The Initial Design of Data Sharing Infrastructure for the Critical Zone Observatory

Ilya Zaslavsky[1], Thomas Whitenack[1], Mark Williams[2], David Tarboton[3], Kim Schreuders[3], Anthony Aufdenkampe[4]
[1] San Diego Supercomputer Center
[2] University of Colorado at Boulder
[3] Utah State University
[4] Stroud Water Research Center
zaslavsk@sdsc.edu, twhitenack@sdsc.edu, markw@cutler.colorado.edu, dtarb@usu.edu, kim.schreudres@usu.edu, aufdenkampe@stroudcenter.org

Abstract—**The Critical Zone Observatory (CZO) program is a multi-institutional collaborative effort to advance scientific understanding of environmental interactions from bedrock to the atmospheric boundary layer across scales and disciplines. To create a comprehensive hydrogeochemical portrait of experimental sites the observatories collect large volumes of data. Publishing, analyzing and archiving these data in a consistent and integrated manner across all CZO sites is challenging due to the inherent heterogeneity in data collection and processing techniques. We present the initial design and a prototype of the CZO data sharing infrastructure. While each CZO site maintains its own data management system, the integrated infrastructure design specifies formats and protocols for presenting the information on CZO web sites, where it can be browsed by users as well as automatically harvested into a centralized data system. The latter validates, archives and converts the data into standards-compliant data services, which can be consumed by various client applications.**

Keywords—environmental observatory; CZO; cyberinfrastructure; hydrology; information integration

I. Introduction

The CZO project [1] integrates data from several earth science disciplines in order to describe and model complex physical processes in the critical zone. Typical research scenarios involve accessing both geochemical samples and hydrologic time series of water quality and water quantity within experimental watersheds, relating the dynamics of differently measured parameters, modeling soil nutrients under different topographic, geologic, hydrologic and vegetation conditions, analysis of fluxes across watershed boundaries, etc. While closely connected research teams have been successful in such cross-discipline analysis and modeling, accomplishing such integration at a higher level, across CZO sites and spatio-temporal scales, faces several interoperability challenges. They stem, in particular, from differences in information models used in different disciplines and by different research groups to describe observations, differences in data representation and access, and discrepancies in metadata and their semantics. For example, the geochemical community has been developing infrastructure for managing geochemical sample information and created a standard XML schema encoding for geochemical datasets named EarthChem XML [2, 3]. The hydrologic research community, via the Consortium of Universities for the Advancement of Hydrologic Science, Inc.'s Hydrologic Information System (CUAHSI HIS) project, has been creating a service-oriented system for sharing hydrologic observations [4, 5], and proposed a canonical data model for hydrologic observations [6] encoded as Water Markup Language [7]. Large scale cross-observatory systems are being developed within the Long Term Ecological Research Network [8], the National Ecological Observatory Network [9], and several other NSF-supported earth science projects. Common cyberinfrastructure challenges of earth science observatory projects have been summarized in [10].

The CZO program is a relatively new large-scale observatory effort, which allows the CZO information network design to leverage the experience and cyberinfrastructure components developed in the neighboring projects. It currently includes 6 observatories: the Boulder Creek CZO (led by the University of Colorado at Boulder), the Christina River Basin CZO (University of Delaware), the Jemez River and Santa Catalina Mountains CZO (University of Arizona), Luquillo CZO (University of Pennsylvania), the Southern Sierra CZO (University of California, Merced) and the Susquehanna Shale Hills CZO (Pennsylvania State University). Research agendas of each site are different, yet several cross-cutting topics and data needs have been identified, in particular with management of hydrologic time series; water, soil and rock samples; spatial data including LiDAR; and meteorological variables. This justifies development of a CZO-wide data sharing infrastructure, to enable uniform publication, discovery and retrieval of data collected across sites.

Despite differences in research foci and scope, the experience of large environmental observatory cyberinfrastructure projects suggests multiple common requirements and infrastructure issues; they have been addressed in the literature [e.g. 10, 11, 12, 13]. Specific requirements of the CZO-wide data management system derive from the unique role of the CZO program as an evolving cross-disciplinary multi-site effort. They can be summarized as follows:

- Reliance on standards for data exchange adopted in research communities comprising the CZO program.

- Leveraging domain data systems, synthesizing information management experience and software from CZO partners and neighboring earth science disciplines (CUAHSI [4, 5], EarthChem [2, 3], CZEN [14], NCED [15], LTER [8], etc.)
- Uniform data modeling, data description and formatting practices, to ensure that the published data can be unambiguously interpreted and their provenance can be traced.
- CZO research teams maintain their own data management systems, while sharing data via a centralized publication system that is scalable and extensible to additional data types and research sites.
- Evolving the integrated data system towards better standards compliance and cross-CZO integration without burdening individual CZO sites.
- Availability of CZO data both in a human-readable form at individual CZO web sites as well as via web services from the central CZO data repository.

This paper presents the details of the original design of the CZO-wide data publication and sharing system developed in response to these requirements, and describes its main components.

II. The Vision of the CZO Data System

The CZO project is enabling access to a variety of data types required for modeling physical processes in the critical zone, including geochemical, geophysical and hydrologic observations, spatial data and field measurements. For some types of data, uniform protocols and formats for data and metadata exchange have been established and agreed upon within respective communities, while other domains see a wide variety of approaches to data representation and description. Therefore, we consider CZO data interoperability at several levels (Fig. 1).

At the first level, different types of CZO resources (files, services, downloadable data folders, etc.) are registered at a CZO data portal, with Dublin Core metadata, so that these resources can be browsed or queried by title, contributor, spatial location, thematic category and similar fields as defined in the Dublin core standard, and subsequently invoked or downloaded to a user's workstation.

At the next level, the resources have common semantics (a set of shared vocabularies for variable names, methods, units, features of interest, measurement medium, qualifiers, censor codes, etc.) which ensures that, once the resources are discovered and downloaded they could be easier interpreted and integrated.

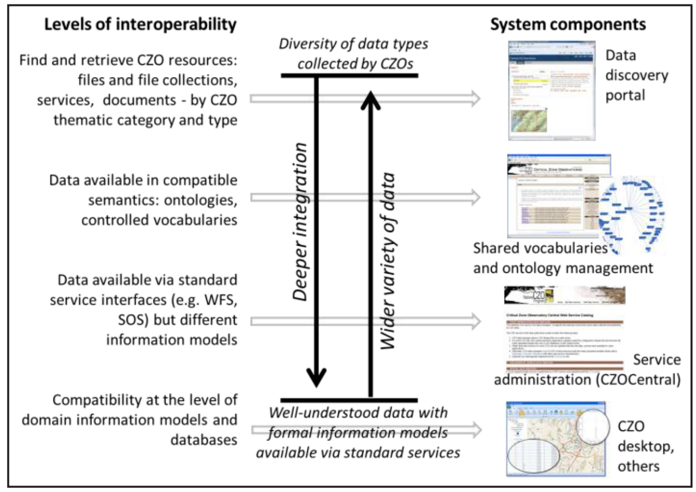

Figure 1. Levels of interoperability and corresponding components of the CZO data system

Further, resources of certain types may become available via standard service interfaces, such as those developed by the Open Geospatial Consortium (OGC), so that they can be accessed from standards-aware client applications.

Finally, at the fourth level the data become available via standard services and in standard encodings that reflect domain information model, to enable a much wider range of operations across different compliant sources.

Different types of data considered by CZO support different levels of interoperability, and, therefore, rely on different system components. For example, soil samples, gridded data, flux information are currently registered as resources with minimal metadata and made available via the data discovery portal, while their semantic consistency is recommended by a set of shared vocabularies but is not currently enforced, and standard information models are being developed. Hydrologic observations, on the other hand, represent the type of data that is made interoperable at all four levels within the CUAHSI HIS project. In the current design, the CZO data infrastructure leverages HIS components and generally follows Service Oriented Architecture (SOA) for publishing, indexing and accessing hydrologic observations as implemented in the HIS project.

The CZO data system design follows the general SOA "publish-find-bind" pattern, with the additional requirements described above. In particular, these requirements affect the "publish" component which is represented as two interlinked modules: publishing CZO data at individual web sites as human-readable ASCII files, and their harvesting and republishing as standard-compliant web services at the central CZO data repository/archival site (Fig. 2). The system components supporting CZO data interoperability at different levels are described in the following section.

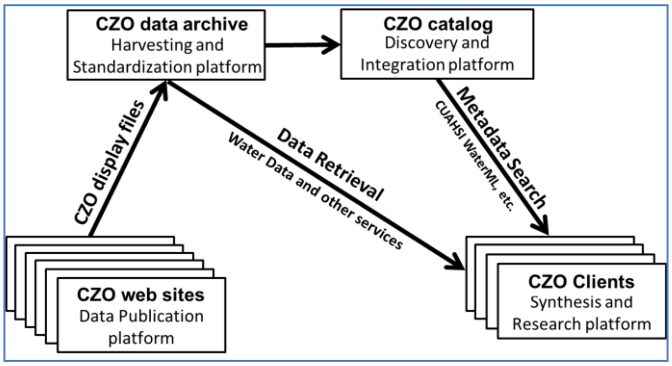

Figure 2. A high-level view of service oriented architecture patterns in the CZO data publication and sharing system

III. The Prototype: Main Components and Interfaces

This overall design is further detailed in Fig. 3. Data published at each of the six CZO web sites following an agreed upon ASCII format (display file format, described below), are automatically harvested into a centralized data repository housed at the San Diego Supercomputer Center (SDSC), validated against shared vocabularies and parameter ontology and archived in a set of databases established for each CZO. Upon harvest and validation, standard CZO data services are automatically updated to include the new data. The CZO data services become available in a range of standard formats: for hydrologic observations these are CUAHSI WaterOneFlow services, which transmit data according to the WaterML 1.x specification, and Web Feature Services (WFS) specified by the Open Geospatial Consortium, which are used to exchange time series catalog information. The services are registered and indexed in the CZO Central's service registry, and can be discovered via a CZO Data Portal, which is compliant with OGC's Catalog Services for the Web (CSW) standard. The standard services can then be consumed by various applications, as well as registered in cross-project domain registries such as CUAHSI HIS Central (for hydrologic time series) or the EarthChem Data Portal (for geochemical data). CZO data products generated from the published data will be made available via the same CZO Central services and portal. In addition, we envision collaboration with the NSF-supported DataNet program on long-term preservation of CZO data.

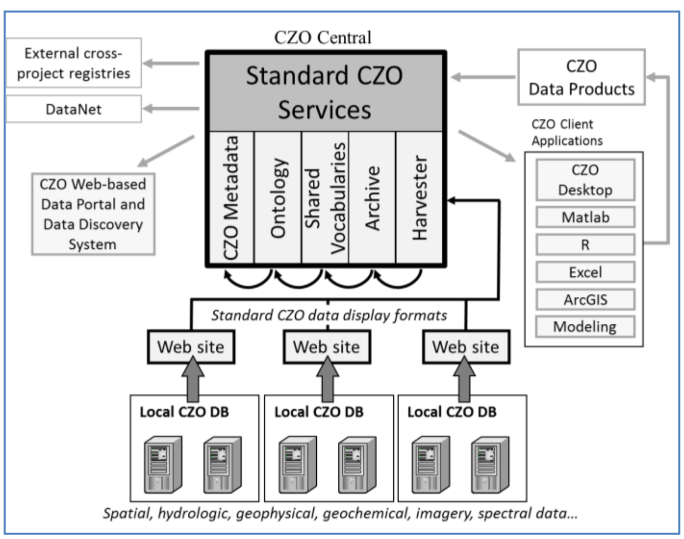

Figure 3. Main components of the CZO data publication workflow

Below we describe several key components of this design, which is now implemented in a prototype data sharing system.

A. *CZO Display File Format*

The CZO display file format for hydrologic time series has the following key features:

- The format is based on the information model adopted from CUAHSI Observations Data Model (ODM) [6]. At the same time, it incorporates several information model enhancements made necessary by CZO data collection practices, including multiple types of named vertical offsets (e.g. "upper canopy", "lower canopy"), support for data loggers collecting information from groups of sampling locations, and a more flexible definition of a data series as any logical grouping of observations defined by data publisher.
- The ASCII format of the display files is both human- and computer-readable, and is uniform across the CZO sites.
- Display files include a configuration file (specifying which files shall be regularly harvested from a CZO web site), sites file, methods file, series metadata file, and a data file. In a typical scenario, each configuration file housed by a CZO will point to single sites and methods files, and to one or more series metadata files, each of which would reference one or more data files.
- The display data file closely follows a common data logger format, to minimize re-formatting at CZOs. It encodes the following characteristics of each observed value: location (where the observation took place), date and time (when), the attribute measured (what), the measurement method (how), and the responsible investigator (who). Details of each of these characteristics are encoded in the sites, methods and header (series metadata) files.
- While initially focused on hydrologic time series, the display file format is extensible to other types of data, in particular geochemical samples. Further, metadata display files (configuration, sites, methods, data series) may reference binary data files if appropriate for certain data types (e.g. spatial data, grids).

B. *CZO Central Catalog and Web Services*

The CZO Central model generally follows the organization of the centralized components of the CUAHSI Hydrologic Information System [16] and extends it to accommodate the specific CZO data management requirements: managing centralized rather than distributed collection of ODM databases and supporting harvesting and validation of display files. New or updated display files are being retrieved from each of the six CZO web sites into the CZO central data repository (currently configured to re-harvest new data automatically every week or manually by request from a data manager). The harvesting triggers updates of respective ODM instances installed at the CZO Central for each site, along with validation of the display files configured by CZO data managers. Through the CZO Central's online interface, data managers can browse harvesting logs and correct errors if necessary. Once the central ODM instances are updated, the time series metadata are harvested into the CZO Central time series catalog, which makes the data from all CZO sites discoverable by a range of spatial, temporal and semantics-based requests.

The data in each ODM are available via a standard set of water data services compliant with the WaterML 1.x specification [7]. For each CZO hydrologic observation network the services include the following standard methods: GetSites, GetSiteInfo, GetVariableInfo and GetValues. Once harvested into the central catalog, metadata from all CZO sites become available via requests that return time series information based on spatial, temporal and attribute-based requests (GetSeriesCatalogForBox), site information (GetSitesInBox), information about services (GetServicesInBox, GetWaterOneFlowServiceInfo), as well as information about searchable concepts and their hierarchy (GetOntologyTree, GetSearchableConcepts, GetWordList), and mapping between variables and concepts (GetMappedVariables). Compatibility with CUAHSI HIS at the level of services and information models makes it easy to integrate CZO data with data available from over 70 government and academic observation networks available through CUAHSI HIS Central. This enables easier validation of CZO-collected data against hydrologic observations made at USGS, EPA, and possibly collocated stations from other networks, and additional data interpolation/imputation processing.

CZO data managers login to the CZO Central administration interface to edit service metadata for their sites: abstract, contact information, recommended citation, data access policy, icons/logos, etc. (Fig. 4), request harvest of their published display files into the central system, and examine the harvesting logs. In addition, data managers can use the CZO Central application to associate variable names with concepts in the ontology of hydrologic terms developed by CUAHSI. Establishing this association enables data discovery based on thematic categories. The CZO Central web site is central.criticalzone.org.

C. CZO Data Portal, and compliance with OGC services

Besides making the time series metadata available via WaterOneFlow and CZO Central web services, the CZO Central application also generates WFS services for each CZO network, which list time series available from each site, and their metadata. These services are automatically registered in the CZO Data Portal, a custom application of the ESRI open source GeoPortal Server [17]. With this application, the registry of CZO services becomes available via standard OGC Catalog Services for the Web (CSW) interface, which makes them accessible from a variety of OGC-compatible client applications, and enables federation with other CSW catalogs, such as the CUAHSI HydroCatalog. Sample search results in

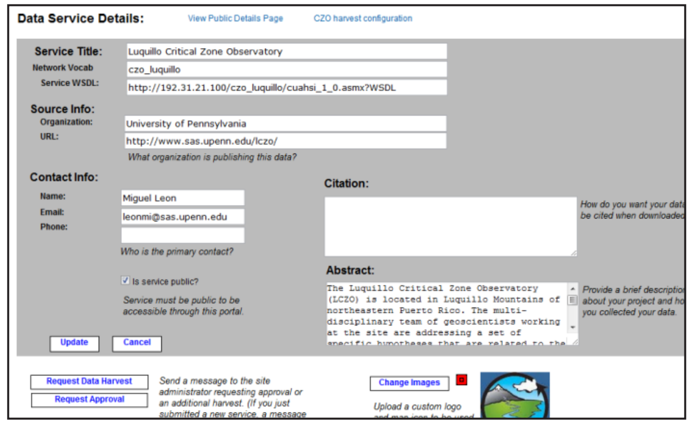

Figure 4. A fragment of a CZO service management and metadata editing web page at CZO Central

the CZO Data Catalog, federated with the HydroCatalog at CUAHSI, are shown in Fig. 5.

In addition to registering water data services to the CZO Data Portal, the harvesting application automatically adds display files retrieved from CZO web sites, to the same central CSW catalog, thus enabling full text search over the content of registered metadata files, and data file download directly from the portal application.

One of the key roles of the CZO Central and the CZO Data Portal is to expose CZO data via standard OGC-compliant service interfaces, and evolve these interfaces once new specifications are adopted. With respect to hydrologic data, an essential new standard is WaterML 2.0, which is being developed under the aegis of the Hydrology Domain Working Group of the OGC and the World Meteorological Organization [18]. At the time of writing, this specification, after being

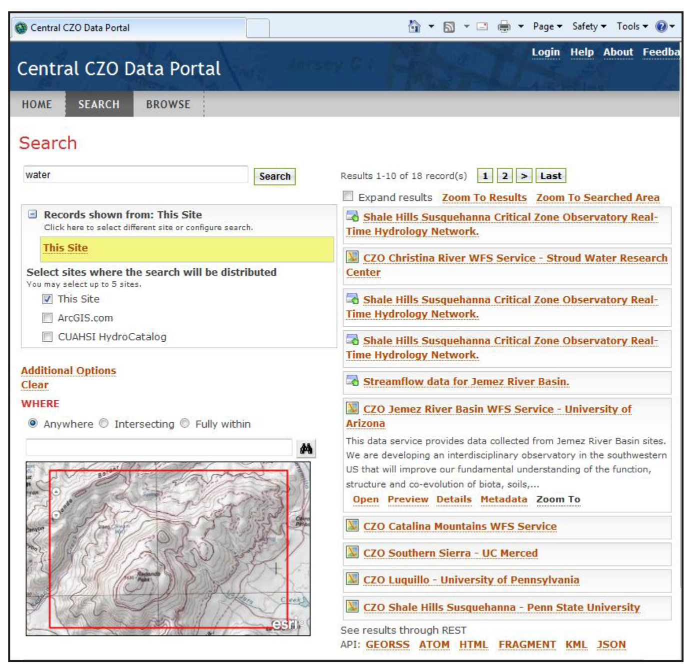

Figure 5. The search page of the CZO data discovery portal

approbated and refined through OGC Interoperability Experiments, is entering the OGC standardization process, and is expected to be approved by the end of 2011. The first version of WaterML 2.0 is a profile of the OGC/ISO "Observations & Measurements" [19] model and specifies time series encoding for hydrologic data. Thus encoded time series data will be transmitted over OGC Sensor Observation Service (SOS) 2.0 interface, initially alongside WaterML 1.x/WaterOneFlow services, and eventually replacing them for both CUAHSI HIS and the CZO data system.

D. CZO Shared Vocabulary Registry

Another key component of the CZO data system is a collection of controlled (shared) vocabularies, also inherited from the CUAHSI HIS ODM controlled vocabulary submission system [20] but extensible to other types of data collected by CZO sites. These vocabularies, available via a web interface and via web services, are used to establish semantic conventions within the CZO system, ensuring that terms describing key metadata elements are well defined, unique and unambiguous, which, in turn, supports cross-CZO attribute-based data discovery. The web interface for the shared vocabulary system allows data managers to browse the vocabulary content, and propose additions and edits, while the web service API is used by the CZO Central's harvesting application to validate submitted metadata for compliance with the vocabularies. The following vocabularies are moderated by the system: variable names; methods; units; value type (e.g. field observation, model output); sample type (physical medium from which the sample is taken); data type (e.g. average, continuous, cumulative); data level (processing level or quality control level); spatial reference (projection and datum, based on EPSG [21]); censor code (e.g. not-censored, non detect); qualifier code (e.g. approved, provisional); vertical datum. If a particular term is missing from any of the vocabularies, data managers can submit it via the web interface; once the term is considered and accepted by vocabulary curators it becomes part of the master list of approved vocabulary terms. The web site for the CZO shared vocabulary registry is sv.criticalzone.org.

IV. Discussion and Conclusion

The initial effort to design and build an integrated CZO data system prototype has achieved several important goals: the CZO sites have converged on a uniform data publication model and a display file format convention, enhancements to the original information model for hydrologic observations have been developed and tested, the initial centralized data system has been built to share and integrate data from all CZO sites, and each CZO has started publishing the data through the system. Most importantly, the system has been designed and developed in close collaboration with data managers from all CZO sites, taking into account differences in data types, metadata organization and data publishing practices established at each site.

While following the CUAHSI HIS architecture, the CZO data system presents a new publication model, which reflects specific requirements of the CZO cross-site and cross-domain data integration. The key advantages of this model include:

- Individual CZO sites are responsible for maintaining their own data systems and are not required to install and maintain additional software (e.g. a HydroServer, which represents the data publication platform within CUAHSI HIS), which may not fit with the existing software environment or skill set of local data management personnel. Developing an ASCII export into the display file format usually presents a lesser problem compared with the need to manage additional software.
- The data publication and sharing model preserves the autonomy of individual research sites, which reflects the level of autonomy of investigator teams in this large and complex project, and thus does not violate the established relationships and practices of the CZO virtual organization.
- The burden of compliance with evolving standard service interfaces and encodings is on the central data management system, rather than on individual CZO sites, where research and data management work can remain focused on science objectives of each site.
- The developed display file format serves a dual purpose: it presents the data in a human-readable form on CZO web site, and at the same time supports automatic harvesting of the data into the central data repository.
- The publication model is extensible to other types of data (raster data, GIS layers, geochemical data, soil profiles, geophysical data, photos, etc.), once respective information models and metadata profiles are agreed upon.

At the same time, these advantages underscore the core drawback of the publication model: it introduced a new exchange format, which needs to be governed and further developed as CZO needs evolve – rather than passing the governance burden to standards organizations such as OGC. Being a text format, it provides limited options for content validation of the display files – which is to some extent compensated by extensive content validation as the files are harvested into the CZO Central repository.

The described CZO information system prototype creates new opportunities for critical zone environmental observatories to publish and discover data and integrate them in new types of cross-CZO data-intensive analysis and modeling that were not possible or too time-consuming before. While the system is at an early development stage (at the time of writing, only about 15 million hydrologic observations collected by CZO sites are available via web services, and about 70 resources are registered in the CZO Data Portal), the volume of data is growing. The prototype demonstrated that a scalable data sharing infrastructure for environmental observatories can be built by leveraging and integrating service-oriented approaches and cyberinfrastructure components developed in neighboring Earth science disciplines, while careful consideration is given to the specific requirements of the CZO research community, in particular: information modeling needs; standards compliance and semantic consistency; and distribution of data management roles and responsibilities between individual sites and the central archival, cataloguing, and services system.

ACKNOWLEDGMENT

We are grateful to data managers from the six CZO sites for their contribution to the development of the integrated CZO data system and for useful discussions: Matej Durcik (University of Arizona), Chi Yang (University of Colorado at Boulder), Otto Alvarez and Xiande Meng (University of California, Merced), Miguel Leon (University of Pennsylvania), Brian Bills and Jennifer Williams (Pennsylvania State University), Charles Dow and Melanie Arnold (Stroud Water Research Center). U.S. National Science Foundation support (awards EAR-0724960, DEB-1027341) is gratefully acknowledged. 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

[1] National Critical Zone Observatory (CZO) Program, www.criticalzone.org (last accessed June 1, 2011).

[2] The EarthChem Portal, www.earthchem.org (last accessed June 1, 2011).

[3] K. Lehnert and S. Vinayagamoorthy, "Geoinformatics for Geochemistry (GFG): Integrated digital data collections for the earth and ocean sciences", Proceedings of Geoinformatics 2007—Data to Knowledge. USGS Scientific Investigation Report 2007-5199, U.S. Geological Survey, Reston, VA, pp. 32-34, 2007.

[4] D. G. Tarboton, J. S. Horsburgh, D. R. Maidment, T. Whiteaker, I. Zaslavsky, M. Piasecki, J. Goodall, D. Valentine and T. Whitenack, "Development of a community hydrologic information system," 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, ed. R. S. Anderssen, R. D. Braddock and L. T. H. Newham, Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, July 2009, p. 988-994.

[5] D. G. Tarboton, D. R. Maidment, I. Zaslavsky, D. P. Ames, J. Goodall, and J. S. Horsburgh, "CUAHSI Hydrologic Information System 2010 Status Report," Consortium of Universities for the Advancement of Hydrologic Science, Inc, 2010, 34 pp., http://his.cuahsi.org/documents/CUAHSIHIS2010StatusReport.pdf (last accessed June 1, 2011).

[6] J. S. Horsburgh, D. G. Tarboton, D. R. Maidment and I. Zaslavsky, "A Relational Model for Environmental and Water Resources Data," Water Resour. Res., 44: W05406, 2008.

[7] I. Zaslavsky, D. Valentine and T. Whiteaker, "CUAHSI WaterML," OGC 07-041r1, Open Geospatial Consortium Discussion Paper, 2007, http://portal.opengeospatial.org/files/?artifact_id=21743 (last accessed June 1, 2011).

[8] The US Long Term Ecological Research Network, www.lternet.edu (last accessed June 1, 2011).

[9] National Ecological Observatory Network, www.neoninc.org (last accessed June 1, 2011).

[10] Cyberinfrastructure for Environmental Observation Networks (CEON) Workshop Report, held February 25 & 26, 2008 at The National Science Foundation, Arlington, VA. http://feon.wdfiles.com/local--files/start/Feb2008WorkshopFinalReport.pdf (last accessed June 1, 2011).

[11] J. Dozier and W. B. Gail, "The emerging science of environmental applications", In "The Fourth Paradigm", T. Hey, S. Tansley and K. Tolle, Eds., Microsoft Research, Redmond, WA, pp. 13-29.

[12] M. Lehning, N. Dawes, M. Bavay, M. Parlange, S. Nath and F. Zhao, "Instrumenting the earth: next-generation sensor networks and environmental science", In "The Fourth Paradigm", T. Hey, S. Tansley and K. Tolle, Eds., Microsoft Research, Redmond, WA, pp. 45-51.

[13] J. S. Horsburgh, D. G. Tarboton, D. R. Maidment and I. Zaslavsky, "Components of an environmental observatory information system," Computers & Geosciences, 37(2): 207-218, 2011.

[14] Critical Zone Exploration Network, www.czen.org (last accessed June 1, 2011).

[15] National Center for Earth Surface Dynamics, University of Minnesota, http://www.nced.umn.edu/ (last accessed June 1, 2011).

[16] T. Whitenack, I. Zaslavsky, D. W. Valentine, "HIS Central and the hydrologic metadata catalog", Eos Trans. AGU, 89(53), Fall Meet. Suppl., Abstract IN51A-1142, 2008.

[17] ESRI Geoportal Server, http://geoportal.sourceforge.net/ (last accessed June 1, 2011)

[18] Open Geospatial Consortium – Hydrology Domain Working Group, http://external.opengis.org/twiki_public/bin/view/HydrologyDWG/ (last accessed June 1, 2011).

[19] Observations and Measurements – XML Implementation, OGC Document 10-025r1, http://portal.opengeospatial.org/files/41510 (last accessed June 1, 2011).

[20] J. S. Horsburgh, D. G. Tarboton, M. Piasecki, D. R. Maidment, I. Zaslavsky, D. Valentine and T. Whitenack, "An integrated system for publishing environmental observations data," Environmental Modelling & Software, 24(8), pp. 879-888, 2009.

[21] European Petroleum Survey Group, Geodetic Parameter Dataset, http://www.epsg-registry.org/ (last accessed June 1, 2011).