INTEROP: A Community-driven Scientific Observations Network to achieve Interoperability of Environmental and Ecological Data

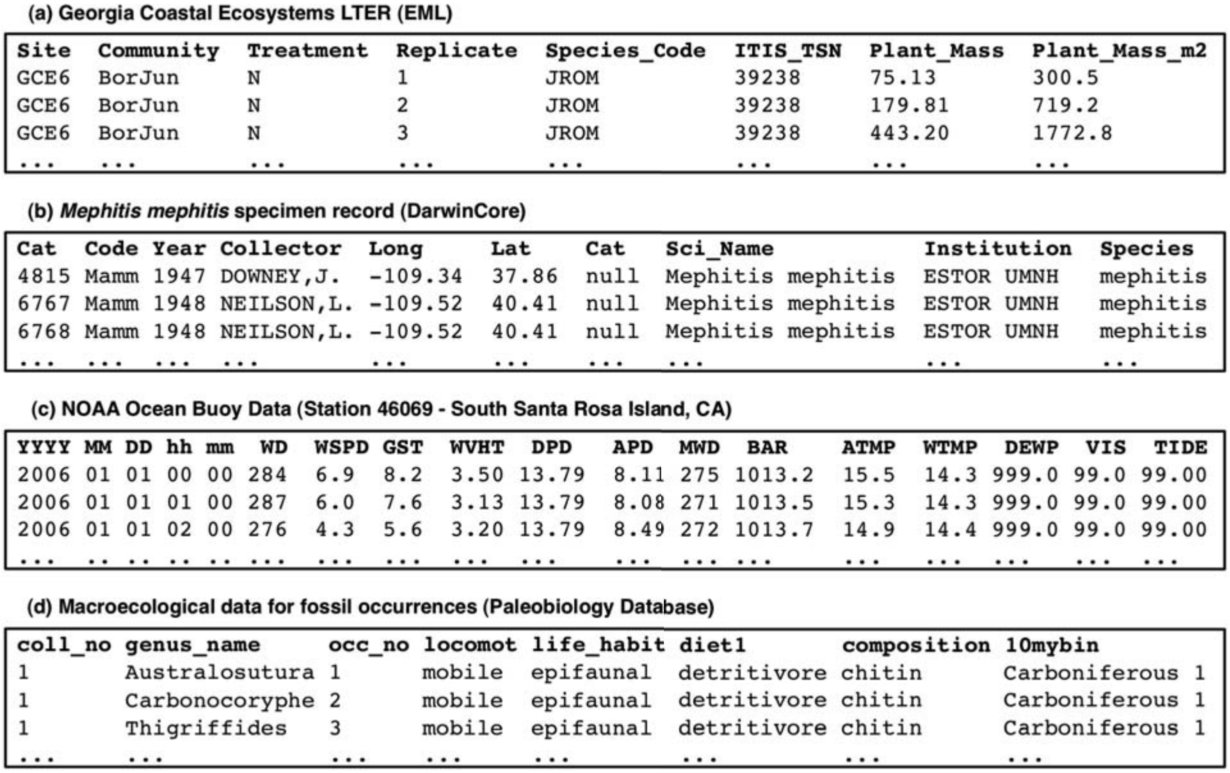
PI: *Mark Schildhauer*, National Center for Ecological Analysis and Synthesis, UC Santa Barbara (*lead organization*). **Co-PIs**: *Shawn Bowers*, UC Davis Genome Center; *Corinna Gries*, Global Institute of Sustainability and CAP/LTER, Arizona State University; *Phillip Dibner*, Open Geospatial Consortium Interoperability Institute; *Deborah McGuinness*, McGuinness Associates. **Senior Personnel**: *Luis Bermudez* and *John Graybeal*, Monterey Bay Aquarium Research Institute; *Josh Madin*, National Center for Ecological Analysis and Synthesis, UC Santa Barbara.

Intellectual Merit. Advances in environmental science increasingly depend on information from multiple disciplines to tackle broader and more complex questions about the natural world. Such advances, however, are hindered by data heterogeneity, which impedes the ability of researchers to discover, interpret, and integrate relevant data that have been collected by others. A recent NSF-funded workshop on multi-disciplinary data management concluded that interoperability can be significantly improved by better describing data at the level of observation and measurement, rather than the traditional focus at the level of the data set. That is, for systems to interoperate effectively, the scientific community must unify the various existing approaches for representing and describing observational data. A community-sanctioned, unified data model for observational data is thus needed to enable interoperability among existing data resources, which will in turn provide the necessary foundation to support cross-disciplinary synthetic research in the environmental sciences. The investigators propose the *Scientific Observations Network* to initiate a multi-disciplinary, community-driven effort to define and develop the necessary specifications and technologies to facilitate semantic interpretation and integration of observational data. The technological approaches will derive from recent advances in knowledge representation that have demonstrated practical utility in enhancing scientific communication and data interoperability within the genomics community. This effort will constitute a community of experts consisting of environmental science researchers, computer scientists, and information managers, to develop open-source, standards-based approaches to the semantic modeling of observational data. Subgroups of Network experts will also engage in extending this core data model to include a broad range of specific measurements collected by the representative set of disciplines, and a series of demonstration projects will illustrate the capabilities of the approaches to confederate data for reuse in broader and unanticipated contexts.

Broader Impacts. There is currently fragmentation among the environmental science subdisciplines, such that each is typically working to meet its own, internal data access and integration needs, without considering how data interoperability could be achieved more broadly through collaboration with researchers and technologists from other fields. By bringing together scientists from representative environmental disciplines, knowledge engineers and conceptual modeling experts, and specialist information managers working within these domains, we hope to initiate a new crosscutting network to derive consensus on technology strategies for achieving data interoperability. This will be accomplished by retaining the momentum of prior NSF-funded activities that have identified a clear path forward for dealing with data interoperability, recommending that the broader community develop and ratify a unified model for scientific observation onto which current and future data models can be superimposed. Key to the success of the proposed network will be outreach to the broader environmental science communities and stakeholders through a number of meetings and community-focused workshops. These activities will directly engage a diverse group of community members, allowing the broader community to contribute requirements and use cases, provide feedback on proposed approaches, and participate in community-building activities (such as ratification of a core data model). Education will also be key to project success and will be supported through a number of activities including student participation in network meetings and a workshop dedicated to training students, postdoctoral scientists, and researchers on the models and approaches developed through the network.

1 Introduction and Network Objectives

Interest in data sharing and interoperability within the environmental research community has grown rapidly in recent years. This increased interest is due in part to the investigation of complex ecological and environmental issues at broad geographic and temporal scales, which typically require the integration of data from multiple research efforts [MBH+97, ABW+04, EOH+06]. These synthetic analyses rely on the effective discovery and processing of data from many independent research projects, i.e., uncoordinated studies that often focus on restricted thematic issues and spatiotemporal scales (Figure 1). These traditionally small, focused studies, however, form invaluable information resources that collectively enable the investigation of important new research questions in complex, biologically diverse communities and ecosystems. Integrated data sets can represent much better measures of potentially critical environmental variables, enhance sample sizes for increased statistical power, and permit examination of issues at broader spatiotemporal scales than any individual study by itself.



(a) Georgia Coastal Ecosystems LTER (EML)

Site	Community	Treatment	Replicate	Species_Code	ITIS_TSN	Plant_Mass	Plant_Mass_m2
GCE6	BorJun	N	1	JROM	39238	75.13	300.5
GCE6	BorJun	N	2	JROM	39238	179.81	719.2
GCE6	BorJun	N	3	JROM	39238	443.20	1772.8
...	...	...	...	...	...	...	...

(b) *Mephitis mephitis* specimen record (DarwinCore)

Cat	Code	Year	Collector	Long	Lat	Cat	Sci_Name	Institution	Species
4815	Mamm	1947	DOWNEY,J.	-109.34	37.86	null	Mephitis mephitis	ESTOR UMNH	mephitis
6767	Mamm	1948	NEILSON,L.	-109.52	40.41	null	Mephitis mephitis	ESTOR UMNH	mephitis
6768	Mamm	1948	NEILSON,L.	-109.52	40.41	null	Mephitis mephitis	ESTOR UMNH	mephitis
...	...	...	...	...	...	...	...	...	...

(c) NOAA Ocean Buoy Data (Station 46069 - South Santa Rosa Island, CA)

YYYY	MM	DD	hh	mm	WD	WSPD	GST	WVHT	DPD	APD	MWD	BAR	ATMP	WTMP	DEWP	VIS	TIDE
2006	01	01	00	00	284	6.9	8.2	3.50	13.79	8.11	275	1013.2	15.5	14.3	999.0	99.0	99.00
2006	01	01	01	00	287	6.0	7.6	3.13	13.79	8.08	271	1013.5	15.3	14.3	999.0	99.0	99.00
2006	01	01	02	00	276	4.3	5.6	3.20	13.79	8.49	272	1013.7	14.9	14.4	999.0	99.0	99.00
...	..	..	..	..	...	...	...	...	...	...	...	...	...	...	...	...	...

(d) Macroecological data for fossil occurrences (Paleobiology Database)

coll_no	genus_name	occ_no	locomot	life_habit	diet1	composition	10mybin
1	Australosutura	1	mobile	epifaunal	detritivore	chitin	Carboniferous 1
1	Carbonocoryphe	2	mobile	epifaunal	detritivore	chitin	Carboniferous 1
1	Thigriffides	3	mobile	epifaunal	detritivore	chitin	Carboniferous 1
...	...	...	...	...	...	...	...

Figure 1. Examples of observational data sets of potential use in larger synthetic studies: (a) data from a large-scale LTER experiment examining the effects of increased nitrogen on terrestrial communities; (b) specimen collection records for the skunk, *Mephitis mephitis*, with location information; (c) ocean buoy data for the Santa Barbara Channel, representing a number of different physical observations; and (d) macro-ecological data about trilobite genera from the Carboniferous Period, with accompanying trait (descriptive) information.

It is a major challenge, however, to effectively discover and integrate environmental data, due to the extremely broad range of data types, structures, and semantic concepts used, and the relatively few standardized methodologies to constrain the various ways in which these data are collected (i.e., data heterogeneity). Moreover, environmental data are widely distributed, with few well-established repositories or standard protocols for archival and retrieval. These factors make the discovery, access, and integration of such data sets, and other potentially relevant supportive data, a labor-intensive task. Solutions requiring a single, one-size-fits-all database schema to represent the broad array of environmental data are generally accepted as impractical, if not impossible [JSR+06]. Regardless of underlying technology (e.g., relational, object-oriented, or hierarchical), such schemas cannot capture the variety of structures and concepts routinely found in the data of these large and diverse research communities.

Metadata standards such as the Ecological Metadata Language [EML], Darwin Core [DwC], and the Geography Markup Language [GML], among others, represent important first steps for improving our ability to discover and access environmental data, but are generally limited in terms of their ability to provide detailed descriptions of data content, which is typically expressed using simple keywords or plain-text metadata fields. More flexible and powerful mechanisms for capturing the semantic richness of data are needed—including structured, semantic descriptions of data variables and their inter-relationships within a data set—to develop a coherent system for effectively managing environmental information, supporting data discovery, and automating data integration.

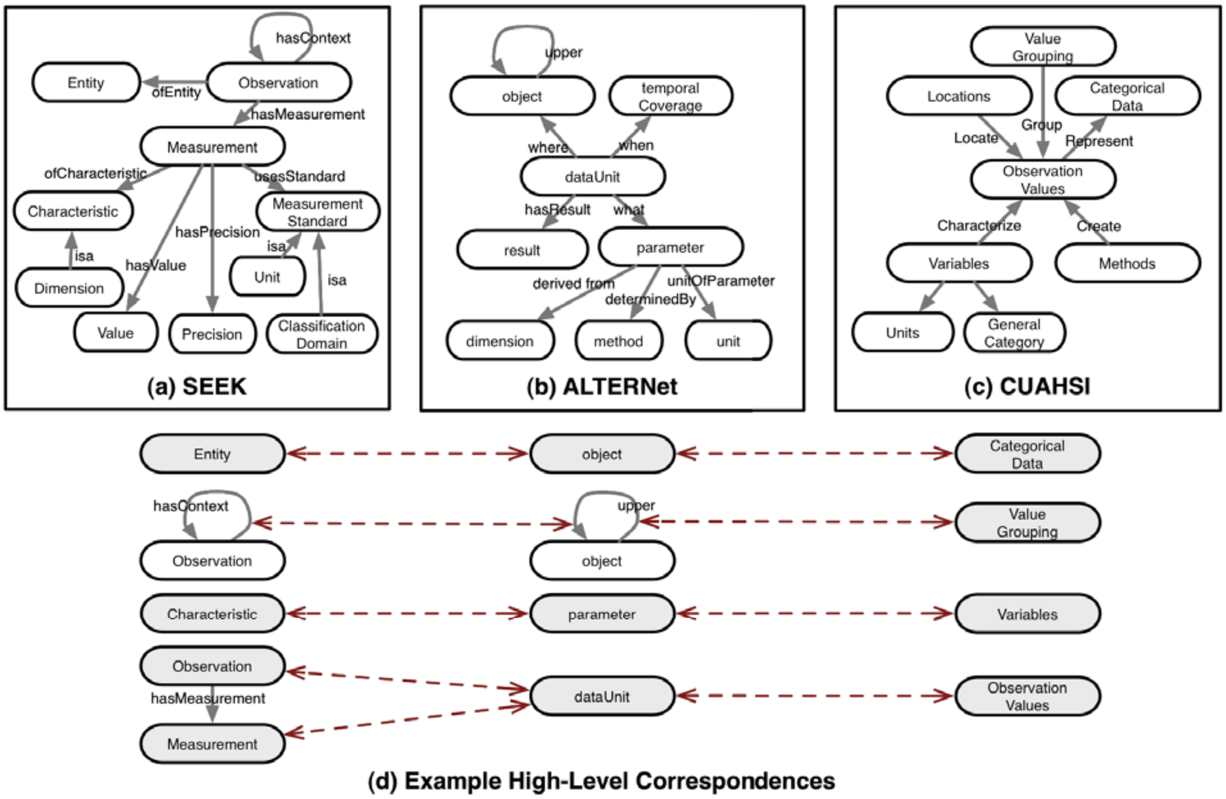
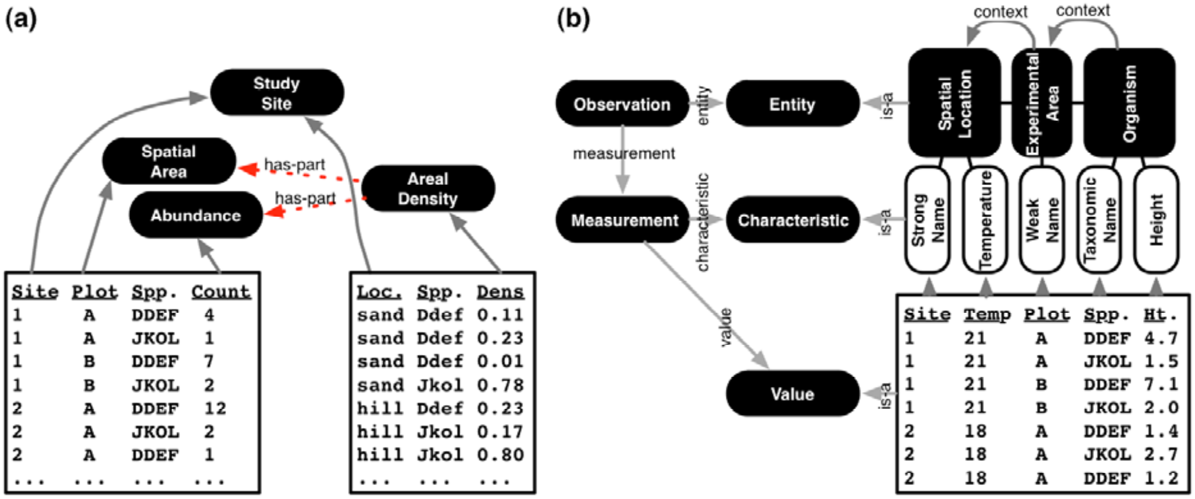


Figure 2. Basic representation of core concepts and relationships in the (a) SEEK [MBS+07], (b) ALTERNeT [SM03], and (c) CUAHSI [THM07] models for observational data. Ellipses represent concepts and grey arrows represent relationships (ontology properties) between classes. Panel (d) illustrates several high-level correspondences among the models (dashed arrows).

A number of recent efforts have adopted alternative approaches for enhancing data interoperability and interpretation using more general conceptual models based on *scientific observations* (Figure 2). Ideally, such a model could flexibly represent any type of measurement that might be found in a research context, while allowing specialized interpretations and semantic subtleties to be captured as disciplinary extensions of the general framework. One major approach towards this goal involves the use of formal ontologies to "superimpose" semantic interpretations onto data sets (Figure 3) [JSR+06, MBS+07, LLB+03, LLB+06]. This approach provides the flexibility for multiple interpretations of an observation, represented in disciplinary-specific terminologies, and permits flexible restructuring of data as specialized needs arise. Such ontological approaches to data description offer a number of advantages over more traditional representations, e.g., relational schemas—where data tends to be tightly constrained by both the modeling language and the typical storage framework [SMJ02].

Approaches built around the notion of scientific observation include, to varying levels of detail, representations of measurements, units, measured traits, measured values, observed entities, and so on. Each approach, however, proposes a slightly different model of these "core" concepts, reflecting the specific needs of their particular domain (see Table 1). Moreover, each new approach, instead of adopting a previous model, typically creates an entirely new representation from scratch. Standardizing upon a ***core conceptual data model*** for representing scientific observations would greatly benefit these existing and future knowledge representation efforts.[1] In particular, a standard ontology-based model, akin to the various ontology standardization efforts in bioinformatics (e.g, the Gene Ontology [GO00, ABB+00, BR04]) and biomedicine (e.g., the Unified Medical Language System [HL93, BDA+03]), would provide a common basis for developing, extending, and applying the highly specialized terminologies required for describing data relevant for environmental research. A community must be organized around this effort to reduce the "babel" of scientific dialects that currently impede effective data integration. Indeed, the risk of multiple, non-interoperable solutions are likely if coordination and communication are not achieved early on in these knowledge formalization efforts.



Site	Plot	Spp.	Count
1	A	DDEF	4
1	A	JKOL	1
1	B	DDEF	7
1	B	JKOL	2
2	A	DDEF	12
2	A	JKOL	2
2	A	DDEF	1
...	...	...	...

Loc.	Spp.	Dens
sand	Ddef	0.11
sand	Ddef	0.23
sand	Ddef	0.01
sand	Jkol	0.78
hill	Ddef	0.23
hill	Jkol	0.17
hill	Jkol	0.80
...	...	...

Site	Temp	Plot	Spp.	Ht.
1	21	A	DDEF	4.7
1	21	A	JKOL	1.5
1	21	B	DDEF	7.1
1	21	B	JKOL	2.0
2	18	A	DDEF	1.4
2	18	A	JKOL	2.7
2	18	A	DDEF	1.2

Figure 3. Simplified illustrations of "superimposing" (semantically annotating) data sets (boxes) with ontology concepts (ellipses). (a) Although column names differ in two data sets, they map to the same concept (e.g., "Site" and "Loc." attributes are both StudySites). Further, a single attribute in one data set may conceptually relate to multiple attributes in a second data set. For example, a "Plot", which implicitly has an area, and a "Count" together suggest "Areal Density", thus making the two data sets semantically compatible. (b) The SEEK ontology prescribes a more structured approach (i.e., conceptual data model; see Figure 2) for describing observational data, which includes the "Entity" observed, the "Characteristic" measured, the resulting "Value", the "Measurement Standard" or "Physical Unit" used (not shown), as well as contextual relationships among atomic observations that make up the data set [MBS+07].

[1] We use the terms 'conceptual model' and 'data model' interchangeably to generally denote a set of intuitive information modeling constructs and associated operations for describing and accessing information via the model.

Network Objectives

We propose to address interoperability issues in the environmental sciences by building a network of practitioners, the ***Scientific Observations Network*** (hereafter, shortened to the *Network*), which will initially include over two dozen key environmental researchers, computer science experts in knowledge representation and conceptual modeling of data, and scientifically-trained information managers, working together to build generic, cross-disciplinary interoperability solutions for scientific data. The primary goal will be to advance the interoperability of data in the environmental sciences by (1) developing a core data model to unify the burgeoning number of domain-specific models for observational data, within (2) a semantic framework that allows for open-ended, but rigorous descriptions of the details and nuances of scientific terminology. This will be accomplished via (3) the coordinated development and eventual ratification (via an international standards body) of a core data model for observation and measurement based on open-standards for data exchange over the Internet, and then (4) developing discipline-specific extensions for this core model. The Network will also (5) develop prototype software applications to demonstrate the utility of these approaches with respect to data interoperability within and across environmental science disciplines. A final directive for the network will be to envision and propose a mechanism for sustaining these community efforts beyond the duration of the proposed project.

Table 1. Representative efforts for modeling observations within the environmental sciences.

Organization	**Short description of observational data modeling approach**
SEEK	The SEEK extensible observations ontology (OBOE) focuses on capturing the essential information about observations required to comprehensively discover and integrate heterogeneous ecological data. [MBS+07, OBOE, WMG06]
NatureServe	The NatureServe Observational Data Standard focuses on developing an XML Schema for specimen-oriented survey data to improve data aggregation and sharing within and between organizations. [ODS]
ALTER-NeT	The European ALTER-NeT Ontology, CEDEX, focuses on developing an object-oriented data system for cataloguing observational ecological data while retaining semantic information to aid data discovery and analysis. [SM03, SSM05]
SPIRE	The Spire initiative focuses on developing domain-independent, general-purpose ontologies to enable annotation of the contents and structure of existing ecological databases with an initial focus on taxonomy and food web issues (ETHAN). [P07]
OGC	The OGC Observation and Measurement Standard focuses on developing a generic framework for representing all aspects of observation and measurement data. [C06]
VSTO	The Virtual Solar-Terrestrial Observatory focuses on building ontologies for interoperating among different existing meteorological and atmospheric metadata standards. [MFC+07]. VSTO also incorporates the SWEET ontologies [R04, SWEET]
TDWG	TDWG is developing a "meta-model"to integrate biodiversity observations with specimen data by identifying similarities between these two data types, determining whether existing standards suffice to describe them, and if not, developing the additional concepts needed for clarification. [TAG]
ODM	The CUAHSI Observations Data Model and associated relational database focus on storing hydrologic observations data in a system designed to optimize data retrieval for integrated analysis of information collected by multiple investigators. [THM07]

This Scientific Observations Network will represent the first step towards building a community of multi-disciplinary scientific practitioners, dedicated to extending the interoperability of data across their individual domains using advanced knowledge representation methods. The vision is that this community will grow to support additional participants representing broader environmental and earth science areas, as the benefits of this type of technical collaboration become apparent. The Network represents an opportunity to begin constructing true, crosscutting technology solutions in the service of data interoperability within the environmental sciences.

The remainder of this proposal is organized as follows. Section 2 further describes the background and rationale of the Network. Section 3 describes the proposed organizational structure and activities of the Network. Section 4 describes the education, outreach, and training activities of the Network, which will be key to Network success. Finally, Section 5 describes the management plan for the Network, and sustainability issues beyond the duration of the project. Additional material related to key personnel, current activities, and letters from key stakeholder groups are provided as supplementary documentation.

2 Background and Rationale

Observational data are defined broadly as the outcomes of acts of measurement using particular protocols within the context of any objective and potentially replicable activity. Examples include data from survey or monitoring efforts, controlled experiments, and sensor-derived measurements. In each case, the basic or atomic notion of an observation represents the outcome of some measurement taken of a defined attribute or characteristic of some "entity" (e.g., an organism "in the field", a specimen, a sample, an experimental treatment, etc.), within some context (possibly given by other observations). Every observation entails the measurement of one or more discrete properties of some real-world entity or phenomenon. As a result, a data model that focuses on the structure of observations can richly model the fundamental semantics of the scientific measurements that are being made. For example, a fundamental model of observations ideally allows one to discover data based on the entities that were observed and on the context in which an observation was made [MBS+07]. In addition, such a model would enable the construction of a variety of data integration services that can mediate the wide range of differences among observational data that might be relevant to a particular study.

Traditional approaches used in environmental science have focused on entire data sets as the fundamental unit to be managed. Data sets are constructed for a number of reasons. For example, they typically encapsulate some of the semantic relationships among observations, such as a shared set of methods, nesting within spatial or temporal hierarchies, and participation in a shared experimental design. Thus data sets can be seen as a mechanism for "optimizing" the storage of metadata, in which associated observations are assumed to uniformly "inherit" the metadata provided at the data-set level. This metadata is often not explicit, however, and not easily inferred from the data alone, which is compounded by the fact that data sets are often structured for specialized purposes, e.g., for use in a particular analysis or to simplify the process of collecting data. Any new model of scientific observations should preserve this fundamental aspect of modeling data sets—that there are collections of observations that must be packaged together to be understood and used within their proper context.

Many groups have realized, however, that the data-set model and traditional metadata approaches have significant limitations, which are mainly related to their inflexibility with respect to packaging observations in new combinations for new scientific purposes. These limitations become obvious as the scale of integration among studies increases. Synthetic studies frequently involve the collation and subsetting of observations from many diverse source data sets into derived data sets. Consequently, it becomes increasingly difficult to track the provenance of individual data points within derived data sets; and the need to model the more atomic observations within these various data collations becomes critical to validate their appropriate application from a scientific perspective.

Ontologies are one enabling mechanism for providing more comprehensive data description, discovery, and integration [JSR+06, MFC+07, R04]. Ontologies can offer formal representations of domain knowledge via the terminology (concepts) used within the domain and the properties and relationships among concepts [BCM+03]. Although ontologies and conceptual modeling have been used for some time, it is the recent emergence of the Web as a standard conduit for information exchange that has generated excitement over the potential of ontologies for "layering" an enriched semantics on top of Web content (including scientific data). The promise of the Semantic Web [BHL01] is based on the notion that a standardized syntax for expressing ontologies can be promoted throughout the Web, such that generalized applications can be built that are capable of parsing and reasoning with these ontologically-enriched documents.

Several research efforts are currently investigating the utility of W3C standard ontological approaches for enriching the semantics of scientific data (e.g., SEEK, SPIRE, GEON, VSTO). These efforts are converging on the use of OWL (the "Web Ontology Language") [MvH04], which is based on description logics [BCM+03] and supports a "natural" representation for formalizing terms. However, while formal languages such as OWL provide a means to capture ontologies, the quality of the realized ontology will determine its utility for assisting in data interoperability. Additionally, as the number of ontologies and their included terms increase, organizing these into a coherent framework becomes increasingly complex, as recognized within the biological community [BR04, SK05]. Thus, a standard, well-defined core model for describing scientific observations that is expressed in OWL can provide a number of benefits to current and future projects, including: (1) the ability to adopt a community-driven core ontology, allowing efforts to be focused on developing high-quality and relevant domain-ontology extensions, which can lead to improved data discovery and integration; (2) a common data model for facilitating the interchange of observational data, providing greater levels of system interoperability; and (3) an open, non-proprietary approach based on Semantic Web standards, providing a variety of freely-available tools for OWL (e.g., Protégé, SWOOP, Jena, and Pellet, among others).

Results from Prior NSF Support

Investigators on this project have been instrumental in a number of prior NSF-funded projects directly related to the proposed Scientific Observations Network. This work includes hosting and organizing community workshops on data interoperability, the development of standard metadata languages and ontologies, and software tools that use these standards for managing environmental data. Below we briefly highlight relevant results from these projects.

A Workshop for Advancing a Unified Model for Observational Data in the Ecological and Environmental Sciences. $50K, **Schildhauer**, Jones, **Madin** (UCSB), **Bowers** (UCD), Kelling (Cornell), Sugarbaker (NatureServe), held July 9-11, 2007, NSF Award #0733489.

The Scientific Observations Network proposed here will directly build upon this recently held, NSF-funded workshop to discuss the various data models and ontologies used within the environmental sciences for managing observational data. Workshop participants included over twenty-five researchers, informatics specialists, and computer scientists representing various environmental-science disciplines, projects, and organizations. The models discussed by workshop participants included SEEK's Extensible Observation Ontology [MBS+07, OBOE], SPiRE's Evolutionary Trees and Natural History Ontology [P07], NASA's Semantic Web for Earth and Environmental Terminology [R04, SWEET], CUASHI's Observations Data Model

[THM07], LTER-Europe's Classes for Environmental Data Exchange [SM03, SSM05], NatureServe's Observation Data Standard [ODS], OGC's model for Observations and Measurements [C06], the TDWG Ontology [TAG], and the ontologies developed as part of the VSTO project [MFC+07], among others (see Table 1). A primary goal of this workshop was to determine whether there was sufficient accord among the various conceptualizations of "scientific observations" across groups to relate and unify these various models

As part of the workshop, participants determined areas of overlap among existing observation models, discussed requirements of a common data model, and defined core data-interoperability capabilities enabled by a common model. Broad consensus was reached among workshop participants that defining a unified core obesrvation model was both possible and could offer significant benefits for data interoperability, including:

- **Improved reusability:** Projects can directly adopt and leverage a core observations data model instead of developing their own, effectively *ad hoc* approaches. Reuse of the core model will in turn lead to greater opportunities for interoperability between current and future projects and systems.
- **Structured Approaches for Extensibility**. Instead of developing a monolithic approach to capture all facets of environmental data, the core model can narrowly focus on the fundamental aspects of representing scientific observations, allowing specific projects to extend the model to address their particular application requirements (e.g., by adding new facets of measurements and enumerating the various domain-specific types of entities and characteristics being observed), while still maintaining interoperability.
- **Implementation-Independent, Open Standards**. Different projects require and use different implementation architectures and technologies. A core data model based on open, technology-independent languages (e.g., OWL and its corresponding XML serialization syntax) can ideally enable well-defined and non-proprietary interchange of information between disparate systems without dictating the technology that projects must use.
- **Enhanced Capabilities for Discovery and Integration**. Current approaches for representing scientific observations are driven by specific use cases and application functions. A core model driven by a well-defined set of generic operations over observation data can both help to support these specific uses and lead to improved interoperability and more robust data-management support. Participants identified a number of enabling operations, which include both ontology-based data discovery and a variety of integration functions such as unit conversion, context resolution, statistical summarization, and so on.

The Scientific Observations Network will continue and further the momentum of this workshop by directly engaging with these and other efforts within the environmental science communities for managing observational data, and will directly build upon workshop recommendations and outcomes to provide wide-scale and needed data interoperability support.

ITR: Enabling the Science Environment for Ecological Knowledge (SEEK), $12.2M (collaborative), 10/1/02–9/30/07, Michener (0225665, UNM), Reichman, Jones, **Schildhauer** (0225676, UCSB), Ludäscher, Rajasekar (0225674, UCD/UCSD), Beach (0225635, U Kansas). The SEEK project has produced numerous software products (e.g., Kepler, EarthGrid, and OBOE) and computer science advances that underlie many of the approaches proposed by this Network.

OBOE (the Extensible Observation Ontology) is an ontology that is designed to provide a linkage between concepts drawn from science domains (e.g. terms such as "biomass" or "population density"), and observations contained within scientific data sets—whether these refer to the data object as a whole, or some component of the data, such as an attribute (column) or cell (individual value) [JSR+07, BBJ+05, MBS+07]. OBOE is a formal ontology expressed using OWL and was influenced by existing ontology approaches such as SWEET [R04] and DOLCE [GGM+02], and controlled vocabularies such as the CSA/NBII Biocomplexity Thesaurus [CSA]. OBOE is designed to capture the semantics of scientific data, including observation and observation context, sampling hierarchies, and complex units. The basis of OBOE is the formalization of Observation, which defines an event in which a real-world Entity is observed by taking one or more Measurements of its particular Characteristics (Figures 2a and 3b). For quantitative measurements, OBOE also describes the Unit system that is used and the Precision of the measurement. OBOE is designed to be easily extended through specialized domain ontologies, and provides structured and explicit extension points where new domain ontologies can be added. These include new domain-specific Characteristics, Entities, Context relations between observations, and Units and classification Measurement Standards.

KDI: A Knowledge Network for Biocomplexity: Building and Evaluating a Metadata-based Framework for Integrating Heterogeneous Scientific Data (KNB), $3M, 9/15/1999-2/28/2005, Reichman, Jones, **Schildhauer**, Brunt, Willig, and Helly, Award #9980154. The KNB project produced a number of ecological data management applications, including EML, the Morpho metadata editor, and the Metacat metadata repository system.

The **Ecological Metadata Language** (EML) is a community-developed metadata content standard for describing the physical and logical structure of scientific data and the context in which data were collected [JBB+01, MJ02, EML]. EML has surfaced as the *de facto* standard for documenting ecological data at hundreds of field stations, including the Long Term Ecological Research Network (LTER) in North America, the Organization of Biological Field Stations, Kruger National Park in South Africa, the Partnership for the Interdisciplinary Study of Coastal Oceans (PISCO), and many others. EML describes structural aspects of a data set, such as its physical format and logical model, the spatial, taxonomic, and temporal extent of the data, and descriptions of experimental design and methods. Like many other metadata standards, EML expresses data semantics mostly in terms of natural language descriptions thus limiting its applicability for machine-based reasoning that relies of ontologies. Software tools, including Morpho, Metacat, and more recently the Kepler scientific worklow system, support the EML standard for describing, storing, searching, and retrieving environmental-science data.

NSF OCE: Collaborative Research: Marine Metadata Initiative. $1.1M, 7/1/2006-6/30/2009, **Graybeal** (PI), Chavez, Wright, Award #0607372. This project continues and enhances the MMI (Marine Metadata Interoperability) project. The MMI project [GB06, BGA06, BBB+06, WWG+05, GBB05] is focused on (1) addressing specific interoperability needs of MBARI marine scientists, (2) educating marine scientists about direct benefits and tools available to them for improved collection and organization of metadata, and (3) collaborating on the creation of reference metadata ontologies and software implementations that demonstrate how metadata techniques can help investigators in their daily research. This project includes the Coastal Atlas Interoperability Workshop, training 30 attendees in ontologies and controlled vocabularies; the Sensor Metadata Interoperability workshop; the OOSTethys interoperability demonstration; and the OGC Oceans Interoperability Experiment. Project participants are currently developing an infrastructure to support semantic interoperability, using ontologies and services to be provided by MMI. Dr. Luis Bermudez, senior personnel on this proposal, is also the technical lead on the Marine Metadata Initiative.

SCI: SEI +II: Towards a Virtual Solar-Terrestrial Observatory. $1.1M, 10/1/2004-9/30/2007, Fox (PI), Solomon, Middleton, **McGuinness**, Award #0431153. The VSTO project [MFC+07] is developing technology for enabling scientific virtual observatories, with an emphasis on providing semantic data integration tools and services for accessing diverse observational data on solar atmospheric physics and terrestrial middle and upper atmospheric physics, theoretical models, and analysis programs. A number of OWL-based ontologies have been developed through the VSTO project (including the VSTO ontology) to support these aims, as well as services utilizing semantic-web technologies for searching and accessing data based on parameter, date-time, and instrument semantics.

3 The Scientific Observations Network

The Scientific Observations Network will consist of environmental scientists, computer scientists and information technologists, collaborating to develop specifications and tools for multidisciplinary data interoperability, based on the semantic modeling of observational data. As shown in Figure 4, the Network will consist of four subgroups. Each subgroup will focus on meeting the needs of distinct user groups, including data providers, data consumers, information managers, and informatics tool developers. Subgroup 4 will ensure that the overall data interoperability goals of the Network are met and will demonstrate utility of the approaches through a set of prototype applications. The members of each subgroup will consist of two project co-leaders and about a half-dozen working-group participants. The subgroups will work closely with each other and with members from the broader informatics and scientific communities through a series of community workshops (Figure 5). The co-leaders and project PI will serve as an initial executive committee for the overall Network, to maintain consistency of vision, strategically adjust project priorities, and assist with overall coordination of the effort.

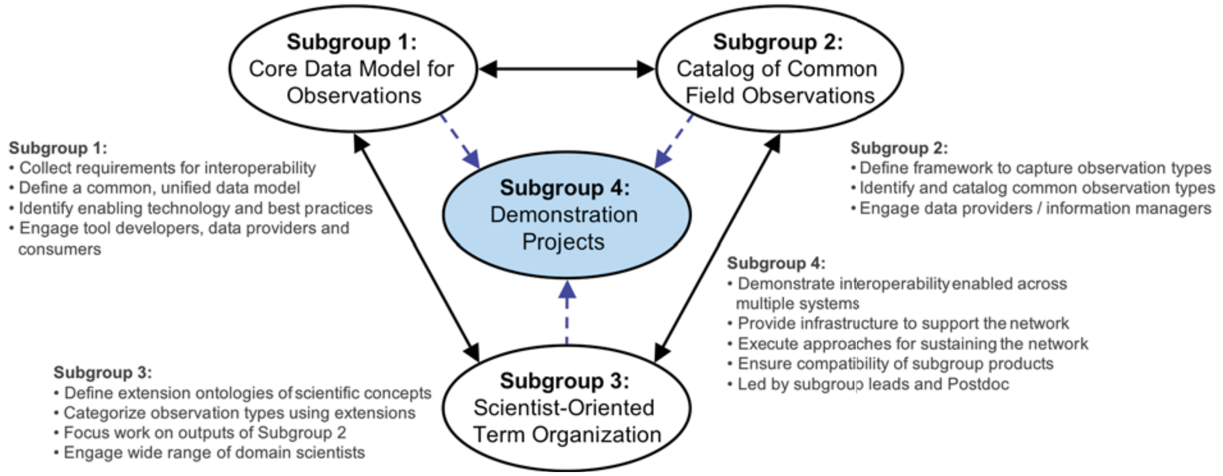

Subgroup 1:
Core Data Model for Observations
Subgroup 2:
Catalog of Common Field Observations
Subgroup 1:
• Collect requirements for interoperability
• Define a common, unified data model
• Identify enabling technology and best practices
• Engage tool developers, data providers and consumers
Subgroup 4:
Demonstration Projects
Subgroup 2:
• Define framework to capture observation types
• Identify and catalog common observation types
• Engage data providers / information managers
Subgroup 4:
• Demonstrate interoperability enabled across multiple systems
• Provide infrastructure to support the network
• Execute approaches for sustaining the network
• Ensure compatibility of subgroup products
• Led by subgroup leads and Postdoc
Subgroup 3:
• Define extension ontologies of scientific concepts
• Categorize observation types using extensions
• Focus work on outputs of Subgroup 2
• Engage wide range of domain scientists
Subgroup 3:
Scientist-Oriented Term Organization

Figure 4. The primary subgroups of the Scientific Observations Network. Subgroups 1-3 will be responsible for: developing the core data model; developing a broad catalog of terms used to describe observations in multiple scientific disciplines; and extending the core model and catalog terms with scientifically meaningful concept hierarchies and interrelationships; respectively. Subgroup 4 will be responsible for ensuring compatibility among subgroups and developing prototype projects to demonstrate data-interoperability solutions enabled by subgroup products.

Network Organization

Subgroup 1: Core Data Model for Observations. The primary goal of this subgroup is to define a common, unified, and extensible model for representing scientific observations and measurements. This model will serve as the basic framework for enabling data interoperability across data repositories and systems. The first two tasks of this subgroup will be (1) to compare and contrast the observation models used in current systems, and (2) to collect requirements for a common observations data model. Thus, this subgroup will directly build on the outcomes of the NSF observation workshop described in Section 2. This subgroup will then use this information to define a formal, shared model for representing scientific observations and measurements. Working closely with Subgroup 4, Subgroup 1 will implement the model (e.g., using OWL), identify enabling technologies in support of the model, and define best practices for using the model to represent and exchange observational data. The participants of this subgroup will include key members of projects developing systems that explicitly provide a data model for observations and measurements (including those identified above), and involve information managers, computers scientists, and integrative domain scientists who rely on a broad range of data.

Primary Deliverables: *(1) An analysis of approaches employed by existing informatics systems for representing and describing scientific observation and measurement data; (2) A formal and community-driven core data model for capturing scientific observations and measurements that is both extensible and can be used to enable data interoperability among existing and future systems; and (3) A concrete representation (e.g., in OWL) of the core data model that can be used by applications and for data exchange.*

Subgroup 2: Catalog of Common Field Observations. The goals of this subgroup are (1) to identify a diverse corpus of observation and measurement types used in existing data sets and repositories, and (2) to create a catalog of these types using the core observations data model defined by Subgroup 1. Initial work will focus on collecting a wide range of examples and use cases for Subgroups 1 and 3 (see below). These use cases will span multiple scientific disciplines and data repositories. The subgroup will then use the core observations model produced by Subgroup 1 to define an open and shared catalog of common observation types. This catalog will serve both as a set of examples for the demonstration projects of Subgroup 4 (see below) and will be made available through the Network's website for use by other projects (e.g., by providing a set of "standard" observation types that can be reused for describing and documenting data). The participants of this subgroup will include a variety of domain specialists and information managers who have extensive experience working with specific but broadly useful types of environmental data.

Primary Deliverables: *(1) An analysis of the actual types of observation and measurement used across a variety of scientific disciplines and repositories; and (2) A representative and diverse catalog of observation and measurement types expressed according to the core observation data model that will be published through the Network's website.*

Subgroup 3: Scientist-Oriented Term Organization. The goal of this subgroup is to define ontology extensions to the core data model for organizing observations according to scientifically meaningful concepts (e.g., ecological community, sedimentation, biological invasion, etc.). This subgroup will initially focus on developing relevant concept hierarchies for classifying the specific observation types identified by Subgroup 2. The ontologies created by this subgroup will be designed with particular demonstration applications in mind. Examples include providing (1) meaningful navigation hierarchies to scientists for finding and organizing relevant observation data, (2) improved data discovery and search capabilities for managers and interested public, and (3) unified terminologies and semantic relationships across disciplines for use in data integration. Another major task of this subgroup will be to compare and contrast existing ontology approaches for organizing scientific concepts (e.g., SWEET [R04], OBOE [MBS+07], ETHAN [P07], CEDEX [SM03], and the VSTO ontology [MFC+07], among others) and for representing abstract knowledge (e.g., DOLCE [GGM+02], SUO [PN02], [S99]), with the goal of creating more flexible approaches that can accommodate a wider range of scientific observations. The participants of this subgroup will include scientists focused on the philosophical and theoretical foundations of their domain, computer scientists with experience in knowledge representation and reasoning, and domain scientists with experience in using diverse data.

Primary Deliverables: *(1) An analysis of existing approaches for organizing and classifying scien-*

tific concepts; (2) A set of extension ontologies for the core observations data model that can provide a flexible, useful, and scientifically-meaningful basis for the organization of observations and measurements; and (3) Categorization of the observation and measurement types identified by Subgroup 2 according to the extension ontologies.

Subgroup 4: Demonstration Projects. This subgroup has a broader charge than the others, and its members will include each of the subgroup co-leaders as well as the Network postdoctoral scientist. Additional participants will be added to ensure that adequate expertise is present relative to informatics tool development, information management, computer science, and the targeted areas within environmental science. The major goals of this subgroup are (1) to ensure that the approaches developed by Subgroups 1-3 are compatible, (2) to define and implement a series of prototype tools based on the products of the subgroups, and (3) to identify approaches for sustaining the Network. The purpose of the prototypes is to demonstrate and evaluate the effectiveness of the subgroup approaches and to serve as example software applications that can be used directly or extended by other projects. One of the primary demonstration projects will be to employ the core observations data model to enable data interoperability and sharing between two or more existing information frameworks represented by project participants (e.g., see Table 1). The specific series of demonstration applications will be selected based on discussion and input from the broader community (see Figure 4). Other possible prototype applications are:

- Tools for discovering observation and measurement types catalogued by Subgroup 2, including different prototypes for navigating and querying observation types and leveraging the ontology extensions of Subgroup 3.
- Tools for adding new observation and measurement types to the catalog, determining if equivalent observation types are present in the catalog, and notifying users of similar or overlapping observation types. These prototypes will be primarily focused at data providers and information managers.
- Tools for clarifying or creating linkages between the scientific terms developed by Subgroup 3 and data, e.g., to indicate how specific observations inform certain scientific phenomena, to identify which observation types are predictors of, or impact certain scientific phenomena, to clarify the compositional structure of observations (from abstract to concrete; atomic to composite), and so on. These prototypes will primarily assist specialist domain scientists.
- A web-based registry of data sets, where each data set is described in terms of its observation types, leveraging the catalog defined by Subgroup 2. This application will allow projects to easily register data sets, e.g., given by system-specific identifiers or URLs, with a public repository hosted on the Network's website. The search prototypes described above could be deployed here to allow users to discover data sets based on observation types. This prototype will primarily benefit data providers and consumers.
- Development and prototype implementations of a standard API and protocol for binding observations and measurements to the core data model. Through such a protocol, users (typically via client applications) could uniformly access the observations and measurements stored in otherwise heterogeneous and disconnected distributed data repositories. This technology will be similar to the popular Distributed Annotation System (DAS) [DAS], which is supported by hundreds of bioinformatics data sources to provide uniform access to genome and protein sequence annotations, but targeted at environmental-sciences data and annotations. These prototypes will be primarily focused at tool developers and information managers.

We will design all demonstration prototypes for easy adoption and extension by other projects, e.g., by developing reusable services (i.e., web services) and employing open-source licensing. We will also leverage existing applications, e.g., for managing and working with ontologies (e.g., Jena, Protégé, Swoop), when appropriate. Additionally, this subgroup will setup and manage the basic infrastructure to support the Network, including the Network website, Network mailing lists, source-code control for software development, and so on. Finally, this subgroup will identify and implement strategies to ensure sustainability of the Network, e.g., by promoting the community-driven approaches developed by Network subgroups, and by working with existing organizations (e.g., TDWG, OGC, ISO, W3C) to ratify the core observations data model as a community standard.

Primary Deliverables: *(1) Setup and management of supporting Network infrastructure including the public website and project collaboration tools; (2) A series of well-defined demonstration projects and software prototypes that exercise the data interoperability enabled by the deliverables of Subgroups 1-3; and (3) A plan for proceeding with the standardization of the core observations model, and sustaining and supporting Network products and activities, through an established organization and coordination with the broader scientific and informatics communities.*

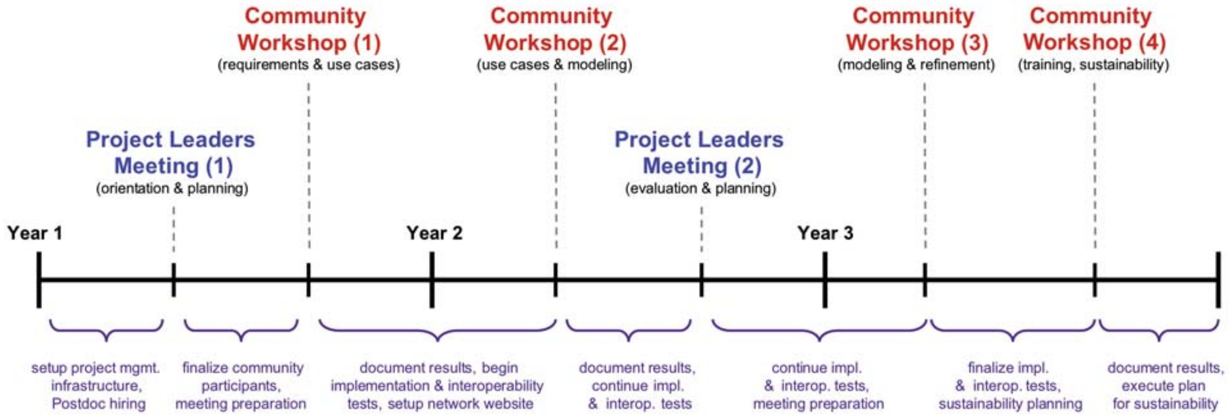

Community Workshop (1)
(requirements & use cases)
Community Workshop (2)
(use cases & modeling)
Community Workshop (3)
(modeling & refinement)
Community Workshop (4)
(training, sustainability)
Project Leaders Meeting (1)
(orientation & planning)
Project Leaders Meeting (2)
(evaluation & planning)
Year 1
Year 2
Year 3
setup project mgmt. infrastructure, Postdoc hiring
finalize community participants, meeting preparation
document results, begin implementation & interoperability tests, setup network website
document results, continue impl. & interop. tests
continue impl. & interop. tests, meeting preparation
finalize impl. & interop. tests, sustainability planning
document results, execute plan for sustainability

Figure 5. Schedule of meetings, workshops, and activities for the duration of the project. Two project leader meetings are planned for project orientation, management, evaluation, and planning. Four larger workshops with invited community members are planned to address the Network objectives. Between meetings, project leaders and postdoctoral researcher will work on Network products and planning.

Network Activities

The overall activities of the Network over the three year duration of this project are shown in Figure 5. To enable broad community engagement in the Network, we will host four separate community workshops over the duration of the project. These workshops will consist of project and community members, including informatics specialists involved in environmental data management, computer scientists, domain scientists, and representatives from the larger environmental informatics communities. The goal of these workshops will be to: (1) collect and analyze requirements for each of the Network subgroups; (2) develop comprehensive use cases related to data interoperability and the core observations data model; (3) actively engage community and project participants in modeling activities and discussion of proposed data models and representations; (4) evaluate and refine the subgroup deliverables; (5) discuss and plan Network sustainability solutions; and (6) engage the community through training and demonstration of Network products. We will provide travel support for twenty-four participants at each community workshop, which will be organized into four parallel Subgroups (described above). The last community workshop, which is scheduled at the end of year 3, will focus specifically on training participants to use the products developed by the Subgroups during the course of the project, and to present and discuss plans for sustaining the Network (see Section 5).

Two additional project-specific working meetings will also be held. The first meeting will occur at the beginning of the project and consist of the initial Network participants. This meeting will primarily focus on Network planning and organization issues surrounding the first two community workshops. The second meeting will be held at the mid-point of the project, and will focus on evaluation of the Network products and effectiveness of community engagement. This meeting will also include planning activities for the remainder of the project.

As shown in Figure 5, project members will perform various activities between meetings, including document results and outcomes from community workshops, develop and refine subgroup deliverables, and engage the broader community on standardization and sustainability planning. This work will be performed with the help of a full time postdoctoral researcher funded through this project, who will work closely with the various subgroup leaders.

4 Broader Impacts

The Scientific Observations Network will bring together a unique mix of experts from a number of traditionally "isolated" sub-disciplines within the environmental sciences, as well as computer science and informatics, collaborating in a focused effort to communicate, develop, and promote the use of standard tools and approaches for unifying observational data and enhancing data interoperability. Although the Network's focus is the environmental sciences, the proposed mechanisms for achieving interoperability will apply more generally to any field in which observational data are collected (e.g., sociology, genomics, epidemiology). Initiating such a broad partnership will be essential in generating a community-sanctioned, "ratified" standard for data interoperability; development of generic software applications and support materials based on these standards; and creation of a framework for effectively promoting and supporting Network efforts beyond the duration of this INTEROP.

While the impacts of the Network are primarily intended to serve the scientific research and information-management communities, we believe these will also democratize access to data, by allowing colleagues at smaller, liberal arts or minority serving institutions, K-12 educators, as well as policy managers and the interested public, to more effectively search for and understand environmental data. Often interested groups are precluded from accessing scientific data due to difficulties navigating domain-specialized terms. An ontological approach will enable searching for data based on more familiar terms. Similarly, beneficial impacts should accrue for educational exercises that are based on finding and analyzing "real" scientific data. Finally, we are sensitive to under-representation issues, and have tried to address these at the PI and senior personnel levels, and will continue to do so when selecting Network participants.

Education, Outreach, and Training

As suggested above, the successful operation of this Network would in itself represent a major outreach accomplishment. The six meetings shown in Figure 5 will be comprised of participants strategically chosen for their skills and interests in the relevant research areas, but also in terms of their broader involvement in initiatives confronting data interoperability challenges (e.g., hydrology [Cleaner/CUAHSI], biodiversity sciences [TDWG and NatureServe], ecology [NCEAS and LTER], aquatic sciences [GLEON], geospatial sciences [OGC], atmospheric sciences [VSTO], and sensors and oceanography [MBARI]). In addition to the diversity of interests represented by the project leaders, attached letters of collaboration indicate the broad-ranging representation expected within this Network.

The project leaders will also present Network progress and approaches at professional societies' annual meetings (e.g., Ecological Society of America, American Geophysical Union, American Society of Limnology and Oceanography, and Biodiversity Information Standards Working Group). Furthermore, included in the travel support for community workshops will be support for three graduate students (one on each of the focused Subgroups outlined in Section 3), and several co-leaders will pursue supplemental funding within the REU program, as the demonstration projects will provide excellent training opportunities. The final community workshop will focus on demonstrations and training. It is expected that 14 students attending this workshop will carry away knowledge that will seed adoption and development of Network approaches.

A postdoctoral researcher will provide strong continuity on the project, while receiving a unique form of training through involvement with a diverse group of scientists. Our vision is that this individual will become a new type of scientist, in the role of domain knowledge engineer, possessing close familiarity with many common types of observational data from the environmental sciences, but also informed about emerging formal approaches using knowledge representation to support data interoperability.

Finally, NCEAS supports 800-1000 visiting scientists per year involved in roughly 40+ Working Groups engaged in synthetic environmental science projects. Outcomes of the Network will become part of standard orientation for NCEAS Working Groups and residents. Moreover, instruction in the availability and capabilities of the Network tools will be provided to participants as a mandatory component of engaging in activities at NCEAS.

5 Network Management Plan

An executive committee chaired by Mark Schildhauer and comprised of the co-PIs (Bowers, Gries, Dibner and McGuinness), senior personnel (Bermudez and Madin) and the postdoctoral researcher will manage the Network. The executive committee possess extensive experience managing cross-institutional research projects such as the Scientific Environment for Ecological Knowledge (SEEK), the Knowledge Network for Biocomplexity (KNB), the Open Geospatial Consortium Interoperability Institute (OGCii), the Virtual Solar-Terrestrial Observatory (VSTO), the Marine Metadata Initiative (MMI), and information management for the Central Arizona Phoenix Urban LTER. The executive committee will convene first to finalize community meeting participant lists, divide participants into subgroups (described in Section 3), and define meeting goals and scheduling (Figure 5). The executive committee will have the primary charge of ensuring overall compatibility and consistency of vision among the subgroup activities.

Two members of the executive committee will lead each Network subgroup (Figure 4)—Subgroup 1 by Bowers and Dibner, Subgroup 2 by Gries and Bermudez, Subgroup 3 by Schildhauer and Madin, and Subgroup 4 by McGuinness and the postdoctoral researcher. Momentum will be maintained between meetings by these project leaders, each of whom has been allocated funds to undertake tasks such as documenting workshop results for publication, pursuing opportunities for observation model ratification with a standards body, and development and implementation of demonstration projects. Prior to the first executive committee meeting, the postdoctoral researcher will be hired and the communications infrastructure will be set in place, which will include a dedicated Web site, an online wiki, a shared source code/document repository, video conference calling facilities, email lists, and an internet relay chat (IRC) for daily multi-person instant messaging.

Network sustainability

A major directive for this Network will be to develop and begin actualizing mechanisms for sustaining these community efforts beyond the duration of the proposed INTEROP funding. The products of this effort will not represent an endpoint, but rather a starting point for further possibilities including growing community participation, constructing more comprehensive, accurate, and capable ontologies, and developing more powerful applications. In addition, there will be a need for technical assistance, ontology curation, and development of manuals, tutorials and other outreach mechanisms, in support of the expected products of the Network. As part of this effort, we will specifically look at and evaluate similar sustainability approaches used in the bioinformatics and biomedicine communities related to ontology curation and maintenance (e.g., the OBO Foundry and the GO Consortium), and explore related activities (e.g., organizing academic conferences or symposia dedicated to research and development issues in managing observational data and ontologies) to foster continued participation of Network participants and tasks.

The executive committee will be responsible for articulating strategies to enable a robust, long-term effort in support of advanced semantic approaches to data interoperability. Given the current proliferation of individualized approaches to ontology construction and development of semantic applications, a coordinated effort such as this one will be necessary to optimize opportunities not only for data interoperability, but also to enhance the consistency in ontology construction, and the generality and capabilities of applications built to use them.

References

[ABB+00] Ashburner M, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, G Sherlock. 2000. Gene ontology: Tool for the unification of biology. *Nat. Genet.* 25, 25–29

[ABW+04] Andelman SJ, CM Bowles, MR Willig, RB Waide. 2004. Understanding environmental complexity through a distributed knowledge network. *BioSciences* 54(3):240-246.

[BBB+06] Bermudez L, P Bogden, E Bridger, G Creager, D Forrest, J Graybeal. 2006. Toward an Ocean Observing System of Systems. In *Oceans'06 MTS/IEEE-Boston*.

[BBJ+05] Berkley C, S Bowers, MB Jones, B Ludaescher, M Schildhauer, J Tao. 2005. Incorporating Semantics in Scientific Workflow Authoring. *Proceedings of the 17th International Conference on Scientific and Statistical Database Management*. IEEE Computer Society.

[BCM+03] Baader F, D Calvanese, D McGuinnes, D Nardi, P Patel-Schneider (eds). 2003. *The Description Logic Handbook: Theory, Implementation, and Applications*, Cambridge University Press, New York, NY

[BDA+03] Browne AC, G Divita, AR Aronson, AT McCray. 2003. UMLS language and vocabulary tools. *AMIA Symposium Proceedings*, 798.

[BGA06] Bermudez L, J Graybeal, R Arko. 2006. A Marine Platforms Ontology: Experiences and Lessons. In *Semantic Sensor Networks Workshop*.

[BHL01] Berners-Lee T, J Hendler, O Lassila. 2001. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Sci. Am.* 284, 34-43.

[BLN+06] Bowers S, B Ludäscher, AHH Ngu, T Critchlow. 2006. Enabling Scientific Workflow Reuse through Structured Composition of Dataflow and Control-Flow, *IEEE Workshop on Workflow and Data Flow for Scientific Applications* (SciFlow), Atlanta, GA.

[BML+06] Bowers S, TM McPhillips, B Ludäscher, S Cohen, SB Davidson. 2006. A Model for User-Oriented Data Provenance in Pipelined Scientific Workflows. *Intl. Provenance and Annotation Workshop*, LNCS 4145: 133-147

[BML07] Bowers S, TM McPhillips, B Ludäscher. In press, 2007. Provenance in collection-oriented scientific workflows. *Concurrency and Computation: Practice and Experience*.

[BMW+07] Bowers S, TM McPhillips, M Wu, B Ludäscher. 2007. Project Histories: Managing Data Provenance Across Collection-Oriented Scientific Workflow Runs. *Workshop on Data Integration in the Life Sciences*, LNCS 4544:122-138

[BR04] Bard, J, SY Rhee. 2004. Ontologies in Biology: Design, Applications, and Future Challenges. *Nat. Rev. Genet.* 5, 213-221

[C06] Cox S (ed.) 2006. Observations and Measurement. Open GeoSpatial Consortium, Inc. OGC-05-087r4

[CNF+06] Cushing JB, NM Nadkarni, M Finch, ACS Fiala, E Murphy-Hill, L Delcambre, and D Maier. In Press. Component-based end-user database design for ecologists. *Journal of Intelligent Information Systems*, in press.

[Cleaner] Collaborative Large-scale Engineering Analysis Network for Environmental Research. http://cleaner.ncsa.uiuc.edu/home/

[CSA] CSA/NBII Biocomplexity Thesaurus. http://thesaurus.nbii.gov/

[CUAHSI] Consortium of Universities for the Advancement of Hydrologic Science. http://cuasi.org

[DAS] Stein LD, S Eddy, R Dowell. Distributed Sequence Annotation System (DAS/1) Specification, Version 1.53. http://www.biodass.org/documents/spec.html

[DwC] The Darwin Core Standard. http://wiki.tdwg.org/DarwinCore

[EML] Ecological Metadata Language (EML) Specification. http://knb.ecoinformatics.org/software/eml/.

[EOH+06] Ellison AM, LJ Osterweil, JL Hadley, A Wise, E Boose, *et al*. 2006. Analytical webs support the synthesis of ecological datasets. *Ecology* 87:1345-1358.

[FMM+06] Fox P, D McGuinnes, D Middleton, L Cinquini, A Darnell, J Garcia, P West, J Benedict, S Solomon. 2006. Semantically-Enabled Large-Scale Science Data Repositories. In *Proc. of the Intl. Semantic Web Conf.*, LNCS 4273:792-805

[GB06] Graybeal J, L Bermudez. 2006. When Hydrospheres Collide: Lessons In Practical Environmental Ontologies. *Intl. Conf. on Hydroscience and Engineering*.

[GBB05] Graybeal J, L Bermudez, K Brinks. Improving Ocean Research (Meta) Data Management. ORION Newsletter 2:1-3.

[GEON] The GeoSciences Network (GEON). http://geongrid.org

[GGM+02] Gangemi A, N Guarino, C Masolo, A Oltramari, L Schneider. 2002. Sweeting Ontologies with DOLCE. *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web* (EKAW), LNCS, 2473.

[GLEON] Global Lake Ecological Observatory Network. http://www.gleon.org

[GML] The Geography Markup Language (GML), Version 3.1.1. http://www.opengis.net/gml.

[GO00] The Gene Ontology Consortium. 2000. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* 25, 25-29

[HL93] Humphreys BL, DAB Lindberg. 1993. The UMLS project: making the conceptual connection between users and the information they need. *Bull Med Libr Assoc*, 81(2):170-177.

[JBB+01] Jones, M, C. Berkley, J. Bojilova, M. Schildhauer. 2001. Managing Scientific Metadata. *IEEE Internet Computing*, vol. 5, no. 5, pp. 59-68.

[Jena] Jena Semantic Web Framework, HP Labs, http://jena.sourceforge.net/

[JSR+06] Jones MB, MP Schildhauer, OJ Reichman, S Bowers. 2006. The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Ann. Review of Ecol., Evol., and Syst.* 37:519-544.

[LLB+03] Lin, K., B. Ludäscher, B. Brodaric, D. Seber, C. Baru, and K. A. Sinha, Semantic Mediation Services in Geologic Data Integration: A Case Study from the GEON Grid, In Geological Society of America (GSA) Annual Meeting, volume 35(6), November 2003.

[LLB+06] Ludäscher, B., Lin K., Bowers S., Jaeger-Frank E., Brodaric B., Baru C., Managing Scientific Data: From Data Integration to Scientific Workflows, GSA Today, Special Issue on Geoinformatics, 2006.

[LBH+06] Ludäscher B., Altintas I., Berkley C., Higgins D., Jaeger-Frank E., Jones M., Lee E., Tao J., Zhao Y. 2006. Scientific Workflow Management and the Kepler System, *Concurrency and Computation: Practice & Experience, Special Issue on Scientific Workflows*, 2006.

[MB05] McPhillips TM, S Bowers. 2005. An approach for pipelining nested collections in scientific workflows. *SIGMOD Record* 34(3): 12-17.

[MBH+97] Michener W.K., J.W. Brunt, J.J. Helly, T.B. Kirchner, S.G. Stafford. 1997. Non-geospatial metadata for the ecological sciences. *Ecol. Appl.* 7:330-342.

[MBL06] McPhillips TM, S Bowers, B Ludäscher. 2006. Collection-Oriented Scientific Workflows for Integrating and Analyzing Biological Data. *Workshop on Data Integration in the Life Sciences*, LNCS 4075: 248-263

[MBS+07] Madin J, S Bowers, M Schildhauer, S Krivov, D Pennington, F Villa. (In press, 2007) An ontology for describing and synthesizing ecological observation data. *Ecol. Inform.*

[MFC+07] McGuinness D, P Fox, L Cinquini, P West, J Garcia, J Benedict, D Middleton. 2007. The Virtual Solar-Terrestrial Observatory: A Deployed Semantic Web Application Case Study for Scientific Research. In the proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-07). Vancouver, British Columbia, Canada, July 22-26.

[MJ02] McCartney, P. and M. B. Jones. 2002. Using XML-encoded Metadata as a Basis for Advanced Information Systems for Ecological Research. *Proceedings of the 6th World Multi-Conference on Systemics, Cybernetics, and Informatics* (SCI 2002). Orlando, Florida.

[MvH04] McGuinness D, F van Harmelen. 2004. OWL Web Ontology Language Overview. World Wide Web Consortium (W3C) Recommendation. February 10, 2004. Available from http://www.w3.org/TR/owl-features/.

[OBOE] The Extensible Observation Ontology (OBOE). http://ecoinformatics.org/ontologies/observation-0.1.0

[ODS] Observational Data Standard, Version 1.0. NatureServe and TDWG (ed.) http://www.natureserve.org/prodServices/pdf/Obs_standard.pdf

[OGC] The Open Geospatial Consortium, Inc. http://www.opengeospatial.org

[OWL] Michael K. Smith, Chris Welty, and Deborah L. McGuinness. OWL Web Ontology Language Guide. World Wide Web Consortium (W3C) Recommendation. February 10, 2004. Available from http://www.w3.org/TR/owl-guide/

[P07] Parr, CS. (in review, 2007) ETHAN: the Evolutionary Trees and Natural History Ontology. *Ecol. Inform.*

[Pellet] Pellet OWL Reasoner, University of Maryland, http://www.mindswap.org/2003/pellet/

[PN02] Pease, A, I Niles. 2002. IEEE Standard Upper Ontology: A Progress Report. *Knowledge Engineering Review*, Special Issue on Ontologies and Agents, Vol 17.

[R04] Raskin R. 2004. Enabling Semantic Interoperability for Earth Science Data. Final Report to NASA Earth Science Technology Office (ESTO), Technical Report, Jet Propulsion Laboratory (http://sweet.jpl.nasa.gov)

[S99] Sowa, J. F. 1999. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. PWS Publishing Co., Boston.

[SEEK] The Science Environment for Ecological Knowledge (SEEK). http://seek.ecoinformatics.org

[SK05] Soldatova LN, RD King. 2005. Are the current ontologies in biology good ontologies? *Nat. Biotechnol.* 23, 1095-1098

[SM03] Schentz H, M Mirtl. 2003. MORIS: A universal information system for environmental monitoring. In *Environmental Software Systems*, vol. 5, Springer

[SMJ02] Spyns P, R Meersman, M Jarrar. 2002. Data Modelling versus Ontology Engineering. *SIGMOD Record* 31(4): 12-17.

[SSM05] Schleidt K, H Schentz, M Mirtl. (2005) Overcoming the multiple islands of ontologies. In *Proc. of the Intl. Conf. on Informatics for Environmental Protection* (Part 1), pp. 342-246

[SUMO] SUMO. http://ontology.teknowledge.com

[SWEET] SWEET. http://sweet.jpl.nasa.gov/ontology

[TAG] TDWG Ontology, TDWG Technical Architecture Group. http://wiki.tdwg.org/twiki/bin/view/TAG/

[TDWG] Biodiversity Information Standards (TDWG). http://www.tdwg.org

[THM07] Tarboton DG, JS Horsburgh, DR Maidment. 2007. CUAHSI Community Observations Data Model (ODM) Version 1.0 Design Specifications, (http://www.cuahsi.org/his/docs/ODM1.pdf)

[UMLS] The Unified Medical Language System (UMLS), http://umlsinfo.nlm.nih.gov/

[WMG06] Williams, RJ, ND Marinez, J Goldbeck. 2006. Ontologies for ecoinformatics. *J. Web Semant.* 4:237-242

[WWG+05] Wright D, S Watson, J Graybeal, L Bermudez. 2005. Making scientific data sets easier to find, access, and use. *Trans. of the American Geophysical Union* 86:522-525.