Project Summary

Intellectual Merit: With recent initiatives within many scientific communities toward trans-disciplinary, synthetic research using both new and existing data resources, our ability to share, discover, interpret, and integrate data are paramount to scientific progress. Indeed, new scientific advances are dependent on the synthesis of observations from multiple measurements, at multiple scales, across scientific disciplines, across environmental observatory or other experimental sites, and from multiple sources. We are now at a point where our ability to collect data far outstrips our capabilities to analyze it using existing technologies and where the inadequacy of tools available for describing and sharing data leads to heterogeneity in the way data are organized, described, and encoded that hinders its discovery and interpretation. Several systems have emerged within geoscience communities for sharing earth observations, including the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS), EarthChem, the Integrated Ocean Observing System (IOOS), and, more recently, the Critical Zone Observatory (CZO) Integrated Data Management System (CZOData) and the Data Observations Network for Earth (DataONE). While these systems have increased the availability and reduced the heterogeneity of earth observations within geoscience domains, deficiencies remain because they describe, encode, and publish data differently. This limits the ability of all of these systems to unambiguously describe observations in a way that they could be interpreted by scientists from outside the domain. Here we propose to develop a community information model and supporting software to extend interoperability of discrete, feature based earth observations derived from sensors and samples and improve the capture, sharing, and archival these data. The information model will be designed from a general perspective, with extensibility for achieving interoperability across multiple disciplines and systems that support publication of earth observations. Data capture is a critical point in the data life cycle and we will develop tools to support, aid, and encourage reliance on the information model during collection and analysis to ensure that the information model enhances scientists' ability to work with data during analysis and at the same time capture metadata critical for later sharing and publication. Our multidisciplinary, community-focused effort will build consensus about the elements of this information model, conceptual implementations that specify schemas for both data storage and communication, and prototype physical implementations using diverse data use cases from existing repositories and observatories to demonstrate how this advanced information model can support federation of earth observations data across multiple data publication systems within the geosciences.

Broader Impacts: The proposed information model and software tools are consistent with the architectures of multiple existing cyberinfrastructures in the geosciences, but enhance domain-specific information models and encodings in a way that will assist data publishers in sharing data, enhance the semantic and syntactic consistency of data from different geoscience domains, and increase the cross-domain discoverability, accessibility, and integration of earth observations for data consumers. Experts from within and outside geoscience communities involved in our design workshops will assist in defining requirements, specification of use cases, design, and prototype development. Design workshops will also create exciting opportunities for student involvement. Several existing, NSF-supported efforts and communities, including hydrology and the CUAHSI HIS, solid earth geochemistry and EarthChem, and CZO and its CZOData system will directly benefit from this work. Our project team is uniquely poised to integrate the proposed work into these systems to better enable community members in sharing and integrating data, as well as demonstrate a basis for sharing earth observations across research sites and observatories that can be used as a model outside of these communities. The proposed work is also perfectly timed to contribute to data interoperability efforts of the Open Geospatial Consortium's (OGC) Hydrology Domain Working Group, which, with integral participation from project members, is now developing international standards for representing and encoding earth observations data. This connection will ensure that our work is impactful on an international scale within a forum that includes participation from academics, government agencies, and commercial software vendors. Finally, developments of this project will serve as a demonstration for how earth observations from multiple domains and cyberinfrastructures can be integrated in support of NSF's EarthCube cyberinfrastructure initiative.

Developing a Community Information Model and Supporting Software to Extend Interoperability of Sensor and Sample Based Earth Observations

1. INTRODUCTION

The last two decades have seen a dramatic increase in environmental data availability. Early websites published data for an individual agency or project with no established standards or protocols for describing, encoding, or delivering the data. Most data on the Internet were not easily discoverable, were not described with sufficient attribute information to facilitate their interpretation, and heterogeneity in the publication protocols, encodings, and semantics of data from multiple sources made it difficult to combine data from multiple sources within a scientific analysis.

More 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 (HIS) (Tarboton et al., 2010), the Critical Zone Observatory Integrated Data Management System (CZOData) (Zaslavsky et al., 2011), the Integrated Earth Data Applications (IEDA) and EarthChem system (Lehnert et al., 2011; Lehnert et al., 2004; 2009), and the Integrated Ocean Observing System (IOOS) (De La Beaujardiere, 2008). These systems are built using the principles of service-oriented architecture (SOA) 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 considerable progress, it still takes a knowledgeable investigator considerable effort to discover and access datasets from multiple domain-specific repositories for a synthetic analysis because of inconsistencies in the way the different domain systems describe, encode, and share data. Members of the team proposed here are integrally involved in each of the cyberinfrastructures mentioned above, and we are acutely aware of the shortcomings that currently exist. We have come together in this project to address these shortcomings and extend our ability to seamlessly share and integrate data across the geoscience disciplines.

First, data structures used by existing domain cyberinfrastructures are often insufficient to store or describe the entire range of earth observations. For example, data structures and encodings used by the CUAHSI HIS contain the necessary metadata to describe time series of *in situ* observations made at point locations such as streamflow gages and weather stations. However, they are inadequate for water quality or solid earth geochemical samples taken in the field and analyzed later in a laboratory because existing method and sample descriptions do not contain all of the needed metadata and are not extensible to allow, for example, important data structures such as sample fractions and sub-sample parent-child relationships. Conversely, the EarthChem system contains the necessary metadata elements and structures to effectively describe observations derived from *ex situ* analysis of geochemical samples, but is not well structured to support time series of observations from *in situ* sensors. Yet, there are many research scenarios that require efficient integration of these data types across different domains of observational earth science. For example, understanding a soil profile's geochemical response to extreme weather events requires integration of hydrologic and atmospheric time series with geochemical data from soil sample fractions collected over various depth intervals from soil cores or pits at different positions on a landscape. Similarly, understanding spatial and temporal patterns in suspended sediment fluxes, sources, and associated contaminants in response to land use and climate change requires close integration of hydrologic time series with a variety of geochemical data analyzed in different laboratories on separate sample fractions (e.g., acid extract of fine sediments for heavy metals, solvent extract of whole water for organic contaminants, dried filter for suspended solids concentration). Currently, integrated access and analysis of data for such studies are hindered because common characteristics of observational data, including time, location, provenance, methods, and units are described using different constructs within different systems. Integration requires multiple syntactic and semantic translations that are, in many cases, manual, error-prone, and/or lossy. Standardizing such descriptions of common characteristics within a common observations information model would lead to more reliable data integration.

Second, there remain deficiencies in the ability of existing systems to unambiguously describe and encode observations in a way that they can be discovered, accessed, and interpreted by scientists from outside the domain. New, trans-disciplinary research programs such as those in the CZO Network and Water Sustainability and Climate observatories (WSCs), where a large number of scientists from multiple disciplines are working collaboratively in collecting and analyzing diverse datasets, will rely on effective data management systems that enable both efficient data capture and foster the ability of scientists both within and outside of these projects to share, discover, access, and integrate the broad range of earth observations that they collect for specific analyses. Work is still needed to achieve consensus and work toward community standards-based approaches for describing and encoding earth observational data across domains to enable effective interchange of information.

We propose to develop a unified earth observations information model and supporting software infrastructure, with broad involvement of the geoscience community in examining and refining our work. The information model will address deficiencies in data interoperability both within and among existing geoscience cyberinfrastructures. The software infrastructure will include storage, transfer, and catalog encodings of the information model and additional software tools aimed at improving the capture, validation, verification, sharing, and archival of earth observations data. The broad nature of geoscience domains, and the even broader context of environmental observations data in general, necessitate that we focus our efforts. In this project we will focus on the domain of spatially discrete, feature-based earth observations resulting from *in situ* sensors and from environmental samples and sample fractions, as well as data products directly derived from them. This is consistent with the data holdings of the systems on which we have worked and for which we seek interoperability. We distinguish our focus from that of gridded or continuous datasets, for which different data models and encodings have emerged, particularly the "NetCDF-CF-OPeNDAP" stack and related Unidata Common Data Model (CDM) (Hankin et al., 2010b; Nativi et al., 2008). We will, however, draw upon and extract useful concepts for our work from these data models and interact with the community advancing them to take advantage of opportunities for alignment. Tarboton and Lehnert serve on the Unidata Policy Committee and will use these connections to continue this engagement. We will also focus on the environmental domain strengths of our project team, which are observations from the Earth surface critical zone (aquifers, sapprolite, soils, and the ecosystem they support), coastal and inland waters, and solid earth observations that extend into the oceans.

Developing a community information model across geoscience disciplines and cyberinfrastructures will create a greater degree of interoperability among existing systems and data resources and will better support data management at each step in the data life cycle, from producers to consumers. Our project team is poised to do this in the context of a number of existing systems, and, in turn, provide the necessary foundation to support trans-disciplinary, synthetic research in the environmental sciences. For data consumers, this means a substantially improved level of discovery, access, and integration of earth observations across geoscience data repositories to support their research. For data producers/publishers, the advanced earth observations information model and supporting software infrastructure that we propose to develop will enable them to share their data using community accepted, interoperable standards. We will bring together scientists and technologists from disciplines within and outside the geosciences to design and develop the high-level information model and then to build standards-based technology implementations for data storage, archival, publication, cataloging for discovery, and exchange over the Internet that can be adopted by existing geoscience cyberinfrastructures.

2. OBJECTIVES

Our overarching goal is to create an information model that is ***integrative*** and ***extensible***, accommodating a wide range of observational data and aimed at achieving interoperability across multiple disciplines and systems that support publication of earth observations. Our team comprises a high level of expertise, bringing together a group of investigators who have successfully developed and implemented information models for operational data systems such as CUAHSI HIS, EarthChem, and IOOS. This team will be able to apply long-term experiences with a wide variety of observational data to

generate the next generation information model that will, for the first time, allow a diverse range of geoscience observations to be consistently shared, discovered, accessed, and interpreted. In order to ensure that our developments can be readily adopted and used by existing and emerging systems with immediate benefits to the geosciences community, we will develop software tools that support, aid, and encourage reliance on the information model during data collection and analysis. The tools will take advantage of the information model to enhance scientists' ability to work with the data during analysis, while at the same time capturing metadata critical for later sharing and re-use of the data. Our community-focused effort will have the following four objectives (activities and outcomes are detailed in Section 5):

1. Development of a community information model for spatially discrete, feature based earth observations.
2. Engagement of geoscience communities in the design of the information model
3. Implementation of the observations information model in encodings for data storage, archival, transfer, and for cataloging metadata.
4. Deployment of data publication prototypes using data from CZOData, CUAHSI HIS, EarthChem, and IOOS.

3. BACKGROUND AND RATIONALE

Observational data are fundamental to the geosciences. The Open Geospatial Consortium (OGC) Observations & Measurements (O&M) standard (Cox, 2010) provides a definition of observations:

"An observation is an act associated with a discrete time instant or period through which a number, term, or other symbol is assigned to a phenomenon. It involves application of a specified procedure, such as a sensor, instrument, algorithm, or process chain. The procedure may be applied in situ, remotely, or ex situ with respect to sampling location. The result of an observation is an estimate of the value of a property of some feature."

While there are many properties of observations that are common across the various types of observational data acquired and used within the geosciences, each domain also presents observation types that are unique. Data structures have been built to support the most common types of observations within a specific domain, without consideration of the broader context of available observations across domains, leading to substantial syntactic and semantic heterogeneity in observational data representations. A domain-agnostic definition of observations, such as the one above, provides the basis for a higher-level information model that can support observations of many types from many domains.

An information model is a representation of concepts, relationships, constraints, rules, and operations that specify the semantics of data for a chosen domain of discourse (Lee, 1999). At its simplest level, an information model defines the domain's entity types and their properties, relationships, and allowed operations on the entities. In a relational database implementation, entities become tables and their properties become table columns. The choice of columns in each table, along with associated rules for populating them with data, essentially define metadata requirements and opportunities. Table structure and relationships between tables govern the types of queries that are possible for loading, extracting and analyzing the data. The information model behind a data system is thus critically important to the effectiveness and interoperability of the cyberinfrastructure.

More generally, an information model provides a sharable, stable, and organized structure of the information requirements for a domain context, without constraining how that description is mapped to an actual implementation in software (Fulton, 2006). There may be many mappings of the information model. Such mappings are called data models, irrespective of whether they are object models, entity relationship models such as those used by relational databases, or XML schemas. The fundamental elements within the information model are based on the domain of discourse to be described and how the model will be used – e.g., to support data discovery or data storage. For example, a rich set of descriptive metadata about the variables that were observed and the context within which an observation was made is fundamental for both discovering and interpreting observational data (Madin et al., 2007).

In practice, the information model must have a number of physical implementations that enable effective interchange of information (Figure 1). This is especially true in the case of SOAs where datasets are shared through loosely-coupled services deployed on heterogeneous computer systems using different software technologies. Semantic and syntactic heterogeneity are major hurdles to be overcome, especially across data types and scientific domains (Beran and Piasecki, 2009; Horsburgh et al., 2009; Hankin et al., 2010a). Because many systems already have their own existing data structures and/or database implementations, one solution to achieving interoperability between systems is to agree upon a common information model to which data held within the existing systems can be mapped. The common element, then, across systems and the services that they provide is the information model, with physical implementations within various file systems and databases for data storage, within XML schemas and file formats for data transfer, and within web service interfaces that provide access to data. Indeed, overcoming heterogeneity and achieving interoperability within SOAs depends on a common information model and well-defined interfaces and data encodings that implement it.

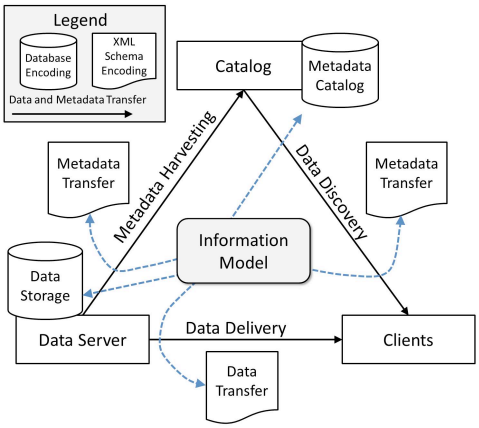

Figure 1. The information model is central to SOA and specifies the information requirements and semantics of data encodings for storage and transfer of data and metadata.

This proposal focuses on development of a community information model for earth observations and its implementation within existing geoscience domain cyberinfrastructures to improve interoperability of data across these systems. It will build on and draw from the information models of existing geoscience cyberinfrastructures developed by members of the project team, who are coming together on this project with the objective of an information model that increases interoperability across these systems. As the architects and developers of these systems, our project team is intimately familiar with their strengths, their deficiencies, and opportunities for creating greater interoperability both within and across them. Indeed, the work proposed here is only achievable in the context of the existing domain cyberinfrastructures on which we have worked. In the following section we provide relevant details of the existing systems, including information about the information model used by each, their deficiencies, and opportunities for addressing them via the proposed work.

4. EXISTING GEOSCIENCE CYBERINFRASTRUCTURES

CUAHSI HIS – Over the past 8 years, Tarboton, Zaslavsky, and Horsburgh have worked on the CUAHSI HIS, which has achieved tremendous success in advancing the interoperability of hydrologic observations made at monitoring points through the development and standardized use of the Observations Data Model (ODM) (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., 2009) 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.

Despite the success of the CUAHSI HIS for hydrologic time series measured at fixed geographic points, its underlying information model and implementations (e.g., ODM, 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. Each of these limitations would be addressed by the proposed information model.

EarthChem – Since 2005, the EarthChem project, led by Co-PI Lehnert, has developed and operated a suite of data systems and services for solid earth geochemical data that are acknowledged worldwide as a leading resource for sample-based analytical data and include the EarthChem Portal that provides a central access point to data in federated databases, the Geochemical Resource Library as a repository and publication agent for geochemical datasets, and Geochron. One of EarthChem's main achievements has been the community-driven development of metadata standards for geochemical data through workshops, the Editors Roundtable (http://www.earthchem.org/editors), and collaboration with other geochemical data systems (Lehnert et al., 2007). EarthChem has developed templates for investigators to format their data and assemble metadata according to these standards. Lehnert has also led the development of standards for the identification, registration, and documentation of physical samples in the geosciences, creating the International Geo Sample Number IGSN and the System for Earth Sample Registration SESAR (www.geosamples.org; Lehnert et al., 2005; Lehnert et al., 2011).

EarthChem data collections employ a modified version of the data model that was developed for the PetDB and GEOROC databases (Lehnert et al., 2000), and that has been adopted by various other geochemical databases such as MetPetDB (Spear et al., 2009) and CZchemDB (Niu et al., 2011). In order to develop interoperability among geochemical databases and allow users to seamlessly discover and access data in distributed systems, EarthChem developed an XML schema for sample-based geochemical data that is now used as the standard data transfer protocol for databases in the EarthChem federation.

EarthChem's information model contains all information necessary to describe geochemical samples, subsamples, sample fractions, the observations derived from them, and provenance of the data. However, the EarthChem information model does not accommodate time series-based information because all data in the EarthChem information model are related to a discrete sample. The proposed information model will allow EarthChem to overcome this deficiency, broadening the application of the entire system and substantially enhancing interoperability with other systems.

CZOData – Developed over the past two years by Zaslavsky, Horsburgh, Tarboton, Aufdenkampe, and others, the prototype CZOData system has focused on publishing hydrologic observations collected at CZO sites and leverages SOA approaches and software components developed by the CUAHSI HIS 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 CZO site data managers to publish their data for public access. The Display File format is based on the CUAHSI HIS ODM information model. Once published at an individual CZO web site, Display Files are automatically harvested into the CZO Central Data Repository at the San Diego Supercomputer Center (SDSC). The harvested data are then validated against shared vocabularies and a variable ontology, archived in a set of ODM databases established for each CZO, and then published via standard CUAHSI WaterOneFlow web services that transmit data according to WaterML 1.1. CZO shared vocabularies are also adapted from the CUAHSI HIS ODM controlled vocabulary management system and establish semantic conventions within the CZO system.

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 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. *This comprehensive, common model is yet undefined and is, in fact, the goal of this proposal*. Enabling data integration at the information model and storage level would enable a more flexible data publication infrastructure and additional types of cross-domain database queries.

The entire proposed project team has a pending proposal, led by M. Williams (U. Colorado Boulder) and A. Aufdenkampe, for the second phase of development of the CZOData system. We are awaiting award notification, and if funded, the new CZOData project would provide substantial synergisms with the work proposed here. The new CZOData project does not provide for new information model development, but primarily CZO specific software and tools for interoperability. Those CZOData project products would be significantly enhanced by the proposed new information model.

DataONE – The Data Observation Network for Earth (DataONE) is part of NSF's ongoing DataNet (Sustainable Digital Data Preservation and Access Network Partners) program. PI Horsburgh is a member

of DataONE's Core Cyberinfrastructure Team and leads the Data Integration and Semantics Working Group, which is developing strategies for implementation of semantic technologies that enhance data discovery and integration. The observations information model proposed here is of immediate interest to this work. 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 SOA consisting distributed "Member Nodes" on which scientific data are published and "Coordinating Nodes" that provide data discovery and other services.

In its first phase, DataONE is treating data as opaque objects and does not require a specific format for submitted data (DataONE, 2012). 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.

IOOS – U.S. IOOS is a federal-regional partnership enhancing the nation's ability to collect, deliver, and use data and information needed better understand our oceans, coasts and Great Lakes. Central to IOOS is the presence of a Data Management and Communication (DMAC) subsystem capable of delivering real-time, delayed-mode, and historical data for *in situ* and remotely-sensed physical, chemical, and biological observations, as well as model-generated outputs (De La Beaujardiere, 2008; IOOS, 2010). Co-PI Mayorga leads DMAC efforts for the Northwest Association of Networked Ocean Observing Systems (NANOOS). NANOOS DMAC coordinates with national DMAC efforts and with regional providers in the integration of coastal monitoring data across a wide range of sources, platform types, and instruments, and provides data management, access, and visualization services to a cross-section of stakeholders in the region (Mayorga et al., 2010). NANOOS is currently engaged with the national IOOS program and other partners in a project to clarify, refine and extend pilot O&M-based IOOS data interoperability systems. However, the focus remains largely on sensor observations. The information model proposed here would greatly enhance the integration of sample-based data with sensor observations, and would facilitate "summit-to-sea" integrative assessments of watershed freshwater exports with coastal impacts such as eutrophication and pollutant runoff.

Other CI Efforts – There are related efforts within several disciplines to overcome heterogeneity and advance interoperability of observational data. For example, the Scientific Observations Network (SONet), of which PI Horsburgh is a participant, is an NSF-supported INTEROP project that is focused on creating a network of researchers within the ecological, environmental, and computer sciences working toward defining and developing the necessary specifications and technologies to facilitate semantic interpretation and broad interoperability of scientific data (Scientific Observations Network, 2012). Another example is the OGC O&M standard (Cox, 2010). O&M as an observations information model has also gained a lot of traction in many different scientific domains for representing earth observations and now has several XML profiles serving domains in geosciences – e.g., WaterML 2 (Open Geospatial Consortium, 2012), GeoSciML (CGI, 2012), SoilML (Montanarella et al., 2010), etc. Our team's participation in SONet and our experience in profiling O&M in the development of WaterML 2 (Co-PI Zaslavsky is a co-chair of the OGC/WMO Hydrology Domain Working Group (HDWG)) and in IOOS marine data implementations (Co-PI Mayorga is a key member of the national team finalizing an O&M sensor profile for IOOS) will enable us to incorporate the best aspects of each of these efforts into our work. In fact, it is probable that the work proposed here will substantially influence or guide those efforts, so the exchange of ideas will benefit all.

5. PROJECT ACTIVITIES

We propose a tightly integrated, collaborative, and iterative set of activities necessary to achieve our overarching goal of creating an information model that accommodates a wide range of observational data and that is capable of enhancing interoperability across multiple disciplines and systems. These activities, the challenges that each will address, our innovative approach, and expected outcomes are described in following sections. Central to our work will be collaborative development of the information model

(Objective 1), with feedback from the community (Objective 2). Implementation of storage, transfer, archive, and catalog encodings of the information model (Objective 3) and deployment of data publication prototypes for test datasets from our respective domains (Objective 4) will demonstrate how the new information model can improve SOA functionality of existing geoscience cyberinfrastructures.

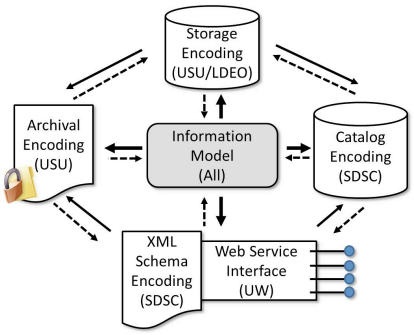

Figure 2. Integrated information model development and prototyping.

We expect tight integration and iteration between project tasks. The information model will determine the structure of prototype encodings (Figure 2, large solid arrows), and the requirements for each of the prototype encodings will feed back into the design of the information model (dashed arrows). We also expect iterative development among the prototype encodings. For example, the storage encoding will influence both the archival and catalog encodings, each of which may impose requirements on the design of the storage encoding. At each step in our design and prototype development process, we will test the model and encodings with diverse, complementary geoscience datasets that span the domain of discrete, feature-based earth observations. These include: 1) *Aufdenkampe* – geochemical source "fingerprints" of suspended sediment fractions collected from streams at different hydrograph stages, seasons and land uses within the Christina River Basin CZO; 2) *Horsburgh* – a diverse set of hydrologic time series, water quality (both *in situ* time series and *ex situ* samples), and weather and soil time series measured within the Little Bear River Experimental Watershed; 3) *Lehnert* – a subset from the CZChemDB database, which contains CZO observations derived from solid earth geochemical samples; 4) *Mayorga* – the PRISM Puget Sound marine dataset that contains water-column profile observations of physical and chemical characteristics from *in situ* sensors and *ex situ* water samples at fixed stations. These data use cases will define the requirements for and challenge the ability of the information model and its physical implementations to unambiguously describe, encode, and transmit earth observations within a SOA.

Objective 1: Development of a Community Information Model

Task 1.1. Development of the Core Observations Information Model and Extensions

Challenges: The information models of current geoscience cyberinfrastructures representing discrete, feature based 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. Challenges lie in how to represent the core elements of both sensor and sample-based observations using common semantics, while enabling extensions for representing the different data types and domain specific subtleties.

Proposed Activities: Horsburgh and Tarboton will lead the entire project team in development of the earth observations information model, adding extensibility aimed at achieving interoperability across multiple disciplines and systems that support publication of earth observations. Development will proceed in multiple phases. In Phase 1, we will define the scope of the information model through: 1) using our experience in developing existing geoscience cyberinfrastructures over the past several years to jointly identify the needs of existing systems; 2) examining our data use cases and those of participants we invite to our design and prototyping workshops (discussed below); and 3) eliciting input from community members through surveys, informal communications, and through a review of considerable feedback that we have received over the past several years as we have deployed our existing systems for use by the community. The scope of the information model will define the types of data to be addressed (e.g., point time series, water quality samples, solid earth geochemistry samples, etc.) and the functional use cases that the information model must support (e.g., data storage, transfer, archival, cataloging, data reuse, etc.).

In Phase 2, we will define the information requirements needed to meet each of the functional use cases for each of the identified data types. Again, we will lean on our past experience, which has shown

that each functional use case has its own information requirements. An important new focus of this phase will be to examine use cases for cross-domain observation data catalogs, including harvesting, indexing and searching metadata, which will be required to achieve interoperability across observation repositories. We will examine the identified information requirements to extract the "core" information common for all observations and will separate information specific to particular data types or functional use cases into extensions. This will enable us to define a "core" information model, with extensions for particular data types and/or use cases, providing extensibility for addition of new data types or extensions in the future (Figure 3). We will focus on identifying information requirements of the data and functional use cases, but will define an information model that can be extended by users, giving them the necessary flexibility to add their own attributes to describe their data.

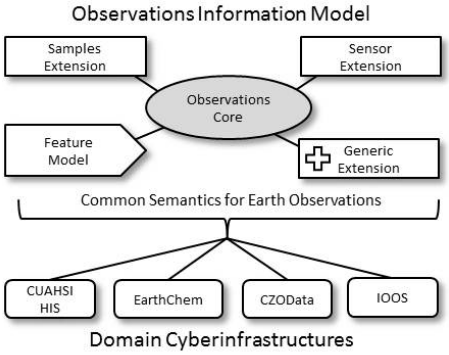

Figure 3. The core observations information model and extensions will provide common semantics for earth observations shared by multiple geoscience cyberinfrastructures.

Finally, we will transform the information requirements into a conceptual model that is independent of any physical implementation. We will begin this work using OGC O&M as the unifying information model, although we will consider other emerging information models such as Unidata's CDM. This is important for several reasons: 1) O&M is already being used within geoscience domains; 2) O&M is an international standard, with existing supporting software infrastructure and industry support; and 3) our resulting work can feed directly into efforts to extend and profile O&M such as those of the OGC HDWG who have recently finished profiling O&M for *in situ* point time series of hydrologic observations in WaterML 2 and who are now poised to begin work on the representation of water quality samples. Where needed, we will use elements from our existing information models (e.g., CUAHSI HIS ODM and WaterML for hydrologic time series, EarthChem and EarthChem XML for solid earth geochemical samples, etc.) and from the community input and targeted design workshops we will hold to profile and/or extend O&M to meet the information requirements identified in Phase 2. The conceptual information model will be captured using a formal modeling language such as UML (from which implementations can be automatically derived) to document the design, but will be translated into several forms that can more easily be communicated to members of the scientific community and that lend themselves to encoding within the various physical implementations that are planned.

Expected Outcomes: An observations information model for spatially discrete, feature-based earth observations that is compatible, where possible, with the OGC O&M standard.

Task 1.2. Linking Observations to the Geo-Environment

Challenges: Within current geoscience cyberinfrastructures there is inconsistency in the way that the area or volume to which an observation applies, or geospatial support, is represented and in the way the environmental feature of interest is represented. For example, the CUAHSI HIS associates all observations with a point location. It is left up to a data consumer to determine that a monitoring point lies on a particular river reach, measures the outflow of a particular catchment, is downstream of another monitoring site, and is located near a weather station – yet all of this information is needed by an investigator who is building a model of the watershed and takes precious time to develop. In another example, a point location may identify where a solid rock core sample was physically collected, but it does not convey information about the rock formation from which the core was collected or the geographic area over which observations from that particular core are expected to be representative. To address this problem, the EarthChem system maintains linkages between samples and geological features, but this type of association is not consistently applied across geoscience cyberinfrastructures.

Proposed Activities: As part of our information model development, we will identify the information requirements for representing the location at which observations from geoscience disciplines were made and for representing the domain features that the observations represent. Observations 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 of those geographic features could be linked not only through their physical proximity on a map, but also through specific relationships encoded within feature attributes that could then be queried for data discovery and that would aid in data interpretation. O&M has similar concepts of "sampling features" (e.g., stream gages, groundwater wells, etc.) and "features of interest" (e.g., stream reach, aquifer, geologic formation) that will be useful for this purpose, but there remains significant work in defining whether these concepts are adequate for representing the geospatial support for geoscience observations and how they should be used to do so. Defining a geospatial data framework that can serve as the context for observational data is a reasonable next step – and particularly important for integration of the results of this project with EarthCube (see below). Here, we will build on existing geospatial datasets and data models such as the National Hydrography Dataset (Horizon Systems Corp., 2010), Arc Hydro (Maidment, 2002), Arc Hydro Groundwater (Strassberg et al., 2011), and others. Locating observational data within such a geospatial "fabric" provides a way to evaluate complex geospatial queries such as those posed above and enables an explicit linkage between observations and the earth features that they represent.

Expected Outcomes: An improved information model for geospatial features of interest upon which observations are made.

Objective 2: Engagement of Geoscience Communities in the Design of the Information Model

Challenges: Engagement activities will identify the perspectives and needs of both data publishers and data consumers and will enable us to incorporate community feedback into our design and development process.

Task 2.1. Design and Prototyping Workshops

We will hold two design and prototyping workshops within Year 1 and a third interoperability challenge workshop (Task 4.1) in Year 2. In addition to our immediate team, we have budgeted travel and participant support to invite and host a diverse group of both domain scientists and cyberinfrastructure experts at these workshops, enabling us to elicit requirements, present preliminary designs, and receive immediate feedback from both data publishers and data consumers. We will rotate the location of these meetings (USU, SDSC, LDEO) to broaden participation of students and related colleagues.

Task 2.2. Education and Training

Our coordinated educational activities will help train the current and next generation of scientists by directly funding a full time PhD student at USU to work on the project and by targeting young CZO scientists and CZO data managers to participate in the design and prototyping workshops. The PhD student will travel to the workshops and have the opportunity to work closely with all project team members. As part of a recent NSF EPSCoR cyberinfrastructure award, PI Horsburgh will be developing a graduate-level course in Hydroinformatics, which will include educational modules about the management and representation of observational data based on results from this project. Last, the anticipated funding for a second CZOData project by our team will fund several workshops for CZO scientists and data managers to refine data use cases, metadata requirements, and controlled vocabularies, offering complementary educational and training opportunities.

Objective 3: Implementing the Information Model

Task 3.1. Implementing the Information Model for Data Storage and Capture

Challenges: Although multiple geoscience cyberinfrastructures use relational database management systems (RDBMS) for data storage, the lack of a common information model means that no common data storage schema exists that is cross platform and works with multiple RDBMS. Existing storage schemas also lack extensibility that supports adding new data types or new attributes or informational elements to

existing data types. Additionally, data capture is a critical point in the data life cycle, yet there are few tools available for data producers and managers to work flexibly with relational observation data models.

Proposed Activities: Horsburgh, Tarboton, and Lehnert will develop an implementation of the observations information model as a relational database schema. This implementation, which we will call ODM 2.0, will be generic for use within any RDBMS, will build upon ODM Version 1.1 (Horsburgh et al., 2008), and will be tested using our data use cases within the data publication prototype deployments described below. The relational database approach, has several advantages: 1) relational databases are easy to deploy using mature and robust commercial and open source RDBMS; 2) our experience with ODM 1.1 has shown that domain scientists and non-IT experts can understand and work with relational schemas that provide a guide for metadata requirements and help scientists create more fully described datasets; 3) a structured, relational schema simplifies development of related web services and data management software; and 4) the cyberinfrastructures within which we will develop our prototype implementations (i.e., CUAHSI HIS, EarthChem, CZOData, IOOS) already use RDBMS technology for data storage.

We will document the specifications for the relational schema and create and publish scripts for creating ODM 2.0 databases within Microsoft SQL Server and the open-source PostgreSQL to ensure that implementation is possible within multiple RDBMS on multiple computer platforms. We will adopt the core and extension structure of the observations information model in our storage implementation to create an extensible relational data model that supports *in situ* sensor data, *ex situ* water quality and solid earth geochemistry samples and fractions, and other data types within the defined scope of the information model so that data publishers can implement the core and only those extensions that are needed for their data use cases. We will adopt, where possible, ideas from the Environmental Data Model (Beran et al., 2008), which, after a review of ODM 1.1, suggested potential mechanisms for adding to it improved geospatial support, provenance tracking, data versioning, and extensibility of supported data types.

We will develop software tools to support, aid, and encourage reliance on the information model during data collection and analysis, ensuring that the information model and associated tools enhance scientists' ability to work with the data during analysis, while at the same time capturing metadata critical for publication and reuse. We will modify the existing suite of ODM software tools (e.g., ODM Data Loader (Horsburgh, 2011a), ODM Streaming Data Loader (Horsburgh, 2011b), and ODM Tools (Horsburgh, 2011c)) to ensure that our project team and other data managers can effectively work with the relational schema that we create.

Expected Outcomes: Documentation and specifications for a relational data model for implementation within any RDBMS (ODM 2.0). Scripts for creating ODM 2.0 databases within Microsoft SQL Server and PostgreSQL. Modified versions of the ODM Data Loader, ODM Streaming Data Loader, and ODM Tools.

Task 3.2. Implementing the Information Model for Data Transfer

Challenges: The XML schemas currently being used by geoscience domain cyberinfrastructures to encode observations data for transfer over the Internet are currently inconsistent, as are the web service interfaces used to publish data for both sensor and sample-based data. No cross-community standards have been adopted, and interoperability of data transfer among systems is limited.

Proposed Activities: Zaslavsky will develop an XML schema encoding of the observations information model. While following the information model developed under Task 1.1, the schema will rely on OGC baseline encoding standards, in particular re-using compatible constructs from OGC Geography Markup Language (GML) (Portele, 2007), the O&M XML encoding (Cox, 2011), and WaterML 2 (Open Geospatial Consortium, 2012). In addition, the schema and the developed use cases will be presented to appropriate OGC working groups, in particular the HDWG and the Earth Systems Science Working Group (ESSWG). This will help obtain feedback from an international group of experts in standards development and thus enhance the information model and the data exchange schema. At the same time, it will ensure that the cross-domain use cases developed within this proposal are considered in

refining existing and emerging international standards, widening the impact of the project. The HDWG is discussing a potential new Interoperability Experiment (IE) focused on water quality, including sample data addressed by the U.S. agency-led Water Quality Exchange (EIEN, 2012), to continue a series of IEs focused on improving information exchange for key hydrologic use cases. To date, surface water and groundwater IEs have been completed, and, as a result, a refined version of the WaterML 2 specification has been submitted for consideration as an international standard. Leveraging the OGC IE mechanism ensures a wider examination and testing of the information model developed, in particular as the WaterML 2 specification is extended to accommodate water quality sample data.

Mayorga will develop prototype web service interfaces over the ODM 2.0 relational model, allowing data stored in ODM 2.0 databases to be accessible via web service requests that implement the XML encoding described above for data transfer. Dr. Mayorga has co-led the development and implementation of standards-based web service software for data transfer for NANOOS in coordination with IOOS. The OGC Sensor Observation Service (SOS; Na and Priest, 2007) is the relevant OGC standard web service interface for publishing sensor-based observations, but Dr. Mayorga will use his expertise, previous SOS server code developed by him for NANOOS as open source Python software, and related, third-party open source components, to develop a cross-platform, web service software prototype. This code will be broadly compatible with OGC SOS and related OGC Sensor Web Enablement (SWE) standards, but will enable modifications necessary for the interoperability demonstrations of this project.

Expected Outcomes: An XML schema implementation of the information model for encoding observations for transfer over the Internet. An OGC-compatible web service interface deployment for publishing observational data stored in the relational schema and transmitted using the XML schema.

Task 3.3. Implementing the Information Model for Metadata Cataloging

Challenges: Each of the repositories with which we are working currently organize their data discovery metadata differently and support differing data discovery queries, reflecting different discovery requirements and expectations of respective domains. Additionally, when working across repositories, the granularity of datasets differs. For example the primary dataset granularity of the CUAHSI HIS is a time series, whereas EarthChem is focused on sets of observations derived from individual samples.

Proposed Activities: Zaslavsky will develop a metadata catalog implementation of the information model that addresses the complexities of cross-repository, cross-data type, and cross-domain cataloging of metadata for the purpose of supporting data discovery. The schema of the catalog will be based on an entity/object representation of the observations information model, with development of specific extensions required for cataloging metadata of multiple organizations, data types, and services. For example, the catalog must maintain information about the service from which observations can be downloaded. Datasets of different types from each of our data publication prototype deployments (described below) will be indexed within the metadata catalog prototype to test our catalog design and to provide an opportunity for feedback into the information model development. This work will build upon the existing observations metadata catalog and data discovery web services of the CUAHSI HIS, but will create opportunities for us to explore new, cross-data type and cross-repository data discovery mechanisms that are enabled by the observations information model and prototype catalog that we develop – e.g., discovery by common metadata fields, spatial location, common association with domain features, by observation type, and through full-text metadata search. The catalog prototype will use standard service interfaces such as the OGC Catalog Services for the Web (CSW) (Nebert et al., 2007) and OpenSearch (Clinton, 2012), in addition to the legacy CUAHSI HIS interface.

Expected Outcomes: A catalog database implementation of the observations information model populated with metadata describing the data within our data publication prototype deployments.

Task 3.4. Implementing the Information Model for Data Archival

Challenges: The geoscience cyberinfrastructures that we have built employ varying degrees and methods for data archival. For example, data are shared via the CUAHSI HIS by storing them in a database and implementing the WaterOneFlow web services to publish them on the Internet. While this

makes the data publicly accessible, the CUAHSI HIS in not currently a permanent archival system in the sense that if the service is discontinued, the data that it serves would no longer be available. In addition to providing dynamic, web service-based access to their data, publishers of earth observations also need the ability to deposit finished datasets within permanent data archives where they can be assigned persistent identifiers and become citable in the same way we currently cite journal publications.

Proposed Activities: Horsburgh will develop an implementation of the observations information model designed for archival of earth observations data within systems like DataONE and EarthChem. The basic construct of the archival encoding will consist of two related and self-contained, platform independent files, one containing the observations and their value-level metadata, and the other containing a scientific metadata file describing the contents of the data file. Mechanisms for packaging multiple data files within a single collection will also be considered, including the BagIt File Packaging Format (Boyko, 2009) and the Open Archives Initiative's Object Reuse and Exchange (OAI-ORE) format (Lagoze, 2008). An archival format will enable members of our community to more easily use publication and archival systems such as DataONE to broaden the impact of their data and to ensure its long term preservation and access.

Expected Outcomes: Specifications for an encoding of the information model for use in archiving spatially discrete, feature based earth observations within archival systems such as DataONE.

Objective 4: Deployment of Data Publication Prototypes

Task 4.1. Demonstrating Improved Discovery, Access and Integration of Earth Observations

Challenges: Although we have constrained the scope of our work to spatially discrete, feature-based earth observations, data within this domain are still incredibly diverse and complex. Our data use cases demonstrate this complexity and will serve to adequately test the prototype data storage schema, transfer web services and XML schema, and catalog implementations of the information model.

Proposed Activities: In Year 2, we will hold a data interoperability challenge workshop. At this workshop, project members will "flex" the information model and its prototype storage, transfer, and catalog implementations to explore challenges in deploying the prototypes using our data use cases. For example, we will load multiple datasets into a database that implements the storage schema. We will examine data encoded using the XML schema and evaluate the signatures and methods of the web services. We will load metadata from multiple datasets into the catalog schema to begin exploring challenges and opportunities in enabling cross-data type discovery queries. Our testing and evaluations will be made from the perspective of potential cross-cutting science questions and synthetic analyses that would seek to discover and integrate data across the systems we are working with. This workshop will serve as a near-final step in the design of our information model and prototype implementations, resulting in feedback to tune our design and illustrating the successes and potential shortcomings of our work.

Following the workshop, Aufdenkampe, Lehnert, Mayorga, and Horsburgh will finish loading data from our data use cases into databases that implement the storage schema (Task 3.1) and will publish the databases by deploying instances of the publication web services (Task 3.2) described above. These data publication prototypes, which will be deployed at SWRC, LDEO, UW, and USU, will provide a distributed test bed within which we can further test our information model and demonstrate how diverse data from multiple systems can be shared in a consistent way. Zaslavsky will also finish loading metadata from each of the data publication prototypes into the catalog prototype (Task 3.3) to enable further testing of the catalog implementation. The data publication prototypes will serve as a role model for how our work can be formally adopted within the SOAs of existing geoscience cyberinfrastructures. We will use what we learn from these prototypes to further refine the information model and supporting software.

Expected Outcomes: A project design and prototyping workshop focused on data interoperability challenges. Four deployments of the proposed data publication infrastructure using existing geoscience datasets. Feedback to the development of the information model and its encodings.

6. PROJECT MANAGEMENT

Our team builds on experience with existing cyberinfrastructure projects and forms a new collaboration across these projects from different disciplines. USU will be the lead institution, with Horsburgh serving as the Lead PI. Table 1 lists project participants and their roles and responsibilities. Biweekly teleconferences will be held using low-cost web, audio, and video conferencing, with more regular communication on an as-needed basis. The timeline of proposed activities (Table 2) shows substantial overlap between tasks because of the coordinated co-development and iteration between tasks.

Table 1. Project roles and responsibilities.

Utah State University (USU) – Jeffery Horsburgh and David Tarboton
• Overall project leadership and management. • Lead design of observations information model representing hydrologic observations • Develop hydrologic and water quality data use cases • Lead the design and development of storage and archival encodings and associated tools • Manage and maintain the project open development website and version control system • Develop a data publication prototype for the Little Bear River Experimental Watershed
Columbia University, Lamont-Doherty Earth Observatory (LDEO) – Kerstin Lehnert
• Participate in observations information model design representing geochemical observations • Develop geochemical data use cases • Participate in the design of storage and transfer encodings • Develop a data publication prototype for EarthChem
Stroud Water Research Center (SWRC) – Anthony Aufdenkampe
• Participate in observations information model design representing CZO data types • Participate in the design of storage and transfer encodings • Act as liaison with the CZOData System and CZO PIs • Develop data publication prototype for the Christina River CZO
University of California San Diego, San Diego Supercomputer Center (SDSC) – Ilya Zaslavsky
• Participate in observations information model design from metadata cataloging perspective • Develop use cases for cross-domain metadata catalogs, including indexing and data discovery • Create an XML schema implementing the enhanced information model • Develop a prototype metadata catalog implementation of the information model • Act as liaison to the OGC Hydrology Domain Working Group
University of Washington (UW) – Emilio Mayorga
• Participate in observations information model design representing marine/coastal observations • Investigate NetCDF-CF-OPeNDAP technologies to inform storage and transfer encodings • Develop OGC-compatible service interfaces over the ODM 2.0 relational model

Table 2. Proposed timeline.

Project Activities	**Year 1**				**Year 2**			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
1.1. Development of information model								
1.2. Linking Observations to the Geo-Environment								
2.1. Design and Prototyping Workshops								
2.2. Education and Training								
3.1. Development of Data Storage Schema & Tools								
3.2. Development of Data Transfer Schema & Services								
3.3. Development of Catalog Schema and Prototype								
3.4. Development of Data Archival Schema								
4.1. Deployment of Data Publication Prototypes								

6.1. Metrics for Success and Sustainability

The sustainability and success of our proposed efforts relies on the ability of our project team to design an observations information model that has broad applicability and to demonstrate its usefulness by implementing prototype systems of supporting storage, transfer, catalog, and archival functionalities for selected test datasets taken from existing geoscience cyberinfrastructures. The focus of this project is not to develop a new architecture that must be maintained, but rather to develop a new information model, encodings, and software that bring greater functionality to existing systems. Our project team is uniquely poised to integrate the proposed work into the CUAHSI HIS, EarthChem, and CZOData, and possibly IOOS, where the enhanced data and system interoperability will immediately benefit multiple geoscience communities. These larger, longer-term cyberinfrastructure projects will thus continue to maintain and enhance our products. Additionally, our positioning and timing is perfect to communicate our advances to other large national and international efforts. PI Horsburgh co-leads the DataONE Integration and Semantics Working Group and participates in the Scientific Observations Network (SONet), both of which are seeking a robust new information model for enhanced data discovery and integration. Likewise, Co-PI Zaslavsky co-chairs the OGC HDWG, which has become a primary international standards setting body and which is just beginning to consider the development of data transfer standard encodings and web services for water quality samples. Thus it is likely that our achievements will influence emerging efforts and standards for sharing earth observational data and leave a lasting legacy.

Our project team is committed to an open development model for all aspects of this project. We will use existing, publicly available source code repositories such as the CUAHSI HIS HydroServer CodePlex website for our work. The HydroServer CodePlex website is managed by PI Horsburgh and employs a subversion source code repository, community discussion forums, a formal issue tracker, and facilities for releasing source code, software deployments, and documentation.

7. INTEGRATION WITH EARTHCUBE

While NSF's EarthCube initiative is still in its formative stage, we envision that it will include 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. Data analysis, visualization, modeling, and synthesis will merge to create a system for knowledge exploration and management. For such a system to work, it must reuse data from existing geoscience cyberinfrastructures, and, within EarthCube, data must be represented in an interoperable way. This demands a powerful information model. Indeed, several EarthCube White Papers articulated this need (e.g., Hooper, 2011; Rew, 2011; Hibbard et al., 2011; Horsburgh and Tarboton, 2011; and others). The information model we develop and our prototype physical implementations (e.g., storage and exchange schemas) will advance ideas for how discrete, feature-based earth observations from sensors and samples are represented within EarthCube. Our innovations in adding extensibility, generality, and cross system interoperability to the information model will add to the knowledge base that EarthCube can draw upon. Our information model will serve as a candidate for Earthcube to adopt for this class of data as well as advance concepts that could be beneficial for other classes of data that EarthCube will need to accommodate.

8. BROADER IMPACTS

Our proposed multidisciplinary, community earth observations information model and supporting software infrastructure represent an opportunity to transform the way scientists capture, manage, work with, and share their data. The need is great. Existing geoscience cyberinfrastructures are all limited in their extensibility and interoperability by the lack of a robust, shared observations information model such as we propose. Furthermore, the timing is ripe for our proposed effort to have a lasting legacy in existing and future geoscience cyberinfrastructures. Our work will directly impact data management and sharing with the hydrologic, critical zone, solid earth geochemistry, and ocean science communities, as well as influencing international standards development through our connections with OGC. We believe that well

described and interoperable earth observations will drive new innovations in synthesis and modeling in the geosciences (including water, energy, and material balances; as well as weathering/soil processes) based on extensive data already being collected nationally by organizations like the United States Geological Survey and by academic research scientists at intensive research sites such as CZOs

9. RESULTS FROM PRIOR NSF SUPPORT

Tools for Environmental Observatory Design and Implementation: Sensor Networks, Dynamic Bayesian Nutrient Flux Modeling and Cyberinfrastructure. D.K. Stevens, **J.S. Horsburgh**, **D.G. Tarboton**, and N.O. Mesner, CBET 0610075, $356,000, 11/1/06 – 10/31/09. This project established an environmental observatory test bed in the Little Bear River in Utah, USA. Observational infrastructure established includes continuous discharge, water quality, and weather monitoring sites connected via a real-time communication network (Horsburgh et al., 2010a). Continuous streamflow and water quality data were coupled with periodic and event based water quality samples to estimate fluxes of water quality constituents using turbidity as a surrogate (Spackman Jones et al., 2011) and to estimate the impact of sampling frequency on estimates of water quality constituent loads (Jones et al., 2011). Cyberinfrastructure was also developed and prototyped to support integrated environmental observatory information systems (Horsburgh et al., 2011).

Collaborative Project: Facilities Support: Earthchem: Advancing Data Management in Solid Earth Geochemistry, K.A. Lehnert, EAR-0522195, $ 1,062,311, 9/15/2005 – 9/14/2011. This grant funded the development of EarthChem as a cyberinfrastructure for solid earth geochemical data. Major achievements of the project include: the *development of EarthChemXML* as a standard data transfer protocol for publishing geochemical data; *development of EarthChem Portal*, including tools to access, visualize, and analyze geochemical data from partner databases; *growth of the EarthChem Federation* of partner and affiliate databases that make their data available at the EarthChem Portal, which (as of Fall 2011) is serving >16 million geochemical data values for ca. 690,000 samples; *development of the Geochemical Resource Library* as a repository for geochemical datasets with services for data submission, DOI® registration of datasets via DataCite, and long-term archiving; *development of new data collections* such as the Deep Lithosphere Dataset; and community-based *development of standards for data reporting* through workshops and the Editors Roundtable. Since October 2010, EarthChem is operated as part of the NSF-funded data facility Integrated Earth Data Application (IEDA).

GeoInformatics: CUAHSI Hydrologic Information Systems, D.R. Maidment, **D.G. Tarboton**, **I. Zaslavsky**, J. Goodall, and D.P. Ames, EAR 0413265 and EAR 0622374, $1,156,059, 4/1/04 – 3/1/08 and $4,500,000, 1/15/07 – 1/15/12. The CUAHSI HIS project and cyberinfrastructure serve as a foundation for the work proposed here. In particular, the HIS project developed a SOA for sharing hydrologic data (Horsburgh et al., 2009; 2010b) that employed ODM for data storage (Horsburgh et al., 2008) the WaterML XML schema as a data transfer encoding (Zaslavsky et al., 2007). The CUAHSI HIS has been deployed at WATERS Observatory Test Beds and at numerous other sites. HIS Central catalogs metadata for web-accessible hydrologic time series for nearly 2 million measurement points in the U.S. Additional publications include: Ames et al. (2009), Beran and Piasecki (2008; 2009), Beran et al. (2009), Castronova and Goodall (2010), Goodall et al. (2008), Maidment et al. (2006; 2009), Maidment (2008a; 2008b; 2009), Piasecki and Beran (2009),Tarboton et al. (2009), and Zaslavsky and Maidment (2011).

Christina River Basin Critical Zone Observatory: Spatial and temporal integration of carbon and mineral fluxes: a whole watershed approach to quantifying anthropogenic modification of critical zone carbon sequestration. EAR-0724971 (2009-2014). Lead PIs. D. Sparks & **A.K. Aufdenkampe**. The two-fold goal of the CRB-CZO is to: 1) create a community resource for studying critical zone processes through sampling, sensing and cyber- infrastructure, and 2) test a set of hypotheses on the whole-watershed links between water, mineral and carbon cycles over a range of modern and historical land uses. The large project team includes 4 new post-docs, 11 new graduate students, a full time sensor engineer and 17 faculty at several institutions including **Zaslavsky** at SDSC. Results to date have been presented in six peer-reviewed publications, more than a dozen invited talks and dozens of other presentations.