Cyberinfrastructure for Environmental Observation Networks (CEON) Workshop Report

held

February 25 & 26, 2008

at

The National Science Foundation

Arlington, VA

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The participants travel was funded under the following NSF grants: EAR-0326064, EAR-0622374 (CUAHSI), EAR-0323309 (Earthscope – IRIS), DEB-0236154 (LTER), DBI-0735106 (NEON), OCE-0418967 (OOI), ATM-0618200 (UCAR), and CBET-0533513 (WATERS) along with the Joint Oceanographic Institution Ocean Observing Initiative, the United States Geological Survey (USGS), and the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.

August 8, 2008 - Version 1.1

Table of Contents

Participants

Peter Backland	UCAR	Organizational Breakout
James Brunt	LTER	Technology Breakout, Scribe
Rob Casey	Earthscope	Organizational Breakout, Report Editor
Peter Fox	UCAR	Standards Breakout
Corinna Gries	LTER	Organizational Breakout
Richard Hooper	CUAHSI	Standards Breakout, Report Editor
Tim Kratz	LTER	Standards Breakout
Barbara Minsker	WATERS Network	Organizational Breakout, Moderator/Scribe
Jim Myers	WATERS Network	Technology Breakout, Report Editor
Marc Nodell	NEON	Standards Breakout
Mark Parsons	AON	Organizational Breakout, Report Editor
Marshall Peterson	NEON	Technology Breakout, Moderator
Oscar Scholfield	OOI	Standards Breakout, Moderator/Scribe
Gus Shaver	AON	Standards Breakout
Alex Talalayevsky	OOI	
Frank Vernon	OOI	
Jake Weltzin	USA-NPN	Organizational Breakout
Bruce Wilson	USA-NPN	Technology Breakout, Scribe
Ilya Zaslavsky	CUAHSI	Standards Breakout

<u>NSF Observers</u>

Patrick Clemins	NSF – BIO/DBI	Report Compilation
Elizabeth Blood	NSF – BIO/DBI	
Henry Gholz	NSF – BIO/DEB	
Bruce Hamilton	NSF – ENG/CBET	
Douglas James	NSF – GEO/EAR	
Dan Lubin	NSF – OD/OPP	
Peter McCartney	NSF – BIO/DBI	
Stephen Meacham	NSF – OD/OCI	
Kay Shedlock	NSF – GEO/EAR	
Kevin Thompson	NSF – OD/OCI	
Shelby Walker	NSF – GEO/OCE	

Introduction

Purpose

The purpose of this workshop is to promote collaboration in cyberinfrastructure design, implementation, and maintenance between environmental observation networks (EONs) funded by the National Science Foundation (NSF).

Expectations

The participants of the workshop were tasked with three main goals:

- Identify the common cyberinfrastructure requirements between the observatories;
- Identify modes of interoperability and coordination between the observatories;
- Assess the process of cyberinfrastructure development from research and development, through production, to operations and maintenance.

It is also expected that a report will be generated from the workshop outlining the topics discussed and any action items that might come from the discussions.

The discussion between the EONs should take place on three different levels: (1) overall goals of the combined EONs, (2) similarities between the science questions each EON is trying to answer, and (3) the cyberinfrastructure needed to accomplish those shared goals and individual science questions. An EON Strategic Plan could result from this discussion which would help policymakers form a roadmap for the evolution of EONs.

Vision

The environmental observation networks comprise distributed, yet interconnected networks spanning local, regional, and global scales that allow scientists to study a range of high priority processes which have been identified by a number of related environmental science communities. Cyberinfrastructure constitutes the integrating element that binds all of the distributed observatories into a coherent system. Ideally, the environmental observatories' cyberinfrastructure should enable a federated system of observatories, laboratories, classrooms, and facilities that realizes national scientific goals. The vision of the environmental observation networks' cyberinfrastructure is to provide any scientific user with a system that enables simple and direct use of resources to accomplish their scientific objectives which would not be possible without the linked infrastructure. This vision includes direct access to instrument data, control, and operational activities described above, and the opportunity to seamlessly collaborate with other scientists, institutions, projects, and disciplines using streaming and historical data.

Common Challenges

After brief introductions, each observatory gave a short presentation on their science objectives, cyberinfrastructure architecture, and cyberinfrastructure challenges. These presentations are available at the workshop's wiki site, http://roadrunner.lternet.edu/drupal/. The observatories then rotated among four breakout sessions where the participants discussed common cyberinfrastructure challenges. A list of common cyberinfrastructure challenges, grouped into three main topics, was generated by these breakout sessions.

1. Technological Challenges
   - Adaptable cyberinfrastructure to respond to evolving science
   - Need for a tailored product which provides trustworthy and understandable information
   - Enable PIs to submit well described data
   - Network interoperability
   - Timeliness of data delivery
   - Secure delivery of data
   - Accessibility of data and tools by a diverse user base where not all are experts
   - Support of multi-disciplinary research
   - Verification of properly labeled data / Metadata standards
   - Maintenance and reporting of provenance of data and derived products
   - Exponential growth of data, scalability
   - Establish a raw data center for the collection of data
   - Create infrastructure for a single logical data storage location
   - Storage and management of highly dimensional data
   - Registration of data in four-dimensional space-time
   - Authentication of system users to support multiple levels of access (read only, comment, edit, delete)
   - Capture of data context including disturbance history and other historical aspects

Many of the technological challenges relate to data accessibility, scalability, and interoperability. Researchers want to be able to access stored data quickly and securely, and through an easy-to-use interface. There is also a desire for the cyberinfrastructure to scale easily to larger data storage requirements, multiple geographic locations, and higher dimensional data. Finally, the data should be well labeled with standard metadata so it can be used with various tools, each of which requires data in a specific format.

There is a need for increased communication between the EONs and with the computer science community. There is not a well-established dialogue between these groups to inform cyberinfrastructure design decision makers about the available technologies and what other EONs are using to solve their cyberinfrastructure issues. Regardless of the technologies used, data access needs to be open and unrestrictive because it is often used in ways not initially anticipated. Finally, as these technologies are developed, they need to be integrated into the core infrastructure so that it is available for use by the wider EON community.

How these technologies are implemented has definite implications for future scalability, administration of the services, and economic model. The design should be flexible enough so as not to limit implementation choices. Collection of services at a supercomputing center is one option. This implementation model limits dynamic load adjustments, but generally has better administration and interoperability because all services are maintained by a single staff. A more distributed implementation where services are implemented at various locations, possibly at each EON's computing center, can respond to load demands more dynamically but requires better planning and more communication for administering the services and ensuring interoperability. Whichever implementation model or combination of the two is chosen, there needs to be a virtualization of the storage and computing capabilities of the system to ensure ease of use and accessibility regardless of location. Global Earth Observations and the Data Management Group were specifically mentioned as organizations that have well developed cyberinfrastructure implementation models. There is interest in other potential models for cyberinfrastructure architecture and development.

For those EONs that are in operation or at advanced stages of cyberinfrastructure implementation, there is a need to balance ongoing operations against moving to new technologies. Moving to new technologies often requires a significant investment and those funds are not always available to operating EONs. In addition, users of the EON's cyberinfrastructure may not have the resources to update to new technology along with the EON, thus abandoning that user or forcing the EON to continue to operate legacy systems.

2. Organization / Societal Challenges
    - Governance structure for shared resources and collaboration
    - Community buy-in to resources
    - Incentives and culture change to make data sharing popular
    - Rules vs. Incentives in terms of data sharing and community buy-in
    - Ethical concerns about shared data
    - Societal contributions, Community/Citizen science
    - Promotion of sustained communication between EONs
    - Keeping human expertise

The organizational challenges mostly relate to the culture change required for the acceptance of sharing data and infrastructure. While rules can be made to force the sharing of data and reuse of cyberinfrastructure components, incentives for sharing and adherence to standards could work equally well. Increased communication between EONs needs to be promoted as well as a review of ethical use of data contributed by other researchers.

3. Standards
    - Standards body
    - Sustainability / Maintenance
    - Ontology for environmental and ecological variables
    - Mission planning and optimization

The final set of challenges dealt with the standards that were required for the interoperability of shared data and infrastructure. This includes the creation of a standard ontology as well as versioning standards for metadata and software tools as they adapt to meet the changing needs of the community. An organization which oversees these standards will need to be created to administer the standards documents.

After some discussion, the participants decided to forgo the discussions planned for the morning of the second day and instead work in breakout groups on each of these three main areas to generate well-developed plans of action. The results of these breakout sessions are presented in the next section.

Breakout Sessions

Technology

Rationale and Recommendations

It is widely agreed that there are a variety of cyberinfrastructure needs which span the existing and developing environmental observation networks (EONs). There is also a clear science-driven need for interoperability across observatories. Thus, coordination across observatories and between observatories and related mission agencies is a clear imperative. However, there are a number of social, political, and technical barriers to creating universal solutions and it is important that coordination efforts balance the savings of common approaches against the coordination and opportunity costs in requiring standard approaches and tools. Thus, the technology breakout group recommends an approach combining

- ongoing technical exchanges between observatory cyberinfrastructure efforts and with the larger EON and cyberinfrastructure communities,
- development of standard software interfaces and data abstractions to support interoperability in areas such as data access, security, workflow execution, modeling, and provenance, and
- identification of commodity services whose specification can be standardized across observatories and whose acquisition can be made through open competition.

In making these recommendations, the Technology Breakout group considered both the benefits and costs of coordination and the drivers of and barriers to coordination. And, as a leading example, we define a cross-observatory content management service to provide standardized secondary data storage capabilities for disaster-recovery purposes which could also provide a mechanism for basic inter-observatory data exchange and inter-observatory research. In conclusion, the group recognized that the degree to which efforts should be coordinated must ultimately be driven by the value that can be delivered to the observatories and their scientific users. We believe that an appropriate route is to begin with something which is a clear need across the EONs, implement this in a way which quickly provides value to the EONs, and then leverage this technical and social platform as the foundation for developing additional services and tools, as well as for including other EONs. Taking a lead from the agile development community, the technology implementation working groups should focus on continuous delivery of working functionality to the EONs, with designs and standards that support evolution and refactoring as technologies mature and needs are better understood.

Understanding the Cyberinfrastructure Economy

Cyberinfrastructure for scientific research is being developed in an amazingly complex environment where

- raw computing, data storage, and network capacity are all doubling within the lifetime of typical development efforts,

- commercial and open source software development efforts targeted at large markets dwarf those in the scientific market,
- funding is split between computer science research, cyberinfrastructure R&D efforts, domain-driven pilot projects, and, increasingly, ongoing operations and maintenance, and
- cyberinfrastructure is being developed in a staggered temporal pattern with more than one generation of technology between pilot projects and production facilities and between efforts in different disciplines.

Within this maelstrom, domain researchers partner with available IT researchers to gain needed expertise and to leverage existing software while IT researchers enter such partnerships to gain intellectual and funding drivers for their own programs. New projects continuously make choices between extending existing software and taking advantage of new software techniques and tools to achieve next-generation capabilities.

Into this mix, observatories bring a new challenge – creating and sustaining continent-spanning infrastructure designed to support distributed multi-disciplinary research communities over decades. This scale brings new attention from the commercial sector and a new emphasis on techniques to build well designed, but not over-engineered, systems that can flexibly support research and education needs as cost-effectively as possible. Exacerbating the problems caused by the rapid pace of technological change is the lack of appropriately trained personnel (e.g. IT-savvy domain researchers, domain-savvy IT researchers, and science-savvy systems-engineers and development teams) and the lack of sustainable career paths for such professionals. Because large-scale observatories are an emerging phenomenon[1] and the broadly inclusive concept of cyberinfrastructure is new,[2] the creation of observatory cyberinfrastructure involves culture and career changes and innovative leveraging of emerging tools and multiple funding sources.

Together, these factors drive the community as a whole towards sub-optimal collaboration and coordination on cyberinfrastructure, but they also place real limits on the degree to which coordination will be cost-effective. In many ways, the historical lack of funding for cyberinfrastructure leads to close coupling and co-dependence of specific domain/IT project teams; domain scientists acquire infrastructure through a combination of joint proposals and a willingness to participate in computer science research. There is little incentive to find off-the-shelf solutions that while cost-effective at meeting current needs would ultimately reduce the total infrastructure funding available and limit access to knowledge about next-generation tools. The observatories, due to their similarity, size, and classification as infrastructure, have a potential to change these dynamics. As a federation, the observatories have the range of experience and expertise to compare, contrast, and evaluate multiple technologies and approaches which in turn, can lead to the definition of open interfaces that enable the identification of both common components

[1] Craine et al. 2007. Annual Report of the National Solar Observatory. Arlington, VA: National Science Foundation.
[2] Atkins et al. 2004. Revolutionizing Science and Engineering Through Cyberinfrastructure. Arlington, VA: National Science Foundation.

and areas where projects have unique needs and interests. In addition, a federation of observatories possesses a measure of buying power and can instantiate a process for developing technical specifications for which service providers would be willing to compete. Current trends in service-oriented computing towards hardware as a service (HAAS) and software as a service (SAAS) present a fertile environment for innovative approaches to scientific computing as a commodity. Done well, with modern concepts such as service-oriented architecture, workflow, and content management, this could develop a 'cyberinfrastructure market' where the benefits of one-to-one partnerships for cyberinfrastructure are outweighed by the benefits of being able to integrate best-of-breed tools and to market individual technologies across observatories.

It is critical to recognize, though, that some of the circumstances of the observatories make some level of independence in cyberinfrastructure development the best way to provide value. The different start times and operational horizons for the various observatories is one of the more significant circumstances; newer technologies often promise increased scalability, enhanced functionality, and reduced maintenance costs yet, for projects in operations, may also require significant new investments in hardware, software, and staff training. In addition, an upgrade can potentially be disruptive to ongoing research. Disciplinary differences in culture, domain conceptual models, and the maturity of relevant observational and modeling capabilities can also make different approaches to cyberinfrastructure the most effective path.

Coordinating Observatory Cyberinfrastructure

These considerations drive the group recommendations mentioned previously. Ongoing technical exchanges and the identification of standard interfaces and data abstractions will improve the observatories' ability to be educated consumers of cyberinfrastructure and will start to open the infrastructure for more incremental advances and enable competition between cyberinfrastructure providers at the level of individual components and add-on capabilities rather than full systems. This work will also reduce the effort that will be required for domain researchers and computer science researchers to work across observatories in pursuit of their research interests. Technical exchanges and standardization efforts may best be pursued within broader forums (e.g., AGU, ESIP, OGC), but some additional coordination between observatories to speak with a common voice regarding interests and priorities would be important in making practical progress.

In discussing the potential for standards across the observatories at the workshop, the group noted that there are aspects of cyberinfrastructure in which the observatories are driving the development of new functionality, and others, such as security mechanisms and data back-up for disaster recovery where they essentially follow current best practice as defined in the larger cyberinfrastructure/IT community. These latter areas are particular ripe targets for standard interfaces and commodity solutions.

The commodity solution as discussed was clearly distinguished from an undesirable "lowest common denominator" approach. The style of interface represented by the Pluggable Authentication Module (PAM) and Java Authentication and Authorization Services (JAAS) specifications was cited as a way to provide a standard interface for commodity solutions (for user authentication in this case) without over-specifying the solution or requiring exactly the same solution for all observatories. In the case of PAM and JAAS, cyberinfrastructure using these interfaces can be easily configured to work with authentication modules providing simple username/password, Kerberos, Grid certificate, or another authentication mechanism.

Commoditizing Observatory Data Recovery Cyberinfrastructure

One area where a clear potential for coordinated work emerged was in consolidated data management services, specifically related to off-site back-up storage for disaster preparedness. This capability is a need for virtually any EON and is a function which is outside the discipline focus of the EONs. It is also an area far enough removed from the public face of the EONs that a cost-effective third party solution shared across observatories is likely to be welcomed. While there are other technology areas which are potential near-term opportunities, such as identity management, data provenance, portal technologies, metadata tools, and data discovery, our assertion is that starting with data management services provides both a technical and social foundation upon which collaborations in these other technical areas can be based.

The initial focus for Consolidated Data Management Services would be the secondary storage or backup data center functionality with is critical to EONs, particularly those of national scope, since any single data center will experience some amount of down time. This secondary storage would enable EONs to have off-site copies of primary data, initially for disaster recovery purposes. The services could target primary observational data rather than derived or higher-level data products, and the EON data centers would retain ownership of the data. As these services mature and multiple service providers are available, EON data centers could potentially choose to use these services for derived products and/or use them as the primary storage for the data with the EON providing only the domain-specific interfaces to that data in ways that best serve the particular community for that EON.

While the specific scope of such services will need to be defined by the EONs, the breakout group focused on a model that would standardize geospatial and temporal metadata and indexing while allowing variation in other aspects of metadata. It should be easier to reach agreement on a simple model in which data is tagged with geospatial and temporal coordinates and a list of observational parameters (e.g. temperature, stream flow) that were recorded in contrast with defining an overall EON ontology. Further, standardization of this core would be sufficient to support the simple queries needed to recover data after a disaster and would enable enough interoperability to enable some cross-observatory research. Such a model should not limit the additional metadata that individual EONs might wish to store or is required by a specific file format and each EON could decide what level of additional metadata is appropriate for their community. This type of approach is often termed “content management” and is widely used in business to manage highly heterogeneous collections. The specification effort would need to define standard programming interfaces and/or data transfer protocols. The effort would also need to include standardization of language related to service levels and quality of service guarantees that could be used to classify software/service offerings that meet the interface standards, thereby forming a basis for standard evaluation criteria across the EONs.

Given that the scope of such an effort can be relatively well defined and the end goal of enabling commodity disaster recovery storage provides a clear limit on the functionality required, the breakout group believes that there is a near term opportunity here with a clear value proposition. Thus, it should be possible to define a scope, schedule, and budget for the creation of standards in this area and to implement those standards to support disaster recovery in the EONs which would in turn allow decisions to be made as to how the effort can be funded. Given the ubiquity of the need for disaster recovery services, there is the potential for funding from the EONs, NSF, mission agencies (NASA, USGS) and foundations (Moore, Google). Commercial software providers might also be interested in contributing to such an effort. Adoption of such a solution would probably require up-front commitments by at least a subset of EONs to assure a minimal market for solution providers. Further adoption could then be encouraged through mechanisms such as peer-review; any disaster recovery plan proposed could be compared against this default.

Conclusion

Cyberinfrastructure development for the EONs is a complex undertaking affected by a wide range of socio-technical factors, many of which lead to unnecessary duplication of effort and some of which limit the extent to which standardization is practical. The working group unanimously agreed that there is significant value to be had in enhancing the level of coordination across EONs from where it is today and identified three levels of interaction ranging from general technical discussion, to coordination of component interfaces, to the standardization of commodity services needed across EONs that would encourage coordination with potentially measurable impacts on the capital and operations and maintenance costs of the EONs. The area of data back-up for disaster recovery was seen as a particularly “low-hanging fruit” that would be technically feasible, socially acceptable, and of significant value on a relatively short time horizon and thus could serve as an example which could be replicated in other areas of infrastructure over time.

Organization

The organization breakout group recommends that a Federation of Environmental Observation Networks (FEON) be created with the goal to coordinate and advance activities of emerging and established scientific EONs that are focused on observing and understanding Earth systems on local to global scales. The need for coordination across the EONs is driven by common scientific and infrastructure requirements. The EONs share science interests that require common standards and interoperability to enable researchers to easily access data from all EONs. For example, many of the EONs are exploring the effects of global climate change on environmental systems which will require common access to climate change data and forecasts. NEON is proposing STREON (Stream Observatory Network) to assess the effects of human activities on freshwater ecosystems, a topic that WATERS Network is also interested in. EONs also share common technology needs and unnecessary duplication of technology development and integration efforts should be avoided with the aim of reducing the operations and maintenance costs of maintaining cyberinfrastructure in the face of rapidly evolving technologies.

Addressing these needs will require near-term investment in a framework to foster coordination and enable such long-term benefits. The primary focus of FEON would be on cyberinfrastructure (including sensors) to support the scientific and education agenda across the EONs. Figure 1 shows a proposed organizational structure for FEON wherein each EON would nominate representatives to a steering committee which would launch long-term working groups and short-term task forces as needed. The proposed role of each component of FEON is described in more detail below followed by a discussion of the next steps toward creating FEON.

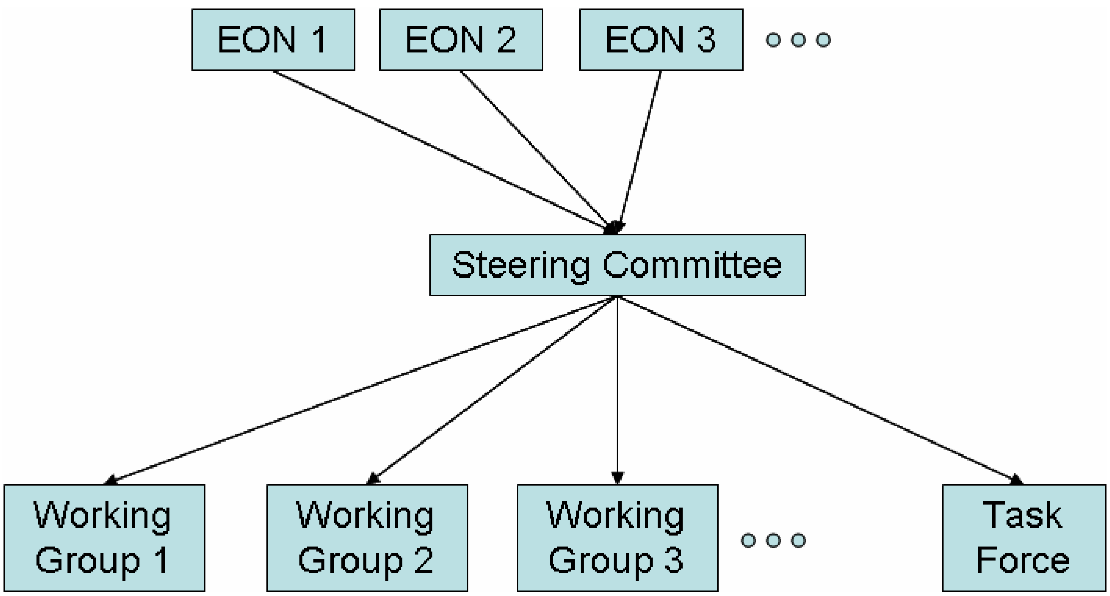

Figure 1. Proposed organizational structure for FEON.

The steering committee members would be nominated by each EON and should be senior-level representatives who understand both the science agenda and the cyberinfrastructure requirements. They would oversee the coordination activities and serve as liaisons between FEON and each of the EONs. Initially, they would set up and populate the working groups and any short-term task forces needed. In the long term, they would coordinate working group activities and approve new membership applications, task forces, and working groups as needed. They would also foster cross-observatory interaction by organizing cross-cutting sessions at existing conferences and hosting an annual EON meeting for more in-depth exchange across the observatories on cross-cutting scientific, education, outreach, cyberinfrastructure, and management issues (e.g., site and deployment issues). The annual meeting could include brainstorming ideas for addressing current challenges, sharing of lessons learned, and visioning of future collaborative opportunities. This group could serve as a unified voice to provide advice on EON activities to NSF (e.g., identifying common cyberinfrastructure requirements for the Office of Cyberinfrastructure or identifying needs for the cross-directorate Environmental Research and Education program), industry, and other national and international organizations (e.g., Global Earth Observation System of Systems, or GEOSS, http://www.epa.gov/geoss/).

The working groups would pursue specific tasks related to cross-cutting technical, scientific, and education issues, setting up task forces as needed. Tasks could include identifying common cyberinfrastructure technology needs, publishing best practices, helping to evolve best practices to standards, and organizing training workshops and tutorials. Each working group would have representatives from the EONs who are interested in the topic as well as external experts appropriate to the particular topic. Working groups related to cyberinfrastructure would also benefit from including community users of the cyberinfrastructure in their membership. To ensure that the working groups are productive, each should have a committed chair and part- to full-time support staff whose expertise and level of effort would be appropriate to moving the tasks forward as efficiently as possible. These staff would coordinate regular remote working group meetings as well as in-person meetings at the annual FEON meeting where strategies for future activities would be identified.

This organization should regularly interface with the computer science developers that can provide cyberinfrastructure for EONs. FEON needs to push their cyberinfrastructure requirements, based on their scientific goals, to the developers so that the resulting cyberinfrastructure is a catalyst for scientific discovery. Inclusion of research goals in each EON's design document and annual reports could help to inform the computer science community as to the needs of the EONs. Annual meetings and representation from the computer science community in appropriate working groups could also strengthen this interface.

To develop this type of organization, the first step would be to further develop this vision through interactions with the broader community and NSF as well as to examine potential organizational models (e.g., the Open Geospatial Consortium, http://www.opengeospatial.org/). The next step would be to form the steering committee which would in turn develop a strategic plan, identify needs for initial working groups, plan meetings and workshops, and identify and pursue proposal opportunities to support the activities. NCAR has offered to help with coordination, travel, and the set up of Web tools for the initial activities, under their environmental cyberinfrastructure coordination project. Each of the EONs should budget for limited staff time to participate in the initial activities, but efforts should be made to obtain funding for more extensive activities as soon as possible. Funding should also be pursued for a "network wrangler" who would keep abreast of developments in all of the observatories and foster information and data exchange. In the longer term, as the activities mature, a more permanent program office is envisioned with a full- or part-time director, ideally a well-known and respected member of the scientific community, and funding to support the activities of the steering committee and working groups. The program office could become an incorporated consortium able to receive funding directly or could remain as a project supported by an existing entity.

Standards

The vision of the environmental observation networks' cyberinfrastructure is to provide any scientific user with a system that enables simple and direct use of resources to accomplish their scientific objectives regardless of which observatory maintains the resource. This vision includes direct access to instrument data, control, and operational activities described above, and the opportunity to seamlessly collaborate with other scientists, institutions, projects, and disciplines using streaming and historical data. A critical issue that will be required to achieve this goal is defining and developing a strategy to provide a common set of standards that will allow for interoperability among these developing and legacy observation networks.

Developing a set of standards among these distributed systems is critical for data interoperability, which requires a contract for publishing, discovering, accessing, and integrating data and services in a reliable and scalable fashion. Standards are a necessary component of a service-oriented architecture capable of self-growth without additional investment because if outside developers comply with the standards and write applications against published schemas, they can expect that they can be easily integrated and re-used in any standards compliant environment. These standards are also extremely important for data provenance to develop a broader community that values the collection and open sharing of data as much as the synthesis of the data.

Development of such standards will require the creation of a common vocabulary to allow semantic mediation/mapping. This often involves developing a seamless means to register data between diverse scientific communities that do not have a common lexicon. This is can be especially daunting for integrating legacy systems where the systems have been evolving for decades. Therefore, establishing standards will require a flexible system, developed bottom-up, which establishes semantic standards driven by the functional needs spanning the diverse scientific communities. It is likely that knowledge encoding will be required to develop an informed, and evolving, ontology. This ontology will likely be hierarchal and contain taxonomic attributes. A grass-roots approach would allow the ontology to be mapped and data gaps to be inferred from the perspective of the science goals of each EON.

Much can be learned from expertise in the geophysical community. One example is learning how organizations deal with data quality. Developing a common ISO-standard for marking data with quality metrics and provenance will be critical to ensuring data quality. Additionally, community expertise can assist in developing the governance issues for security requiring authenticated access to data and services. Finally, past experience suggests this diverse community must apply their own experiences to defining a data publishing policy. Such a policy can be built into the software infrastructure and outline the shared principles of data publication. This publishing policy would cut across all observatories and would evolve in time. Other examples of necessary standards learned from the experiences of CUASHI HIS can be found in the Past Successes Appendix including standards governance, data discovery, ontologies, data integration, and protocol standardization.

Future Directions

Despite failed attempts in the past, now is the time for the EONs to create an organization, like FEON, to foster collaboration. Past attempts were not broad enough in participation to enable large scale collaboration between the networks. EONs at various stages of maturity can benefit from participation in FEON. Observatories in the planning stages can learn from lessons that others have gained previously and because of the fast moving pace of technology, mature networks can learn about new technologies that might be utilized in their next cyberinfrastructure update. An organization like FEON will allow the EONs to demonstrate the science that they perform and enable as well as communicate their needs and barriers as a community.

This report will be presented at the NCAR workshop on Cyberinfrastructure for Environmental Observations, Analysis and Forecasting: A Cyberinformatics Forum and the FEON steering committee will be selected at this workshop as well.

Potential Funding Sources

There are a number of NSF programs which support multi-disciplinary collaboration, especially in the area of cyberinfrastructure. The Research Coordination Networks in Biological Sciences (RCN) program's goal is to foster interactions among scientists to advance a field and encourage novel networking strategies. The Office of Cyberinfrastructure (OCI) has a number of programs which support the acquisition, development, and provision of cyberinfrastructure resources that are applicable across multiple disciplines.

Current grants to the member EONs could also be utilized. Investment in collaborative activities would be beneficial to the EONs because of the potential cost savings in the operation of shared cyberinfrastructure. Funds that were originally budgeted for the creation of their own cyberinfrastructure could be pooled with other observatories to create a shared resource. The cost of operation for that shared resource could then again be shared across the observatories, leaving more funds for scientific endeavors.

Appendices

Cyberinfrastructure Links

National Research Council (NRC). 2006. Toward an Integrated Arctic Observing Network. Washington, DC: National Academies Press.

The AON Cooperative Arctic Data and Information Service. http://www.eol.ucar.edu/projects/aon-cadis/.

Study of Environmental Change (SEARCH). 2005. Study of Environmental Change: Plans for Implementation During the International Polar Year and Beyond. Fairbanks, AK: Arctic Research Consortium of the United States. 104 pp. http://www.arcus.org/search/downloads/SIW Report FINAL.pdf

Past Successes

Expandable Service Oriented Architecture

One of the goals of CUAHSI HIS is to develop a sustainable system that can grow on its own through contributions by community members not formally associated with or funded through the project. Several factors made involvement of third party developers easier. These factors, and the lessons learned are summarized below:

- Understanding of community needs and research scenarios, through a series of user surveys and a range of other feedback channels, is critical for wider acceptance of project results as they allow us to both tune the infrastructure components to user needs and demonstrate that the community has a say in the direction of the project;
- The core services provide significantly improved access to large volumes of data that are in wide demand;
- There is a clearly defined information model, data exchange protocol, and service contracts which are all tuned to the semantics used in the community;
- Developed a fairly straight-forward and well-documented workflow for adding new data to the system along with training and support;
- Buy-in and collaboration with critical data providers at the federal level (USGS, EPA, NOAA) was important. Attaining the current level of collaboration was neither easy nor straight-forward because the agencies would not provide their data catalogs, low-level access to their systems, or expertise until we demonstrated that we can harvest the catalogs ourselves and add value to their system;
- The goals and the architecture of the information system are well defined and the role of each component is fairly clear. This makes adding components or "building into" the system possible;
- CUAHSI HIS was among the first to embrace web services and SOA in this field and introduce other stakeholders to them;

- Focus on client applications that demonstrate integration of data access services in a way that was not be possible before (e.g. ontology-aware search in Hydroseek). At the same time, supporting quick addition of user data to the national hydrologic map;
- Using ontologies and ISO metadata standards to make data semantics and data access easily interpretable to outside developers;
- Coordination of our protocol development (WaterML) with international standards bodies (OGC) which lets us share the development with a wide cross-disciplinary audience of cyberinfrastructure developers and engage in a series of interoperability demos (e.g. a recent GEOSS demo, a planned OGC Water Resources Interoperability Experiment);
- Entrepreneurial activities such as marketing, outreach, publications in trade journals, presentations at trade conferences and workshops, collaboration with companies that have large market penetration, and working with COTS client software were necessary to engage potential external developers;
- Extensive documentation and a workbook on using the services lowers the learning curve for external developers;
- BSD licensing of the code makes it easy for companies to work with us;
- Quick move to providing core services at a production level was necessary to convince external developers to write clients that access the services.
- Having an ambitious but narrow goal (building a comprehensive portrait of hydrologic observations history for the entire country through integration of distributed data sources with easy discovery, access, and analysis/modeling interfaces and client applications), which implies community participation and promises to have a transformative effect on the domain (i.e. instead of painfully assembling locally-downloaded disparate data into model inputs and connecting your models to huge and independently managed external data repositories.) made our infrastructure attractive for outside use;
- The above goal is perhaps worthy of a MREFC (as the recent WATERS meeting showed). However, focusing on the core services allowed us to approach it with 1.75FTE at SDSC and several additional fractions of an FTE at partner universities providing hydrologic expertise. While this level of funding is inadequate for providing production services (not to mention that at the moment we see a lot more low hanging fruit than we can grab), focusing on the core services allowed the system to move to its current state even at this low funding level.

Experience with external groups adopting CUAHSI HIS cyberinfrastructure components to date includes:

2006:

- CUAHSI HIS web services are discussed on the BASINS mailing list as a new way to access hydrologic data. The list is mostly used by hydrologists and developers outside academia;
- NCDC develops ASOS web services following WaterOneFlow service signatures and WaterML.

2007:

- MOU with USGS; USGS is developing WaterML-compliant GetValues service;
- GLEON uses an early version of CUAHSI HIS ODM to develop their own database schema (VEGA);
- Phoenix LTER is developing MySQL-based ODM and Java-based WaterML-compliant web services;
- A Google Earth based client for CUAHSI web services is developed at CSIRO, Australia;
- Deployment to 11 hydrologic observatory test beds.

2008:

- KISTERS develops WaterML-compliant web services over their database for a client;
- MapWindow open source GIS develops WaterOneFlow parsers;
- Florida, Texas, and Idaho use ODM and WaterOneFlow web services to provide access to state data repositories. New Jersey is considering the same.

<u>Standards Development</u>

As an example of community experience, the importance of extracting past experience can be highlighted in the following considerations that have been gleaned from the extensive experience gained in the CUAHSI HIS project.

1) Standards Governance: There should be a body regulating standards use which includes identifier governance, protocol development and implementation, maintenance and publication of vocabularies and ontologies, and persistence implementation. It may be useful to have a cross-EON activity to integrate governance structures (including the issues above) across EON projects.

2) Data Publication and Discovery: CUAHSI HIS has experience in developing both syntactic and semantic standards for data publication and discovery. On the structural and syntactic sides, there is a canonical information model for observations data expressed in a relational schema (ODM) and as an XML schema (Water Markup Language). For publication, observational data are loaded into ODM or exposed via WaterML-based web services, and an observation data catalog is assembled. At the data discovery phase, the catalogs are queried from web-based or desktop clients using GetSites, GetVariables, GetSiteInfo, and GetVariableInfo calls, while data are retrieved via GetValues calls. We expect that such observational data modeling and management experience may be applicable across EONs since this type of data is widely used.

3) Ontologies: On the semantic level, CUAHSI HIS has developed a parameter-based ontology and a set of controlled vocabularies used to systematize ODM content. At publication, observational data catalogs are required to conform to

controlled vocabularies, while parameters are tagged with ontology concepts from the parameter ontology. During discovery, users can use ontology terms to search for sites and ontology terms are resolved to the actual parameters used in the datasets. Prefixing site and parameter identifiers with local vocabulary names (e.g. NWIS:06000 stands for total nitrogen as available in NWIS) ensures globally-unique IDs. While variable semantics are certainly different across EONs, the CUAHSI approach to systematizing it may be useful to explore in other contexts. Further, it would be useful to explore cases where data are integrated across domains, and the potential for semantic conflicts exist (i.e. hydrologic models accessing data from ocean, soil, vegetation, atmospheric domains).

4) Data Integration: Integration of data from different data providers, including federal agencies, state and local agencies, and local projects. The information model for observational data in CUAHSI is developed such that it can accommodate data from different data providers, including federal agencies, state and local agencies, local projects, and individual PI-driven projects. This requires intensive and ongoing collaboration with federal agencies (USGS, EPA, NCDC), on standard methods for data discovery and access. This experience can be extended to other EONs where significant data repositories are being assembled by diverse providers.

5) Protocol Standardization: The main goal of the protocol standardization within CUAHSI HIS was capturing the semantics of hydrologic observations in a form that would create the least barrier of entry for hydrologists. Compliance with international standards, such as the O&M specification from OGC, is another goal. WaterML development, in particular, is an attempt to find a balance between standardizing water data exchanges as a profile of a larger more generic standard versus elucidating common semantics used in the community; between having a large and generic standard specification or a fairly small, rigid, and easily parseable standard. Much attention has also been given to metadata ISO compliance. It would be useful to list and compare standard protocols used by different EONs to see which protocols enjoy widespread use versus areas where further standardization is needed.

Draft FEON Cyberinfrastructure Steering Committee Terms of Reference

Charge

1. Develop a strategic plan for a collaborative FEON cyberinfrastructure (CI).
2. Identify common CI needs and determine how relevant CI elements can be shared across EONs.
3. Identify and implement means to gain efficiencies and reduce duplication of effort in developing CI for the different EONs, including common standards and interoperability protocols.
4. Identify and implement means to reduce CI operation and maintenance costs and enable greater system flexibility and adaptability to technological change.
5. Identify mechanisms that enhance interdisciplinary science and scientific collaboration, such as cross-EON data discovery.
6. Review initial FEON membership and determine criteria for which EONs should participate in FEON. Periodically consider new membership applications.
7. Determine appropriate representation and coordination with relevant international and interagency initiatives such as the Global Earth Observing System of Systems (GEOSS).
8. Promote CI professional development and collaboration within the various EONs.
9. Establish long-term working groups and short-term task forces with focused mandates to address specific CI coordination activities. Specifically address how these activities will be supported.

Membership

Each environmental observation network focused on observing and understanding Earth systems on local to global scales should designate a representative to the Committee. The initial EONs represented will most likely be those involved in this whitepaper and then the Committee can consider broader membership.

It is essential that both science and CI perspectives are represented across the Committee, and that the members are committed to the success of the Committee. Members need not be PIs from EON projects, but they must have sufficient knowledge, background information, and decision authority to serve effectively. Each EON can determine the term of office for their representative.

Leadership

A Chair and Vice-Chair shall be elected by the Steering Committee members from their membership. The office of Chair and Vice-Chair is subject to rotation on a two-year basis.

The Chair and Vice Chair should be committed to ensuring the work of the Committee is done. The Chairs are expected to coordinate the activities of the group, to run the meetings, and to make sure the appropriate topics are dealt with in a timely fashion. This includes developing the agenda based on input from others. The Chair does not necessarily moderate the meetings or define rules for running the meetings, but the Chair does need to ensure it gets done.

Decisions and Authority

The Committee should decide on its decision making structure, but a consensus-based approach is recommended with each EON receiving equal weight. The Committee cannot override the programmatic requirements of the individual EONs.

Accountability

Each EON should budget a portion of their funding to contribute to the workings of the Steering Committee and relevant working groups. One quarter of an FTE and associated meeting travel should be more than appropriate. Individual EONs should report their coordination activities to relevant managers and funding agencies.

The Committee should make publicly available minutes of their meetings and records of their decisions and rationale in a timely manner.

Structure and Logistics

The Committee should develop their own schedule, working groups, and methods of operation, but an annual meeting with periodic teleconferences is recommended.

Working groups and task forces may communicate more frequently. They may be able to do most of their work in a virtual space but should plan to meet in person on a periodic basis, perhaps in association with other meetings or conferences.

As part of their strategic plan, the Committee should identify if they will need logistical support over time—e.g. a Committee secretary or a Program Office.

At their first meeting the Committee should:

1. Determine how decisions will be made (e.g., majority vote, consensus).
2. Elect a Chair and Vice-Chair.
3. Review their terms of reference and clarify as necessary.
4. Consider whether there are major gaps in the membership and whether other EONs should be invited in the short-term.
5. Begin to develop a strategic plan and identify initial working groups. Two working groups have been suggested:
    a. A Working Group to identify and develop methods for sharing technologies.
    b. A Working Group to develop methods to enable data discovery across the EONs.

Agenda

<u>Day 1</u>
8:00 Breakfast
8:30 Introductions
9:00 Expectations of Workshop (NSF Program Directors)
Identify common requirements of cyberinfrastructure (CI)
Identify modes of interoperability and coordination between observatory projects
Assess the cycle of CI development from R&D to Production to Operations
9:30 Observatory Overviews (Observatory Representatives)
3 slides each on Science Overview, CI Architecture, CI Barriers/Challenges
10:15 Break
10:30 Observatory Overviews (Observatory Representatives)
3 slides each on Science Overview, CI Architecture, CI Barriers/Challenges
12:00 Lunch
1:00 Rotating Breakouts
Identify common CI implementation requirements and problems
5 Rotations at 30 minutes each
3:30 Break
3:45 Breakout Summary
Rank common CI implementation requirements and problems
5:00 Adjourn

<u>Day 2</u>
8:00 Breakfast
8:30 Breakouts
Discuss most pressing CI problems and
10:15 Break
10:30 Directed Discussion
12:00 Lunch
1:30 What's Next?
Peter Backlund's workshop series
Create/disseminate report
Use contacts and momentum to continue discussion
3:00 Adjourn