that categorize the variables being measured. The ontology has been developed to support concept based search. The ontology and controlled vocabulary components of the information model have been developed to provide semantic consistency of the terms used in metadata and to support search and discovery based on these semantics. A web site collects and manages community additions and edits to controlled vocabulary content to allow dynamic growth of this content while encouraging semantic consistency across the user community.

The architecture shown in Figure 2 has evolved as an approach for sharing hydrologic observations data that is general and open to allow broad participation. The HydroServer software stack is not the only entry point for data publishers. Anyone can publish data using web services that deliver data in WaterML format and thus have their data become part of this system. Similarly the HydroCatalog and HydroDesktop functionality is not limited to the software we have developed. The definition of standard functionality for transmission of information to and from a catalog provider enables others to establish their own catalogs. HydroDesktop is our prototype client for consumption of web service based hydrologic data, but this does not preclude others from establishing their own clients. We envisage that EarthCube will work in a similar way with multiple servers, clients and catalogs interacting using standard services for the exchange of information. Defining and agreeing upon the services to achieve the needed functionality is a critical part of the problem.

Vision for EarthCube

Our vision is an integrated computing environment that provides a seamless flow of data between sensors, repositories and models in a highly interactive and socially networked setting, accessible from a broad range of computing devices (mobile, pad, laptop, desktop, HDTV). Data analysis, visualization, modeling and synthesis merge to create a system for knowledge exploration and management. Our

work with CUAHSI HIS has helped glimpse some of what we feel is possible. Consider the following example (Figure 3). Numbers in the text that follows refers to numbers on the figure. Data from multiple observers and/or instruments (1) is fed or streamed (2) into a server where they are published in a standard format, e.g. as a web service or generalization of such. Metadata is harvested and catalogued (3) in a Catalog Server. A researcher working using a desktop client discovers the data using search services supported by the catalog (4) and accesses it using generalized web services (5). Analysis is enabled by virtue of the data being in a known format with sufficient metadata for analysis and unambiguous interpretation. Web services here may be broader than SOAP or REST services and extend to any technology for transmission of the information between components (e.g. OPeNDAP http://opendap.org/ or iRODS http://www.irods.org/). We are at this stage not concerning ourselves with technical details. The client may be desktop software or server based software in the cloud on a High Performance Computing (HPC) system accessed through a portal or gateway supporting client like discovery (6) and access (7) functionality. The analysis may be simple visualization and synthesis or may extend to quite elaborate modeling, that if gateway/portal based, may use HPC compute functionality (8). The result is a "work product" consisting of derived data or analysis output together with some narrative or graphical interpretation that represents new knowledge/information. This is "uploaded" (9) to an EarthCube server (or if already in the server, access settings are adjusted) to share the results. Another user, e.g. a colleague alerted to the work product through social networking, discovers (4 or 6) and accesses the work (5 or 7), adding his/her further analysis and republishing an enhanced work product (9). After iterating among a group that may be small or large, a paper may be generated and submitted for publication with the data attached. Reviewers then scrutinize and approve the work after which it is set to be immutable and becomes published and catalogued as an archival work product. All of this happens relatively seamlessly. The software and hardware know the data formats and information is presented to users at the high level of abstraction most relevant for their analysis. The drudge work of interpreting and reformatting data is completely automated, just as it is for a teenager loading a photo or video from her phone into Facebook or YouTube. It just works.

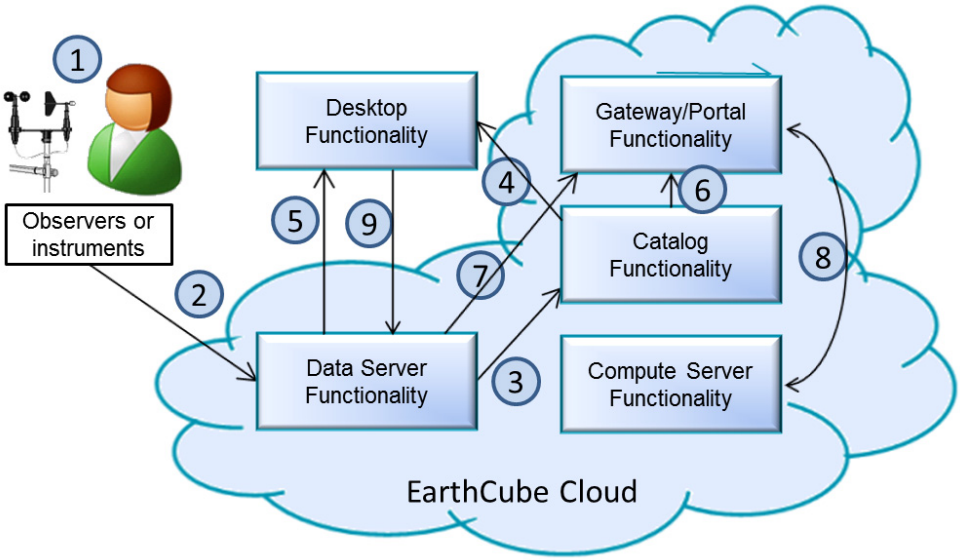


Figure 3. Data analysis and publication use case.

This paradigm extends beyond just data to include any "work product" that researchers may share and interact around, including, but not limited to models, visualizations, aggregate information etc. Furthermore, the paradigm needs to support sustaining these work products over time through evolving and disruptive technologies.

Community based governance model

We do not think that such a system as described above for the geosciences will be able to be constructed top down with a fixed a-priori design and functional specification. Perhaps a large corporation such as Google, Apple or IBM could do this, working in a closed software ecosystem. For the research community we feel that the system "EarthCube" will emerge organically from the incremental development of capability to support interoperability. This will require a community process and community governance using an open development model. Much has been written about open development and open source software development methods (e.g. Fogel, 2005, http://www.oss-watch.ac.uk/). Some key elements are:

1. Establish and document a governance model for development activities
2. Establish the tools needed to support community interactions and functioning
3. Fully exercise an open development method approach to actively engage partners and contributors

The purpose of a governance model is to guide the management of the project in a clear and transparent way. A governance model is needed at the beginning of the project to encourage and welcome contributors. Any potential contributor needs to know how their contribution will be handled. There are several examples governance alternatives available to evaluate (e.g., http://www.oss-watch.ac.uk/resources/governanceModels.xml).

However, the volunteer approach of common open source open development projects may not be up to the complexity of an enterprise the size of EarthCube. It will likely be necessary to involve and work with consortia, such as CUAHSI that already has community infrastructure and trust provides an ideal vehicle for governance and policy setting. The CUAHSI HIS has been markedly strengthened by the community backing lent to it by being of and for a community of research universities. Different disciplines within the geoscience umbrella have developed different governance and community representation mechanisms, e.g. Critical Zone Observatories (CZO), Community Surface Dynamics Modeling System (CSDMS), EarthChem (http://www.earthchem.org/). A broad governance mechanism will need to be established that brings these together into a formally recognized geoscience-wide governance entity.

Conceptual CI Architecture

There are several layers of functionality required for the EarthCube architecture that are illustrated in Figure 4. First and foremost a community governance and policy setting framework is needed to serve as a foundation for how the system is built and governed. Next up, the data models are critical. Data models define how information is organized and represented, and can enhance or inhibit the analysis that is required. Data models are initially abstract, but then are realized in terms of specific

implementations and standards. Adoption of standards is critical for wide acceptance and to facilitate the contributions of diverse and potentially competing entities. Standards provide a framework upon which hardware can be acquired and software configured. Both data and compute servers will be required and these will need to support persistent data storage in databases and file systems that potentially span multiple servers. Next is the data transmission layer that facilitates the loading and movement of data and supports the analysis, modeling and visualization functionality that is exposed to users. In an ideal system the users (scientists) should hardly have to be concerned with the lower 5 layers. It should be CI that just works. Yet defining and developing the capability of these layers is critical to the functionality that is enabled in the analysis layer. An innovative aspect that we (and others) envisage for EarthCube is the blending of collaboration and analysis using social networking functionality. The collaboration layer represents this and serves as the setting where the intellectual work and contributions occur. This is where knowledge is generated, synthesized and and persisted in scientific publications. The concept of publication here is generalized, drawing upon the ideas of the 4th paradigm (Hey et al., 2009) involving data and models being integral to publication and knowledge preservation.

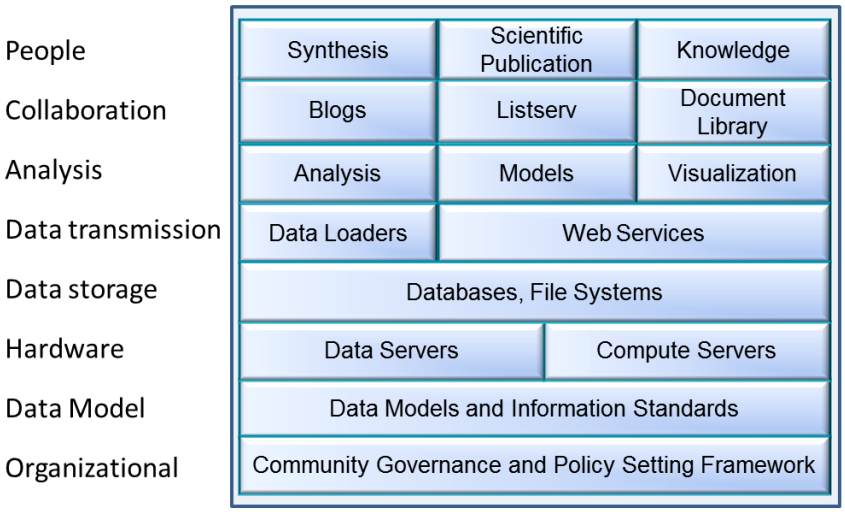


Figure 4. Potential EarthCube Conceptual Architecture.

The prototype functionality of the CUAHSI HIS has developed, albeit in prototype form and in the context of a narrower disciplinary scope, aspects of the lower six layers of this stack. The conceptual way that things are organized in HIS may be a useful starting point for EarthCube. Our experience suggests the need for distinct architectural elements for server functionality, catalog functionality and desktop/client functionality regardless of how or where it is implemented. These may be all on a single server accessible from the web, a collection of distributed servers that interoperate, in the "cloud", or relying on considerable local "desktop" resources. The triangle organizational framework depicted in Figure 1 and Figure 2 holds for the functionality regardless of which of these is used.

Design Process

As in the Governance model, modern software design is best served by a bottom-up approach in large software development projects. This differs from the traditional top-down (e.g. “waterfall”) model to

software design which assumes that requirements can be fully specified at the onset of development; never vary throughout the development lifecycle; and done in the sequence of requirements analysis, design, coding, testing, and delivery (Tucker et al., 2011). In Earthcube-related research, the varying nature of scientific discovery can dynamically change software design requirements throughout development. Our vision for Earthcube therefore is to use the more modern Agile software development methodology (Schuh, 2004) and Scrum framework (Schwaber, 2004; Gorakavi, 2009) to enable geographically distributed teams to collaborate in open source software development. Agile software development cycles require that the user be engaged continuously allowing new features to be added from user-provided scenarios and test cases throughout the development (Tucker et al., 2011). Sound software engineering including modular architecture and sound configuration management are central to successful open source software development projects (Fitzgerald, 2011). Standard source code management practices using features like versioning, branching and tagging should be used in a continuous integration process. Standard source code management tools coupled with time-based release management strategies will continuously deliver code fixes and new functionality to users (Fitzgerald, 2011).

Testing is also important. Stress testing of various portions of the Earthcube infrastructure under different file count, file size, networking, and federated data and model use case loads across a representative wide-area heterogeneous environment is necessary to ensure that the infrastructure on which Earthcube will be implemented maintains not only robust functionality, but performance as well. A heterogeneous distributed testing cluster tied to a continuous integration facility is necessary. Regular testing should automatically be launched checking out the latest version of code from the continuous integration facility, and run simultaneously on cluster nodes set up for testing various resource and federation configurations. Once stand-up has been successfully completed, which includes a scratch build and installation, testing scripts should be checked out and executed for various features, work and data loads as well as client application programmer interfaces (APIs). These testing scripts need to be generated from a testing matrix, where each dimension represents a possible change in the system: platform, network topology, server configuration, system features and select clients APIs.

Agile software development methodology with robust software engineering practices ensures that increasing numbers of users have the lowest possible barrier to contributing to a globally shared development effort resulting in fewer defects (i.e. as compared to top-down development), more innovation, and project sustainability that results from successfully-implemented open source mechanics.

Operations and Sustainability Model

Operations and sustainability is integrally related to governance and the business model adopted for EarthCube. Business model options include:

- Commercial
    - Software for sale (e.g. Microsoft)
    - Funded by advertising revenue (e.g. Google)
    - Funded by device or appliance sales (e.g. smartphone)

- Service Facility (e.g. commercial cloud services or fee based journals)
- Government (sustained essential infrastructure)
- Open source voluntary (e.g. LINUX, Open Office)
- Open source service based (e.g. Red Hat)
- Charitable beneficiary

Each of these have their advantages and disadvantages. Commercial options may not be consistent with the research focus of EarthCube, but the Internet was initially noncommercial, but now is dominated by and a significant infrastructure for commerce. A key factor in sustainability, whatever the model is user demand. EarthCube should not be a requirement imposed on Earth Scientists, much like data management plans are perceived to be at present. The benefits of using EarthCube must overwhelm the adoption inertia. The functionality accessible by entering data into the system must incentivize users to participate. It must be better to join rather than stay out.

While some flavor of open development model is anticipated, added to it needs to be mutually beneficial industrial, commercial, agency and organizational partnerships. The CUAHSI HIS has benefited from synergies created with the USGS (who publish streamflow data using CUAHSI developed systems) and ESRI (a commercial Geographic Information System Software developer) who has adopted elements of CUAHSI data publication into their online data sharing portal. EarthCube will need to foster these sort of synergies and be a software ecosystem to which many can contribute.

Libraries and Archival Journals are where scientific knowledge has traditionally been held. EarthCube will need to involve and integrate with the digital library and journal publishing community to develop models for preservation of the new forms of knowledge that EarthCube fosters.

Conclusions

This paper has reviewed the CUAHSI Hydrologic Information System and suggested architectural, organizational and design and operation elements for EarthCube that derive from this work and vision. Many of the HIS concepts developed in the context of hydrology have a generality that could apply broadly in EarthCube. There is a fundamental need within the geosciences communities for new, scientific methods to organize and utilize observational data that overcome the syntactic and semantic heterogeneity in data from different experimental sites and sources and that allow data collectors to publish their observations so that they can easily be accessed and interpreted by others. Generalizing the HIS model, requirements are: (1) Capability for **data storage** in a Community Data Model; (2) Capability for **data publication** from a server; (3)Capability for **data access** through internet-based data services using a consistent data language and format; (4) Tools for data **access and analysis**; (5) Capability for **data discovery** through thematic keyword and place based search functionality; (6) Capability for **integrated modeling and analysis** combining information from multiple data sources. The importance of data integration was underscored by the Committee on Opportunities in Hydrologic Sciences (1991, p265) who wrote "Advances in the hydrologic sciences depend on how well investigators can integrate reliable, large-scale, long-term datasets." This applies to other domains too, and to the integration of information among domains. EarthCube has an important contribution to

make in this regard. Achieving the functionality outlined will require a high level of interoperability among contributing data and modeling systems.

Beyond technical aspects, the organizational and community aspects are important, and in some sense more enduring. In the case of HIS, building partnerships with major federal and state agencies to incorporate their data into the system and ingrate with data provided by multiple academic partners was important.

We anticipate that EarthCube, like HIS will be a federated system linking data from multiple providers. As such, data availability and quality does depend on it being maintained by the provider. Developing standards and best practices to ensure and quantify data quality and uncertainty will be important. Any EarthCube data model should provide capability to document data sources, methods and quality controls, while limiting the barrier to entry into the system.

The combination of HIS capabilities has created a common window on water observations data for the United States unlike any that has existed before. We hope to participate in the development of EarthCube to help create and even broader window into the Earth system.

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

References

Fitzgerald, B., (2011), Open Source Software: Lessons from and for Software Engineering. Computer. **44:** 25-30. http://doi.ieeecomputersociety.org/10.1109/MC.2011.266.

Fogel, K., (2005), Producing Open Source Software: How to Run a Successful Free Software Project, O'Reilly, 192 p, http://producingoss.com/.

Gorakavi, P. K., (2009), Build Your Project Using Scrum Methodology, http://www.asapm.org/asapmag/articles/A3_AboutScrum.pdf, accessed 6/6/2010.

Hey, T., S. Tansley and K. Tolle, (2009), The Fourth Paradigm, Data-Intensive Scientific Discovery, Microsoft Research, Redmond, Washington, 283 p, http://research.microsoft.com/en-us/collaboration/fourthparadigm/.

Horsburgh, J. S., D. G. Tarboton, D. R. Maidment and I. Zaslavsky, (2008), "A Relational Model for Environmental and Water Resources Data," Water Resour. Res., 44: W05406, http://dx.doi.org/10.1029/2007WR006392.

National Research Council Committee on Opportunities in the Hydrologic Sciences (COHS), (1991), Opportunities in the Hydrologic Sciences, Editor, P. S. Eagleson, National Academy Press, Washington, D.C., http://www.nap.edu/catalog.php?record_id=1543.

Schuh, P., (2004), Integrating Agile Development in the Real World, Charles River Media.

Schwaber, K., (2004), Agile Project Management with Scrum, Microsoft Press, Redmond, WA, 192 p.

Tomlinson, R., (2003), Thinking about GIS, ESRI Press, Redlands CA, 283 p.

Tucker, A., R. Morelli and C. de Silva, (2011), Software Development: An Open Source Approach CRC Press, 398 p.

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