

Final Summary: Cyberinfrastructure for Environmental Observations (CEO)

*Peter Backlund, Peter Fox, and Greg Guibert.*¹

Introduction

The NCAR Cyberinfrastructure for Environmental Observation (CEO) project was conducted by the National Center for Atmospheric Research between 2006 and 2010, supported by a grant from the National Science Foundation (NSF).

The overall goal of the project was to arrange a series of workshops and meetings that would stimulate discussion of cyberinfrastructure (CI) issues and challenges across NSF-funded environmental observing projects. This document is intended as a brief, high-level summary of the project. It represents the distilled thinking of the core project team about key challenges and opportunities for NSF as it considers future CI and environmental observing projects. More detailed description of the project, including workshop reports and descriptions of issues identified by participants, can be found on the project website at www.cyberobservatories.net.

Project Description

Primary Objectives: The community of researchers and developers involved in “cyberinfrastructure for environmental observations” is large, diverse, and dispersed, with a broad range of stakeholders from multiple disciplines and domains. The primary objectives of NCAR’s NSF-funded CEO project were to facilitate communication among these individuals, groups, and institutions and to increase collaboration and synergy across observatory efforts. Through workshops, reports, and a clearinghouse website, NCAR has sought to prime ongoing discussion on community-wide issues in order to develop a common conception of the challenges, solutions and opportunities in this rapidly developing area of science.

This project built on the previous work of many in the observing and CI communities to articulate, address, and develop solutions to the institutional and structural challenges associated with collaboration, but differed in one key respect from most previous efforts. We intentionally structured our project as a facilitated conversation driven by evolving participant interests and needs rather than attempting to promote any specific objective or outcome. This organic structure allowed issues of primary importance to community

¹ Peter Backlund (NCAR) was a co-leader of the CEO project from its inception to completion. Peter Fox, who moved from NCAR to Rensselaer Polytechnic Institute (RPI) in 2009, was also co-leader of the project from beginning to end. Greg Guibert (NCAR) participated in the project in 2009-10.

members to be self-identified and addressed among a forum of peers with common interests. The goal of the activities was not to generate a consensus for action or to develop an actionable set of recommendations, but was instead an open dialogue among participants from a wide range of scientific and technical disciplines, representing both large and small observing networks funded by NSF, other federal agencies, and third-party or private sector initiatives.

The project steering committee comprised of: Chaitan Baru (SDSC, 2006-2010), David Fulker (Unidata/UCAR and retired, 2006-2010), Kerstin Lehnert (LDEO, 2006-2010), David Maidment (U Texas, 2006-2009), Barbara Minsker (UIUC, 2006-2008) and John Orcutt (SIO/UCSD, 2006-2010) advised the project leaders on topical aspects of the activities including scope and community needs, as well as participating in many of the working activities. The steering committee met three times in person and several additional times by teleconference for consultation.

The interactions of many of the steering committee members on observing system CI actually date back to a major cyberinfrastructure meeting hosted by NCAR in 2002. Several members commented that their willingness to contribute to the CEO process was heightened by what they saw as a chance to facilitate an extended and in-depth community interaction around a set of important challenges and opportunities that were identified in the 2002 meeting.

Statistics and Methodology: NCAR and its partners conducted four workshops as part of this project between 2006 and 2010, along with a number of smaller discussions on the margins of other meetings. The first two workshops focused on CI issues in specific observing projects, while the second two were larger-scale fora focused on broader coordination mechanisms and issues for CI across multiple observatories.

- Our first workshop, conducted October 3-4, 2006, addressed issues and opportunities associated with the development of the NSF Ocean Observing Initiative.
- Our second workshop, on March 28-30, 2007, addressed issues and opportunities associated with the NSF Arctic Observatory Network (AON).
- Our third workshop, on May 5-7, 2008, focused on examination of a concept for a federation of environmental observing networks to promote coordination among NSF-funded observing efforts.
- Our fourth and final workshop, on May 17-19, 2010, was again focused on issues of large-scale coordination.

Taken as a whole, more than 250 people comprised of researchers, technicians, academics, teachers, museum curators, social scientists and CI professionals attended these events.

The first two workshops combined participants from the subject observing projects with a range of CI practitioners and experts from different observing projects. The agendas were built around the goals of the subject observing systems, with plenary presentations and

breakout discussions focused on identifying key issues and applicable lessons learned from other development efforts.

For the final two workshops, an effort was made to solicit input and participation from as broad a cross section of CI stakeholders as possible. Meeting notification and invitation was broadcast widely on listservs, e-newsletters, and through individual outreach efforts. Additionally, interested participants were central in developing the workshops agendas. Organizers used 3 or 4 framing questions to focus input from participants but the responses formed the foundation for discussion and presentation at the events themselves. In both 2008 and 2010, plenary sessions were used to prime discussion on overarching issues, with breakout groups focusing on more detailed discussion of specific issue areas, such as data management or standards. Additional breakouts, again derived from themes self-generated by participants, were then used to crosscut conversations, seeking to explore multiple dimensions of these complex issues.

The 2008 participant list is illustrative of the diverse attendance and wide range of interest and expertise. Of the nearly 75 participants,

➤ **45** were “domain experts” from a wide range of scientific disciplines, such as:

- Arctic science
- Atmospheric science
- Biomedical informatics
- Civil and environmental engineering, hydrology
- Climate change, both regional and global
- Environmental science
- Geological science , geography, cartography
- Geospatial Science, GIS
- Ocean/Marine Sciences
- Paleoclimatology and Paleoceanography
- Phenology Networks
- Plant ecology
- Polar research
- Solar science
- Management science

➤ **16** were from computer science–oriented academic departments or centers

➤ **5** were representatives from government agencies

➤ **4** were education and outreach experts from museums, libraries, and public schools.

The workshops resulted in some specific, tangible products aimed at building foundational capacity for the CI community. For example, the website www.cyberobservatories.net was created to serve as focal point and clearinghouse for community information and activity. It is currently hosted by NCAR and will continue to be a repository for the community beyond

the funding life of the NSF grant. Additionally, each of the two last workshops hosted at NCAR resulted in meeting reports that summarized the discussions. The reports, as well as many of the associated plenary presentations from 2010, are publically available on the website.

We see the admittedly intangible networking and general community building that occurred through these events as the most important outcome of the project. Through these efforts over the last five years, there is now a significantly greater shared understanding of the technical issues in observing system CI, the social issues inherent in interdisciplinary work that integrates CI with domain sciences, the challenges and opportunities associated with collaboration, and generally a better sense of ownership from individual institutions and projects towards potential community-wide solutions. Voice was given to a broad range of stakeholders, who in turn have been able to identify prospective collaborators, develop new partnerships across networks, and co-explore solutions to common problems. While the conversations and discussions will continue into the future and challenges will continue to be debated and addressed, these workshops have helped build cooperative capacity in a diverse community and have provided opportunity for solutions to form organically from within.

Participant Concerns

Although there were substantial differences among our workshops and forums (i.e., composition of the participants and designed focus), several general themes emerged frequently throughout the course of discussions and sessions.

Given the purpose of the overall CEO project, it is not surprising that the most prevalent discussion topic at all the workshops was mechanisms for coordination and organization. Initial discussion focused on the challenges of collaboration between CI practitioners and domain scientists in the context of specific observing systems. Later workshops focused more on the overall challenges and opportunities of organizing a large number of observing systems into a unified program structure or network.

A number of organizational structures and paradigms have been presented and discussed, including the concept of a Federated Earth Observing Network (FEON). The FEON concept was discussed at our third workshop in 2008 (and was the main topic at a separate, targeted NSF workshop as well) and garnered significant interest, only to lose momentum as potential participants began to disagree over issues such as inclusiveness versus exclusiveness, ability to mandate standards and solutions, and potentially disruptive overlap with existing federations of observing projects such as the Earth Science Information Partners (www.esipfed.org). The question that remains is how to move both projects and coordination mechanisms from competition to cooperation with minimum disruption of effective working arrangements.

In the past several years, interest has been increasing in looser, more organic, non-hierarchical, 'grassroots' management philosophies. Central to the issue of form and

function of a coordinating organizational structure are concerns about loss of innovation, imposition of non-optimal standards across observatories, and the overwhelming challenge associated with centrally coordinating a vast array of networks with different domain specializations, ontologies, cultures, and reward structures. Yet participants also recognize significant commonalities across observing efforts and expressed considerable concern over the danger of excessive reinvention and duplication of effort. Coordination and communication are seen as part of the solution, and are valued if they can be accomplished with minimal costs (both time and money) and a light touch. Many observing programs feel that their CI is very underfunded and marginal, and are very wary of any coordination concept that seems to impose additional unfunded tasks upon them. A smaller number consider themselves well-funded, but are concerned about the possible imposition of standards and/or methods that could be sub-optimal for them.

At the heart of many of the challenges identified during the series of workshops, were concerns around data. Almost every aspect of data collection, curation, dissemination and availability represented specific hurdles to the effective coordination and integration of observing system CI. The ability to archive vast quantities of data, ensure its accessibility at meaningful timescales for an array of known and unknown user groups, and provide sufficient provenance throughout the use-chain were issues of primary importance. The scale, scope, and timeframe of data needs across different sciences and end-use applications made it difficult to address specific solutions during the workshops themselves. But there was broad agreement that data collection and curation are areas ripe for inter-observatory coordination and the development of (loose) standardizations.

Consideration of standards often intersected with the discussions surrounding data and organizational issues. The development and use of standards was generally felt to be one of the most compelling and obvious reasons to coordinate. However, the *imposition* of standards on networks and observatories to *encourage* the collaboration was deemed detrimental and in opposition to the generally accepted best practice of allowing standards to evolve in response specific use purposes and user preferences. The utility of standards encompassed a wide range of CI related issues, including common ontologies, observational hardware and instrumentation, data storage and provenance systems, and the ability to expand the reach of networks to include new users such as K-12 educators and citizen scientists.

Cutting across all of the specific issues associated with CyberInfrastructure coordination and management lays a fundamental conflict between the rate of change within the community and technology more generally and the need to preserve and sustain legacy systems. The speed of innovation and the development of exciting new tools and technologies have the potential to greatly enhance efforts to coordinate and integrate disparate observatories and networks. However, many of the observing projects require substantial, sustained investment in existing infrastructure and data archiving to effectively meet their scientific objectives and as a result are unable to take full advantage of potentially cost-saving innovations. New observatories, many designed explicitly with multi-disciplinary observation goals, have some implicit advantages in the design and use of new technologies but the challenge of integrating with historical networks, whose robust

data represents a valuable reservoir of scientific knowledge, remains. This underlying tension was not the specific focus on any one session or presentation throughout the series of workshop but nevertheless represents a significant organizational characteristic and cultural determinant among individual observatory projects that partially defines their enthusiasm for coordination with their peers.

Key Technical Factors and Developments

Challenges: A consistent theme across each of the workshops was the fundamental disconnect between the timescale on which environmental observatories are conceived, proposed, funded, developed, operated and evaluated, and the awareness, understanding and application of rapidly changing information and communication technologies.

Due to the path that observatory CI developments have taken over the last ~10 years, many groups/ projects have had to adapt, experiment and develop immediate and practical solutions often at the expense of keeping pace with state of the field development, especially with numerous open-source offerings. More importantly is that the particular skill sets and experiences of observatory CI software engineers bias development choices as they often do not have the time (or mandate) to collaborate with other projects or utilize shared cyberinfrastructure.

Over the last decade, commercial interests have entered the observational data and CI arena. These include large corporations (e.g. Microsoft, Google) as well as consulting and development companies. On the whole, there has been a very positive contribution to observatory developments, encouraged by the NSF. However, there is an uneven level of engagement between observatory CI projects and potential commercial partners. There are several factors involved: constraints on philanthropic involvement of businesses beyond exploratory phases of projects, lack of a sustainable model of collaboration, distrust in some sectors of the research community in partnering and relying on commercial (for profit) entities, as well as cultural norms/ incentive differences. As a result, CI projects often develop customized software and services, often instead of leveraging open-source software, which in most cases, comes with costs that are similar to in-house developed software.

The final technical challenge involves balancing the tension between research and operations in observatories, especially when designing and scoping CI development. CI efforts often find themselves having to adopt older but more stable software approaches and applications to meet operational needs. While this approach is desirable within a project, it often meets with opposition and poor adoption when confronted with the demanding and fast changing needs of user communities who use numerous systems on a daily basis to do their work. The heavy weighting of risk aversion over being risk capable is seen to impact observatory CI project in many ways that cannot be overcome, usually as a result of observatory schedules, and resource constraints, i.e. insufficient fund and personnel allocations to CI.

Opportunities: Based on both practical implementation experience and increasing maturity and availability of representation and encoding content standards, many observatory CI developments have or are beginning to explicitly deal with domain vocabularies. These vocabularies take the form of encoded ontologies in one or more of the WWW Recommended languages (Ontology Web Language, Simple Knowledge Organization System, Resource Description Framework). Increasingly, these vocabularies are being made available over the web for others to use, in essence promoting the use and re-use of common vocabulary definitions as well as mappings among vocabularies. Several observatory projects are implementing such capabilities to address important functional needs, such as data discovery and access, as well as data use, e.g. integration. With these advances, it has become clear that the observatory CI community has a nascent opportunity to coordinate both vocabulary development and use to promote semantic interoperability whenever and wherever possible.

Due to increasing experience with service and resource oriented architectures using Web and Internet technologies, many observatories are functioning as virtual observatories (broadly defined). This evolution is in service of a variety of functional needs of both specialist and non-specialist users, with the latter including the increasing attention to citizen science. Environmental observatories are also increasingly aware that it is not just their data and information products that are of interest to users but those in combination with external providers, both of data but also of services, e.g. for analysis, fusion, or visualization. The capability to allow providers external to the observatories to provide services for the observatory user means that the CI development requirements are reduced, i.e. better leveraging of community developments. The big opportunity for observatory communities then shifts to the identification and adoption of international and/ or community standards for virtual observatory interoperation.

Despite tensions between ‘build versus adopt’ and ‘research versus operations’ noted above, there are now many successful examples of repurposing of CI applications and toolkits developed in environmental and atmospheric observatory communities. Robust community and agency standards such as netCDF (network Common Data Format), HDF (Hierarchical Data Format), DAP (Data Access Protocol), and vocabularies such as CF (Climate and Forecast) and the GCMD (Global Change Master Directory) to name a few, are in widespread and very effective use. Adding to these are international standards arising from ISO and the OGC (Open Geospatial Consortium). A major opportunity for environment CI communities exists in being able to learn about and adopt useful and functionally mature software. Such opportunities may be taken advantage of by regular exchange forums (conferences/ workshops) at which substantial community participation (i.e. the practitioners) can meet and converse in detail.

Overarching Findings

Many interrelated challenges make it difficult to coordinate CI across NSF environmental observing programs. The most fundamental barrier is that the NSF projects with whom we interacted, and most similar projects funded by other agencies,

were not conceived and funded as components of an overall system, but are rather stand-alone efforts with their own goals, priorities and timelines. The primary sponsors have different levels of risk tolerance and aversion. Some projects are short-term, some long-term, and some indefinite. Aversion to risk and differing timelines combine to inhibit the co-development and co-deployment of common CI, even when projects have similar needs and requirements, such as archiving data. Some projects are large-scale efforts with substantial funding, other are much more resource constrained. Project leaders repeatedly stated that their sponsors are not willing to make key “critical path” functions dependent, or partially dependent, on budgets and decision processes in other projects. And the technical experts within observing programs have varying attitudes about the relative desirability of dependability versus innovation.

The heterogeneity of observing programs and projects within NSF and other agencies argue against a tightly coupled approach to NSF environmental observing CI.

A single unified management or program structure does not appear desirable, especially when the continued development of virtual observing methods and other advances are making agreement on common data formats and technical standards less important. It is clear that domain science goals and priorities should drive the development, deployment, and operation of research observing system CI. It is more important that observing systems be tightly coupled to their domain science communities than to each other. The question is how to maintain and improve domain-CI links, domain-domain links, and CI-CI links at the same time.

The idea of a looser federation of NSF observing programs was discussed extensively in multiple meetings organized by CEO. The concept initially generated significant enthusiasm, but has not yet managed to attract a critical mass of supporters. The main areas of disagreement concerned an exclusionary versus inclusionary approach to membership and the degree to which a federated approach would or should demand adherence to specified standards and methods. There are also concerns about the relationship of a separate, NSF-centric federation and the existing Earth Science Information Partners (ESIP), which includes many NASA, NOAA, USGS, EPA and DOE projects (and not until recently, NSF projects, e.g. DataNET). On the one hand, many participants in the CEO process felt that a separate NSF federation would be duplicative and even divisive, arguing that coordination should not be organized along sponsorship lines but rather should be driven by needs and common technical challenges. On the other, many participants in ESIP expressed concern about the scalability of the ESIP model and the danger of ESIP effectiveness decreasing as it is overwhelmed by the addition of numerous new projects.

Continued periodic discussions will help grown and nurture the CI for environmental observing community. Despite the heterogeneity of programs, projects, and sponsors noted above, the last decade has seen substantial growth and maturation of a multidisciplinary technical community focused on CI for environmental observations. Practitioners in this developing field want to interact and work together. They face many common challenges over the lifetime of an environmental observing project, and the same differences in project maturity and timelines that make tight coordination difficult have also created a situation where mature projects can share lessons with less mature projects

in time to make a difference. There is an increasing pool of experts with experience in multiple projects across disciplines and sponsors.

Simply providing regular opportunities for this community to interact is very likely to result in many benefits and efficiencies over time, and NSF should continue to support such interactions. This can either be accomplished through continued support of ad hoc workshops and meetings, such as those organized as part of the CEO project, or through support of a regular annual forum. The danger of the ad hoc approach is that it may not prove sufficient. The danger of a regular annual forum is fitting another annual meeting into an increasingly crowded annual calendar. A bi-annual meeting might thus be more appropriate, especially if coupled with consistent support for CI for environmental observing sessions at AGU, AMS, ESA, and other regular scientific meetings. Another concept worthy of exploration would be periodic large meetings that include NSF programs, ESIP participants, and other relevant projects.

Supporting projects or individual investigators to work in the space between observing efforts could help bridge gaps and stimulate the development of alternative approaches for data analysis, data management, and integration of observations and modelling. Funding a group or groups that are linked to, but separate from multiple projects, can enable development of and experimentation with innovative methods without threatening project budgets, schedules and milestones. Such an entity would also help increase the bandwidth of ongoing technical discussion and communication across projects, again without imposing significant new tasks on existing project staff. At the 2010 Forum, the last in the series, participants expressed considerable enthusiasm for the concept of an “Observational CI Synthesis Center” to promote innovation and the sharing and adaptation of CI tools and methods. As conceived during the forum, it would serve as a clearinghouse for best practices, solutions and innovations, perhaps with an “open code repository” to support sharing and repurposing, and as a virtual community gathering space where groups could convene for specific integration projects or conduct targeted “hack-a-thons” to explore alternative approaches to well-defined CI/domain problems.

Another area that is ripe for cross-observatory CI is long-term shared archival of data from NSF-funded observing systems. Differing schedules and access requirements make it difficult for current projects to co-develop shared archives. But the need for data archival and preservation extends well beyond the planned lifetime of current projects. As many observers have pointed out, observational data are essentially unique and increase in value over time. NSF is now funding the acquisition of very large and extensive data sets. The need for preservation of and easy access to these data is essentially permanent, and a shared long-term facility would be the most cost efficient means of satisfying this need, extracting the greatest possible value from the NSF investments, and providing the widest possible benefit to the larger scientific community. The NSF DataNET programs are now emerging as a potential partner in this activity.