Jun 16th, 9:00 AM - 10:20 AM

Toward Integrated Environmental Modeling Using Research Data Infastructures

Jeffery S. Horsburgh
Utah State University, jeff.horsburgh@usu.edu

International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA,
Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Toward Integrated Environmental Modeling Using Research Data Infrastructures

Jeffery S. Horsburgh

Utah Water Research Laboratory, Utah State University,
8200 Old Main Hill, Logan, UT 84322-8200
jeff.horsburgh@usu.edu

Abstract: Anticipated changes to climate, human population, land use, and urban form will alter the hydrology and availability of water within the water systems on which the world's population relies. Managing water resources, as well as maintaining associated capacity to provide ecosystem services (e.g., regulating flooding, maintaining instream flow during dry periods, cycling nutrients, and maintaining water quality) will require better information characterizing both the natural hydrologic system and human mediated hydrologic systems. The next generation of integrated environmental models will seek to provide this information but requires better access to data at the spatial and temporal scales relevant to decision making. However, significant challenges lie in coupling model components and in matching data to required model inputs and outputs at spatial and temporal scales appropriate for decision making. This paper describes the availability of data within research data repositories, including the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System, the DataONE system, and the Critical Zone Observatory integrated data management system (CZOData), and its availability to and relevance for integrated environmental modeling.

***Keywords**:* Research data; CUAHSI HIS; DataONE; Integrated environmental modeling

1. INTRODUCTION

Many cyberinfrastructure systems in the Geosciences are now at a state of functionality that a savvy scientist could, with effort, find and access datasets from multiple disciplines for a synthetic analysis. However, this process still requires the use of multiple software systems and techniques, and there remain inconsistencies in the way different systems describe, encode, and share data that make integration difficult. Yet, many research scenarios require integration of data types across different domains of observational earth science. For example, advancing understanding of the functional behavior of watersheds and encoding it in the next generation of predictive hydrologic models requires synthesis of observations at multiple scales, across disciplines, across observatory or other experimental sites, and from multiple sources. This is a problem of data fusion, and the manner in which data are organized, encoded, and described either enables or inhibits their scientific analysis. Emerging today is an era of new data collection using new observing systems that recognize the spatial and temporal heterogeneity of earth processes. The focus is on precisely representing earth environments with data and advancing our understanding of their functional behavior (both natural and built) in efforts to improve our predictions for operational and management purposes. Indeed, we are collecting more data than ever before.

So, why is it still so difficult to synthesize emerging data resources with existing data to advance our understanding of earth processes? First, we are not sharing data as freely as we could, in many cases because we lack the expertise or resources to do so effectively. Despite our best intentions, many of the datasets we collect are destined for obscurity. Emerging data repositories are doing a tremendous job of increasing the availability of environmental datasets and better supporting the long tail of scientific data (Heidorn, 2008); however, many datasets are still shared in primitive formats that are hard to find, difficult to interpret, and do not express the knowledge and insights of the data

collector that could be applied to the next study that uses the data. Additionally, there are few data "savvy" scientists because we are not training them to be successful in data intensive research environments. As a result, our current system for publishing scientific knowledge contains only a fraction of the data we collect and, subsequently, the knowledge we have gained. Better infrastructure is needed for the full range of scientific activities, including data capture, curation, analysis, and publication.

2. RESEARCH DATA REPOSITORIES

Recently, a number of cyberinfrastructures have emerged within the geosciences for sharing earth observations data, including the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (Tarboton et al., 2009), the Critical Zone Observatory Integrated Data Management System (CZOData) (Zaslavsky et al., 2011), the Integrated Earth Data Applications and EarthChem system (Lehnert et al., 2011), the Integrated Ocean Observing System (De La Beaujardiere, 2008), the Data Observation Network for Earth (DataONE), among others. In general, these systems are built using the principles of service-oriented architecture and rely on standard data encodings and, in some cases, standard semantics for classes of geoscience data. The focus of these systems is on publishing or sharing data on the Internet via web services in domain specific encodings or markup languages. While these systems have made tremendous progress in making data available, they are still domain specific, and it takes a knowledgeable investigator considerable effort to discover and access datasets from multiple repositories for a synthetic analysis because of inconsistencies in the way the different domain systems describe, encode, and share data. These inconsistencies are roadblocks in integrated environmental modeling, where accessing data from multiple repositories is essential. The following sections describe three such repositories as background for the discussion that follows.

2.1 The CUAHSI HIS

Over the past several years, the CUAHSI Hydrologic Information System has advanced the interoperability of hydrologic observations made at monitoring points through the development and standardized use of the Observations Data Model (Horsburgh et al., 2008), which was encoded for data storage in a relational database, translated into an XML schema for data transfer via web services (WaterML and WaterOneFlow, respectively; Zaslavsky et al., 2007), and used to structure a central metadata catalog database that supports data discovery services (Whitenack et al., 2010). The information model is also supported by a set of controlled vocabularies (Horsburgh et al., 2014) that promote semantic consistency in the language used to describe observations. The system currently provides web service access to over 70 government and academic water observation networks, presenting over 5.2 billion data points for 1.96 million measurement sites in the U.S.

2.2 DataONE

DataONE is part of NSF's ongoing DataNet (Sustainable Digital Data Preservation and Access Network Partners) program. DataONE is poised to be the foundation of new, innovative environmental science through a distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data. DataONE seeks to ensure the preservation and access to multi-scale, multi-discipline, and multi-national science data and is developing a services oriented architecture consisting of distributed "Member Nodes" on which scientific data are published and "Coordinating Nodes" that provide data discovery and other services.

2.3 CZOData

Developed over the past several years, the prototype CZOData system has focused on publishing hydrologic observations collected at Critical Zone Observatory sites and leverages services oriented architecture approaches and software components developed by the CUAHSI Hydrologic Information

System project (Zaslavsky et al., 2011). The current prototype uses a group of ASCII files that follow a Display File format created to provide a simple way for Critical Zone Observatory site data managers to publish their data for public access. The Display File format is based on the information model represented in the CUAHSI Hydrologic Information System Observations Data Model. Once published at an individual Critical Zone Observatory web site, Display Files are automatically harvested into the Critical Zone Observatory Central Data Repository at the San Diego Supercomputer Center. The harvested data are then validated against shared vocabularies and a variable ontology, archived in a set of Observations Data Model databases established for each Critical Zone Observatory, and then published via standard CUAHSI WaterOneFlow web services that transmit data according to WaterML 1.1. Critical Zone Observatory shared vocabularies are also adapted from the CUAHSI Hydrologic System's Observations Data Model controlled vocabulary management system and establish semantic conventions within the Critical Zone Observatory system.

3. REPRODUCIBLE SCIENCE VERSUS INTEGRATED MODELING

The discussion of the CUAHSI Hydrologic Information System, DataONE, and CZOData is included above to illustrate to some degree two points: 1) there are still many inconsistencies among domain-specific research data infrastructures and opportunities for harmonization, and 2) existing geoscience research data infrastructures are currently focused more on data sharing and publication than they are on supporting integrated environmental modelling.

Data sharing and publication are important in ensuring reproducible science. Scientists wish to publish their data with their results to ensure that other scientists could reproduce their work. Current research data infrastructures generally support this use case. Scientists can deposit results of their analyses into a repository and share them with the world. However, the next scientist, who is developing an integrated environmental model and hopes to use data from existing research data infrastructures to fill requirements for input and forcing data may have difficulty finding what they need because: 1) next generation environmental models are becoming more interdisciplinary, requiring inputs from multiple repositories; 2) products deposited into research data infrastructures are project/study specific and the researcher may struggle to determine whether data is appropriate for a new use; 3) many models require inputs at spatial and temporal scales not found in research data infrastructures (but that may be found elsewhere in agency or government data repositories). This mismatch between current functionality of research data infrastructures and what is required to support integrated environmental modeling illustrates several opportunities for increasing the utility of research data infrastructures in supporting integrated environmental modeling.

4. TOWARD INTEGRATED ENVIRONMENTAL MODELING USING RESEARCH DATA INFRASTRUCTURES

Next-generation cyberinfrastucture, such as that envisioned for the National Science Foundation's EarthCube effort (http://www.earthcube.org), must present an integrated view of the best available data that holistically represent the earth's hydrologic and environmental systems. Such a system would support complex queries such as:

> *"I am interested in modeling the effects of aquatic nutrient concentrations on stream metabolism. Show me locations where high frequency stream discharge, water temperature, and dissolved oxygen data have been collected in second order streams for which samples of nitrogen and phosphorus have been collected and that are within one mile of a weather station that measures solar radiation."*

Such queries are commonly posed by modelers who are seeking environmental settings with the rich datasets required to test their latest model formulations. However, queries like this are impossible using existing geoscience data infrastructures. Creating this new capability will require seamless presentation of and access to data across existing repositories. Here, I describe four major needs to achieve this.

4.1 A Common Observations Information Model

There is a need for agreement about the information needed to describe observational data so they can be discovered and interpreted by investigators who did not collect the data. The information models of current geoscience cyberinfrastructures representing earth observations are currently not interoperable because they represent common informational elements of observations in inconsistent ways and they lack the extensibility required for supporting additional data types and informational elements. Semantic and syntactic heterogeneity across repositories and data types are major hurdles to be overcome. The CUAHSI Hydrologic Information System is one demonstration of how a common information model used for data storage (the Observations Data Model), transfer (WaterML), and cataloging, along with a community vocabulary for describing hydrologic observations made possible unprecedented discovery and access to observations collected at fixed monitoring sites (Figure 1).

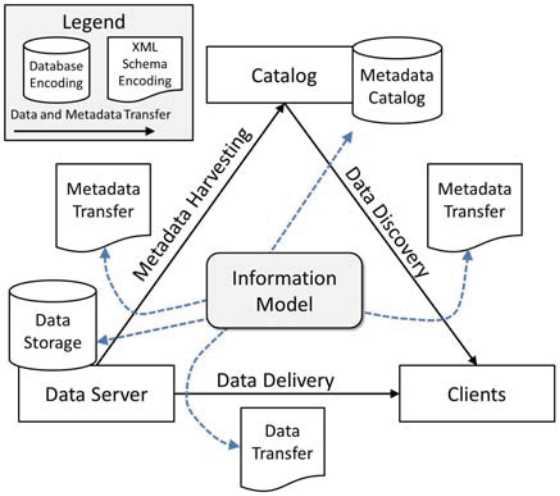

Figure 1. The information model is central to services oriented architecture of the CUAHSI Hydrologic Information System and specifies the information requirements and semantics of data encodings for storage and transfer of data.

However, Despite the success of the CUAHSI Hydrologic Information System for hydrologic time series measured at fixed geographic points, its underlying information model and implementations (e.g., the Observations Data Model, WaterML 1.1, etc.) lack: 1) adequate structures to fully describe some types of observations derived from *ex situ* analysis of field samples, subsamples and sample fractions, as well as other data types used commonly in the geosciences; 2) the ability to represent observations made on other geometries (e.g., average precipitation over a watershed); and 3) extensibility that would enable it to easily accommodate additional data types or metadata attributes.

In its first phase, DataONE is treating data as opaque objects and does not require a specific format for submitted data (DataONE, 2014). Integration and use of observational data retrieved from DataONE could be significantly improved if the format and semantics of data objects deposited into the system conformed to a well-specified observations information model such as the one proposed here. Similarly, because of the lack of an information model that accommodates both time series and sample-based data, a Display File prototype for CZOData based on the CZChemDB (Niu et al., 2011) and the EarthChem information model is under development to operate in parallel with the hydrologic observations prototype. Integration across these parallel prototypes is hampered by the lack of a common information model. New work in this area includes development of a next generation of the CUAHSI Observations Data Model called ODM2 as a profile of the Open Geospatial Consortium's Observations & Measurements standard (Cox, 2010). The focus of ODM2 (preliminary version shown in Figure 2, with larger version and documentation at https://github.com/UCHIC/ODM2) is better integration of spatially discrete, feature based observations from samples and specimens within existing domain-specific cyberinfrastructures and data repositories such as the CUAHSI Hydrologic Information System and Water Data Center, the CZOData system, and the Integrated Earth Data Applications systems.

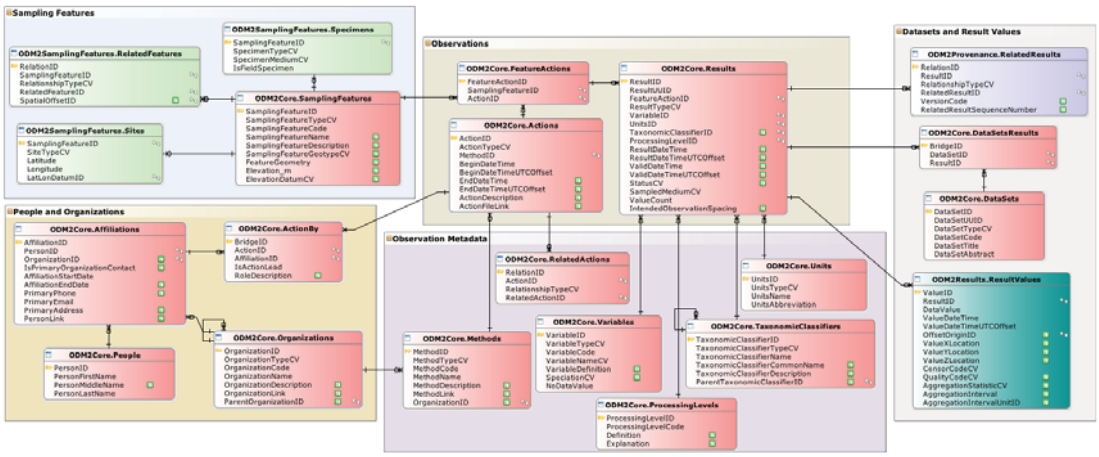

Figure 2. A prototype of the Observations Data Model 2 information model as a profile of the Open Geospatial Consortium's Observations & Measurements standard.

Enabling data integration at the information model and storage level may enable a more flexible data publication infrastructure and additional types of cross-domain database queries. ODM2 represents progress in this area, but there remains a need to achieve consensus and work toward standards-based approaches for representing data of many types and across domains – e.g., geospatial data, time varying fields, etc.

4.2 Linking Data to the Geo-Environment

Within current geoscience cyberinfrastructures, there is inconsistency in the way the area or volume to which an observation applies, or geospatial support, is represented and in the way the environmental "feature of interest" (e.g., Cox, 2010) is represented. The location at which observations were made is often described only by latitude and longitude coordinates. Point coordinates (e.g., for a streamflow gage) rarely capture the geospatial context, support, or feature of interest for the measurements. A data user must determine that the gage lies on a particular river, measures the outflow of a particular catchment, is downstream of another gage, shares its location with a water quality monitoring site, and is located near a weather station – yet all of this contextual information is needed to satisfy queries such as the one above and to determine whether discovered data are appropriate as inputs to a particular model or analysis.

Creating a geospatial data framework that can serve as the context for observational data is a needed next step. Observational data could then be linked to a geospatial feature of interest to which they apply. For example, a stream gage would be represented as a point location, the point would be associated with a line representing a stream reach, and the point and line would be associated with a polygon representing a catchment boundary. Each geographic feature could be linked through specific relationships encoded within feature attributes (Figure 3). Much work has already been done in this area with geospatial data models like the National Hydrography Dataset, Arc Hydro, Arc Hydro Groundwater, Hy_Features, and others. Building this geospatial "fabric" and locating observational data within it will provide a way to evaluate the geospatial components of complex queries such as the one posed above and enables easier data interpretation and use through explicit linkage between observations and the earth features that they represent.

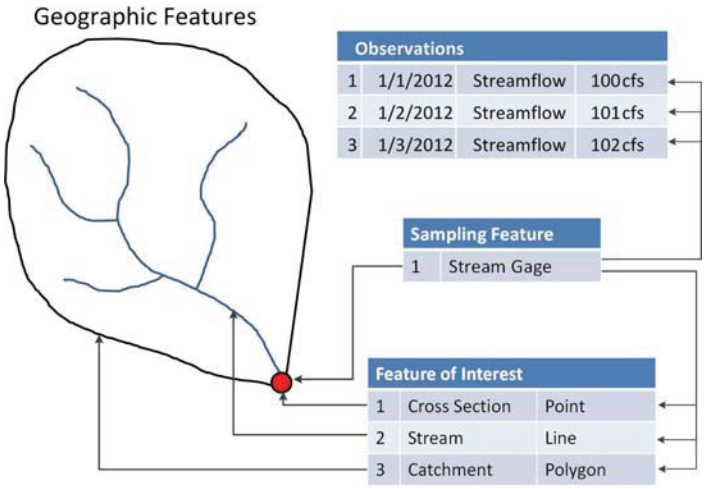

Figure 3. Linking observations to the geo-environment.

4.3 Capturing the Knowledge Content of Data

It is rare for datasets to be annotated with interpretations of what the data mean, how they have been used, conclusions that have been drawn, and appropriateness for given uses. New methods are needed for annotating data with these types of contextual information. A good example of this would be annotating a streamflow dataset with comments that specify when major events such as storms, floods, etc. occur. A trained hydrologist could infer when major events occur from the numeric values in the dataset and by overlaying ancillary datasets such as precipitation, but it would be far easier for a non-hydrologist to interpret a set of abnormally high streamflow values if there was an annotation stating that a hurricane occurred. Some of this could be done through collaborative and social media technologies that enable data users to annotate published datasets, leaving value-added information that can be evaluated by subsequent data users. Additionally, data users could "tag" datasets as fit for a particular purpose – e.g., tagging a dataset as appropriate or useful in particular modeling contexts, which may facilitate discovery by researchers seeking datasets and locations on which they can test their models.

4.4 Integration with Government and Agency Data Repositories

The spatial and temporal extent of data in research data infrastructures will remain limited because the research studies that generate them are limited in geographic and temporal scope. However, a significant contribution that can be made by research data infrastructures is to facilitate access to government and agency data repositories. Realizing the vision of "an integrated view of the best available data that holistically represent the earth's hydrologic and environmental systems" requires integration of important national and even international-scale datasets from which integrated modelers can draw input and forcing data.

One example of successful integration of a research data infrastructure with agency data is the CUAHSI Hydrologic Information System, in which first-of-their-kind web services for publishing hydrologic time series data resulted in development of an international standard for hydrologic data transmission on the web called WaterML. For the first time, the CUAHSI Hydrologic Information System Central metadata catalog and search web services provided integrated search capabilities across multiple national databases (e.g., United State Geological Survey's National Water Information System, United States Environmental Protection Agency's Storage and RETrieval System, United States Natural Resource Conservation Service's SNOpack TELemetry system, and others) and academic repositories of hydrologic data. The software client application HydroDesktop (Ames et al., 2013) now provides unprecedented ability to discover, access, and analyze hydrologic data from

more than 100 different research groups and repositories hosted by the CUAHSI Hydrologic Information System.

5. CONCLUSIONS AND RECOMMENDATIONS

Speeding innovations in synthesis and modeling in the geosciences will require new cyberinfrastructue techniques and tools as well as researchers trained in data intensive science. In this paper, several strategies have been presented for improving the discoverability, interpretability, and interoperability of geoscience datasets stored in emerging research data infrastructures. A more robust information model that links data to the geo-environment and captures the knowledge content of data could support more complex data discovery queries and improve the reusability of data described using the information model. This new cyberinfrastructure is imperative in achieving the type of interoperability needed to realize the full value of data hosted in these disparate repositories and to speed their integration with new models and analyses.

6. REFERENCES

Ames, D. P., Horsburgh, J.S., Cao, Y., Kadlec, J., Whiteaker, T., Valentine, D., 2012, HydroDesktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis, Environmental Modelling & Software, 37, 146-156, http://dx.doi.org/10.1016/j.envsoft.2012.03.013.

Cox, S., 2010. Geographic Information: Observations and Measurements OGC Abstract Specification Topic 20, v2.0.0, OGC 10-004r3, Open Geospatial Consortium, Inc., 49 p., http://portal.opengeospatial.org/files/?artifact_id=41579.

DataONE, 2014. DataONE Architecture Documentation: Reference for implementation of services and tools forming the DataONE cyberinfrastructure, http://mule1.dataone.org/ArchitectureDocs-current/, Last accessed 3/18/2014.

De La Beaujardiere, J., 2008. The NOAA IOOS Data Integration Framework: Initial implementation report, OCEANS 2008, 15-18 Sept. 2008, p. 1-8, http://dx.doi.org/10.1109/OCEANS.2008.5152007.

Heidorn, P.B., 2008. Shedding light on the dark data in the long tail of science, Library Trends, 57(2), 280-299, http://dx.doi.org/10.1353/lib.0.0036.

Horsburgh, J.S., Tarboton, D.G., Hooper, R.P., Zaslavsky, I., 2014. Managing a community shared vocabulary for hydrologic observations, Environmental Modelling & Software, 52, 62-73, http://dx.doi.org/10.1016/j.envsoft.2013.10.012.

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, http://dx.doi.org/10.1029/2007WR006392.

Lehnert, K.A., Carbotte, S.M., Ryan, W.B.F., Ferrini, V., Block, K., Arko, R.A., Chan, C., 2011. IEDA: Integrated Earth Data Applications to support access, attribution, analysis, and preservation of observational data from the ocean, earth, and polar sciences, Geophysical Research Abstracts, 13, EGU2011-13113.

Niu, X., Lehnert, K.A., Williams, J., Brantley, S., 2011. CZChemDB and EarthChem: Advancing management and access of critical zone geochemical data, Applied GeoChemistry, 26, S108-S111, http://dx.doi.org/10.1016/j.apgeochem.2011.03.042.

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

Whitenack, T., Valentine, D., Zaslavsky, I., Piasecki, M., Tarboton, D., Horsburgh, J., Whiteaker, T., Ames, D., Maidment, D.R., 2010. Hydrologic metadata catalog and semantic search services in CUAHSI HIS. In: Francisco Olivera (Ed.), 2010 AWRA Spring Specialty Conference: Geographic Information Systems (GIS) and Water Resources VI, American Water Resources Association, TPS-10-1, ISBN 1-882132-82-3.

Zaslavsky, I., Valentine, D., Whiteaker, T. (Eds.), 2007. CUAHSI WaterML v0.3.0. OGC07-041r1. Open Geospatial Consortium, Inc. 76 pp. http://portal.opengeospatial.org/files/?artifact_id=21743S.

Zaslavsky, I., Whitenack, T., Williams, M., Tarboton, D.G., Schreuders, K., Aufdenkampe, A., 2011. The initial design of data sharing infrastructure for the Critical Zone Observatory. In: Proceedings of the Environmental Information Management Conference, Santa Barbara, CA, 28-29 September, EIM 2011, http://dx.doi.org/10.5060/D2NC5Z4X.