Summary of a Discussion about Cyberinfrastructure For Environmental Observatories held on December 6 & 7 in Arlington, VA.

Introduction

On December 6th and 7th, a group of 21 researchers in the areas of computer science, ecology, hydrology, environmental engineering and ocean science met at NSF to discuss the cyberinfrastructure needs of environmental observatories and the issues associated with developing cyberinfrastructure to support observatories. A number of NSF program officers from various disciplines, together with some outside visitors, listened to all or part of the discussion. The discussion focused on the needs of four observing systems which have been the subject of much planning by environmental researchers in the last few years: the environmental observatories being planned as part of the CLEANER program, the National Ecological Observatory Network (NEON), hydrological observatories, and the Ocean Observatories Initiative (OOI). However, the discussion also considered issues relevant to a broader class of environmental observatories.

At the outset, there was a recognition of several pertinent facts that provided context for the discussions:

- There is a high degree of commonality in cyberinfrastructure needs amongst the major planned environmental observing systems.
- To maximize the return on the substantial investment of tangible resources and effort that the development and deployment of these observing systems will entail, there is a need to make sure that the data coming out of the observatories is available and usable by a diverse community of researchers and educators.
- The design and implementation of cyberinfrastructure for environmental systems require the combined efforts of both IT specialists and environmental researchers.
- It was assumed that the deployment of the four observing systems that were the main focus of this discussion would begin within a window of from 2-6 years from now.

In addition, environmental education was excluded from the scope of the workshop. Some aspects of the cyberinfrastructure needed for environmental education are discussed in a workshop report entitled, Cyberinfrastructure for Geosciences Education, and in some of the NEON planning activities.

Purpose

The spectrum of CI issues associated with environmental observatories is vast. A workshop held at Scripps Institute of Oceanography in 2002 (http://lternet.edu/sensor_report/), provided a very general review of these issues. Some additional relevant material may be found in the report of a workshop on Environmental Cyberinfrastructure that was held at Boulder in 2001 (http://www.ncar.ucar.edu/cyber/). The focus of the recent discussion was narrowed to the technological possibilities and issues associated with developing an information infrastructure that facilitates the greatest use and scientific impact of environmental observatories with particular reference to the four observatory networks listed above. The participants in the discussion were a mixture of people who are experts in the domain science and engineering research issues that the four observatories will be able to address, and people who have been pioneering the

development of integrated information and knowledge management systems for various scientific communities. The goals of the meeting were: to increase communication between environmental researchers and computer scientists, to increase communication on common CI problems that must be confronted in the development of the four observatories, to develop a better sense of what is possible and of where the difficulties in the development of environmental observatory information infrastructure are likely to lie, and to share some ideas on how to foster the development of such infrastructure.

Outline
The emphasis at this meeting was on discussion rather than presentations and this is reflected in the agenda (Appendix A). Participants were encouraged to exchange views by email prior to the meeting.

After the participants were welcomed by David Campbell, the chair of the NSF internal Working Group on Environmental Research and Education (WG-ERE) and Sangtae Kim, the Director of NSF's Division of Shared Cyberinfrastructure, the meeting continued with a series of short presentations on the four planned environmental observatory systems (OOI, NEON, environmental engineering/CLEANER, and hydrological observatories). This was followed by short presentations from three computer scientists that highlighted some of the technologies available for constructing modern scientific information infrastructures as well as some of the development issues associated with them. The morning was completed by a discussion of scenarios for how researchers might use the observatory resources and link them to other available resources such as numerical models and analysis tools.

The afternoon of the first day was taken up with a series of four further discussions led by various participants. The second morning consisted of two discussions that tried to summarize some of the points that emerged during the first day, to examine how to close the gap between observatory needs and IT tools, to consider approaches that span the observatories, and to suggest next steps.

The theme of each discussion and the main points that emerged are summarized below. Copies of the presentations on the individual observing systems and on modern approaches to information infrastructure may be found at www.orion.org and are not summarized here.

Discussion summaries
1. Observatory CI Use Cases

1) Use case: The eruption of a new vent field along an oceanic spreading center. The objectives are to be able to rapidly detect the creation of a new vent field triggered by the evolution of sub-sea floor magma, to localize it, and then to observe the geochemical, biological and physical processes that ensue. The scenario given was that, in a ocean observatory with distributed geophysical, physical and video sensors encapsulated as web services within a shore-side information infrastructure, a multi-node seismic sensor network is configured as a virtual ocean observatory (VOO) within the actual observatory to look for seismic events that might be precursors of vent formation.

This VOO detects and locates an earthquake swarm above a designated threshold, automatically sends an alert to vent monitoring software. This prompts a shore-side control center to increase the sample rate of vent sensors and cameras within a predefined distance of the epicenter of the earthquake swarm. An AUV docked at one node of the observatory is instructed by the onshore control center to survey the swarm area and find the vent plume, providing direct evidence of the eruption. The AUV returns to its home node and uploads plume and photo data. Vent and plume data are assimilated into a plume dispersal model. Based on the results of the dispersal model, proximate water column chemical and biological sensors within the actual observing system are configured as a new VOO to monitor plume dispersal. AUV assets are used to survey the new vent field.

Capabilities highlighted by this scenario included the need for bidirectional communication paths (within sensor networks, between nodes, between users and sensors) and some redundancy, both to cover multiple demands and for robustness. This was compared to the status of ocean observatories as they are currently deployed which are characterized by: static infrastructure, hardwired connections, a one-way communication path from sensor to node to shore-side infrastructure. Once observatories are fixed, it is not easy to reconfigure them but this style of architecture isn't optimal. Now ocean observatory designers are moving towards a web-service oriented, component-based architecture, using platforms and sensor clusters with multiple sensors, providing bindings that allow sensors to be combined into virtual sensors that draw on different components of different sensor clusters.

A strong design driver is the desire to study episodic, transient events: a single event that lasts for a short period of time. This needs CI to help with the detection of events and the ability to mobilize additional resources to the site of the event. This requires real-time capabilities. Other requirements are: the ability to bind together different sensors from different sensor networks, in different geographic regions; data assimilation into analytical and predictive models; adaptive sampling capabilities; mobile assets within the observatory, such as AUVs; and sufficient communications capacity. Challenges include managing data flows, making predictions based on data assimilation in sufficiently short time to be usable to guide observatory responses to events; adequate analysis capability; dynamic reallocation of power and bandwidth; scheduling; monitoring and portraying observatory status and asset availability; store-and-forward local data management capabilities within the observatory; and the overall design of such a system.

The discussion emphasized the standpoint of usability — the need to begin observatory cyberinfrastructure design by focusing on what people really want to do with the data, and how it is going to fit into their workflow. The feeling was that if one wanted to engineer a good information management infrastructure, then one needed to put considerable effort into finding out how people want to use, or might in the future want to use, the data.

2) A second use case considered the combination of ecological data integration and ecological models using workflow orchestration tools to estimate the evolution of the

global distribution of ecological niches for a particular species in response to climate variation.

To accomplish this requires that data be pulled from different sources, compiled and then dynamically fed into models. The data need to be understandable to different users. To make it usable, the interface needs to be common with easily identifiable components and functions. An example of ecological niche modeling orchestrated by the Kepler workflow system was given, implemented as part of the SEEK project. The Kepler system is developing a large user community, but the system now does not currently have long-term funding for development, maintenance and user support. As part of the discussion of this use case several needs were identified:
- The ability to harden the developing CI, and to preserve it so that it is available 10 or 20 years from now.
- Good support for remote user access, including sufficient bandwidth over the "last mile".
- Widely adopted metadata standards, especially after the deployment of complex arrays of sensors.
- The need to have leverage with industry to come up with standards on sensors. The need to link the existing sensors to metadata.
- Smart sensor networks with increased intelligence embedded in sensors and sensor platforms.
- Long-term archives—for some types of data there are already national and international archives of data; however, there are a number of communities without such long-term archives. This is important for environmental data which are often used to examine processes evolving over ranges of time scales that include very long periods and for which new uses often become apparent years after they were originally collected.
- Better capabilities for semantic mediation of data sets.
- Increased usability of data and tools. Most of the current data is available to high-level users who have IT savvy (or support staff) to work out how to access and transform it, but it needs to be available for use by the broader stakeholder community.
- Approaches to data management that facilitate interoperability between data collections and observatory operated by different disciplines.

A number of points were brought out in the ensuing discussion.

- Some participants felt that a significant question was — what data do you throw away and what do you keep? Others felt strongly that as much raw data as possible should be archived.
- There is a lot of emphasis on data collection and not as much on maintaining or distributing it. In terms of resources, an 80/20 ratio between expenditure on data collection and expenditure on data management is probably the appropriate balance. In new observing efforts, there is sometimes a tendency to say "we don't have data yet so let's not worry about data management yet," but one needs to have information infrastructure already in place when the data begin to flow. This implies a need to put energy and resources into information management issues early in the process.

- The integration of environmental data from multiple sources is needed to answer the more complex questions about the environmental dynamics that researchers are asking. Currently there are often incompatibilities between the ways that data are collected and stored in different communities. An awareness of the different ways that the data were collected becomes a prerequisite for their use. Approaches to data integration should be more closely based on the types of science question that people wish to address with environmental data.
- How are we going to realize an environmental information infrastructure that encompasses the different environmental observatories?
- Human complexity will also be a challenge.
- Environmental data integration must go beyond linking together a set of measurements — researchers also need tools that can provide context for data.
- A process should be developed for scoping cyberinfrastructure needs for environmental observatories – a requirements analysis that includes a survey of what capabilities researchers need for different classes of research questions, which of these can be supplied by the tools available today, and what additional tools are needed. A requirements analysis should to be done by first by the different components of the environmental research community, then for environmental research as a whole, and then with the computer scientists. (The NEON community is already doing something similar to this.)
- Requirements analysis efforts can be hampered by the fact that the communities involved are diverse. There is need to build a system of requirements analysis that is flexible enough to be used by people multiple communities. Each observatory community will need to include an assessment of what is needed to make observatory data available for use by researchers in other fields. Some people may say that since you can't anticipate everything, so why not just concentrate in your own field's needs. Although one can't anticipate everything, it's still important to go through this multi-dimensional requirement setting process.
- What problems in information management and integration are simply a question of deploying existing cyberinfrastructure solutions and which are hard problems in computer science? This is an important question to be addressed in requirements scoping. In answer to the question, is it the volume or the complexity of environmental data that is the big issue? The consensus was that it is both. The computer scientists felt that in the face of these real world challenges it isn't clear that computer science solutions actually work in the real world. (This was one of the motivations behind the recommendation that realistic test-beds be developed – see below.)
- NSF should maintain a funding stream for supporting, and updating the infrastructure. This will require thinking "globally" and abstractly about environmental research instead of acting "locally" & specifically to meet the needs of a specific research question.
- Ideally, one should meet the needs of a broader set of communities by generalizing from robust operational solutions developed to meet the needs of a particular community, not from conceptual computer science solutions, otherwise

each community repeats the often substantial expense of translating from a conceptual solution to an operational solution.
- Interagency cooperation needs to be addressed as well to preserve, to maintain and promote access to and integration of environmental data, and to avoid duplication of effort.

2. Networking and security

Phil Papadopoulos identified a number of issues, made a number of suggestions, and asked some provocative questions to open the discussion of networking and security issues. These included:

- The future is very high-bandwidth network connections with parallel resources at end-points.
- EOs should design for hardware that will be available in 5 years and should expect 10Gbs optical or high-bandwidth wireless connections to be possible to fixed instruments in many settings.
- Public-key based certificate systems appear to be the most appropriate basis for security. Certificates can be used to identify people, instruments, and other resources. However, explicit certificate management by users is untenable. An emerging common solution is a grid certificate bank and proxies. EOs can use the same sort of solutions. However, there are still difficulties that need to be sorted out. It is easy to build certificate authorities (CAs) but the challenge is in establishing trust between CAs. Who should be the master CA for EOs or other NSF-supported activities? What happens when a user has multiple identities?
- Identification by itself is not enough. It establishes "who", but not "what resources may the person/instrument use." It's going to be essential to have role-based authorization but there is dearth of tools in this area.
- Another aspect of security is data integrity. How does one validate data in an archive and cope with "bit rot" associated with the storage of data on impermanent media? How does one validate data coming from sensors? How does one provide data provenance for derived data?
- For true data security, end-to-end encryption is needed, together with audit trails. How does one watermark digital data?

In the ensuing discussion, people made a variety of points related to security and also embarked on a discourse on data storage. Some of the points were:

- Using lots of different software together, some open, some not, introduces data integrity challenges.
- One approach is to keep multiple distributed copies of data and periodically check their digital signatures.
- In some fields, e.g. biomedical, it may be important to ensure that general users cannot access audit information or otherwise determine who has accessed what in a data grid. EO networks are potentially different from existing data grids in their scale, complexity, their dynamic structure, their heterogeneity, and fact that they can include embedded systems.

- To balance the coexistence of science goals and security, we need basic research on security in ad hoc networks and tools like lightweight cryptography.
- How does one balance security and accessibility? Who is responsible for providing security packages for open source software? Will grid computing create new vulnerabilities for security?
- It was noted that some of these issues affected many uses of cyberinfrastructure and hence were being actively examined. One such effort is the "Liberty Alliance" aimed at finding broadly usable solutions to the authentication process. Some wondered when tools like ssh would be Liberty-enabled. Some expected the use of biometric identification and single-use passwords to become the norm and noted NIST's work on developing standards in these areas. EOs should be consumers of security technology not innovators.
- There was a suggestion that a cyber-threat analysis was needed for EOs. It was pointed out that the likelihood of attack should be assumed to be 100%.

There was an active discussion on just what data needed to be preserved with several differing points of view. One participant thought that there was no need to keep all raw data, that the total life-cycle costs of data are large, and that instead we should strive to come up with advanced QC and data aggregation techniques. However, others pointed out that, for many types of data, one cannot predict how those data might be used in the future and cited examples of how the subsequent development of new processing algorithms or new science questions greatly enhanced the utility of raw data collected previously. A number felt that raw data is the most important data to keep since derived data can always be regenerated provided the processing algorithms are preserved. One participant noted that compared to the expense of building, operating and using some instruments, keeping data is relatively cheap. The observation was made that, in EOs, the volume of data is not growing exponentially but the ability to store them is.

The group observed that it is easier to invest in technology and machines than in data storage. Investing in new, powerful ways of collecting data is attractive because of the potential for new scientific insights. Often not all of the data collected is examined in detail. What are the appropriate mechanisms for determining the optimum balance of investments between collecting new data, and preserving and analyzing data already collected.

On the issue of strategy for the deployment of networking and other cyberinfrastructure, there was a recommendation that for infrastructure in remote or hard-to-reach locations that are expensive to revisit, one should sink the best assets one can; however, for the rest of an observing system it might be more cost-effective to use just-in-time acquisition. Just-in-time acquisition also enhances risk. A later generation of CI may break something and some examples from the grid environment were cited. Some of this risk could be reduced by using a modular design structure which, while it might not be the most efficient approach, might be more flexible and diminish risk.

And finally, it was thought to be important to demonstrate the value of a security measure to the pursuit of science. Email was suggested as an example of a very valuable class of

software that was relatively insecure and that would probably be less useful if made completely secure.

3. Metadata, semantic information, and ontologies

To help researchers make effective use of data from EOs, there is a need for standard, machine readable, and semantically rich descriptions of data and other resources (like QC or other data processing routines and models) to support machine-based inferencing and reasoning – metadata standards and ontologies. For environmental systems, these should be built by the environmental researcher communities. Metadata standards capture how researchers think about and use data. Ontologies are needed to make the link between scientists' conceptual space and the data themselves. Together they support the use of data by broad communities of users who may not have detailed knowledge of how the data were collected. Hundreds of tables in a relational database is not a data model that is very meaningful to many users. Metadata permits users to find and manipulate data in such a database without needing to understand the internal database structure. Metadata and ontologies are investments in data usability and reusability. It is sometimes useful to keep metadata on the absence of data (for example, so that a data management system can tell when a request is for something that isn't there).

Ontologies are organic things that need to grow as concepts change so it is important to use a layered approach to systems of ontologies for a community; e.g. core/base ontology, community ontologies, personal ontologies. SWEET was given as an Earth science example. The top-down approach to ontologies has been tried and found not to work. Letting people add to ontologies in a loose fashion doesn't work either because people forget to think about general structural issues important to the usability of the ontology.

One approach to ontology development is to get input for preliminary development through a workshop process, put out a first version and then get people to augment the concept space. Publishing a base ontology and an interface to the ontology would help people develop the ontology further. However, it is critical to get ontology-building experts to guide the process and essential to have real data available to help focus and test the ontology. It is also helpful to have a clear purpose in mind for the ontology to guide its design. SWEET and UMLS are examples of things that are very large but are not built for a very targeted focus. Some prior attempts at community-based development of ontologies got too bogged down in top-down approach.

The group wondered whether there is a generic component of environmental ontologies that would go across the 4 EOs? It was pointed out that for managing large information infrastructure with many users, it had been found to be useful to have a policy ontology defining concepts of authentication and usage.

In general, it is important to capture experimental context into experimental data holdings. There is a need to automate the capture of metadata, especially metadata about processing pipelines – Kepler does this to some extent.

In developing ontologies, a big initial cultural barrier is that between domain scientists and IT folks. In similar efforts, in other areas, it has been found that, while these barriers can be initially quite high, they come down as IT and domain folks start working together.

Someone asked the question of how one might get scientists to use metadata, particularly to generate it. In response one participant cited examples of activities in which, at first there were questions about why people should put their own experimental notes into metadata for posterity. The developers provided a tool to generate lab reports from metadata that made users lives easier and hence was adopted.

4. Resource registration, collections management and workflow tools,
Much of the ensuing discussion centered on workflows. Inter alia, scientists are looking for tools that allow them to do more easily things they already know how to do. Workflow orchestration systems are examples of just such tools. Useful examples of potential applications are the automatic collection of data in synchronization with a satellite pass and situations in which one instrument, or some processed measured quantity, triggers the operation of another instrument. Several areas in which R&D on workflows is needed were described:

- An issue for observing networks is coordination of the measurement process. For some networks, turning one thing on may affect another. This gets more challenging when one has instruments from multiple investigators on the same network node. If implementing workflows then one can (and should) incorporate the appropriate checks and perform conflict resolution.
- In the BIRN collaboration environment, one can log on and look up workflows and either use archived workflows, modify them, or devise ones own.
- The idea of using Kepler as an operational tool is a good one. However, further development is needed. It doesn't have language to represent pre & post-conditions and ways of characterizing things like proximity and potential for interaction between instruments are needed. Ptolemy has choices of synchronization that might permit trying to extend workflows to time and there are sophisticated scheduling systems like LabView. One question that needs to be addressed is how to take advantage of scheduling systems and integrate them into workflow tools. There is also a need for recursion in workflow orchestration and the ability to shift workflow execution closer to the instrument level.
- One current research question is how to increase the flexibility possible in inputs to and outputs from workflows.

5. Collaborative environments
Barbara Minsker made a presentation on collaborative environments, describing NCSA's vision in this area and citing the electronic notebooks used by NEES investigators as an example. She noted that many web-based community portals work well because they have a limited set of tasks that can be centralized. For environments of larger scope like EOs with 100s and 1000s of users, web-based central portals might not be sufficient. Instead distributed portals and distributed collections of tools supplied by them may be

the way to go. Dr. Minsker cautioned that as groups get larger, it is harder to produce collaborative tools that work as well as what scientists already do. She felt that, to be successful, CI development efforts in this area must involve

- prototypes subjected to rigorous testing
- using a user-centered design methodology
- early adopters must be supported by IT developers with good communication skills
- systems must scale to many users and be highly reliable and highly usable
- the IT part must have scheduling, project management and communication that support rather than hinder physical infrastructure deployments
- Some social issues are important for success:- e.g. ensuring professional recognition for work done in developing CI.

In the ensuing discussion, people talked about a variety of collaborative environments.

- R and Matlab are powerful environments with which researchers are familiar and it would be better to augment them rather than trying to replace them, even if this approach is inelegant from a computer science perspective.
- Accessibility to HEC is an example of the need for a better collaborative environment. Many people need HEC and don't use it because of barriers to entry, e.g. the cost of learning how to access centers, learning how to adapt algorithms to highly parallel architectures, queuing delays).
- Even though now [system XYZ] is very effective, it was difficult to use when it was first introduced. This produced a negative backlash which built a barrier to later adoption by users. Moral: should only let prototype code be used by friendly users.

There were some comments on software development projects in general that were intended to make sure that unrealistic expectations were not created for EO CI. These included: A software project is seldom under-budget and early. We still don't have good ways of estimating the costs of developing software, making it robust, and maintaining it. It was felt that industrial software development was a better model for the development of EO CI than that typical in academic software development.

6. Closing the gap between observatory needs and IT tools & approaches that span the observatories

The discussion was really an attempt to identify crucial gaps and, in some cases, ideas about how to bridge them. It included some general observations and a strong feeling that the development of cyberinfrastructure for environmental observatories would be much more successful if tackled collaboratively across the observatories.

The general remarks included:

- Observatories can change the grand vision of how environmental researchers approach science. With them, researchers can ask and answer different questions. One sees examples of this happening in astronomy – changes in the way data is available and how people can interact with them are changing how they ask questions.

- It is important to think in terms of dealing with objects – not raw data but data + methods. We need to ensure long-term preservation of both data and methods.
- There is a critical need to agree on data object models for environmental data (e.g. OpenGIS, which some expect to be around for a while). A common vocabulary is needed for base data models so that data can be easily exchanged between them and so that it is possible for one person's measurement to be consumed by others in different fields.
- One of the unsolved challenges in data fusion is working in space-time. It would be useful to attempt to bring together different ways of representing spatial and spatio-temporal data; for example, working out how to combine data stored in HDF, NetCDF and GIS data models.
- NSF should encourage PIs to publish data as they publish research. And NSF should fund such data publication. At the moment there are few good general methods to publish data and more development is needed.
- Scalability of infrastructure is something that has to be designed for from the beginning.
- The complexity of software development should not be underestimated.
- Working within fixed budget is different from an Apollo program.
- EO CI also has to deal with data coming from other sources and simulations outside of the EOs themselves.
- It is not clear that an effective funding model for EO CI is yet in place.

Points raised in the discussion of the need to work together were:
Data management proposals are often not very attractive in the peer review process. A key is to articulate a big vision of what the community would like to do but can't currently do – a stretch goal e.g. knit together all environmental observing systems.

There is a growing competition between people with IT tools to offer. Who are scientists and engineers to listen to?

There isn't enough justification for having many similar efforts – NSF needs to support the coalescence of groups and encourage them to pursue a less self-interested approach. One participant from industry made the analogy to how proposals for software development projects were evaluated in a particular large company, in terms of market risk, technical risk and organizational risk. His evaluation of the environmental research communities' readiness to develop EO CI was that the signals for market risk (will there be user demand) were green, for technical risk, yellow, but for organizational risk (having the necessary collaborative team with the appropriate skills whose members are working together effectively towards a well understood common goal), the signals were flashing red.

One of the most significant gaps identified was the lack of a roadmap for the development of EO CI. It was recommended that the roadmap start with a minimal viable project (a lifeboat project) and then expand from there in a spiral development approach rather than trying to start with a complete design.

Data integration is not something that can be done effectively by environmental scientists working alone. They need to form partnerships with computer scientists. Similarly computer scientists cannot develop data integration alone. Involvement of the environmental research community is crucial if the results are to meet the needs of environmental research and be usable and robust. The development of such partnerships involve a number of social issues that must be successfully addressed: the partnership must be win-win for all partners, including professional recognition in areas that typically use different metrics for such recognition. Teamwork is needed to ensure end-to-end functionality and effective CI but the development must also involve creative challenges for those involved with the team.

To encourage communities to come together to jointly develop EO CI, it would be helpful to have scientific visions that cross the different observing systems.

ITR has been transformative in fostering interdisciplinary collaborations to develop cyberinfrastructure tools. There is a continuing need for things like ITR. It is also necessary to address usability – e.g. the hardening of tools developed under ITR.

Some suggestions for how to move forward were described. [Some of these are also discussed in Section 7.]

- Common shared test-beds – These permit development to move forward; provide tangible success stories that can convince skeptical researchers of the value of EO CI; help overcome the cultural barriers between different communities; make it easier to formulate shared visions for environmental observatory science.
- More interdisciplinary community discussions meetings – In particular a joint observing system workshop to explore the shared visions of the environmental research communities.
- To move forward, there need to be opportunities for communities to work together.

7. The path forward

The discussion focused on identifying two types of activities: "low-hanging fruit" – short-term, relatively inexpensive activities with a high potential impact, and "higher-hanging fruit" – activities that would require more time and resources.

Low-hanging fruit

- Begin a joint requirements analysis exercise for the four observatories. (Here the focus is on CI but similar processes could be applied to other observatory technologies, such as sensors, too.)
- Institute regular meetings between observatory planning and development groups.
- Exchange information on best practices.
- Enhance communication to ensure interoperability.

- Glue together, through some supplementary funding mechanisms, the observatory IT teams.
- Initiate projects focused on integrating different technologies in a way that would be applicable to a broad range of environmental research questions but that is guided by a specific question – e.g. a research problem built around HDF/NetCDF/GIS integration, or workflows.
- Develop a shared identity management system that spans communities, across observatory projects.
- Joint work on tools to collect metadata about shared data objects – e.g. extending Kepler to other domains.
- Research projects that demonstrate value of cross-over between EOs.
- Projects that help broaden the understanding of the power of technology development.
- Develop an NSF-sponsored interoperability working group (c.f. IETF).
- Develop consensus on larger transformative science questions that cross EOs. Could also try to develop a "grand vision" for environmental observatories, e.g through workshops.
- A testbed to demonstrate remote collaboration and communication.
- Assessment of data curation needs and the development of plans to meet these.
- Conduct an inventory of available tools and approaches to observatory CI:
    - Need to get a better sense of what is going on in industry,
    - Need to make an inventory of what assets we have (funded activities) and how leverageable they may be.
- Explore developing NSF data & CI policies that promote usability and interoperability – e.g. requirement that data from observatory A should be consumable by observatory B.

Higher-hanging fruit

(The group cautioned that it is important to connect these to the user community to ensure that CI leads to the ability to solve science and engineering problems that the researchers care about.)

- Support a test-bed with specific science goals to address a particular spatial location where one demonstrates the ability to recover all relevant ecosystem, hydrological and environmental engineering information and to integrate these data in the context of ecosystem, hydrological and environmental engineering models.
- Support a similar test-bed that links ecological, hydrological and environmental observations in and around an estuary and watershed to a coastal ocean observatory.
- Projects that are large-scale prototype CI. The rationale for this is that building out helps design process. One area of emphasis could be on prototypes in the data analysis area (with the ability to apply statistical models to data from multiple sources). This topic naturally couples to workflow. The recommended approach is to begin by using existing software and then to find out what else is needed.

- Prototypes of how to bridge CI development for different disciplines but in the context of existing large CI projects.
- Concrete data integration demonstrations within existing projects. The group felt that data integration across fields is a big challenge.
- Peer-to-peer federation of databases and interface development for standardizing tools like GIS and sensors/instruments. This could include work with industry and professional societies.

Related comments included:

- Existing large information integration projects like GEON have had a strong coordinating IT center. For test-beds to succeed they may need something similar and this is not cheap. One suggestion for reducing costs was to try and nucleate a coordinating center for all of the EOs CI.
- There is a need to look at how to manage shared resources, e.g. accounting.
- The four EOs should have one site in common.
- Education and outreach specialists are concerned, amongst other things, about how to best get information out to the general public and get them engaged. There exist both low and high fruit here and education specialists should be involved at an early stage.
- Need to include modeling & assimilation in observatory CI to extend observations to unobserved regions.
- Should use models to "train" data management and archiving processes and to design observatories.
- Support the integration of EO CI development so that it supports instantiation of virtual observatories that cross boundaries between physical observatories
- Environmental observing system science groups should meet to explore the following questions:
    - What are the grand cross-cutting science challenges and visions?
    - Where does it make sense to take a common approach to aspects of EO CI?

Appendix A: Agenda

Discussion Group on Planning for Cyberinfrastructure For Environmental Observatories: December 6 & 7, Arlington, VA (Room 375, NSF)

December 6

8:30 Light refreshments

8:45 Welcome and NSF Perspective — Dave Campbell & Sangtae Kim

9:00 Charge/Agenda — Coordinating group

9:15 **10 minute briefings on OOI, CLEANER, HOI, NEON**
Bob Detrick, Jim Bonner, Jon Duncan, Bill Michener
Very brief overview of scope of each observing system

10:05 Discussion

10:20-10:30 Break

10:30 **Introduction to CI topics**
Chaitan Baru, Bertram Ludaescher, Michael Welge

11:10 **Observatory CI Use Cases**
LEAD: Bill Michener & Bob Detrick
RESPONDENTS: Cherri Pancake & Matt Jones

12:30-1:30 Working Lunch

1:30 **Observatory CI Challenge 1: Networking and security**
LEAD: Phil Papadopoulos
RESPONDENTS: David Maidment & Art Sanderson

2:30 **Observatory CI Challenge 2: Metadata, semantic information, and ontologies.**
LEAD: Danielle Forsyth
RESPONDENTS: Jim Bonner & Bertram Ludaescher

3:30 Break

4:00 **Observatory CI Challenge 3: Resource registration, collections management and workflow**
LEAD: Chaitan Baru
RESPONDENTS: John Helly, Leonard Krishtalka

5:00 **Observatory CI Challenge 4: Collaborative environments**
LEAD: Barbara Minsker
RESPONDENTS: Matt Jones

5:30 Summary, what have we neglected, planning for homework.

6:00 Break for dinner (Neighboring restaurant)

December 7

8:30 Light refreshments

9:00 **Closing the gap between observatory needs and IT tools & approaches that span the observatories.**

LEAD: Jim Gray

RESPONDENTS: Miron Livny, Mark Ellisman, Jeff Burch

What software tools/approaches are common to different environmental observatories? Of the necessary tools, which exist now? What are the obstacles to their adoption? What tools would be helpful if they had a little more work? What tools are missing? How much of the design and implementation of information infrastructure for environmental observatories can be shared across observatories?

10:30 Break

10:45 **The path forward.**

LEAD: Bill Michener, Art Sanderson, Yi Chao, Jon Duncan, Michael Welge, Chaitan Baru, David Maidment

What should the communities do next, what should NSF do next?

12:00 Adjourn