# Data Infrastructure for the Critical Zone Observatories (CZOData): an EarthCube Design Prototype

Ilya Zaslavsky[1], Mark Williams[2], Anthony Aufdenkampe[3], Kerstin Lehnert[4], Emilio Mayorga[5], Jeff Horsburgh[6]

## 1. CZOData and EarthCube

CZO is a group of NSF-funded observatories that are dedicated to investigating earth processes in the critical zone, which is the region between bedrock and the atmospheric boundary layer. Comprehensive understanding of the earth processes requires integrating information and knowledge across earth science domains, including hydrology, geomorphology, atmospheric sciences, studies of soil and vegetation, geochemistry and geophysics. *CZOData* is the cyberinfrastructure (CI) supporting integrated analysis and modeling in the critical zone across observatory sites and earth science disciplines. As such, its vision is similar to the vision of EarthCube. We believe that its design ideas and challenges are relevant to EarthCube, in particular the focus on efficient data and knowledge integration across earth science domains to understand, model and eventually manage earth processes. This white paper emphasizes key requirements and design principles of CZOData as an EarthCube prototype. Besides presenting an initial conceptual architecture of CZOData, we consider it from the following perspectives: 1) driving CZOData design from user requirements and patterns of community organization, 2) different levels of interoperability supported by different system components, as needed by different research designs, 3) organization of various NSF-supported CI efforts in the earth sciences into an efficient and sustainable infrastructure with community-based governance, 4) organizing the CZO data infrastructure around community standards for data exchange. Besides CZOData, the paper is motivated by the experience of the authors in the design of several large cyberinfrastructure efforts for the earth science observatories: CUAHSI Hydrologic Information System (HIS), Ocean Observatories Initiative (OOI), the Geosciences Network (GEON), Chesapeake Bay Environmental Observatory (CBEO), the EarthChem system for geochemical datasets and the Integrated Earth Data Applications (IEDA), management of local LTER and CZO research sites, and involvement in community standards development through the Open Geospatial Consortium.

## 2. Key user requirements of CZOData

Typical research scenarios that CZOData must support rely on data fusion across several earth science domains. Analysis of water and chemical fluxes under different snow covered catchments; spatially distributed hydrologic modeling under different snowmelt regimes, and respective flow path changes and geochemical weathering impacts; nutrient dynamics in different topographic settings under long-term climate changes - are example scenarios that require integration of remotely sensed climate and snow cover grids, hydrologic and atmospheric time series measured at stations, geochemical samples,

[1] San Diego Supercomputer Center, UCSD
[2] University of Colorado, Boulder
[3] Stroud Water Research Center
[4] Lamont-Doherty Earth Observatory, Columbia University
[5] University of Washington
[6] Utah State University

geomorphological and topographic maps, and other data collected by the observatories. CI that can efficiently enable such research scenarios is the main goal of CZOData.

Common requirements and governance patterns for integrated observatory systems have been discussed, in particular as part of the NSF Federation of Environmental Observation Networks (FEON) initiative. In the CZO case (and, we'll argue, in the EarthCube case), additional emphasis is on information and knowledge integration across disciplines. Specific needs of the user community, which directly affect cross-disciplinary system design, include:

- Flexibility with respect to diverse foci and research scenarios, as research agendas evolve and incorporate more diverse and voluminous data from other domains, while new scientific data types and respective processing and modeling routines become available.
- Support for several levels of interoperability: from simple discovery of relevant datasets across domains based on standard metadata about location and observed variables, to generation of integrated data products that are normalized (i.e., brought into common spatial and temporal frameworks, consistent with respect to variables and measurement units, etc.), curated and QA/QC-ed, and ready for data mining, analysis, and modeling.
- High quality and transparency of the data and derived products, and easily traceable provenance of integrated data products. These are critical components for both CZOData and EarthCube, since, for efficient knowledge integration in the earth sciences, data and research results from neighboring domains must be trusted to be included in models, which typically implies high quality and known lineage.
- Knowledge integration across different information models, data discovery and access interfaces, and vocabularies as adopted by different domains. As typical examples, domain models and corresponding storage schemas, file formats and data access protocols are different for observations made at ground locations vs. satellite remote sensing; time series from sensors (typically, *in situ*) vs. *ex situ* analytical samples. Integrating these data to support modeling of physical processes is the key issue of CZOData, and would also be a component of EarthCube.
- Support for longer-term data analysis, which requires long-term preservation of observational data and sufficient metadata to interpret changes in observed properties, measurement methods, units, data quality characteristics, etc. (typical with changes in sensing devices) – especially if the data are generated in a different earth science subdomain.
- Integrating data across governance and licensing models, leveraging emerging social networks in the community. The established governance patterns (along agency boundaries, domain boundaries, and within scopes of individual projects) have been too rigid to support efficient knowledge integration. These boundaries, which are exacerbated by different data licensing models, different time frames and modes of decision-making, have been difficult to cross. However, the emerging communication models (e.g., via social networking) and governance models (e.g., standards governance as practiced by the Open Geospatial Consortium) appear to provide a better way to connect people – and hence support knowledge integration – across domain and institutional boundaries.
- Ability to access and analyze data from other domains using familiar applications and interfaces. User loyalty to interfaces and APIs adopted within their discipline is often underrated. Rather than creating new user-focused applications from scratch, incremental enhancements of existing interfaces and the added ability to access and interpret catalogs and data from other domains, appear to better match community expectations.

- Maximally leveraging previous NSF investment in CI for the earth sciences, and integrating different and separately managed projects as CZOData and EarthCube components (a key governance issue). We consider three groups of such projects/facilities:
  - Information systems in different earth science domains: CUAHSI HIS for hydrologic observations; EarthChem for geochemical data; OpenTopography for LiDAR data; UNIDATA for atmospheric fields; LTER Network for long-term ecological data, etc. Each information system has developed community-specific data publishing and sharing solutions. Rather than creating a new system, the integrated infrastructure should incorporate these domain systems, especially those that received significant traction in the community (e.g., the CUAHSI system provides web service access to 5.2 billion observations collected by 75+ observation networks from government and academia; the EarthChem system is an integrated portal serving multiple geochemical analysis datasets.)
  - Large cross-domain CI: high performance computing facilities and tools, general infrastructure for data sharing and long-term data preservation (DataNet, TeraGrid), authentication/authorization (Shibboleth), etc. While these components may not be developed specifically for the earth sciences, the EarthCube would establish procedures and workflows for engaging these advanced tools and data preservation environments in integrated earth science data infrastructure.
  - Data management facilities at observatory and research sites. Even before data management plans became a proposal submission requirement, many NSF-funded earth science research sites, environmental observatories in particular, have maintained computer facilities and have stored, documented and shared collected data. In the CZO project, these sites are the core knowledge generation centers, developed and cultivated by groups of PIs over many years. The sites have maintained data systems tuned to specific research designs, software environment and personnel skills, and not necessarily adapted for cross-domain data sharing. In fact, large-scale data sharing solutions are often perceived as taking time off immediate research tasks. Yet, our experience suggests that building EarthCube "top down" without taking into account needs of these sites or incorporating these facilities as the key components of the EarthCube, won't be successful.

## 3. CZOData design

Interoperability is often poorly defined. Different research designs require different "interoperability levels." Often, it is sufficient for a researcher to discover and download datasets using coarse metadata and do interpretation/analysis offline. Such discovery and download interfaces (both visual and programmatic) can be sufficiently uniform across disciplines as they require minimum metadata. At the other extreme, researchers are expected to interact with specific data models, for detailed analysis, modeling or curation. As these data models, or scientific data types, evolve and diverge, so do supporting data access interfaces. While in some earth science domains, community data models and protocols have been emerging (e.g., CF-NetCDF for atmospheric variables, CUAHSI ODM and WaterML for hydrologic observations), others see a wide variety of approaches to data representation and description, with little convergence or standardization.

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. To accommodate these different needs, we consider CZO data interoperability at several levels (Fig. 1). This grouping of interoperability levels is not exhaustive, but presented here as an

illustration of the layered organization of CZO functionality, and CZO software components that implement it.

- 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 standard, and subsequently invoked or downloaded to a user's workstation.
- At the next level, the resources have defined 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 can be easier interpreted and integrated.
- 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. In particular, observational data in EarthCube may be made available via OGC Sensor Observation Service (SOS) interface, though specific encodings transmitted over SOS would remain different for different communities.
- Finally, at the fourth level the data become available via standard services and in standard encodings that reflect a domain information model, to enable a much wider range of operations across different compliant sources. In some cases, encoding components are shared across disciplines (as several of them are developed as supplication schemas of OGC Geography Markup Language, GML). As described in a white paper by Rick Hooper, the OGC Observations and Measurements (O&M) model may provide a common basic encoding for different observational data. It is already used as the foundation of WaterML 2.0 for exchanging hydrologic data and the Climate Science Modeling Language (CSML) for atmospheric and oceanographic data, while several additional O&M profiling efforts are emerging, such as EarthChemXML and SoilML. This creates a foundation for reconciling different encodings and integrating datasets across domains in a way that is transparent and provides compatible data provenance descriptions.

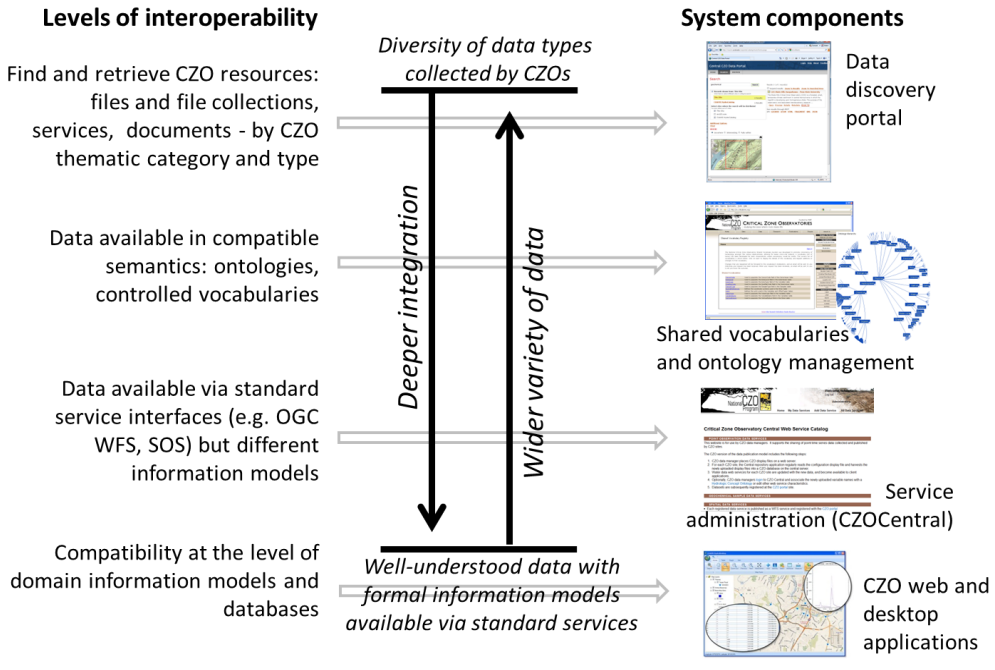


Figure 1. Levels of interoperability and corresponding components of the CZO data system

Different types of data integrated in CZO research scenarios support different levels of interoperability, and, therefore, rely on different system components. For example, soil samples, gridded data, and 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 not currently enforced. Hydrologic observations, on the other hand, represent one type of data that is made interoperable at all four levels within CZOData. In the current design, CZOData leverages components of the CUAHSI HIS and generally follows a Service Oriented Architecture (SOA) for publishing, indexing and accessing hydrologic observations, as implemented in the HIS project. One of the key CZOData design challenges is creating a flexible system, where new ways of measuring and representing environmental data developed by the observatories lead to enhancement of information exchange standards through an open community-based process, and to further evolution of domain CI towards adopting agreed-upon information models, vocabularies and service interfaces.

Figure 2 is an expanded version of the previous figure. It shows the three system layers (research observatory sites, domain infrastructure, and cross-domain data integration and information management). The CZO program 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). These sites collect and analyze large volumes of observations

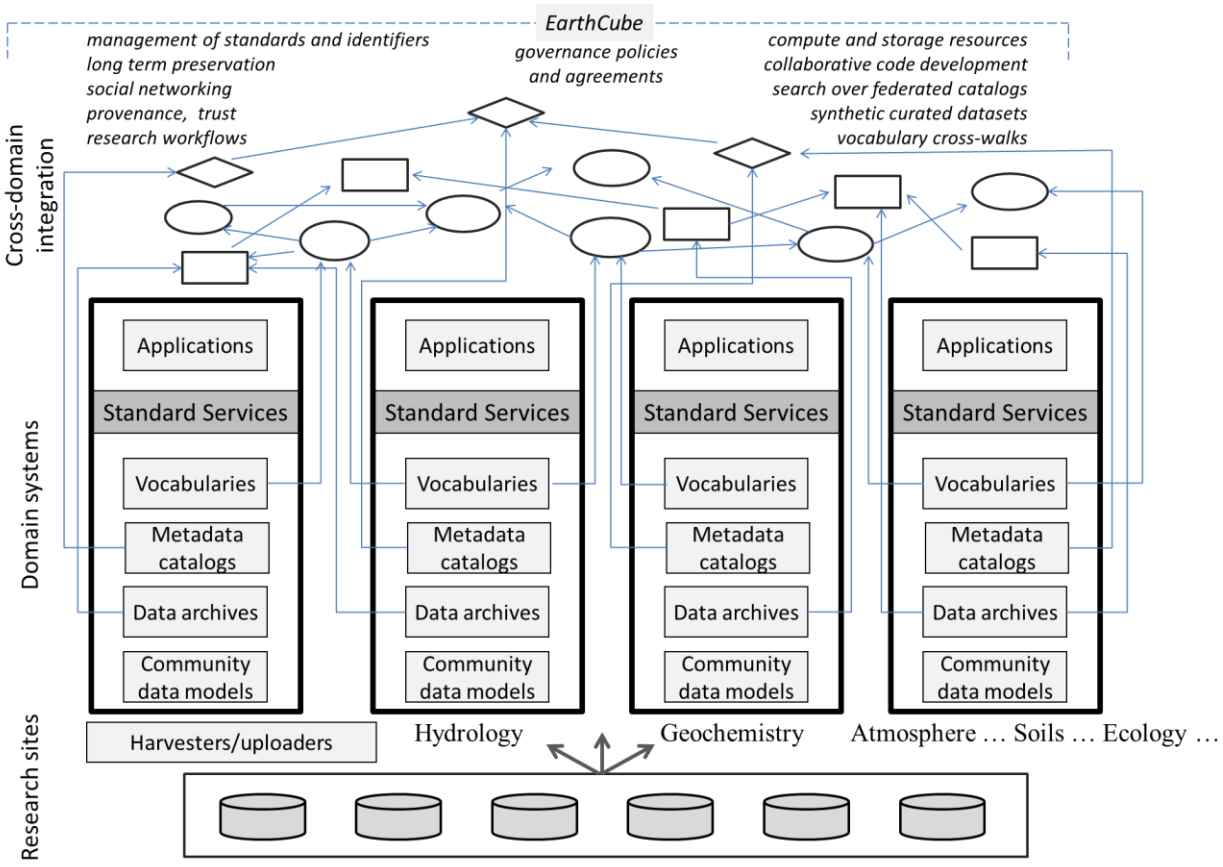


**Figure 2. CZOData vision (extended to EarthCube) as is an integrated information system that includes research observatories generating large volumes of observations, domain systems that publish the data according to community conventions about data models, vocabularies and protocols, and cross-domain knowledge layer that includes federated catalogs, normalized and curated datasets integrating data from domain systems and observatories, vocabulary cross-walks, as well as social networking, governance and compute infrastructure.**

of different types, both in situ and ex situ. The data are published via domain information systems: currently leveraging CUAHSI HIS for hydrologic time series and expanding to include the EarthChem system for managing geochemical samples, eventually extending to other discipline-specific data types and infrastructures. The domain systems manage data and metadata archives aligned with community information models, maintain domain vocabularies, and enable access to data, metadata and vocabularies via standard services. As long as these domain components can be accessed via standard interfaces (in case of CZO, we rely, to a large extent, on OGC standard specifications for data and metadata services, and migrating to SKOS/RDF for vocabularies), they can be integrated within the cross-domain knowledge integration layer, which would include: vocabulary cross-walks that establish correspondences between terms in domain vocabularies and support attribute-based data discovery and interpretation across disciplines; metadata catalogs that organize and federate domain catalogs and enable discovery of resources; integrated datasets that are compiled (and, ideally, curated) drawing data from several domains; persistent identifier management for cross-domain resources; high-performance and cloud-based compute facilities to manage and analyze large composite datasets; collaborative code development ecosystem; and a system of policies and governance ensuring that all CZO components work together and can be configured to address different research issues.

## 4. Community-based governance

A successful EarthCube governance model would reflect a long-term vision and scope of EarthCube, follow essential patterns of community organization, and specify how EarthCube components should be managed, how it interoperates with other infrastructures (including government/commercial), and how it evolves. As its primary role is to better support cross-domain and cross-project knowledge integration in fundamental earth science research, we expect that serving research needs of the community will drive the design of governance structures. Also, given the complexity and multi-level organization of the system, with many collaborating projects and funding arrangements, we expect that EarthCube will adhere to a set of overlapping and evolving governance models, and establish a set of governance policies for resolving potential "jurisdictional conflicts." In particular:

- for observatories and individual research sites, governance arrangements are defined by a PI or a group of PIs, following their contracts with granting agencies. In the NSF model, these are separately funded and fairly independent efforts, with different science goals and scope.
- domain CI efforts present a more complex governance structure. For example, governance of CUAHSI HIS includes many elements: a consortium of universities with a governing board; development teams; a user committee; operational and curation support of HydroServers and the central hydrologic metadata catalog; community consensus process about agreeing on domain models, vocabularies, data access protocols/services, and catalogs, which is leveraging the OGC framework for standardization and consensus building; agreements defining data publishing and data use rights and responsibilities, etc. Since the four key layers in a domain infrastructure (data models, vocabularies, services, catalogs) are different between disciplines, this governance structure won't automatically extend to other domains. Also, each domain CI is a "moving target" as it evolves the four components towards community consensus, at the same time accommodating new scientific data types and ensuing infrastructure components developed at observatories.
- a super-domain (or cross-domain) infrastructure would present many of the same governance components, which define curation and management policies for deriving domain models from

basic models such as O&M; vocabulary cross-walks, catalog federation, standard service interfaces for observations (e.g. SOS), etc.

In a complex governance structure like this, policies need to be developed to orchestrate consensus-building and regulate potential conflicts at the boundaries of EarthCube subsystems. They would define, for example, what should happen when new or enhanced scientific data types and formats are developed and propagated to domain systems and the cross-domain knowledge integration layer; what should happen if a sensor network managed by an observatory needs to be reconfigured by another group (perhaps from another domain) to address their research problems; how data life cycle arrangements are coordinated across domains (since data collected in one domain often provide context for data from another domain).

The experience of OGC points to a successful governance model, where communities of practice (organized as OGC domain working groups, with members representing different organizations) develop and present specifications for data description and exchange protocols. Development of a standard specification is preceded by carefully defining the scope and use cases it would address, describing how it will relate to other existing and proposed standards, and how it can be extended. The proposed specifications are shared with larger OGC membership, who are requested to comment on them, and go through a series of approvals before coming to a final vote. In this way, OGC maintains an open, transparent and formal process for bringing standards to the community and orchestrating their discussion, refinement, compliance testing, approbation in various projects, and eventual community adoption.

## 5. A CZO Use Case

Innovative solutions to process-level problems are needed to understand the biogeochemical evolution of surface waters in seasonally snow-covered catchments. Intensive and extensive research conducted at the catchment scale in the last decade shows that our understanding of the biogeochemical and hydrologic mechanisms that determine surface water quality is not sufficiently mature to model and predict how biogeochemical transformations and surface water quality will change in response to climatic and/or anthropogenic changes in energy, water, and chemicals. In recognition of this problem, the National Research Council has identified as a critical research need an improved understanding of how global change will affect biogeochemical interactions with the hydrologic cycle and biogeochemical controls over the transport of water, nutrients and materials from land to freshwater ecosystems.

A critical advance in addressing this problem was developed by Molotch et al. [2008]. They used remotely sensed snow cover data and a physically based snowmelt model to estimate the spatial distribution of energy fluxes, snowmelt, snow water equivalent, and snow cover extent over the different land cover types within the Green Lakes Valley, Front Range, Colorado, part of the Boulder Creek Critical Zone Observatory. The spatially explicit snowpack model then coupled to the Alpine Hydrochemical Model (AHM) to simulate discharge and hydrochemical fluxes at the catchment scale. Geochemical processes represented within AHM include ion elution from snowpacks, ion exchange, mineral weathering and equilibrium precipitation and dissolution. Nitrogen biogeochemical processes affecting ammonium and nitrate concentrations were represented to handle the acid base implications of nitrogen transformations in the soil solution and surface waters. The coupled model approach significantly improved our ability to model inorganic and organic solute concentrations and fluxes at the watershed scale.

This model approach required integration of field measurements of hydrology, hydrochemistry, climate parameters and remote sensing data to drive a spatially distributed snowmelt model, all combined with geochemical and biogeochemical models. EarthCube infrastructure, once operational, would make evaluation of such models more reliable and transparent, easy to replicate, and less time consuming, while at the same time letting researchers present results in a way that would be immediately useable by other environmental modelers. In particular, having geochemical and discharge information, land cover and snow cover time series, and other required data, in standard formats, with complete provenance and data quality information, would make it easier to reconcile different spatial and temporal resolutions and increase confidence in simulation outcomes. At the same time, the new CI would allow model extension and systematic deployment to all CZO sites and other sites (LTER, NEON, ARS, WEBB, USFS ERF, etc) where snow cover is important.

## 6. Conclusion

CZOData design recognizes that:

- an effective CI needs to integrate existing research observatory sites, domain infrastructures and cross-domain projects and facilities;
- variety in research designs mandates that our integration solutions are flexible and can support different levels of data interoperability – four such levels are described;
- domain infrastructures often implement information models, vocabularies, data access protocols and catalogs following different conventions - managing this diversity in a large dynamic system requires that interfaces to system components are standardized via an open community process;
- each of the domain CIs is a “moving target,” evolving to better community agreements on data sharing for well-understood data yet presenting new scientific data types as measurement technology and our understanding of earth processes improve – necessitating adaptable system design;
- separately managed system components may have different governance models and follow different policies - requiring reconciliation at the CZO-wide level;
- a knowledge integration and governance layer is the key to a successful cross-discipline distributed system – it is designed to not only discover and bring together datasets and metadata across organizational and domain boundaries, but also make it easier for researchers to connect, form communities of practice and share knowledge and interpretations.

We believe that these considerations are likely to be important for EarthCube, as it is being designed to support infrastructure needs of the larger earth sciences community, is supported through multiple interrelated projects and tries to maximally use existing infrastructures.

## 7. REFERENCES


[1] National Critical Zone Observatory program (CZO), www.criticalzone.org.

[2] The EarthChem Project , www.earthchem.org

[3] Tarboton, D. G., J. S. Horsburgh, D. R. Maidment, T. Whiteaker, I. Zaslavsky, M. Piasecki, J. Goodall, D. Valentine and T. Whitenack, (2009), "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.

[4] 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.

[5] 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.

[6] The US Long Term Ecological Research Network, www.lternet.edu.

[7] National Ecological Observatory Network, www.neoninc.org.

[8] 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.

[9] Open Geospatial Consortium – Hydrology Domain Working Group, http://external.opengis.org/twiki_public/bin/view/HydrologyDWG/.

[10] Observations and Measurements – XML Implementation. OGC Document 10-025r1, http://portal.opengeospatial.org/files/41510.

[11] Molotch, N. P., T. Meixner, and M. W. Williams (2008), Estimating stream chemistry during the snowmelt pulse using a spatially distributed, coupled snowmelt and hydrochemical modeling approach, Water Resour. Res., 44, W11429, doi:10.1029/2007WR006587.