SCALABLE INFORMATION NETWORKS for the ENVIRONMENT

SINE

REPORT OF AN NSF-SPONSORED WORKSHOP
SAN DIEGO SUPERCOMPUTER CENTER
OCTOBER 29-31, 2001

ALISON WITHEY
San Diego Supercomputer Center
University of California, San Diego

WILLIAM MICHENER
Long Term Ecological Research Network Office
University of New Mexico

PAUL TOOBY
San Diego Supercomputer Center
University of California, San Diego

SINE Workshop

This workshop was supported by the National Science Foundation under grant DBI-0120071. All opinions, findings, conclusions, and recommendations in any material resulting from this workshop are those of the workshop participants and do not necessarily reflect the views of the National Science Foundation.

Withey, A., W. Michener, and P. Tooby (Editors), 2002. Scalable Information Networks for the Environment (SINE). Report of an NSF-sponsored workshop (San Diego Supercomputer Center, October 29-31, 2001). 65 pp.

Graphic design and page layout by Benjamin Tolo.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

SENSOR NETWORKS

DATA TECHNOLOGIES

SCALABLE INFORMATION NETWORKS FOR THE ENVIRONMENT

BIBLIOGRAPHY

APPENDICES

Scalable Information Networks for the Environment

EXECUTIVE SUMMARY

Findings and Recommendations of the SINE Workshop

Alison Withey
San Diego Supercomputer Center
University of California, San Diego

William Michener
Long Term Ecological Research Network Office
University of New Mexico

Workshop Overview

An NSF-sponsored workshop on Scalable Information Networks for the Environment (SINE) was hosted by the Partnership for Biodiversity Informatics (PBI) from October 29-31, 2001 at the San Diego Supercomputer Center. The SINE Workshop was attended by a diverse group of research scientists, directors of field stations and marine laboratories, and experts in the computational and information sciences that met to discuss the requirements for building advanced environmental networks. These networks, designed to deliver continuous, integrated high-quality data in real or near real time, must be scalable from local to regional and national levels. A multidisciplinary approach, as reflected in diversity of disciplines represented by workshop participants, is seen as essential to resolving the interrelated technical, discipline, and social challenges to building scalable environmental networks.

Important opportunities exist for understanding the Earth system in its full complexity through the application of emerging technologies that can improve data management and delivery; enhance modeling and prediction capabilities; and facilitate communication among environmental sensors, databases, and scientists. This workshop is a first attempt to outline a *scalable national environmental information infrastructure* that meets the needs of scientists working at both local and broader scales, as well as decision-makers, educators, and other stakeholders who require comprehensive environmental information.

Workshop presentations and working group sessions focused on three topics:

- **Sensor Networks:** Building distributed sensor networks, including design and implementation issues.
- **Data Technologies:** Enabling technologies and user requirements for data and information management and delivery.

- **Scalable Information Networks for the Environment:** Scaling components of environmental information networks including data, computers, and people.

Information about the SINE workshop (including PowerPoint presentations) and the Partnership for Biodiversity Informatics (PBI) can be found at www.sdsc.edu/pbi. The complete workshop report is posted as a downloadable PDF file. A limited number of printed copies are available upon request.

Recommendations for Infrastructure Development

1. Data repositories and IT infrastructure: There is an urgent need to establish long-term, stable data repositories and IT infrastructure, including, as examples, integrated distributed archives, data centers, clearinghouses, and other facilities that institutionalize public-domain availability of data holdings.

Scientists, scientific societies, and funding agencies will benefit from partnering in the establishment of *best data management practices,* developing policies that promote data sharing, and creating a national repository for biodiversity and ecological data.

2. Interdisciplinary research: There is a need to improve support for interdisciplinary research that fosters the development of tools and technologies that (a) overcome the significant challenges associated with the extreme heterogeneity of environmental data, and (b) meet the needs of the wide range of users of environmental data. Emphasis should be placed on developing appropriate data and metadata standards.

As an example, progress in geospatial data integration is limited by the lack of interoperability among GIS/cartographic, database, knowledge representation, and visualization data structures, as well as the paucity of comprehensive (nationwide coverage) and interoperable environmental databases (e.g., National Wetlands Inventory, 24K National Hydrological Data, National Vegetation Map, and rare species databases) and the difficulty of discovering critical databases. Workshop participants expressed concern that the length and complexity of the FGDC Geospatial and Biological Metadata specifications may be inhibiting investigators from developing and publishing adequate metadata for environmental data sets. One solution may be *tiered metadata systems* that are better integrated with W3C/RDF technologies and are designed to facilitate use by clearinghouses and information discovery tools.

Continental-scale studies will, at least in part, be based on bringing together information from existing, major regional efforts. Thus, it will be most effective to identify common data, metadata, and other standards that will piggyback on existing standards and conventions, in order to arrive at a "common denominator" for continental-scale studies. It is also important to consider how sensor networks will be deployed at the continental scale. For example, should sensors in the U.S. be distributed uniformly, or in "representative" regions/ecosystems? IT approaches must be able to deal with increased heterogeneity in data formats, metadata schemas, and data quality at the continental scale.

3. Data infrastructure and communication systems: There is a critical need to build capacity in field station, marine laboratory, and shipboard data infrastructure and communication systems. This will yield significant near-term benefits for the scientific research community and help to lay the foundation for developing standards for instrumenting the environment and managing data networks on a larger scale.

4. R&D test beds: There is a need to develop environmental sensor R&D test beds in which new environmental sensor technologies and associated data or network architectures can be deployed and tested. Efforts should focus on research in distributed, self-configuring environmental sensor networks and on developing standards for sensors, platforms, and user interfaces. There is a specific need for self-describing, autonomous sensors that can report their measurements to a data acquisition system (e.g. network) with minimal operator intervention, and that can interoperate with other sensors and data systems in terms of adaptive routing, metadata-based services (such as the existence and status of any given sensor), operating status, location and similar housekeeping functions including reprogramming.

Sensor design and distribution will be driven by a series of parameters determined by the scientific question under consideration. Parameters include but are not limited to: cost; whether data collection is continuous or event driven; spatial and temporal scaling to include interval and extent; whether the data stream is real time; requirements for data reliability, redundancy, and format; whether physical samples must be collected; and the need for QA/QC measures and recalibration.

The design of sensor networks must accommodate investigation of a wide variety of scientific questions, while establishing generic protocols for information sharing among different sensors, networks, and users. Sensor networks need to incorporate *flexibility* in the design of sensor grids and *standardization* in the architecture of information exchange. The balance between flexibility and standardization is an important focus for future investigations. Standardization will both drive down the costs of sensor deployment and ease the integration of sensors and data over space. Clusters of specialized micro-sensors deployed on standard platforms across landscapes will provide the infrastructure needed to build scalable environmental information networks. With the advent of wireless interfaces, sensor clusters will provide bidirectional communication between sensors and users via Internet, without the expense of wired infrastructure. Costs, power requirements, and lack of standardization are the biggest obstacles to building scalable environmental sensor networks.

Sensor networks should be of recursive design, with data collection components repeated for communication and storage. Although there is no single sensor that addresses the diversity of scientific needs, regionalization efforts will be facilitated by the development of Universal Sensor platforms (i.e., incorporating plug and play sensors that address specific questions). The basic unit of the sensor network needs to have a physical layer that interacts with the environment to be measured, recursive storage and node processing, communication among components, and the capacity to change sampling parameters through a *sensor query language*. Networks of these basic units need to incorporate derived processing (detection, identification, and extraction); aggregation mechanisms; information management and archiving capacities; and internetworking. Thus, there is both a logical and a physical change in structure between the *in situ* network and the derived information products to be managed and distributed.

The communication infrastructure is a key constraint on network development since expendable and recoverable sensors in the environment have a high probability of failure due to environmental conditions. The ability to obtain data from *in situ* sensors, "pop-up" platforms (including UAVs or surface drifters) and communications/data pods released from various platforms, requires communications that are reliable, inexpensive, and global. A comprehensive study of what will constitute a sufficient communications architecture is required to enable interoperation among the different and demanding requirements of the rich diversity of terrestrial as well as freshwater, inshore/nearshore/offshore, and surface/submarine environments.

5. Building environmental Knowledge Environments: Knowledge environments represent scientific information and knowledge, including both data and the results of analysis and modeling, in a formal, highly interoperable framework. Creating such environments, which do not yet exist for environmental science, will significantly accelerate scientific research by enabling:

- Researchers to easily and quickly comprehend the context of scientific findings.
- Researchers to more effectively collaborate across disciplines by understanding the semantic

differences among information sources, and integrating these sources.

- The process of science to be captured and represented so that researchers can replicate and elaborate on previous work. Capturing the entire scientific process allows efficient reuse of both data and processing, and will be made possible by new knowledge-integration technologies in conjunction with a substantial cultural shift to a broader view by scientists of their responsibilities for communication and collaboration.

Large, complex data spaces that span the diverse information needed for environmental science will require new techniques for querying, browsing, and visualization. Query systems need to address the extreme heterogeneity of environmental data (e.g. from population ecology to climate to oceanography), including the extreme heterogeneity in syntax, schema, and semantics within subdisciplines. Browsing capabilities based on automated feature extraction and data mining need to be provided for quickly locating information of interest in the complex information landscape. Both query and browsing need to accommodate the distributed nature of environmental information as well as larger, centralized archives. Visualization needs to adapt to the complexity of information and address the differing needs of domain scientists as well as policymakers, educators, students, the news media, and other communities. This includes, for example, the ability to communicate the degree and implications of uncertainty in knowledge when expressing highly refined models of the environment for use in policy-setting situations.

The vast majority of environmental data now collected is still not being captured in a way that makes it available for the analysis of regional and continental scale issues. The infrastructure targeted at environmental data management, communication, and integration at national scales needs fundamental improvements. These include developing resources for building sensor networks for biological systems, automating data acquisition for biological parameters, facilitating easy movement of data and information products among field stations and universities, and creating an integrated national system for accessing all environmental data. A *national environmental data system* will be an important component of such a system, and will include federated access to all of the nation's distributed environmental data sources (including metadata and data) as well as important archival features for preserving data for long-term research.

Recommendations for Education

IT Education is needed at all levels in the environmental community. The degree of IT sophistication "in the trenches" is far below the cutting edge. Today, *data literacy* needs to be a component of every scientist's education. To enable interdisciplinary collaboration among environmental subdisciplines and rapidly-changing IT fields, sustained outreach and continuing/informal education are essential. Funding opportunities should encourage the development of expert advice centers, teaching workshops, distance-learning curricula, interdisciplinary graduate and undergraduate programs, outreach, etc.

Recommendations for Policy

1. Open availability of data: Proactive efforts by NSF - as well as other government agencies, academic institutions, and professional societies that support environmental research - are needed to encourage and enforce open availability of the data created through research.

Mechanisms that should be considered for promoting data sharing include:

(a) **Agency incentives for data sharing:** Using conditions and incentives in research grants and contracts as mechanisms to ensure that research data are made available to the public in a timely way. Financial incentives from research funding agencies can enable adequate attention to be devoted to data management, archiving, and access within the context of individual projects.

(b) **Legal mechanisms for open data availability:** Development in the university community of new legal mechanisms to promote open data availability. Examples of such new legal approaches include general public licenses, copyleft, and data easements.

(c) **Professional rewards/incentives for data management and data publication:** Development of a professional reward/incentive system for data management and data publication activities, especially led by professional societies such as ESA, AIBS, ASLO, etc. This should be accompanied by improved support for electronic journals and clearinghouses.

(d) **Code of ethics:** Development of a code of ethics for data access and use.

2. Public spectrum availability: A reevaluation of FCC guidelines with an eye to making available greater capacity for the environmental data infrastructure. This includes a review of FCC regulations on bandwidth to meet the critical need for public spectrum availability for sensor networks and other scientific uses.

SENSOR NETWORKS

The Design and Implementation of Aquatic and Marine Sensor Networks

John Orcutt, Facilitator
Cecil H and Ida M. Green Institute of Geophysics and Planetary Physics
Scripps Institution of Oceanography
University of California, San Diego

John Helly, Reporter
San Diego Supercomputer Center
University of California, San Diego

Initial Questions

- What are the most promising recent developments in aquatic and marine sensor networks?
- What critical components of the aquatic and marine environments are not adequately sensed with current technologies?
- What R&D activities are necessary for the environmental sciences community to capitalize on the capabilities of aquatic and marine sensor networks?

The Marine and Aquatic Breakout Group discussed these questions in some detail, as outlined below. The first portion of the breakout session was devoted to identifying what makes the marine and aquatic environment unique. Generally, it was agreed that the aquatic environment and especially the marine environment is a highly challenging place to work. Problems not encountered elsewhere to the same degree include fouling, a corrosive environment, high pressures, expensive access, and inclement weather. At the same time, the marine environment comprises more than 70% of the Earth's surface and is integral to some of our most critical environmental problems.

What are the most promising recent developments in aquatic and marine sensor networks?

A variety of new observational systems are being deployed in the oceans and nearshore environment, including:

- LEO-15: off the New Jersey coast
- ARGO: global, upper-ocean temperature sampling
- GoMOOS: Gulf of Maine Ocean Observing System

Planned projects include:

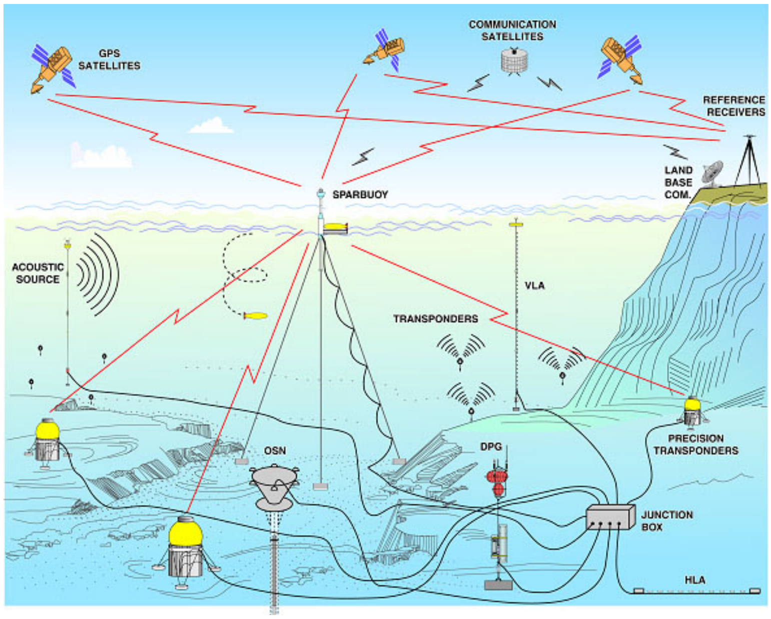
- MOOS: Monterey Bay Ocean Observing System
- DOES: Dynamics of Earth and Ocean Systems, including NEPTUNE and global moorings
- SURANet: Southwestern US coastal network

The development of these marine and coastal observatories has been made possible by a number of technological developments including the miniaturization of electronics and sensors, the rapid development by industry of remotely operated vehicles (ROV), reliable underwater connectors for both electrical and optical connections, continuous advances in underwater housings, and new approaches to communications.

What critical components of the aquatic and marine environments are not adequately sensed with current technologies?
Communications is a critical issue for aquatic and especially marine observations. While commercial systems in some cases provide excellent options for nearshore and inland use (e.g. cell phone coverage and wireless networks), marine observations do not enjoy the same commercial drivers that make terrestrial communications possible. Examples of communications systems presently used in the marine environment include Service ARGOS (France), Iridium (Private/DoD), GlobalStar, and Inmarsat. The longevity of System ARGOS cannot be assumed, the original Iridium provider went bankrupt, and funds are not yet available to maintain the satellite constellation over the long term. GlobaStar is really useful only near land and Inmarsat requires a directional antenna and is quite expensive.

The major issues for marine telecommunications include: longevity; bandwidth; directional antennae, which require tracking and stabilization; "store and forward" systems; latency; and duplex communications.

Agencies are generally more excited about the initiation of new measurements than maintenance over the long-term. For marine and aquatic observations, the infrastructure necessarily includes maintenance of ships and ROVs, as well as the sustained funding of qualified personnel. Critical long-term observations, necessary to answer important questions from climate change to species management, cannot be made without both a major, up-front investment and sustained maintenance.



"As part of its ongoing activities in both the coastal and open oceans, NSF's Division of Ocean Sciences has been working with the academic community to develop an Ocean Observatories Initiative. The effort would provide basic infrastructure for a new way of gaining access to the oceans, by starting to build a network of ocean observatories that would facilitate the collection of long time-series data streams needed to understand the dynamics of biological, chemical, geological and physical processes. Just as NSF supports the academic research vessel fleet for the spatial exploration of our oceans, the system of observatories provided for by the Ocean Observatories Initiative would facilitate the 'temporal' exploration of our oceans."

Testimony of Dr. Rita R. Colwell
Director, National Science Foundation
Before the House Committees on Resources and Science
Hearing on Ocean Exploration and Ocean Observations
July 12, 2001

What R&D activities are necessary for the environmental sciences community to capitalize on the capabilities of aquatic and marine sensor networks?
Further R&D is required in three major areas: *database management, communications, and networking and instrumentation*. Because useful environmental measurements can only be pursued through a consistent, systems-level approach, a balanced R&D program in each of these areas is of equal priority.

Database management is an interesting challenge, largely because with appropriate communications most data can be made available in near-real-time, with a latency of only seconds. In the past, database management in the environmental sciences has had the luxury of time, but this is no longer true. New real-time approaches to data and metadata must be taken, including the ability of instruments to develop as much metadata *in situ* as possible. Significantly, near-real-time data are likely to be of poorer quality than data corrected with the benefit of review and analysis. For example, time is difficult to quantify due to a variety of problems including drift and the loss of reliable references such as GPS for undersea systems. Thus, data corrected after the fact will almost certainly have greater timing accuracy. In this case, what should be done with the original data collected and presumably archived and even analyzed? Reference models of Earth systems may be also be necessary for data comparisons that will reveal when sensors were or are no longer behaving reliably. While it was generally agreed that data collected should be open and immediately available to any interested party, it will be an interesting sociological challenge to develop a broad consensus and practice in this matter. The exponentially increasing rate of access to real-time data, however, demands open data in order to avoid complexity and delays through the imposition of excessive rules on access.

The Breakout Group agreed that all instruments should be designed as IP-addressable devices, individually identifiable in a network. Data compression is seen as important, but the standards are likely to vary from measurement to measurement, and the issue of loss versus lossless compression must be considered in communicating data from a sensor through the network. For example, it would be undesirable for the communications system to induce compression losses in a data stream. Many felt that the ability of the sensor to do on-board computing was important to reduce the amount of information that has to be transmitted. This issue is likely to require the greatest R&D attention.

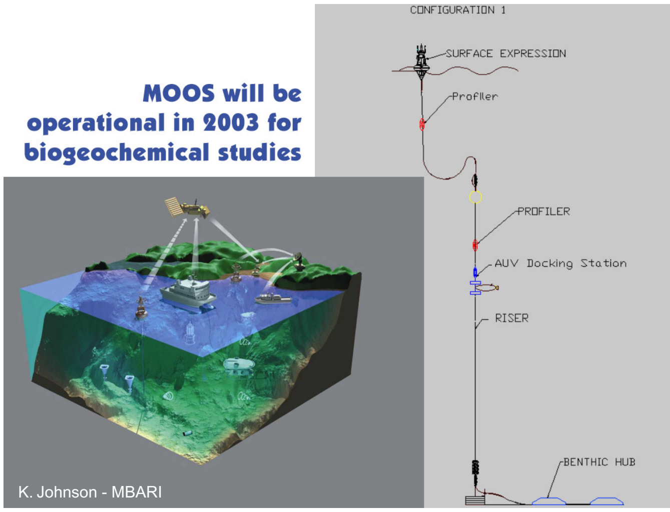
Data formats have traditionally been a matter of contention within scientific communities, and proprietary formats without open specification are particularly onerous. We briefly discussed platform-independent software such as the SDSC Storage Resource Broker [SRB], which provides the following services: federated access to data sets; protocol transparency to diverse and distributed storage systems; location transparency to distributed data sets; and access transparency to remote users.

Heterogeneity in data management systems becomes less important in this context, and there is no longer a need for the data themselves to be centralized into a single, community storage system. Responsibility for data and quality control can thus remain as near to (or far from) the organization responsible for collecting the data as desired. This provides a great degree of flexibility in choosing the degree of decentralization in a data network.

The Breakout Group also discussed communications issues between sensors that are independent of users; that is, how might a sensor network automatically adapt to changes and observations? An example could be the capability of increasing the sampling rate upon the occurrence of an observation. Such a network would undergo an autonomous self-organization that would almost certainly be nonlinear. Network simulation software would be very important in such designs and would help answer the question of *what must be measured at what scales?* Due to weather, networks might have to autonomously adapt to the loss of some component(s) of the primary communications system. Quality of Service (QoS) is an important consideration. For example, how can network and inter-sensor communications be used to ensure delivery of data to priority users? These are important basic research issues that must be answered as the complexity of the observing systems increases.

Instrumentation is a particularly challenging research and development issue. The problems associated with the availability of chemical and biological instrumentation were discussed. Generally, sensors in remote locations must be as power-efficient as possible. Instruments become fouled and must be cleaned. Can this be done remotely, and how can we know when an instrument requires attention? Calibration and drift is a problem for nearly every measurement in the environment. How can drift be detected and corrected? How do these procedures interface with the metadata of the measurement? How can physical data be managed, for example, samples? How can remote vehicles be managed, including mission and navigation and maintaining an overall system clock?

There are fundamental sampling issues that require great attention in network design. For example, can there be a compromise between global, coarse, synoptic measurements and detailed measurements at small scales? More generally, how can observation systems be designed to most effectively promote discovery and exploration of the oceans?


MOOS will be operational in 2003 for biogeochemical studies

K. Johnson - MBARI

The MBARI ocean observing system (MOOS) program includes both mooring-based and cabled-based observatory systems. The mooring observatory system (illustrated here) will provide capabilities to instrument upper water column and benthic locations of scientific interest in various geographical sites. Advanced capabilities will include satellite based bi-directional communications, event detection and response, as well as integration and operation with other advanced platforms including AUV's and vertical profilers.

While all the above issues are of critical importance, environmental observations in the marine and aquatic environment must also deal with a number of significant legal and political issues, including the Law of the Sea, the definition of EEZ's, copyright, data access, international cooperation, and the movement of pollutants and marine life (including exotics) across international boundaries. It is difficult to quantify these challenges in simple terms of information vs. political costs, but practical observational systems must deal with all of these issues.

Recommendations

Initiate a study of marine science data communications requirements.

The marine environment is heterogeneous and vast in both surface area and volume. It requires significantly different approaches to communications for inshore, nearshore, and offshore settings, as well as for surface versus submarine environments. A comprehensive study of the communications architecture for a marine science data network is required to bridge these domains and to enable interoperation between the different and demanding requirements of these dissimilar environments. For example, commercial interests can play an important role for the inshore and nearshore settings, but can provide little help in the distant offshore and submarine environments.

Provide funding for modernization of shipboard data systems.

New technologies have made it possible to achieve significant improvements in the data management of existing shipboard measurement systems, and this will have major and near-term benefits for the entire scientific research community. Such efforts should be the beginning of a long-term effort to develop standards for instrumentation (shipboard and observatory) to facilitate the development of self-describing, autonomous sensors that can report their measurements to a data acquisition system (e.g., network) with minimal operator intervention and are capable of interoperating with other sensors and data systems in terms of adaptive routing, metadata-based services (such as reporting the existence and status of any given sensor), operating status, location, and similar housekeeping functions, including reprogramming. Emphasis should be placed on developing networked sensors with individual IP addresses and Internet operability.

Initiate competition for new ocean-spanning communications systems technologies.
Better communication services are required to support higher data rates from any new classes of sensors. It can be tempting to jump to the conclusion that this relates solely to satellite, wireless, and fiber-optical cable communications, but other platforms can be envisioned such as long-dwelling UAVs (Underwater Autonomous Vehicles), commercial aircraft, and volunteer ships equipped with transponders or other as yet unimagined backbone network platforms.

This communications infrastructure is the key limiter of network development, since expendable and recoverable sensors in the marine and aquatic environments have a high probability of failure due to the harshness of the environment. The ability to obtain data from "pop-up" platforms such as UAVs, gliders, surface drifters, or communications/ data pods released from submerged sensor platforms, requires reliable, inexpensive, and global communications.

SENSOR NETWORKS

The Design and Implementation of Terrestrial Sensor Networks

Robert Waide, Facilitator
Long Term Ecological Research Network Office
University of New Mexico

John Porter, Reporter
University of Virginia

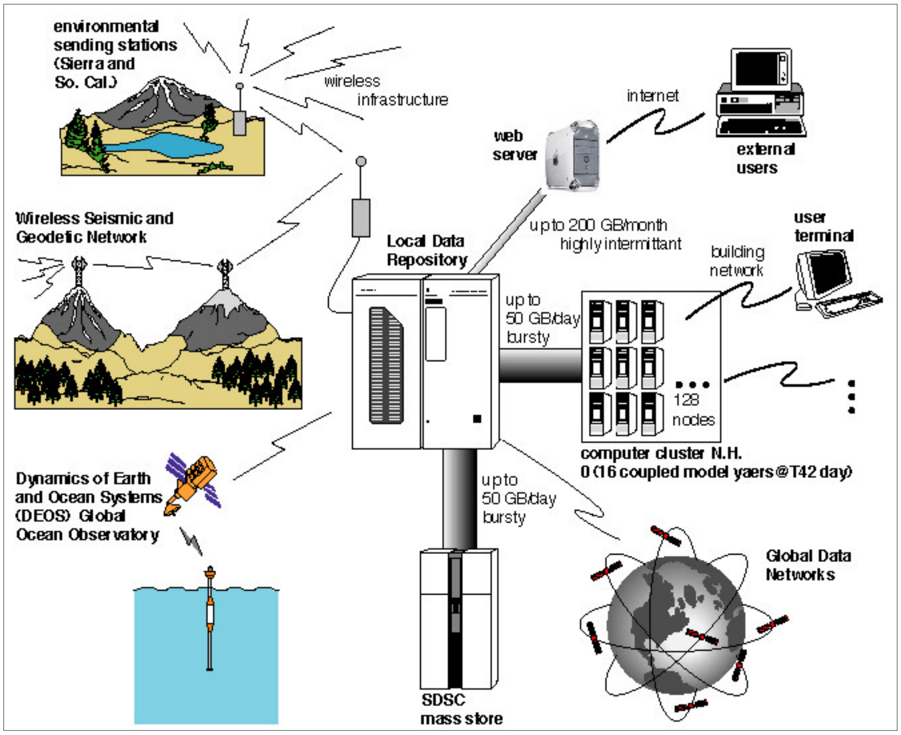
Initial Questions

- What are the most promising recent developments in terrestrial sensor networks?
- What critical components of the terrestrial environment are not adequately sensed with current technologies?
- What R&D activities are necessary for the environmental sciences community to capitalize on the capabilities of terrestrial sensor networks?

In this session, two different approaches, one question-based and the other architecture-based, led to essentially similar descriptions of a terrestrial sensor network. The first approach generated a design that was driven by scientific hypotheses, questions, or models, focusing on the distribution and kinds of sensors and the network needed to connect these sensors. The second approach began by assuming the need for internetworking of information at the broadest level, and constructed sensor networks that facilitated internetworking. Both approaches converged on a network design that emphasized domain-relevant flexibility at the interfaces between sensors and the environment and between the localized sensor network and other networks, while assuming more standardized approaches in aggregating, processing, managing, and archiving information.

What drives the architecture of sensor networks?

Most field biologists begin the design of a sensor network by defining the scientific question that will dictate the attributes of data to be collected. Such attributes include cost; the kinds of processes or organisms to be sampled (e.g. whether mobile or stationary); whether data collection needs to be continuous or event driven; the spatial and temporal scaling needed to include the relevant interval and extent; whether the data stream needs to be real time; requirements for data reliability, redundancy, and format; whether physical samples must be collected; the need for QA/QC measures and recalibration; and other factors. Once these factors are determined, communication among sensors and local processing issues is addressed with the help of sensor and communications experts. Issues of power and efficiency become important in this part of the network design.



This is a schematic outline of the ITR Project ROADNet (Real-time Observatories, Applications, and Data management Network). ROADNet will enhance our capacity to monitor and respond to changes in our environment by developing both the wireless networks and the integrated, seamless, and transparent information management system that will deliver seismic, oceanographic, hydrological, ecological, and physical data to a variety of end users in real-time.

The ROADNet multidisciplinary science and technology team is building upon currently deployed autonomous field sensor systems, including sensors that monitor fire and seismic hazards, changing levels of environmental pollutants, water availability and quality, weather, ocean conditions, soil properties, and the distribution and movement of wildlife. ROADNet scientists are also developing the software tools to make this data available in real-time to a variety of end-users, including researchers, policymakers, natural resource managers, educators and students. The project is funded by the NSF and ONR with matching funds from the UCSD California Institute for Telecommunications and Information Technology [Cal-(IT)2], Scripps and IGPP. Much of the land-based network has already been installed by the SDSC/IGPP HPWREN (High Performance Wireless Research and Education Network) funded by the NSF. For more information see http://roadnet.ucsd.edu/.

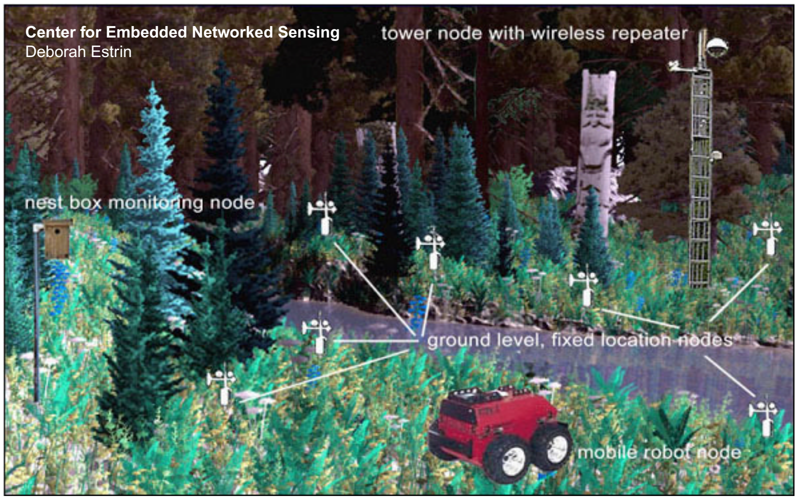
In developing a sensor network, information specialists first concern themselves with derived processing of information collected by sensors, aggregation of data, management of information, and archiving of value-added databases. The types of sensors may not be central to planning information management systems, but attributes of the data generated are. The need to adhere to standardized protocols for the description, storage, and accessibility of data are important in determining the information process components of a sensor network.

Networking specialists then focus on the distribution of processed data among higher-level nodes of a network. Key issues for this group include internetworking, interconnection, and interoperability. The development of interoperability faces challenges stemming from the nature of the data (from simple repeated measurements such as temperature to full motion video distributed across the same platform), and from the range of communication networks involved. An architecture that facilitates network communication must have a variety of communications options built into the system.

General Characteristics of a Terrestrial Sensor Network

The design of terrestrial sensor networks must accommodate investigation of a wide variety of scientific questions, while establishing generic protocols for information sharing among different sensors, networks, and users. Thus, sensor networks need to incorporate *flexibility* into the design of sensor grids along with *standardization* in the architecture of information exchange. The balance between flexibility and standardization is an important focus for future investigations.

Sensor networks should be of a recursive design, with components for data collection repeated for communication and storage. The basic unit of the sensor network requires a physical layer that interacts with the environment to be measured, recursive storage and node processing, communication among components, and the capacity to change sampling parameters through a sensor query language. Networks of these basic units need to incorporate derived processing (detection, identification, and extraction), aggregation mechanisms, information management and archiving capacities, and internetworking. Thus, there is both a logical and a physical change in structure between the *in situ* network and the derived information products to be managed and distributed. The number of iterations of the basic design element that will occur before higher level processing components need to be added may be idiosyncratic to the system and questions under consideration. The capability of re-tasking needs to be built into sensor networks so that new questions or new users can easily be accommodated. Sufficient flexibility in information management needs to be present to allow for the needs of both primary and secondary users of the data. This will include the ability for unanticipated users to overlay data from other disciplines.


Center for Embedded Networked Sensing
Deborah Estrin
tower node with wireless repeater
nest box monitoring node
ground level, fixed location nodes
mobile robot node

Field Experiments at the James Reserve - A Model System

Sensing Infrastructure
Environmental sensors in different habitats.
Multimedia sensors in natural habitats and artificial cavities (nest boxes).
Physiological sensors on trees and shrubs.
Primary nodes for higher level data processing and communications on towers.
Mobile platform for high resolution sensors and tele-robotic operation.

Monitoring ecosystem processes
Imaging, ecophysiology, and environmental sensors
Study vegetation response to climatic trends and diseases.

Species Monitoring
Visual identification, tracking, and population measurement of birds and other vertebrates
Acoustical sensing of birds for identification, spatial position, population estimation.

Realizing the Potential of Terrestrial Sensor Networks

The development of terrestrial sensor networks as envisioned in this workshop will result in a paradigm shift for field biologists. Most ecological research is presently limited by the labor available to collect observations and measurements by hand. The advent of sensor networks with hundreds or thousands of nodes in which initial and derived processing will be accomplished automatically will increase the resolution of ecological data by orders of magnitude. This flood of data will create the need for greatly increased computational power, high-speed connections, sophisticated 4-D visualization techniques, mass archival of data, and data management, navigation, and access tools.

To prepare ourselves for this paradigm shift, ecologists need to begin to evaluate new and developing technologies, create and populate training programs at the undergraduate and graduate level, develop collaborations with sensor manufacturers and national laboratories to create the required new technologies, and participate in joint efforts with experts in sensor technology, communications, information management, and networking to design and implement prototype sensor networks. In the short-term, our most important goal is to initiate the development of such prototype networks, which will serve as test beds for new technologies and training grounds for future generations of scientists.

These prototype networks should focus on implementing the most promising recent developments in sensor networks, identifying needs for the development of new sensors to measure poorly understood processes, and focusing attention on future research and development needs. Specifically, prototype networks should address the elements below.

Promising Recent Developments in Terrestrial Sensor Networks

- IR sensing
- Mass-produced miniaturized sensors, processing and communication
- Satellite communication
- Radar/LIDAR/hyperspectral remote sensing
- GPS
- Ultrawide band radar (ground)

Recommendations

Sensor development needed for events that are not adequately measured.

- Stochastic events
- Sub-surface sensing
- Location: non-GPS (subsurface, sub-canopy etc)
- Sampling of metabolic processes
- Sampling of individual or group stress or "health"
- Species and individual identification on a large scale, including genetic structure
- Emergent ecosystem attributes
- Change
- Ability to instrument and process large areas

R&D activities needed for terrestrial sensor networks.

- Power/energy requirements: demand and supply to support scalable deployments
- Research on sensor design, including reusable or biodegradable design
- Processing architecture
- Mass production of available sensors
- Miniaturization of sensors
- Development of a sensor query and analysis language
- Statistical, modeling, and visualization tools
- Automated image interpretation

SENSOR NETWORKS

Emergent Sensor Technologies

Doug Goodin, Facilitator
Kansas State University

Gregory Bonito, Reporter
Long Term Ecological Research Network Office
University of New Mexico

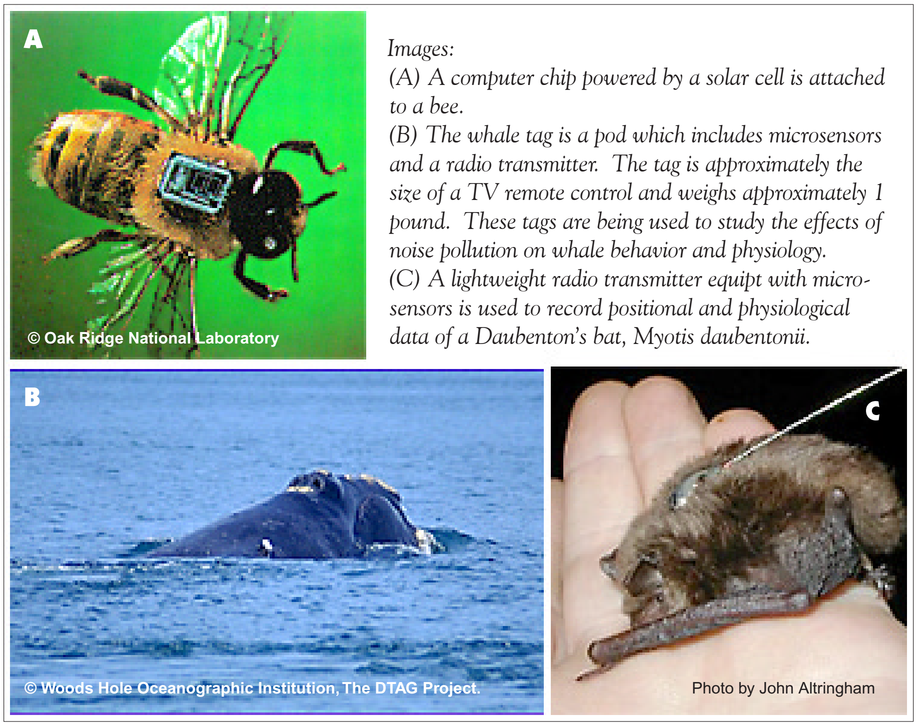
Initial Questions

- What are the most promising recent developments in sensor technologies?
- What critical components of the environment are not adequately sensed with current technologies?
- What R&D activities are necessary for the environmental sciences community to capitalize on the capabilities of sensor technologies?

Emerging Technologies

As sensors become smaller, smarter, and more specialized, the capability for deployment and use of new sensor technology suggests novel approaches to environmental research and data collection [Delin, 2001]. Although these trends are more likely to be driven by research in other fields, e.g. space research [Krabach, 2000], they show great promise for application in field environmental research. Three emerging technological trends are particularly promising: miniaturization, wireless communication, and "smart" sensors.

The trend toward the miniaturization of sensor systems will have a significant effect on how the environment is studied and monitored. "Systems on a chip" technology, for example, may replace chemistry and biology laboratories with portable hand-size instruments used for rapid and sophisticated chemical or biological agent (e.g. DNA, protein) detection and quantification *in situ* [Ho, 2001]. Another example is monitoring standard environmental parameters such as the weather. Despite the development of observation networks for a variety of environmental variables (e.g. weather/climate, solar radiation, rainfall chemistry), coverage is still sparse. For example, the continental United States is represented by fewer than 3,000 permanent meteorological observation sites, a density of less than one station per 10,000 km^2. Sensor miniaturization will enable much denser observation networks. Densities as high as several hundred instruments per km^2 are foreseeable in intensively studies sites. Coverage extent will also be enhanced by sensor miniaturization, since smaller sensors can be deployed in places where current generation sensors will not go. Instrumentation within canopies, underground, and even upon/within individual organisms from bats to earthworms, can be achieved via micro-sensor packages. Micro-sensors can readily be deployed within sensor clusters (i.e. packages of multiple sensors making coordinated observations) and within sensor webs [Nagel, *in press;* Delin, 2001]. These sensor webs may ultimately be reduced to very small size, e.g. "smart dust" [Pister, 1999], while retaining equivalent function to larger sensor clusters. Current research in meteorological and environmental instrumentation is already progressing toward this goal [Nagel, 2000; Delin, 2001]. Cost of miniature sensors could be a limiting factor of their use by environmental scientists. Although mass production of micro-sensors using modern manufacturing processes should drive down production costs, sensor housing continues to be a significant expense. If cost reductions can be achieved along with instrument size reduction, miniaturization will improve existing data collection methods as well as suggesting novel instrument siting opportunities.



© Oak Ridge National Laboratory

© Woods Hole Oceanographic Institution, The DTAG Project.

Photo by John Altringham

Images:
(A) A computer chip powered by a solar cell is attached to a bee.
(B) The whale tag is a pod which includes microsensors and a radio transmitter. The tag is approximately the size of a TV remote control and weighs approximately 1 pound. These tags are being used to study the effects of noise pollution on whale behavior and physiology.
(C) A lightweight radio transmitter equipt with micro-sensors is used to record positional and physiological data of a Daubenton's bat, Myotis daubentonii.

An Ecological In Situ Sensor Resource: a compilation of information on in situ sensors, sensor arrays, and sensor manufacturers.

Sensors are an essential part of scientific inquiry, yet no central sensor resource is currently available to address the sensor needs of the ecological community. Many environmental sensor projects are known only in small scientific circles, and information regarding sensors and their manufacturers are not typically oriented towards the scientific community. To meet this need, a web site, targeted toward the terrestrial and aquatic ecology communities, has been created through a collaborative effort between the Long-Term Ecological Research (LTER) Network Office and the San Diego Supercomputer Center (SDSC). The website includes links to state-of-the-art sensor technologies, sensor manufacturers, and large-scale ecological projects and networks involved in the use of in situ sensors. ***For more information, visit http://www.lternet.edu/technology/sensors/index.html***

Miniaturization also benefits remote sensing. Digital camera and computational technology have enabled creation of small, low power, relatively low-cost multi- and hyperspectral sensor systems which could be deployed on modest aircraft with minimal modification [Price, 2001]. This type of remote sensing system could put powerful airborne imaging technology under the direct control of research groups. This would be an improvement over the current model, where sensor systems are either operated by government agencies or for-profit private ventures.

Along with miniaturization, wireless communications technology holds great promise for environmental sensing [Nagel, *in press*]. Often, the communication infrastructure needed to support instruments in the field (particularly at remote sites) is a significant limiting factor in field research. Remote operation of sensors requires emplacement of wiring, which is prone to failure in harsh environments, or use of *in situ* data logging equipment requiring periodic visits for maintenance and data retrieval. Wireless technology, coupled with the Internet, could replace these cumbersome systems with instruments capable of relaying data to a centralized collection site, and perhaps even directly to the researcher's computer. Wireless communication would also enable bi-directional communication with a sensor web [Delin, 2001], allowing "on-the-fly" sensor programming or retasking. Such programming capability would enhance the adaptability and flexibility of sensor networks. For the dense networks of micro-sensors described above, wireless communication systems are a necessity. Without them, solving control and data retrieval problems would not be feasible. Provision of power continues to be a significant issue for wireless data transfer networks.

As data storage and manipulation technology becomes more compact and powerful, smart sensors will become increasingly common. Smart sensors have the capability for on-board processing of data, hence some data analysis tasks currently carried out offline may become part of the data processing stream. This capability will greatly enhance the effectiveness of data-rich sensor clusters and webs, where the sheer number of sensors multiplies data compression and information extraction tasks [Delin, 2001; Nagel, *in press*]. Smart sensors will have the ability to selectively collect data, i.e. they will be able to discriminate noteworthy events or situations and sense them, while remaining inactive when no meaningful data collection opportunity exists. Smart instruments will also enable automated collection of data based on artificial intelligence or pattern recognition techniques. For example, video or audio sensors capable of distinguishing characteristic shapes and sounds of particular organisms could then selectively collect data about those organisms.

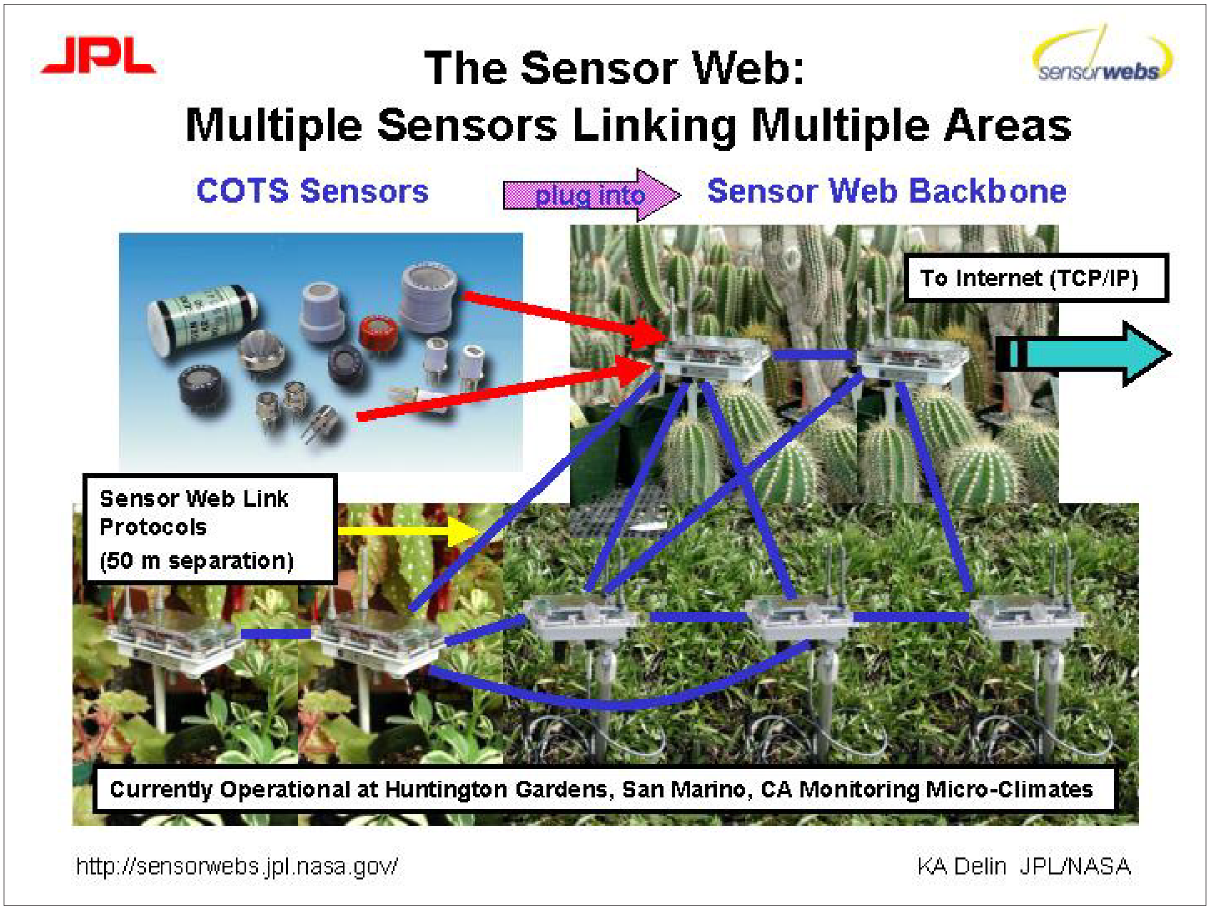
Extension to Unmeasured Variables

Current sensors respond to physical or chemical aspects of the environment. For example, meteorological sensors respond to temperature, humidity, solar radiation flux, and other energy fluxes. While these detectors are quite effective, they are limited to only a few environmental variables. In contrast, humans and other living organisms gather information about their environment through a variety of senses, each utilizing a biological detector evolved to respond to a particular biophysical or biochemical stimulus. Enhanced sensors, capable of emulating biological senses, are opening new windows for observation of the environment. Electronic “noses” and “tongues” now allow ecologists or biologists to directly detect in real time chemicals in the environment that could previously be detected only through lengthy, expensive, and difficult laboratory analysis [Staples, 2000]. Coupled with the miniaturized, wireless sensor technology described above, electronic “noses,” “tongues,” “ears,” and “eyes” could be deployed in sensor webs alongside more conventional instruments, resulting in more robust and adaptable means to observe the natural environment. This technology may be particularly useful below ground, where the opaque nature of this environment makes sensing exceptionally difficult. Small instruments capable of operating outside the range of conventional sensors would greatly benefit the below ground environmental sciences and provide critically needed information.

New sensors to detect and measure properties not accessible to current sensors will utilize smart sensor technology. Thus, electronic eyes and ears will consist not just of audio or video detectors but also analytical components capable of detecting and identifying patterns of sound or vision. Similarly, electronic noses and tongues will be capable of detecting patterns and quantifying hundreds of organic and inorganic chemicals in the environment. This is important in identifying the source and significance of compounds and chemical cues within an ecosystem. These innovations will permit biological sensing at the organism or biota level, instead of the coarser physical/chemical sensing currently in use. Sensory emulation instruments (e.g. gas chromatographs) are available, but their size and cost still limits practical field use. Although faster and cheaper miniaturized portable electronic noses (using micro GC capillaries) are also available, lack of a large market keeps their costs higher than most scientists and federal agencies can afford. Hence, miniaturization holds great promise in adapting these technologies for practical use.

R&D Issues for Next-Generation Sensors

Development of the next generation of sensors should focus on two priority areas: (1) adaptation of existing sensors for field use, and (2) development



JPL

sensorwebs

The Sensor Web:
Multiple Sensors Linking Multiple Areas

COTS Sensors plug into Sensor Web Backbone

To Internet (TCP/IP)

Sensor Web Link Protocols (50 m separation)

Currently Operational at Huntington Gardens, San Marino, CA Monitoring Micro-Climates

http://sensorwebs.jpl.nasa.gov/

KA Delin JPL/NASA

of innovative new sensors. Adaptation of existing sensors offers opportunities for extending current measurement technologies. Miniaturized temperature, humidity, and fluid flow sensors intended for application in laboratory or biomedical applications are already available. With suitable repackaging, these sensors could be used to create the small sensor clusters described above. Bi-directional wireless communication technology (necessary for effective exploitation of new sensor technology) is relatively less developed, but continues to improve [Cook, 2000]. Power requirements are a significant limitation, but the power consumption of these sensors continues to improve [Nagel, *in press*]. The environmental research community in general could greatly benefit from the establishment of a research program emphasizing innovative techniques for useful modifications of current technology and disseminating this information to field scientists.

Longer term R&D initiatives should emphasize development of new sensors including all the features (miniaturization, smart design, wireless communication, sense emulation) described in the previous sections. Standardization of sensors, sensor platform, and software interface between sensors and users is also critically needed. A practical limitation in the development of these sensors is the relatively small size and fragmented nature of the environmental sensor market. While much development research is carried out in universities, government labs, and other non-commercial settings, promising technologies are then transferred to the private sector for manufacturing and marketing. In order to be cost-effective, a sufficiently large market must exist to justify development expenditure by the private sector. In general, field environmental science is too small and specialized a market to attract large-scale private investment, so promising technologies are often not developed beyond the prototype stage, or are made in such low quantities that high cost limits their deployment. A possible solution to this problem lies in the convergence/similarity between sensor needs for field environmental science and the technological needs of larger-market activities such as biomedical applications, defense, and national security. As medical and defense/security applications make increasing use of miniature sensors and exotic detectors such as electronic noses and tongues, an opportunity exists for environmental scientists to “ride the market.” A challenge for the environmental research community will be to work with manufacturers to identify small, feasible modifications to sensors intended for other applications that will allow them to be marketed to the environmental research community as well, resulting in larger markets with little capital investment. Standardization of sensors and sensor platforms may also help bring down the cost of sensors through mass production of interchangeable components, and will increase the availability of sensors and custom-designed sensor arrays. This path allows the same sensor systems to be used by ecologists, federal agencies (e.g., EPA, USGS, NOAA, DoE), and environmental monitoring and restoration companies, widening the environmental sensor market and allowing broader deployment by environmental scientists.

DATA TECHNOLOGIES

Geospatial Data Integration

Karen Stocks, Reporter
San Diego Supercomputer Center
University of California, San Diego

Jim Quinn, Facilitator
University of California, Davis

Initial Questions

- How can information technologies be better used to facilitate integration and synthesis of geospatial data acquired via environmental networks?
- What are the limitations (e.g. intellectual, technical, physical, and funding) to progress in this area?
- What are constructive solutions to overcoming these limitations?

Recent advances have brought exciting changes to the landscape of geospatial information. Remote sensing techniques have created continuous, large-scale coverages of parameters previously sensed only through point-sampling. Affordable, accurate GPS systems have increased the volume of data having good spatial referencing. And upcoming wireless sensors and microchip GPS units show promise for continuing improvements in the future.

Data Issues

Research findings and management decisions will never be better than the data from which they are drawn. Improving the data available for research and management involves creating incentives for data sharing, creating quality base data sets, and focusing research and development funding on new technologies capable of sampling underrepresented data types.

Incentives for data sharing

The largest current limitation on data availability is the lack of incentives for data sharing within the research community. The reigning professional standard of publishing a journal article describing research conclusions generally provides only a text summary of the data. What is needed to facilitate data integration for large-scale, long-term, or multi-factor research is access to fully documented, electronic data sets. It is unrealistic to expect individual researchers to take on the challenge of providing such data sets when it is both unfunded and unrewarded as a professional accomplishment.

Data management and access needs to become a defined and funded part of any proposal that creates new data sets, and funding agencies must make their expectations (including timelines) for this explicit. Information management typically constitutes 10% or more of the cost of commercial R&D, and funding agencies should expect similar resources to be devoted to making geospatial data interoperable and readily shared.

Professional recognition for publishing data is equally important. Methods for crediting data resources that are parallel to literature citation need to be developed. Scientific societies and publishers should be encouraged to follow the GenBank model, requiring that the raw data for any published article be placed in a database and made publicly available after a set period of time. More generally, the development of robust data resources is often a creative exercise fully equivalent to producing journal articles. With the advent of all-electronic professional outlets for publication, it is feasible and would be highly desirable in terms of both professional recognition and traditional quality-assurance, to have peer review of data sets and accompanying metadata equivalent to the traditional review of articles and books.

Finally, there is a need for a formal "code of ethics" for data use covering the issue of how long an investigator can keep a data set proprietary, how intermediate data products (such as Web resources compiled from published data) are credited and cited, etc. Once the expectations are clear, then institutions and funding agencies (and reviewers) can begin to evaluate researchers based on these expectations.

Base Data Sets

Good research and good policy require the creation of high-quality, standard data coverages that are applicable to a broad spectrum of users. Outside of remote sensing, many geospatial data sets are composed of point data measurements (e.g. soil samples, rare species locations). To create useful products, these points must be integrated and interpolated to create continuous views, using models whose assumptions, limitations, and uncertainties are communicated. The assessment and visualization of uncertainty is a particular research and technological challenge for mapped data, particularly when there are repeat measurements. Often the variables measured (e.g. remote sensing "color") are not the variables of true interest ("land use"). Those coverages that do exist, such as watershed delineations and vegetation indices, have proven to be valuable resources. The aggregation of relevant point data is time-consuming and the process of creating a coverage from point data is best done by scientists familiar with the characteristics of the base data sets in collaboration with statistical/analytical experts. Creating standard products properly and making them available in a variety of formats will reduce redundancy and improve decision-making.

A related problem is the availability and appropriateness of "framework data" - the "base layers" used to spatially reference geospatial data from research projects and monitoring. The Federal Geographic Data Committee has recognized a set of framework data sets (elevation, hydrography, roads, etc.) that are essential for landuse planning and related disciplines, and most have complete national coverages or national initiatives to complete coverages. There is less consensus on the "framework" data essential for environmental research (soils? vegetation? land management practices?), and efforts to address these data needs remain fragmented and underfunded.

New Sensor Technologies

Remote sensing technologies can now create large-scale, high-quality maps of a variety of parameters. However, data types that cannot be remotely sensed are still only represented by limited data points. Priorities for the next generation of sampling technologies must include new methods for measurements traditionally taken through *in-situ,* human-mediated, time-intensive point sampling. Promising avenues include computer-aided video identification of species, automated processing of genetic samples, and new acoustic techniques.

Interoperability and Standards

Addressing complex environmental questions requires the integration of data from many resources and the application of multiple informatics tools: GIS, databases, visualization tools, knowledge representations, statistical packages, etc. Current barriers to bringing together heterogeneous data sets and to moving between multiple software platforms form logistical barriers to research progress. While these barriers can be overcome, they require large investments of human effort.

Data format incompatibilities may be partially addressed through standards. Once metadata standards are adopted by the community, this will allow the development of tools that can interact automatically with the metadata. While standards for geospatial data do exist (e.g. Federal Geospatial Data Content standard and Geographical Markup Language), they are not widely used and are not implemented by commercial software packages, in part because they are highly complex. Moreover, the standards address the format of expression but not the actual vocabularies (semantics) used. Much of the power of metadata for information discovery rides on consistent or crosswalked uses of language, which are necessarily tied to particular user communities.

It is recognized that comprehensive documentation of data sets is a worthy goal and that it is unlikely that any single standard will ever suit the plethora of ways in which geospatial data is used. However, the reality is that unimplemented standards are not effective - a data provider creating a small data set that contains location information but is not aimed at geospatial description per se simply will not invest much time in standards compliance without adequate incentives and support.

Software Research and Development

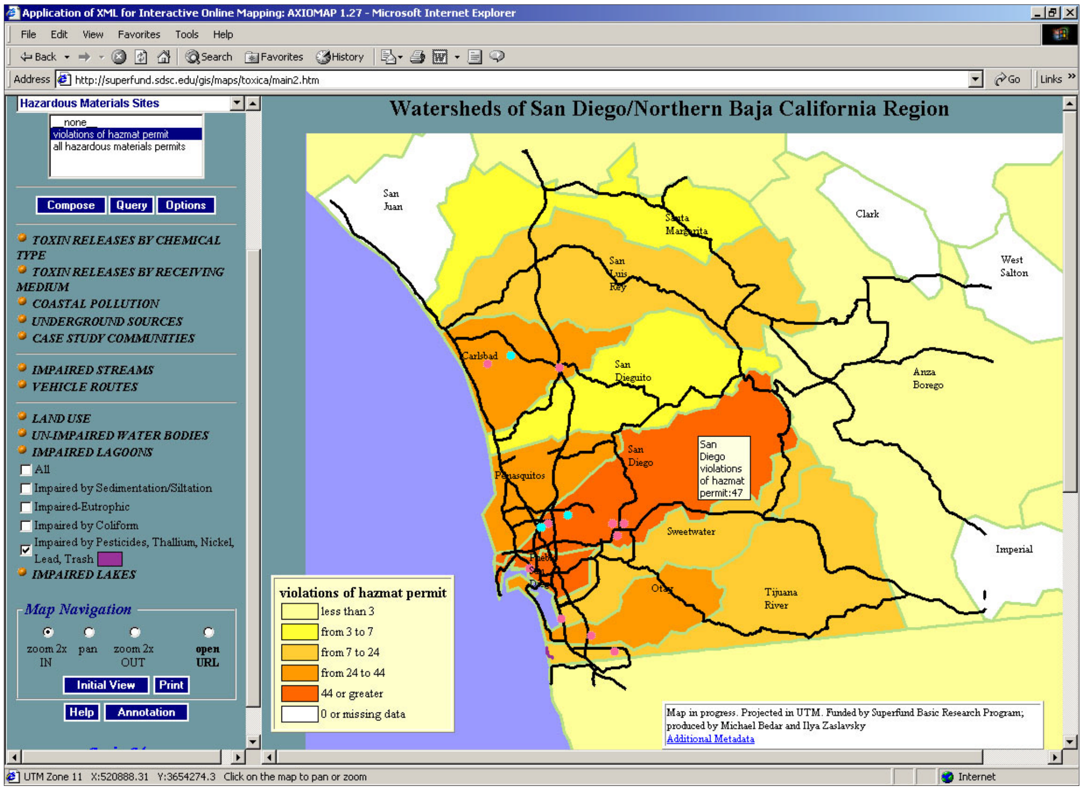
In addition to streamlining the software currently available, new tools and new approaches for working with data are also needed. By streamlining and increasing the capabilities of informatics software, analyzing and processing data can become more efficient and powerful. GIS systems were developed from a cartographic paradigm that does not scale well to today's 4-D data needs: height/depth and time are often poorly represented, and connections to quasi-spatial information are poor. In part this is because the small number of commercial vendors producing mass-market tools cannot respond to the specific needs of small user groups. And when small, specific, individually-built tools are developed to overcome holes in the commercial products, these tools are difficult to integrate with other software and are often not widely available.

The Open Geographical Information System initiative [OpenGIS] partially addresses these issues. However there are still considerable conceptual gaps among the approaches and paradigms of the GIS/cartography/remote sensing community, the visualization community, and the relational database community that need to be bridged to produce an integrated data environment.

It was also noted that there is a spectrum of geospatial data users. While there is a need for powerful and advanced capabilities for leading-edge IT research, there is also a need for user-friendly, easy-to-learn tools for those basic operations common to a broad spectrum of environmental researchers.

Infrastructure

The informatics infrastructure needs to continue to grow and mature in order to support the new data sources and new tools. Overall, advanced geospatial data processing is pushing the computing-power and storage capabilities of the country's infrastructure. Initiatives such as funding for grid-computing projects are welcome additions, and continued support for growing computing systems of a range of sizes is required. Beyond general computing power, there are three specific areas that form, or will soon form, substantial barriers to progress.


Application of XML for Interactive Online Mapping: AXIOMAP 1.27 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://superfund.sdsc.edu/gis/maps/toxica/main2.htm
Go Links
Hazardous Materials Sites
none
violations of hazmat permit
all hazardous materials permits
Compose Query Options
TOXIN RELEASES BY CHEMICAL TYPE
TOXIN RELEASES BY RECEIVING MEDIUM
COASTAL POLLUTION
UNDERGROUND SOURCES
CASE STUDY COMMUNITIES
IMPAIRED STREAMS
VEHICLE ROUTES
LAND USE
UN-IMPAIRED WATER BODIES
IMPAIRED LAGOONS
□ All
□ Impaired by Sedimentation/Siltation
□ Impaired-Eutrophic
□ Impaired by Coliform
☑ Impaired by Pesticides, Thallium, Nickel, Lead, Trash
IMPAIRED LAKES
Map Navigation
● zoom 2x IN
○ pan
○ zoom 2x OUT
○ open URL
Initial View
Print
Help
Annotation
Watersheds of San Diego/Northern Baja California Region
San Juan
Santa Margarita
Clark
San Luis Rey
West Salton
Carlsbad
San Dieguito
Anza Borego
Penasquitos
San Diego
San Diego violations of hazmat permit:47
Sweetwater
Imperial
Otay
Tijuana River
violations of hazmat permit
less than 3
from 3 to 7
from 7 to 24
from 24 to 44
44 or greater
0 or missing data
Map in progress. Projected in UTM. Funded by Superfund Basic Research Program; produced by Michael Bedar and Ilya Zaslavsky
Additional Metadata
UTM Zone 11 X:520888.31 Y:3654274.3 Click on the map to pan or zoom
Internet

Web-Based Interactive Mapping - For Regional Environmental Health Information
(Web EI for San Diego County and Northwestern Baja California)

Web-Based Environmental Informatics (WebEI) is an interactive mapping service that concentrates on integrating and visualizing distributed environmental health data in the San Diego-Baja California border region. Two foci are on issues related to the Total Maximum Daily Load (TMDL) process--an approach to conserving aqueous resources by attending to the total amount of each pollutant that a water body receives--and on community development in the Colonia area of Tijuana.

Available data include impaired waterbodies, watershed boundaries, toxic releases, land use, and soon, health demographics, urban infrastructure (e.g., sewage and power), biodiversity and habitat, and responsible authority. These layers can be overlaid and grouped in various combinations for spatial insight. Users can look in more depth at the issues at work in a particular location by clicking on the point features. As additional data becomes available and integrated into Web EI, it is hoped that this information system will aid in decisions that lead to the sustainable development of the border region.

The first limitation is the lack of organizing elements within online data resources. Data is being served by many groups and at many levels, from individual researchers with desktop servers through field stations, research institutions, libraries, journals, and government agencies from local to federal. While this new data accessibility is an exciting step forward, the hodgepodge of data resources makes it difficult to find a particular data type of interest or to evaluate its quality. Facilities for data clearinghouses/catalogs, tailored search engines, and a method for peer-review and/or user ratings of a resource to indicate its quality are all needed. A key to facilitating the creation of catalogs and searches will be the adoption of metadata standards, including controlled vocabularies, for describing data contents.

Past experience with attempts to establish central repositories for data gathered by individual investigators and programs have not been encouraging. It is likely that authoritative source copies of much of the important geospatial data will remain distributed among thousands of sources and will be somewhat idiosyncratic in content and format, meaning that archival and bandwidth challenges will if anything increase. Mirror sites, portals, and clearinghouses will need robust methods for extracting and validating integratable shared elements from heterogeneous sources, and successively abstracting them, as geographical domains of application increase. Such scalability poses fundamental problems in knowledge representation. It also poses substantial challenges in the sociology of science, since the ability to integrate data requires some consensus in the provider community on the expression of information (semantics and ontologies) in their fields of research.

The second limitation is the lack of long-term data archiving provided in the traditional 3-5 year grant tenure. While national data centers can play a role in this process, enforced data "drop-offs" at the conclusion of a grant tenure will not provide the highest quality data resources. In reality, no data set is ever fully finished, and allowing data authors to have continued access to update and expand their data will improve data quality. It will be crucial to have facilities (and long-term funding) for distributed data centers that allow data management to be kept in the hands of either the authors or the user groups (such as a scientific society or a field station) while still providing a robust framework for data maintenance and access.

The third limitation is bandwidth. Wireless communications paired with micro-GPS and other sensors have ushered in a new era in spatially-referenced environmental sensing. However, the current FCC restrictions on bandwidth are crippling potential applications of sensor networks. Old regulations must be reevaluated in light of current technologies to allow scientific access to bandwidth.

Education

In addition to facilitating interdisciplinary research in geospatial tool development and application, progress in environmental science would be advanced by raising informatics literacy among domain scientists. Just as statistical packages, spreadsheets, and word processors are considered required tools in any scientific domain, environmental scientists today need to have basic familiarity with data management practices and the uses of GIS, database, and visualization software. Efficiently finding, accessing, and using data is intrinsic to the modern process of research and resource management in all fields. Unfortunately, the teaching of geospatial concepts and technologies is particularly fragmented, as important applications cross traditional disciplinary departments. The cartographic conventions underlying modern GIS software have traditionally been taught in geography departments, many of which are struggling and disappearing, and offerings in other departments (optics and remote sensing in Physics, data models in CS, CAD in Engineering, vegetation maps in Biology, geomorphology in Geology) are typically uncoordinated, usually due to institutional barriers to teaching outside one's department or college.

Support for model undergraduate curricula to bring together computer science, geography, and other domain sciences would help provide courses with the appropriate balance of theoretical and applied aspects. Both full courses and IT components integrated into existing domain-science courses are appropriate. The working group also recognized that having people cross-trained in both environmental sciences and informatics (programming, database design, GIS technologies, data server design, etc.) will be critical to future progress and that there is a role for full undergraduate majors or concentrations in interdisciplinary Environmental Informatics. Targeted funding will help institutions develop model curricula that cross traditional departmental lines. At the graduate level, models are needed for facilitating interdisciplinary research through graduate students shared between Computer Science and domain departments.

There is also a need for continuing education for current researchers in environmental sciences. Support for a variety of workshops, distance-learning programs, and related resources to reach current academic and governmental researchers and managers can address this need. In particular, we note that most departments and agencies will be unable to support full-time experts versed in the full breadth of informatics techniques. Thus, there is a need for expert centers offering "consulting-style" advice to projects in managing data, setting up data-access Web pages, integrating data from multiple sources, etc.

Fostering Interdisciplinary Informatics Research

The sections above list many research and development steps that are central to continuing progress in environmental informatics. Critical to all these efforts is a concerted cooperation between computing/information sciences and the domain sciences that use geospatial data. Targeted funding from NSF to support these projects has gone far to foster these partnerships (e.g. BDI and ITR). Further funding support for data integration activities is required for them to continue, but there are also social/institutional barriers that need to be addressed. Interdisciplinary IT work must be professionally rewarded. Tenure decisions, job descriptions, etc. need to recognize the value of this work. Joint faculty and interdisciplinary education programs can help cross traditional departmental boundaries.

Most critical, however, is an equivalent to journal publication for IT work. Peer-reviewed articles are the coin of the realm in academia. They are the measuring stick through which applicants are hired, tenure decisions made, and salaries negotiated. But the development of information systems, data resources, and software tools does not lead to journal publications. There needs to be a mechanism for attaching peer-review status to the actual data product or tools that are produced; scientific societies can take a leadership role in creating a new process for community evaluation of data resources and tools.

Recommendations

Data Recommendations

- Create a data "code of ethics" to cover expectations and timelines for data sharing, methods for crediting intermediate data resources, etc.
- Promote the identification and creation of base data sets for widely-used variables. This includes targeted sampling to fill gaps in data as well as analytical efforts to gather and integrate point data.
- Target funding to develop technologies beyond *in-situ,* human-mediated point sampling, particularly species- and gene-level biological sampling.

Standards Recommendations

- Standardized expression of point data. Space and time are unifying factors that can serve to integrate a large and heterogeneous universe of data that is evolving, if properly applied. There needs to be a simple, standard way to represent x, y, z, and t location *with accuracy and precision estimates* that can be easily implemented in any data set with spatial-temporal components, along with libraries of names and attributes of the entities being temporally and geo- referenced. An example for species data is the [Species Analyst].
- Endorsement of self-describing data formats. There is currently no "common denominator" data format or generally accepted standard. Until that time, the use of self-describing data formats such as [NetCDF] is strongly encouraged to ensure that the information necessary for extracting and understanding the data is always preserved.
- Creation of tiered metadata standards.
- Development of tools and clearinghouses based on metadata standards.

R&D Recommendations

Prioritize development of key software and tools:

- Automated feature extraction and change detection. For very large data sets such as satellite remote sensing, the entire data set cannot be evaluated by a person. Tools are needed to identify and flag "interesting" features to be examined by a researcher.
- Data mining and time-series data analysis tools. Ideally, geospatial and temporal tools need to be integrated for 4-D analysis of data.
- Estimating, visualizing, and appropriately handling uncertainties in values.
- Automated or semi-automated raster/vector data conversion.
- Creating coverages effectively from point data.
- Visualization of high-dimensionality data. ODBC is not sufficient - there is a need for virtual database tools.
- Interoperability functions, particularly for moving between off-the-shelf products; for linking geospatial data with model/simulation output effectively; and for integrating the idiosyncratic, individually-built tools that exist.

Support OpenGIS development:

- Create online workbenches and software for common geospatial operations that are designed for quick learning and ease of use for unsophisticated users.

Infrastructure Recommendations

- Continue growth of computing infrastructure.
- Reevaluate FCC regulations to facilitate scientific use of bandwidth.
- Create metadata catalogs and clearinghouses for data access.
- Define a framework for distributing portions of the national data centers to allow groups interested in a particular type of data to be its caretakers, with long-term, low-level funding provided as long as performance standards are met.
- Develop an initiative on knowledge representation in geospatial environmental data.

Education Recommendations

- Expand interdisciplinary courses, majors, curricula, and workshops for teaching Information Technology applications within the environmental sciences at the undergraduate, graduate, and continuing-education level.
- Create expert-centers to provide data management and analysis advice to the environmental research community.

Interdisciplinary Recommendations

- Create an equivalent to the peer-reviewed publication to foster recognition for data resources and tools.

DATA TECHNOLOGIES

Distributed Data Access and Retrieval

Jim Beach, Facilitator
University of Kansas

Bertram Ludaescher, Reporter
San Diego Supercomputer Center
University of California, San Diego

Initial Questions

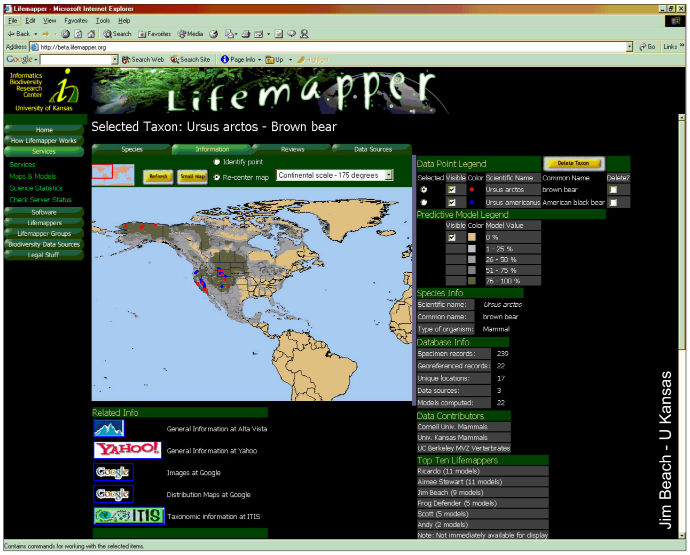
- How can information technologies be better used to facilitate distributed access and retrieval of data acquired via environmental networks?
- What are the limitations (e.g. intellectual, technical, physical, and funding) to progress in this area?
- What are constructive solutions to overcoming these limitations?

Enabling Internet Knowledge Discovery: Beyond Keyword Search and Retrieval

Anyone who has used the Internet for knowledge discovery in environmental biology knows what a bountiful information morass it is. Internet search engines in tenths of a second retrieve daunting numbers of hits from keyword-based queries. In January 2002, a search using Google (www.google.com) for "pacific salmon" generated 325,000 hits, "water chemistry" produced 1,030,000; "fish reproduction" yielded 387,000 linked pages. The overwhelming size of these result sets is only matched by the heterogeneity of linked documents they point to. They include all classes of documents known to man - narratives, research studies, historical essays, data, maps, pictures, sounds, resumes, textbooks, commercial products, as well as regulatory, policy, and educational documents. At the other extreme, a Google search on "pacific salmon AND water chemistry AND fish reproduction" returns exactly four hits: 1) an EPA report to the US Congress, 2) a fisheries management plan for Lake Superior, 3) an encyclopedia of ocean sciences, and 4) a perspectives article on freshwater ecosystems from *Ecological Monographs.*

Now imagine a fisheries biologist with a need to identify the relationship between water quality and spawning rates for pacific salmon. To determine this, she would need to locate formal research investigations that have looked at the impact of physical, chemical, and temperature characteristics of rivers on the physiological and reproductive behavior on any salmonid or other related fish species. She would need to know what variables were under study, be able to review the results, and probably inspect the field data, in order to assess their relevance and suitability to address the research issue at hand.

Clearly, keyword or full-text searching on the Internet is of limited value for enabling such environmental research. On the one hand our biologist has the untenable option of browsing through millions of linked documents; on the other, she could take the traditional approach of using the four perspective/summary documents as starting points into the research literature. The research objectives are manageable with skilled library research and a three-month review of the literature, but manually assembling research knowledge in this way is a slow and costly process. Although searchable publication abstracts and indexes provide shortcuts to the relevant literature, discovering other species models, from *in situ* or artificial *in vitro* studies of water quality and reproduction, is a hit-or-miss proposition, heavily dependent on keyword indexing. Finding and managing her findings with photocopies of research publications of course provides no support for utilizing pre-existing data sets for re-analysis, extrapolation, or the development of predictive models.



Address http://beta.lifemapper.org

Selected Taxon: Ursus arctos - Brown bear

○ Identify point
● Re-center map
Continental scale - 175 degrees

Data Point Legend

Selected	Visible	Color	Scientific Name	Common Name	Delete?
●	☑	●	Ursus arctos	brown bear	□
○	☑	●	Ursus americanus	American black bear	□

Predictive Model Legend

Visible	Color	Model Value
☑	■	0 %
	■	1 - 25 %
	■	26 - 50 %
	■	51 - 75 %
	■	76 - 100 %

Species Info

Scientific name: *Ursus arctos*
Common name: brown bear
Type of organism: Mammal

Database Info

Specimen records: 239
Georeferenced records: 22
Unique locations: 17
Data sources: 3
Models computed: 22

Data Contributors

Cornell Univ. Mammals
Univ. Kansas Mammals
UC Berkeley MVZ Verterbrates

Top Ten Lifemappers

Ricardo (11 models)
Aimee Stewart (11 models)
Jim Beach (9 models)
Frog Defender (5 models)
Scott (5 models)
Andy (2 models)
Note: Not immediately available for display

Related Info

General Information at Alta Vista
General Information at Yahoo
Images at Google
Distribution Maps at Google
Taxonomic information at ITIS

Contains commands for working with the selected items.

Web interface for LIFEMAPPER *(beta.lifemapper.org) a NSF KDI-funded project which uses the Species Analyst distributed search and retrieval network to obtain biological museum specimen data records that it then utilizes in a distributed SETI@ Home-like screensaver architecture to parallelize the computation of species distribution models based on the museum specimen data. Those models are then archived and visualized on the Lifemapper server.*

The emerging "Semantic Web" aims at changing all of that and the way we do science by transforming the process of networked knowledge discovery and retrieval [Berners-Lee, 2001]. Knowledge representation and linking technologies now being developed and deployed in the sciences aim to tame the Internet from an uncontrolled firehose spewing links to hundreds of thousands of documents in milliseconds in response to a simple query, into a rich distributed corpus of contextualized research information, linked by a deep semantic framework with analysis engines. This matrix of semantic relationships will enhance integration and analysis capabilities well beyond today's keyword and full-text search and retrieval facilities to make the Internet a dynamic workbench for ad hoc knowledge discovery and generation. The conceptual mapping of environmental data, information, and knowledge will enable us to expose the deeper foundation of structure and process in natural systems.

Although the infrastructure for the Semantic Web will be standards and protocols that have just recently become the objects of attention (see below), the content and knowledge linking of the Semantic Web will evolve slowly and likely in response to conceptually localized efforts delimited by funding or disciplinary scope.

In addition to the intellectual contributions of the designers and builders, how do we build something that we know has a very high probability of being used? How do we identify and focus on long-term priorities, with our feet in the shifting sands of technology, and continually implement more efficient systems with next year's technology? What is the minimal payoff that should be expected and measured with NSF funding of infrastructure projects?

Challenges in Distributed Access and Retrieval

Data from environmental networks is being collected, transported, stored, analyzed, and disseminated in a highly distributed fashion. Environmental networks can provide the Data Grid under development across the nation [Foster, 1998] with different kinds of environmental sensing capabilities, often combined with real time or near real time accessibility [HPWREN, ROADNet]. Distributed data networks may reflect cached data as well as sensor data and museum data.

One challenge is that from the field, where environmental sensors gather data, to the intermediaries and end users of information, there is an enormous variety of data transport and access demands, data uses, and data users. This absence of a common data and user profile in the environmental sciences community prevents a "one size fits all" approach to distributed data access and retrieval for environmental networks. Seamless distributed data access and interoperability is a challenging goal in the presence of significant heterogeneity of data, infrastructure, and user requirements.

The profile of data usage varies along different dimensions: Technically, data traveling from field sensors through intermediate nodes and different "aggregate states" (e.g., raw data can be transformed, analyzed, annotated with metadata, cleaned, aggregated, and finally stored in a curated digital library or archive) may encounter different bandwidth bottlenecks along the way before it reaches its destination, say a client application on a scientist's laptop. Ideally, dealing with different bandwidths should not be the burden of the end user or even the data provider but should be handled by adaptive software that balances users' needs and available network bandwidth. Parameters and models need to be developed that can describe user demands and usage scenarios. These would address questions and issues such as following:

- How "fresh" and recent should data be?
- How much precision and accuracy makes sense? What sampling rates are adequate?
- How much persistency is needed? For example, does a ring buffer holding one week of data provide enough persistence to guarantee that all relevant analyses and archival requirements are met before data is overwritten?
- How is data quality described, measured, and guaranteed? In particular, if data is automatically published from the field to the Web, how is quality assurance and quality control maintained?
- What access methods will best support users' requirements? Are http and ftp sufficient, or are database languages and APIs (e.g., SQL, JDBC) needed? How about digital library protocols and methods for data access in archived collections?
- How can data from different sources be combined and integrated? When such value-added mediation services are provided, how can the origin and provenance of data be tracked in order to give credit to the data providers?

Below we outline some promising directions toward facilitating distributed access, seamless retrieval, and interoperability of information from environmental networks. Detailed usage models and scenarios describing different types of users (scientists, policy makers, students, etc.) and their requirements will be helpful to determine specific instantiations of the frameworks described.

Information Technology for Data Exchange and Information Integration

Notwithstanding the specific needs of individual communities, the broad goals of seamless distributed access and retrieval from environmental networks are in fact common to many disciplines: Information systems have to be made interoperable such that heterogeneities in platforms, physical location and naming of resources, data formats and data models, supported programming interfaces and query languages, etc., all become transparent to the user. The need for such an interoperable *Grid infrastructure* [Foster, 1998] that can enable new science based on distributed computing, data sharing, and information integration is driving many national-scale projects in several disciplines, e.g. [NEON, NCEAS, LTER, KNB, NPACI, GBIF], as well as international efforts e.g. [GGF, 2001].

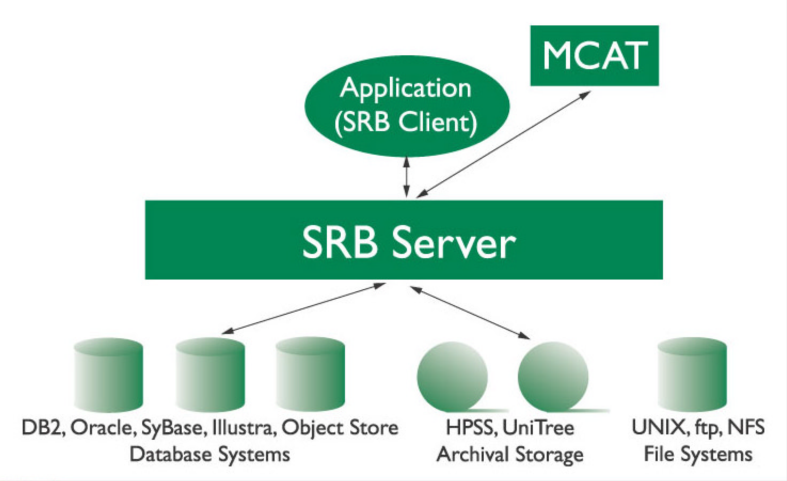
The services provided by such an infrastructure can be roughly classified as: (1) system and data interoperability issues, addressed by *Data Grid Services,* and (2) semantic interoperability and information integration issues, addressed by the *Semantic Mediation Services* of a future "*Knowledge Grid.*"

Data Grid Services

System aspects of interoperability include distributed storage across heterogeneous devices, data transport, access protocols, and distributed computing services. A prominent grid tool that addresses many system aspects is SDSC's Storage Resource Broker [SRB]. Instead of installing your own ftp site or Web server, and worrying about different device drivers, access control, location of distributed storage systems, disks, etc. a user simply becomes a member of a data grid by registering as an SRB user and installing a lightweight client or any standard Web browser. To the end-user, the SRB appears to be a *virtual drive* (the so-called "SRB space") into which environmental data can be put and from which other users (limited to authorized ones, if appropriate) can obtain data. The SRB makes transparent to the user such system aspects as:

- *How to access a specific storage device* (disk, tape, database, etc.). The SRB has an extensive and extensible set of "drivers" (aka "cartridges," "blades," "plug-ins") for storage devices.
- *Where the data set is located.* A user does not have to know or be concerned with the physical location of data sets. The SRB relieves her of having to deal with these details by managing all this information through a metadata catalog (MCAT).

In addition to transparent file access across heterogeneous devices and physical distribution, the SRB also provides solutions to other interoperability problems. For example, in addition to using the SRB as a sophisticated virtual drive (with access control, replica management, support for very large data sets, and other grid capabilities), it can also be used as a *relational data mediator.* By putting data into an SRB-accessible relational database, an *attribute-based* query and mediation mechanism becomes available to the user. This means that a user does not have to be concerned with the detailed structure of relational tables. Instead, the user can pick a set of attributes and search conditions on those attributes, after which the SRB will generate plans that span multiple tables (that may even reside in different parts of the world) and retrieve the desired data.



Client Service Middleware
The Storage Resource Broker (SRB) was developed at the San Diego Supercomputer Center (SDSC) and the National Partnership for Advanced Computational Infrastructure (NPACI) as client-server middleware to provide a uniform interface for connecting to heterogeneous data resources over a network and accessing replicated data sets. SRB, in conjunction with the Metadata Catalog (MCAT), provides a way to access data sets and resources based on their attributes rather than their names or physical locations.

Semantic Mediation Services

There is a recent trend toward "deeper" interoperability and integration of information beyond simple distributed access of data files. First, data sets need to be "wrapped" into a suitable *metadata envelope,* in order to facilitate deeper information integration beyond the data level. The metadata may provide all kinds of descriptive information about the data, including origin and provenance, data quality, accuracy, and last but not least, context information, e.g., the terminology or taxonomy used and a specification of the semantic context within a given domain ontology. Syntactically, metadata should be encoded in XML, the de facto standard for information exchange over the Web. XML is a flexible data format that can encode both regular data (from relational or object-oriented databases) as well as semistructured data (e.g., from system-generated Web pages). By using XML, a large number of tools for storing, querying, and manipulating XML-encoded information then become readily available. W3C standards related to applications of XML such as SOAP (for distributed object access), XML Schema (for modeling XML data), XQuery (for querying XML databases), and XSLT (for transforming XML output into a presentable form) provide a generic interoperability infrastructure based on open standards and tools, and are also employed in the development of grid services. Persistency and archival requirements can also benefit from an XML-based approach, as XML provides largely infrastructure-independent, self-describing means to represent information.

Agreed-upon metadata standards for environmental data are key to the reuse, interoperability, and integration of information. Meaningful links between disparate data are established and become "visible" and manageable to mediation services by using a set of *predefined attributes.* For complex scientific domains that require "semantically deep" dynamic querying of sources from different domains, new approaches such as *Model-Based Mediation* seem promising: In such a knowledge-based approach, the sets of attributes of different metadata standards do not stand in isolation but are mutually related to one another. Relationships between attributes and concepts across standards can be captured by a formalization of those relationships, for example, using logic rules directly [Ludaescher, 2001], or indirectly via the emerging standards developed in the context of the *Semantic Web* [Berners-Lee, 2001] effort, which aims at providing a generic infrastructure for semantic interoperability. The use of widespread, open standards and tools is also likely to positively influence the buy-in of the community, which is essential in order to create the desired high quality data and information content.

Recommendations

- Quality Control and Quality Assurance should be integrated into all aspects of data management, capture, transformation, integration and analysis.
- Maintain emphasis on funding biological informatics, especially collaborations between information technology researchers and biology research laboratories.
- Pay attention to usability and user needs. To enable new research with new kinds of people, the services and applications must be usable, and the NSF should pay close attention to mechanisms that set up feedback loops with the community the architectures serve. Establish a framework for evaluating the usability, use, and impact of evolving architectures and tools.
- Sustainability. Encourage enough labs to this kind of work to reach critical mass. Ensure that this scales socially and professionally. Establish peer review and formal mechanisms for collaboration. Support outreach activities.
- Encourage collaboration with broader, larger activities such as the NSDL.
- To be broadly interdisciplinary with other ESS disciplines, Biology needs to exert its research strengths in addition to a geoinformatics view of the world.
- Semantic Web is an interesting development; research proposals need to track it and build upon it.
- Specifications of user requirements need to be developed in detail and to inform and guide the process every step of the way.

SCALABLE INFORMATION NETWORKS FOR THE ENVIRONMENT

Site to Regional Scaling

James Gosz, Co-Facilitator
Long Term Ecological Research Network

Stuart Gage, Co-Facilitator
Michigan State University

William Michener, Reporter
LTER Network Office
University of New Mexico

Initial Questions

- How can the environmental sciences best employ emerging sensor and information technologies to address critical questions at broader ecological scales (i.e. moving from the site to the region)?
- What are the limitations (e.g. intellectual, technical, physical, and funding) to progress in this area?
- What are constructive solutions to overcoming these limitations?

"The environmental issues confronted in the second half of the 20th century approached the problem from the perspective of stressor, impact and mitigation. The environmental issues of the coming century will be resolved at the system level. Environmental problems within landscapes and ecosystems will, of necessity, be approached from within regional perspectives."

NSF (Bruce Hayden), 1998

A broader regional perspective will require that we expand our spatial and temporal horizons. Important issues include:

- Quantification of net primary productivity
- Land use and land cover change
- Flow of carbon in ocean and atmospheric systems
- Human population effects on ecological processes
- Distribution and abundance of exotic pests in terrestrial and aquatic systems
- Migration patterns of organisms in atmosphere and oceans
- Carbon sequestration by ecosystem types
- Effect of climate change on vegetation distribution
- Changing patterns of crop productivity
- Protection of ecosystems and human populations from terrorist actions

In the following discussion, we define what we mean by a region. Secondly, we present some of the factors that must be considered in developing a scalable regional environmental measurement infrastructure. Third, we focus on the constraints that exist in scaling up from sites to regions. Fourth, we discuss and suggest many of the common measurements that may appropriately be made at a regional scale. Finally, we present conclusions and recommendations for further action to address the future needs to enable scaling from the site to the region.

What Constitutes a Region?

Although regional boundaries are often defined in geopolitical terms, environmental boundaries often are not clear-cut. From an ecological perspective, we define a region as a "dynamic representation of a pattern or manifestation of a process." This definition reflects the view that regional environmental issues are not stable and can vary over time and space. Moreover, we suggest that it is the issue or the parameter to be measured that defines the region. Another way of thinking about scaling in space is to think in the context of sheds (as in watersheds) or "scapes" (as in landscapes). Thus, any given point in space may be contained within numerous regional "airsheds," watersheds, "foodsheds," "smellsheds," and "soundsheds." Furthermore, any point in space intersects with a hierarchy of spatial scales. When placed in this context, scaling from site to region (e.g., watersheds within basins) is one of the significant challenges in the 21st Century. Even a single parameter (e.g., spectral reflectance), is difficult and requires integrating methods and technologies across a range of scales such as from habitat->landscape->region.

Scaling from site to region entails identifying appropriate sensors, designing statistically sound ground validation studies, implementing quality assurance and quality control protocols at each scale, and developing robust techniques for integrating data across multiple scales.

S. Gage - Michigan State Univ.

Implementing a Scalable Information Network for the Environment (SINE)

Implementing SINE requires that we first define the concepts, the applications, and the challenges for a scalable information network for the environment. Questions must be identified as well as the validation information that may be appropriate across temporal and spatial scales. Second, indices of ecosystem function and other biological and physical indicators of environmental change must be developed to identify the appropriate sensors to document changes in the Biosphere. Third, sensors and sensor arrays must be deployed to remotely collect, analyze, and communicate environmental observations from within an ecosystem to one or more receiving sites. It is critical to evaluate our historical means of design for gathering information on processes that occur at regional scales and to develop new thinking about the spatial collection of key information. In addition, designing measurement networks based on hierarchical scales will challenge current computational infrastructures and computational resource management. Fourth, the analytical and communication network must be designed to facilitate the delivery of regional environmental information to the environmental science community (including across disciplines) and beyond to educators, policymakers, the media, and the public. This requires that we address issues related to the management and visualization of data, analyses, synthesis, and the quality and utility of model results.

Consequently, the considerations in developing a scalable regional environmental measurement infrastructure include:

- **Network design (time/space/location)**. As new networks are developed to measure environmental change from site to regional scales, the selection of the position and number of sensors systems in the region must include, among an array of logistical issues, the ability to interpolate between locations.

- **Measurement variables.** Selection of measurement variables should include a suite of measurements types that are universally important to ecosystem function, that can measure change at appropriate scales, and that are comparable between systems.
- **Sensor technology.** Significant advances in sensor technology and automation capacity provide new opportunities to measure ecological variables at rates and times that have not been feasible using historical measurement technologies.
- **Network deployment.** New strategies for the logistical deployment of arrays of sensor systems and decreases in sensor size provide opportunities to increase the density of sensors and communication rates for real time sensing of environmental change.
- **Communications.** Wireless communication technologies have radically increased opportunities and are changing conceptions and designs for real-time sensing in dynamic environments.
- **Operations/maintenance.** Error detection methods, component cost, and self-correcting and calibrating sensor systems can reduce costs of maintaining sensor systems.
- **Information archiving/management.** Storage capacity, cost/availability of on-line storage, and new models of data management and information mining provide new opportunities to capture structure and variation in regional processes.
- **Information analysis and interpretation.** One of the challenges facing the scientific community as we scale from site to region is the need to integrate highly detailed local data into broad scale patterns and processes at the regional level. Typically, this is done with models and broad scale measurements such as satellite imagery.
- **Information delivery/access.** The World Wide Web provides an unprecedented methodology to deliver quality information to the computer screens of the world and must be used coherently to educate the public regarding regional processes and patterns.

Scaling Challenges

There are a number of limitations that must be overcome before environmental monitoring and information networks can be expanded from site to regional scales. These limitations can be categorized as: intellectual, technical, physical, monetary, computational, biological, and industrial.

Intellectual challenges refer to conceptual difficulties that are encountered as we attempt to work at broader scales. There are often major philosophical and scientific hurdles that must be addressed as scales are expanded. Progress and approaches in particular scientific disciplines often reflect the characteristic scales at which the scientists are accustomed to working. Changing the customary scales of study may culminate in the formation of entirely new subdisciplines, as with “landscape ecology” in which the spatial breadth of ecology was greatly expanded along with related tenets and hypotheses. Intellectual limitations may also be associated with the background of the scientists and the difficulties associated with collaboration among scientists from different disciplines. Such multidisciplinary collaborations are often essential for making progress in understanding patterns and processes at broad scales, and can usefully be enabled by education and outreach across disciplines.

Technical challenges are most readily apparent for sensors, sensor arrays, and wireless communication. Sensors with potential applicability to sense the environment that were originally designed for industrial or indoor uses and may not be rugged enough to withstand placement in the environment. Sensors are often used as standalone devices and may not be designed to be integrated with the other types of sensors commonly used in environmental research. Many sensors used in environmental research are not fully automated and require frequent human intervention. Communicating data from remote field sensors to the point(s) of analysis remains a significant problem.

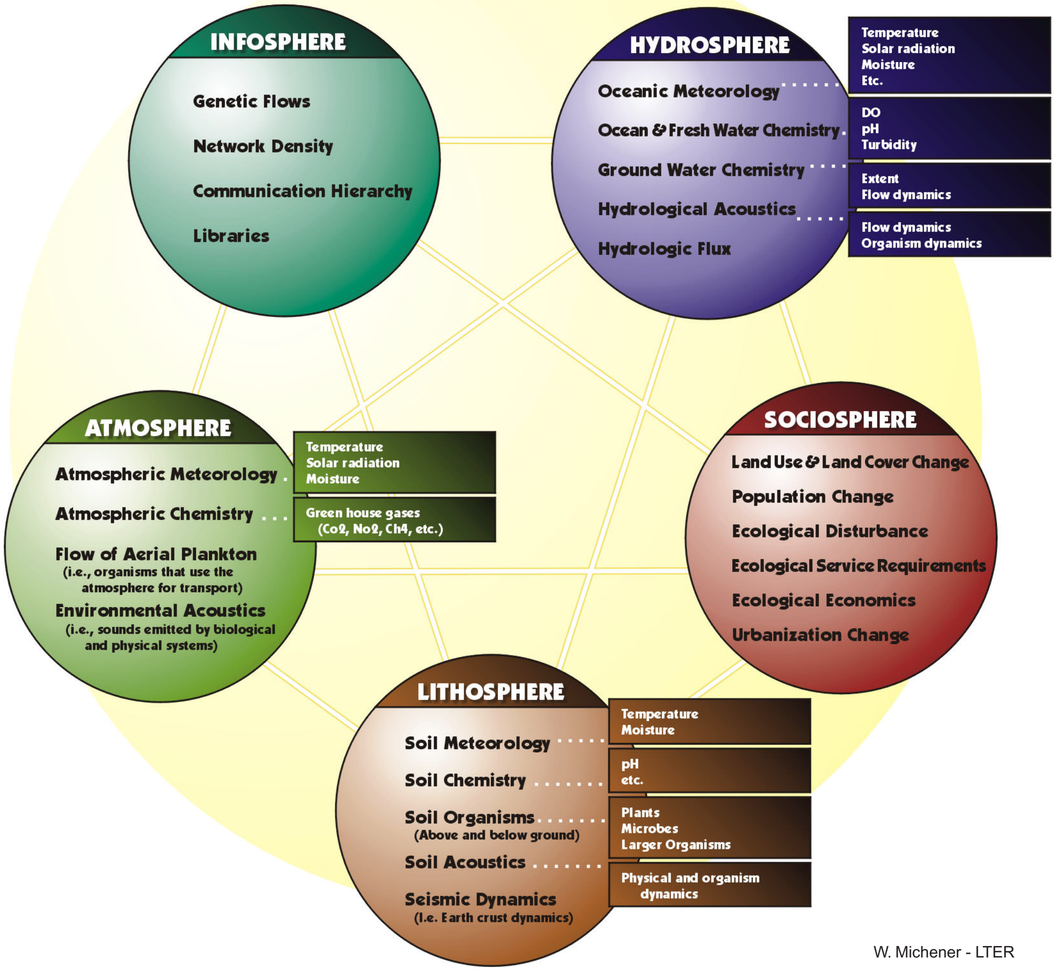
Physical and monetary challenges may also be significant and include the need for space for monitoring, computational, and communication equipment. Communication, maintenance, and calibration costs may represent large expenditures.

Computational challenges are associated with delivering, processing, managing, analyzing, and visualizing the enormous and rapidly-growing volumes of environmental data. Quality assurance and quality control require significant attention, but are often underdeveloped.

There are also significant biological challenges in scaling. The state of sensor technology is rudimentary for measuring many aspects of biological and ecological function. Furthermore, there are often no meaningful indices of what constitutes ecosystem function. It may also be difficult or impossible to monitor ecosystem function with adequate temporal and spatial resolution, and significant difficulties remain for integrating physical and biological data, which are often collected at very different scales of resolution.

Industrial challenges include the need for hardware miniaturization and ease of integration, the need for sensors and other technologies to be adaptable to multiple applications, and strategies for enhanced cost effectiveness.

Potential Variables That Relate To The State Of The Biosphere



W. Michener - LTER

Measures that are Scalable from Site to Region

We suggest that there are identifiable variables that have significant ecological meaning and characterize the function and integrity of ecological systems across scales. We have focused on identifying a broad range of environmental measurements that would be of great value in characterizing ecological and environmental change, including:

- Visual records of ecosystem activity (camera)
- Trapping and counting organisms
- Protein analysis (organism identification)
- Chemical sensing/nose (e.g. CO2, NOX, SO2, CH4)
- Chemical attraction (e.g. pheromones)
- Sonar, microwave, radar detection in the biosphere (e.g. organism movement)
- Sound detection/ear (e.g. organism communication, identification, soil organism activity, storm events, water flow)
- Flux quantification (e.g. energy, water)

Next, we identified those variables that would have broad value for regional pattern characterization associated with the function of the Biosphere (i.e. atmosphere, lithosphere, hydrosphere, sociosphere - human dimensions, and the infosphere).

Conclusion

Scalable Information Networks for the Environment (SINE) have enormous potential for advancing science, public awareness and education, and national and international commercialization. Improved information will depend upon how well we innovate and apply new concepts of remote detection technology, new time-series data collection and analysis, and ecosystem information synthesis. The resulting new information will support policy development and decision-making, as well as public awareness and visualization of the state of the environment and the significant rate of change that is occurring around us.

Several lessons were apparent from the workshop presentations. First, it is possible by properly applying current technology to collect useful biological information at a large scale. Second, a permanent site grid maintained over time provides a meaningful design for spatial time series analysis of the environment. This spatial-temporal information provides a critical modeling and analytical resource to explore scale and to assess risk. Third, patterns of change in biological systems may be highly dynamic, and must therefore be captured at scales and resolutions appropriate to issues facing society. Fourth, the changing nature of the environment is inextricably linked to the human dimension. For instance, political factors are a major component of exotic pest risk assessment.

Recommendations

- Developers of sensors should consider the design of sensors that are frequency, duration, and event-driven. More attention needs to be devoted to developing real-time and smart sensor technologies. Universal Sensor platforms (i.e. for plug and play sensors) are essential for supporting question-driven science.
- Develop descriptions of standard ways to measure a given phenomenon. Such information is needed to facilitate informatics, scaling and management, integration of remote sensing, and modeling. A systems approach to regionalization is clearly needed to encompass the multifaceted complexity (in the environment and across disciplines) in transitioning through the hierarchies of scale.

- Make building capacity in the environmental science community a major focus. Funding is urgently needed to build and enhance the computational and communications infrastructure at field stations and institutions that have the intellectual capacity to design and ask questions at appropriate scales.
- Scientists, scientific societies, and funding agencies must partner to establish best data management practices and policies that promote data and information sharing and establishment of national repositories for biodiversity and ecological data.

SCALABLE INFORMATION NETWORKS FOR THE ENVIRONMENT

Regional to Continental Scaling

Chaitan Baru, Facilitator
San Diego Supercomputer Center
University of California, San Diego

Philip Papadopoulos, Reporter
San Diego Supercomputer Center
University of California, San Diego

Initial Questions

- How can the environmental sciences best employ emerging sensor and information technologies to address critical questions at broader ecological scales (i.e. moving from regional to continental scales)?
- What are the limitations (e.g. intellectual, technical, physical, and funding) to progress in this area?
- What are constructive solutions to overcoming these limitations?

In this section, we address the *why* and the *how* of scaling networks to the continental level. The *why* involves the scientific issues that need to be studied at that scale. The *how* involves the technology issues related to scaling.

Scientific Issues

Identifying continental-scale scientific problems

There is a need to study continental-scale environmental science problems due to their broad impact on important issues such as resource management, community health, food production, bioterrorism, and industrial pollution. Examples of such problems include the spread of the West Nile virus, carbon sequestration, interaction of climate change and disease vectors, and the spread of invasive species. The last topic is of special interest due to the data that are already being collected at both the national and international level, and the economic and environmental impacts of invasive species. Such problems require the integration of information from a variety of sources, e.g. CDC data, bird observation data, mosquito data, and demographics (Census) information, etc.

International aspects

Continental-scale issues cut across national boundaries and introduce an international dimension to this problem. Indeed, regional issues may also have the same character, for example, study of the shared watershed in the San Diego/Tijuana border region. It is important to involve and interact with international partners to address the scientific as well as technological issues in research that spans national jurisdictions.

Industrial partners

Certain classes of environmental problems, e.g. monitoring air and water pollution, are also of great interest to the industrial sector in their efforts to comply with air and water pollution regulations. Thus, we recommend that studies in continental-scale issues should consider identifying industry sectors and "natural" industry partners who would be appropriate collaborators in the SINE effort.

Regional issues

In defining a "region" it is necessary to define the scope more broadly and employ a science-based definition. This can result in dynamic definitions of regions, rather than static, *a priori* political/geographic ones. Thus, a region could be defined based on its "homogeneity," e.g. a watershed or an air quality area may be defined as a region.

While political boundaries often do not correspond to the relevant region for environmental phenomena, they do have practical implications. A given region of the environment may span political boundaries, and as a result the data needed to study the region may come from different political and administrative entities. Thus, the data may well be heterogeneous in format, quality, and accessibility. The scientific results of the same study may have different impacts and importance in different political regions, due to differences in, say, science policy in each region. Indeed, how policy decisions are made and implemented may also vary widely across different political and administrative domains.

Regional-continental interactions

Environmental networks should facilitate regional-continental interactions. Information at the continental scale may reveal something of interest that causes a scientist to focus or "zoom" down to a regional level to better study the phenomena. Conversely, the more detailed information obtained at the regional level may sometimes contradict conclusions reached at the continental level, thus requiring an evaluation of the continental and regional-scale models.

Implementation and Technology Issues

Incorporating data and information from existing efforts

Continental scale studies will, at least in part, be based on the fusion of information from existing, major regional efforts. Thus, in arriving at a "common denominator" or set of standards for continental scale studies it will be most effective to identify common data, metadata, and other standards that are compatible with existing standards and conventions and can "piggyback" on them.

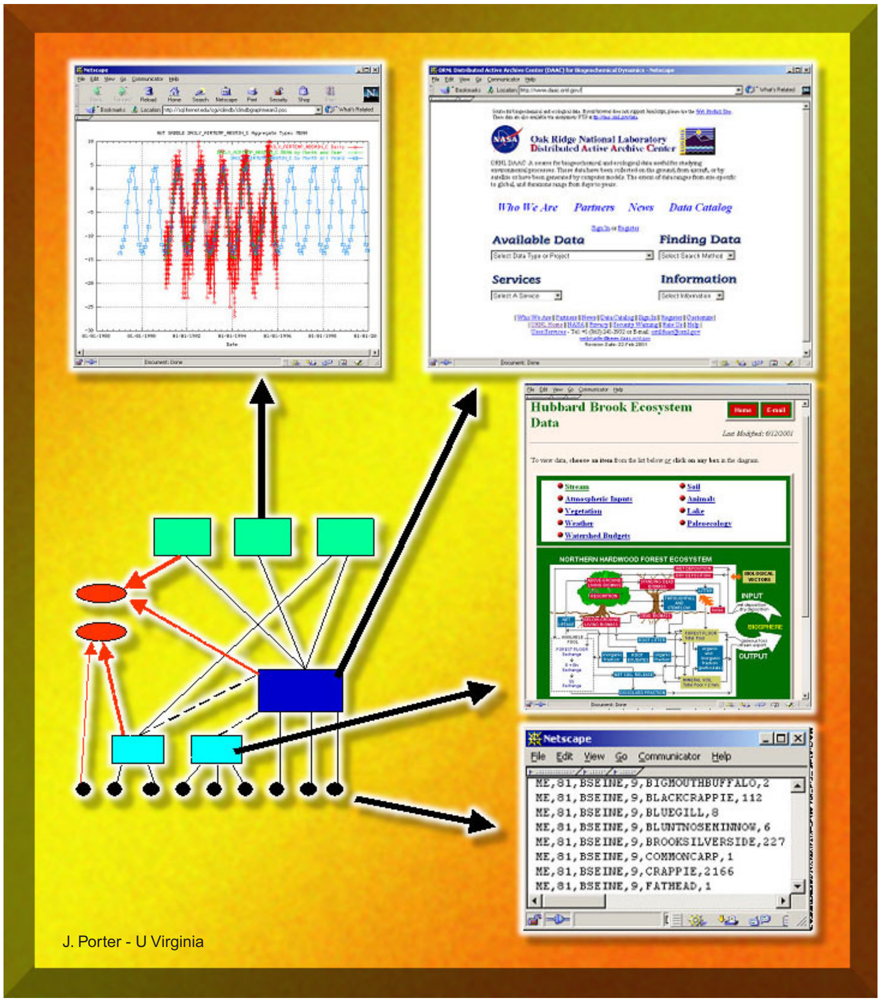
In such a large enterprise, the first step for the various participating parties is to "agree to agree." In terms of data and metadata standards, this means that there should be common agreement on the meta-standards that will be used. For example, the Extensible Markup Language (XML) is an example of a useful metadata standard in this context. Studies at the continental scale could agree to employ XML to encode metadata and, perhaps, data. This provides a basic degree of compatibility. Next, there will have to be common agreement and understanding on the schemas that will be employed to represent and transfer data, and so on. It is very important to initiate early efforts that will focus on defining metadata and data standards to enable the often-fragmented information from these existing sources to be combined and yield its full value.

Deploying continental-scale sensor networks

Combining information from existing regional studies allows the leveraging of existing projects. In addition, it is also important to consider how sensor networks can be deployed at the continental scale for new projects. For example, within a country such as the US, should they be distributed uniformly or in "representative" regions/ecosystems? These factors need to be weighed along with important infrastructure support issues, since deploying sensors at certain locations may be quite expensive (in terms of initial deployment as well as maintenance costs) due to inaccessibility of a region and/or restricted access. Another approach is to exploit existing infrastructure. For example, there is an extensive "network" of schools in the US across the entire country, often with high-speed Internet connections. These schools could be considered as possible sites for deploying sensors. School projects could be formulated around these sets of sensors so that each school provides the basic maintenance of its own set of sensors, thereby creating a powerful national network.

Information integration

Because of the range of disciplines and different types of data sources involved, environmental networks require IT approaches that can deal with issues of integration of information from heterogeneous sources. Continental-scale studies will impose an additional burden on the IT approaches since they will have to deal with further increases in heterogeneity in data formats, metadata schemas, and data quality, despite efforts to establish standards. It is recommended that XML-based standards and XML-based mediation of information be used as the approach for integrating this vastly heterogeneous data (see, e.g. [MIX], Mediation of Information using XML). GIS software should be designed to exploit spatial mediation capabilities so that information from multiple heterogeneous geospatial sources can be integrated into a single map. Another important issue is the ability to combine and integrate data with different accuracies, resolutions, and error characteristics. The mediation system must provide techniques for integrating such information and automatically handling the resulting error propagation across different search, retrieval, and analysis operations. Collaborations with ongoing efforts in this area (e.g. the [GeoGrid] project) will be useful.



```
ME,81,BSEINE,9,BIGMOUTHBUFFALO,2
ME,81,BSEINE,9,BLACKCRAPPIE,112
ME,81,BSEINE,9,BLUEGILL,8
ME,81,BSEINE,9,BLUNTNOSEMINNOW,6
ME,81,BSEINE,9,BROOKSILVERSIDE,227
ME,81,BSEINE,9,COMMONCARP,1
ME,81,BSEINE,9,CRAPPIE,2166
ME,81,BSEINE,9,FATHEAD,1
```

J. Porter - U Virginia

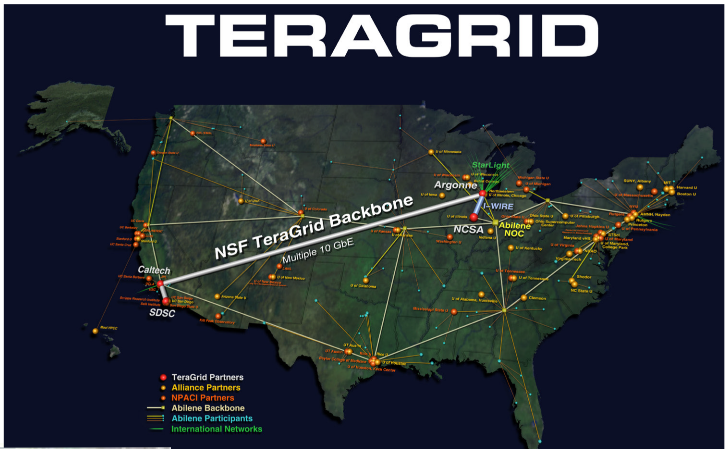
An idealized information system *would allow ready access by scientists (as depicted by the red ovals) to individual data sets and accompanying metadata (black circles: e.g. fish data in lower panel), project databases (aqua rectangles: e.g. Hubbard Brook Ecosystem Database - http://www.hubbardbrook.org/), regional and national databases (navy rectangle: e.g. Oak Ridge DAAC - http://www.daac.ornl.gov), or more specialized value-added databases (green rectangles: e.g. LTER climate database in the left panel - http://lternet.edu), as well as any desired combination thereof.*

A major aspect of information integration is the ability to access data from remote sites. While there are technical challenges that need to be addressed (e.g. database and security technologies), an even more important challenge is related to the policies for data sharing, especially from remote sources. The environmental science community needs to arrive at a consensus. As an initial step this can be done at a subdisciplinary level, if not at the highest level of integration.

Data sharing and archiving

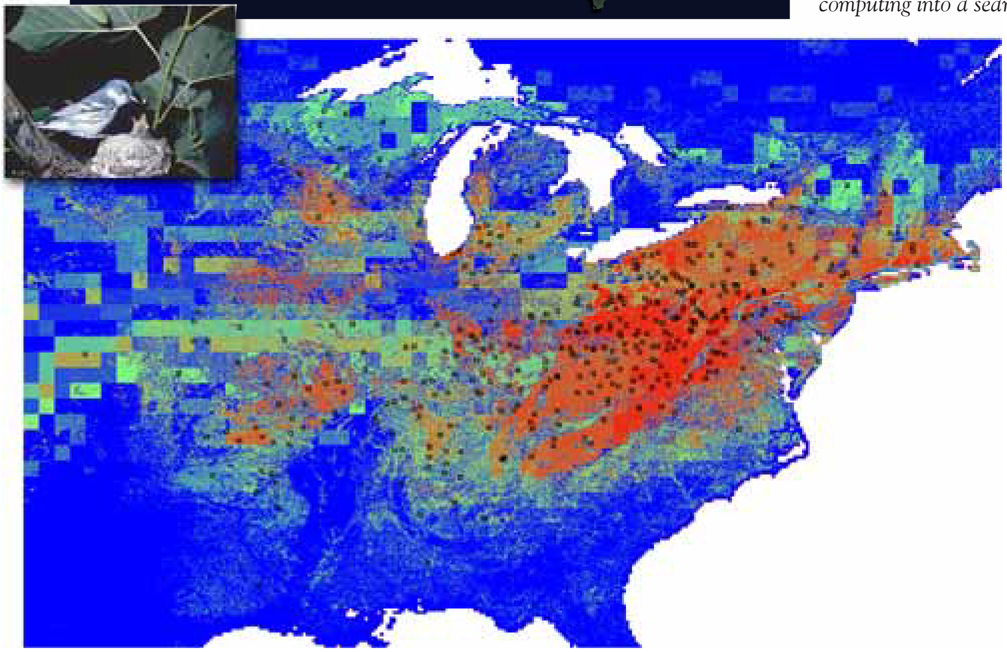
Continental-scale studies depend on data from widely dispersed sources. In addition to data sharing policies and technologies, the issue of data archiving needs to be addressed. For example, it may be useful and necessary to archive not just the results of an analysis but also the source data that was used in the analysis. If the data themselves are being obtained from multiple, distant sources, it will be necessary to arrive at common agreements and procedures for archiving the data as well as the results. Archiving continental scale data may well require the creation of a central repository or data archive. In addition, it would be useful to create an entity such as a *national environmental data archive* (NEDA), which could evolve as a distributed archive that leverages multiple existing archives in various subdisciplines. Another possible model to study is [IRIS], Incorporated Research Institutions for Seismology, which is also moving from a single, central archive model to a distributed archive model.

In general, it will be important to clearly define as early as possible a data sharing policy that is both technically workable and acceptable to the community.



TeraGrid is a multi-year effort to build and deploy the world's largest, fastest, most comprehensive, distributed infrastructure for open scientific research. When completed, the TeraGrid will include 13.6 teraflops of Linux cluster computing power distributed at the four TeraGrid sites, facilities capable of managing and storing more than 450 terabytes of data, high-resolution visualization environments, and toolkits for grid computing. These components will be tightly integrated and connected through a network that will initially operate at 40 gigabits per second. See http://teragrid.org

Recent demonstrations on a prototype TeraGrid have included the WhyWhere application by SDSC's David Stockwell , which combines a massive database of environmental and satellite data, efficient image processing algorithms, and grid-based cluster computing into a search and mapping system that allows biodiversity researchers to answer the question, "Where is it and why?" for any species, anywhere on the globe.



WhyWhere predicted distribution of potential habitat (red areas) for the vulnerable neotropical migrant bird the Cerulean Warbler (Dendronica cerulea) showing the combination of two environmental correlates at different resolutions: average December temperature at 0.5 degree grid cell size, and percent treecover at a resolution of 1 km grid cell size. The National Audubon Society believes the Cerulean Warbler is threatened by fragmentation of forested breeding habitat due to logging and development.

IT Training

A major aspect of scaling from regional to continental networks is the ability to provide access to the latest set of IT tools and training for scientists and technicians who are dispersed across the continent. As the state of the art in IT tools and technology keeps changing quickly, there is a need to keep personnel in the field trained in these latest technologies. For this purpose we strongly urge the creation of a “Data Institute,” which will provide IT expertise to scientific personnel to ensure that they are trained in the latest technologies. In addition, such an institute could also serve the role of archiving important community data sets as well as data and/or tools that individual scientists or groups wish to preserve in the form of a digital library.

SCALABLE INFORMATION NETWORKS FOR THE ENVIRONMENT

Data Sharing, IPR, and Human Dimension Issues

Paul Uhlir*, Facilitator
The National Academies

John Vande Castle, Reporter
LTER Network Office
University of New Mexico

* The views presented here are those of the author and not necessarily those of The National Academies or the National Research Council.

Initial Questions

- What are the critical *human dimension* issues that emerge as we begin to deploy environmental networks in addressing important scientific questions at increasingly broader scales (i.e. moving from site to regional to continental scales)?
- What are the limitations (e.g. intellectual, technical, physical, and funding) to progress in this area?
- What are constructive solutions to overcoming these limitations?

There are numerous legal, economic, and science policy or cultural factors that support open data sharing in the public domain. The public domain in scientific information may be defined as data and information that are ineligible by law to be protected or that are expressly excluded from protection, and that may therefore be disseminated and used without authorization (the discussion here is based on the definition of the public domain in scientific data and information presented in [Reichman and Uhlir, publication pending]). There are three broad categories of public-domain information that are relevant to environmental data sharing. These include: (1) data and databases not subject to protection under exclusive intellectual property (IP) rights; (2) otherwise protectable databases that are expressly designated as unprotected and hence in the public domain; and (3) fair-use exceptions.

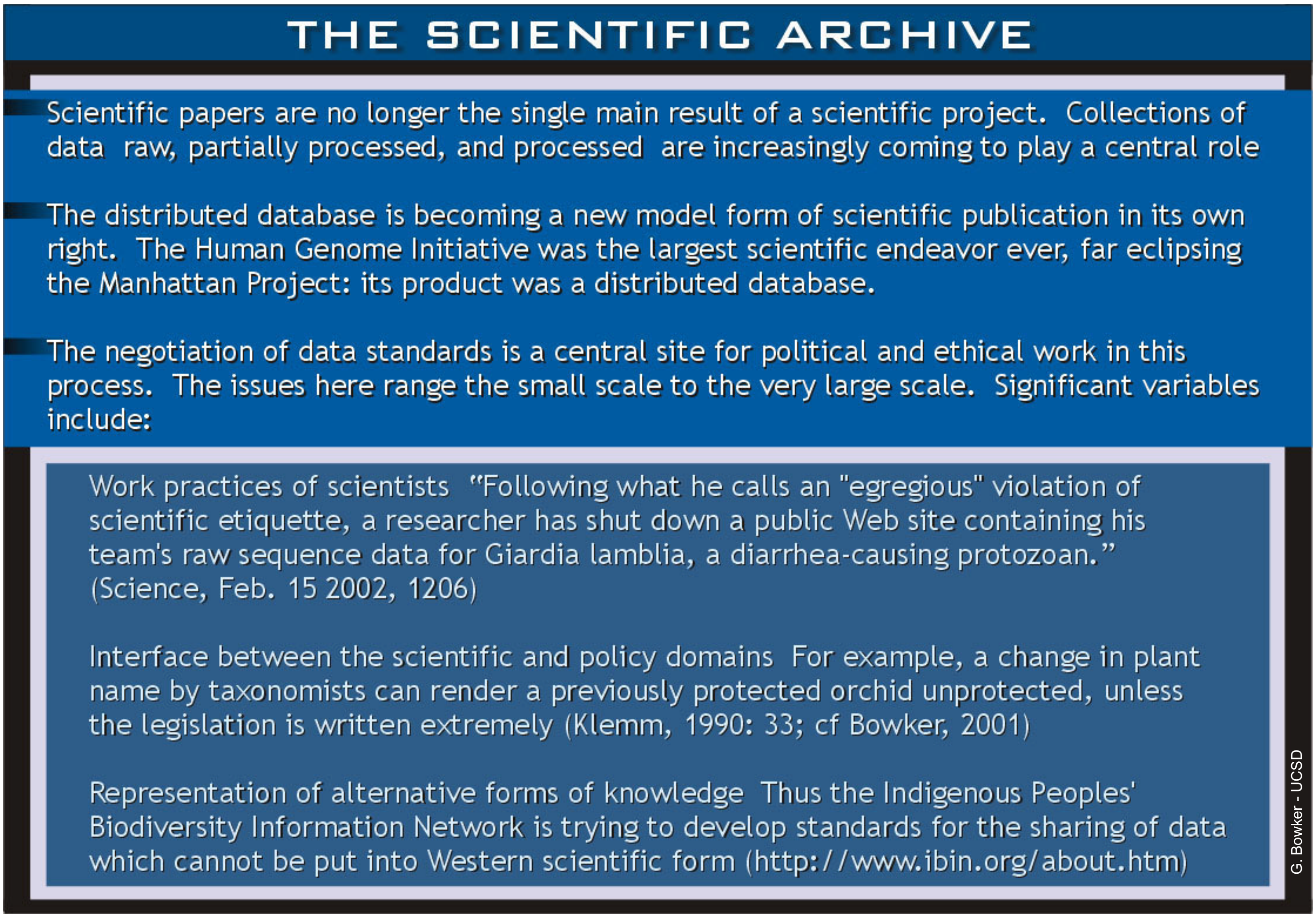
The first category of public-domain information is particularly broad and includes massive amounts of data and other types of information within it. There are three subcategories of public-domain scientific databases that are not subject to protection under exclusive property rights: (a) data that cannot be protected because of their source (i.e., the federal U.S. government and many state agencies); (b) databases for which the statutory period of protection has expired (under copyright law, the life of the author plus 70 years, or under the 1996 European Union Directive on the Legal Protection of Databases, 15 years, with a renewal of protection with each substantial update); and (c) ineligible or unprotectable components of otherwise protectable subject matter (e.g., factual data in databases, or ideas in copyrightable works).

Of these three subcategories, by far the largest and most important in the environmental data context is data and databases created by the federal government and by state governments that have open records laws. The major types of data here are those collected through government environmental satellite and *in situ* remote sensing programs and made available through government data centers and archives. Data and databases created by government agencies or employees are not protectable under copyright or other intellectual property mechanisms, and are subject to public access under the Freedom of Information Act, if they are not made openly disseminated.

The second major category of public-domain information, which consists of otherwise protectable data and databases that are expressly designated as unprotected, is of particular relevance to environmental research conducted in universities and other not-for-profit institutions. This includes data sets created primarily by academics, typically with government funding, who make their data openly available, or deposit their data in public-domain data centers or archives that are either operated by government or with government support. This category is potentially of greatest importance in the area of ecological and biodiversity studies, which are dominated by highly distributed, individual investigators. Unlike the situation in which the government directly produces the data, the data from academic research does not automatically enter into the public domain; it must be actively created rather than passively conferred. If the researcher does not make those data openly available either directly or through some open dissemination mechanism, and the research grant or contract does not stipulate that the data must be made available at some specific point, the presumption is that those data are protectable or proprietary.

There are several economic principles that support the broad dissemination of data resources in the public domain [Stiglitz, 2000]. The first is that basic research and related scientific data have public-good characteristics that make them appropriate to be undertaken as government or government-funded activities. The second is that the government has a well-justified role to play in promoting positive externalities from basic research and data activities. This is particularly true of data made available in an open and unrestricted way through the Internet, which results in a broad range of positive network externalities that are compounded exponentially by the addition of every new user of those data on the Web. Not only are the goals of science greatly enhanced by such open data sharing on digital networks, but there are enormous potential economic and social returns from the broad access and use of those data by individuals and institutions in many different sectors.

Finally, the public domain in scientific data and databases is fully consistent with the U.S. government's “full and open” data exchange policy for collaborative research at both the national and international level. This policy, which arose primarily in the context of geophysical research following the International Geophysical Year in 1957, states that “data and information from publicly-funded research be made available with as few restrictions as possible, on a nondiscriminatory basis, for no more than the cost of reproduction and distribution” (i.e., the marginal cost of the dissemination of data, which, on the Internet, is zero) [NRC, 1997; NRC, 1995]. Moreover, the “full and open” data sharing policy is strongly supported by the non-commercial value system of public-interest government and academic basic research. The values and goals of such research are best served by the maximum availability and distribution of data and research results, at the lowest possible cost, with the fewest restrictions on use, and with the active promotion of the reuse and integration of the fruits of existing research into new research [Reichman and Uhlir, pending].



THE SCIENTIFIC ARCHIVE

Scientific papers are no longer the single main result of a scientific project. Collections of data raw, partially processed, and processed are increasingly coming to play a central role

The distributed database is becoming a new model form of scientific publication in its own right. The Human Genome Initiative was the largest scientific endeavor ever, far eclipsing the Manhattan Project: its product was a distributed database.

The negotiation of data standards is a central site for political and ethical work in this process. The issues here range the small scale to the very large scale. Significant variables include:

Work practices of scientists "Following what he calls an "egregious" violation of scientific etiquette, a researcher has shut down a public Web site containing his team's raw sequence data for Giardia lamblia, a diarrhea-causing protozoan." (Science, Feb. 15 2002, 1206)

Interface between the scientific and policy domains For example, a change in plant name by taxonomists can render a previously protected orchid unprotected, unless the legislation is written extremely (Klemm, 1990: 33; cf Bowker, 2001)

Representation of alternative forms of knowledge Thus the Indigenous Peoples' Biodiversity Information Network is trying to develop standards for the sharing of data which cannot be put into Western scientific form (http://www.ibin.org/about.htm)

G. Bowker - UCSD

These legal, economic, and science policy factors provide a compelling rationale in support of data sharing and the placement of data from government and academic basic research in the public domain. Nevertheless, for data produced in the private sector, there are equally compelling reasons for not sharing data openly and for making such data proprietary. Although commercial, private-sector data activities are largely separate and separable from those conducted by the public-interest basic research sector, there are areas of significant overlap where the respective interests potentially conflict. Obvious instances of potential conflicts arise in the areas of biodiversity research that has both fundamental research and potential valuable biotechnology and pharmaceutical commercial applications. These pressures, which are broadly prevalent across science, are discussed further below.

There also can be a conflict in laws and policies favoring open, public-domain availability of environmental data with other laws and policies seeking to protect legitimate privacy and confidentiality interests. For example, ecologists, systematists, conservation biologists, and geologists, among others, frequently need to be able to keep data they collect confidential. Access to private lands is often contingent on the scientist providing the landowner with a guarantee of confidentiality. Public access to information on locations of rare species can readily lead to their exploitation and loss. Thus, field scientists may face an untenable conflict arising, on the one hand, from both NSF disclosure rules and Freedom of Information Act disclosure requirements and, on the other, the risk of being at odds with professional ethics. In this regard, it is important to note that exemptions from requirements for data release are available in other disciplines. The medical community is protected from requests to release health records of individuals. The archaeology community may keep site locations confidential based on the Archaeological Resources Protection Act of 1979. The Forest Service program for Forest Inventory and Analysis has partial exemption from release of data through the Food Security Act of 1985 (amended 1999). Similar protection is needed for scientists who collect data on private land about rare species. Such specific potential conflicts need to be better understood and anticipated to minimize the negative impacts on all the legitimate competing interests and to resolve them in a fair and balanced manner.

In addition to these fairly specific conflicting motivations for whether to share or not to share research data, there also are broader legal, economic, and policy factors arising from significantly increased intellectual property protections and economic pressures to privatize and commercialize scientific data that are encroaching into government and government-funded public-domain data activities [Reichman and Uhlir, pending]. Intellectual property laws in recent years have become broader, deeper, and longer in their scope and application, substantially reducing the scope of the public domain. For example, the term of copyright protection was extended by 20 years in the Sonny Bono Copyright Term Extension Act of 1998. An unprecedented strong exclusive property right in noncopyrightable databases was created for all Member States and affiliated members of the European Union by the Commission of the European Communities through the Directive on the Legal Protection of Databases in March, 1996. Similar efforts to enact strong legal protection of proprietary databases have been promoted in the U.S. Congress since that time. Perhaps most important, the trend in the private sector to license digital databases has brought about the greatest diminution in user rights. Because contracts for the dissemination of databases only confer rights to use, not purchase, the data, subject to the limitations imposed by the vendor, they bypass the traditional user rights that arose under the "First Sale" doctrine, and frequently override the fair uses available under copyright law [Reichman and Uhlir, 1999]. The legal validity of adhesion contracts (when the customer has no opportunity to negotiate) for information is still unsettled. However, there is an effort to make such adhesion contracts enforceable through the Uniform Computer Information Transactions Act, model legislation that is being promoted by information industry lobbyists at the state level. The licensing of databases, when supported by strong enabling legislation and enforced through digital rights management technologies such as encryption, download restrictions, access controls, and various hardware-based and software-based trusted systems, can remove large amounts of information from the public domain and greatly limit the scope of fair uses of data for scientific research.

These legal developments are being paralleled by economic pressures on both government agencies and universities to restrict public-domain availability of data. Federal science agencies are increasingly being directed to limit online dissemination of public data, and to outsource data collection activities and then license the data back with accompanying restrictions on use and redissemination. One example of this is the Commercial Space Act of 1998, which requires NASA to support private-sector data acquisition for space science and environmental research. Other similar pressures have been placed on Congress and the Office of Management and Budget to require other science agencies, including NOAA, DOE, and USGS to limit data dissemination and to license data from the private sector. Moreover, universities are commercializing the fruits of their research, including publicly funded research, in an effort to generate income to offset rising costs. This results in delays or prohibitions on the release of data and on the publication of research results.

Because of this confluence of legal, economic, and technological motivations to restrict the sharing of data and to reduce the availability of data in the public domain, it is essential for the government and academic scientific community to examine the terms and mechanisms for promoting data availability for research. The increased use of digital networks, data centers, and archives in the ecological and biodiversity communities would help to institutionalize data sharing protocols and promote greater access and use for the benefit of science. Similarly, the research granting agencies need to look at appropriate ways to better encourage and enforce the availability of data collected with public funds. There also have been a number of recent initiatives in the legal, library, and scientific communities to develop new mechanisms to preserve and promote the public domain in data and information [Reichman and Uhlir, pending]. These include efforts to develop public use licenses and copyleft notices that override the presumption of property rights and proprietary restrictions and instead actively confer public-domain status and rights of open access and use in data and information products. Public use licenses, coupled with implementing software, can be used to promote open access to nonprofits, while allowing commercialization efforts in the private sector. Such legal approaches need to be evaluated by the scientific community and applied as appropriate in an effort to offset the countervailing pressures to limit access to and uses of data for research. Finally, there are a number of community norms and cultural attributes - the "human dimensions" - relating to the willingness to share data and the creation of incentives for sharing data that need to be examined and addressed. The recommendations that follow focus on all these factors.

Recommendations

Data-Sharing Recommendations

- There are strong legal, economic, and science policy factors that support open availability and access to government and government-funded environmental data in the public domain; at the same time, the promotion of data sharing for research, education, and other public-interest purposes must nevertheless be balanced against competing proprietary and privacy requirements in certain circumstances.
- The NSF and other government agencies that support environmental research need to encourage and enforce open availability of the data created through that research.
- Mechanisms that should be considered for promoting data sharing include: (1) the establishment of government-supported data centers and archives that institutionalize public-domain availability of the data holdings, and (2) the more effective use of research grants and contracts to ensure that research data are made available no later than the end of the specific research project.
- In the university community, new legal mechanisms such as public use licenses and copyleft notices, need to be developed to promote open data availability in an era of increasing legal and economic proprietary protections.
- At the same time, statutory protection for non-disclosure may be needed for scientists who collect data on rare species or environmental data on private land, and this issue needs to be fully investigated.

Human / Social Factors Recommendations

- With regard to the human dimension aspects for promoting better data management practices and data sharing, it is important to establish effective incentives to promote not only physical infrastructure for long-term data storage and dissemination but also an educational component for training.
- Within individual projects, financial incentives from research funding agencies should be created for data management, archiving, and access. A professional reward system is needed for data management and data publication activities, especially from professional societies such as ESA, AIBS, ASLO, and others.
- Government grants programs should include more collaborative research opportunities for individual projects to include an interdisciplinary component. NSF and other science agencies should enhance multi-Directorate and cross-agency research opportunities integrating IT, education, and social science with traditional discipline research.

BIBLIOGRAPHY

SENSOR NETWORKS

Berger, J., J.A. Orcutt, F.L. Vernon, H.-W. Braun, and A. Rajasekar, 2001. Ocean wireless networking and real time data management, *EOS Trans.,* AGU, 82, F992.

Cook, G., 2000. Broadband spread spectrum wireless extends Internet reach of ISPs and field research scientists (interview with D. Hughes). *The COOK Report on the Internet.* 9(4):1-15. http://www.cookreport.com/.

Delin, K.A, and Jackson, S.P., 2001. The Sensor Web: A New Instrument Concept. SPIE Symposium on Integrated Optics, San Jose, California.

Helly, J., T. T. Elvins, et al, in press. Controlled Publication of Digital Scientific Data, CACM, (accepted October 3, 2000).

Ho, C. K., M. Kelly, M.T. Itamura, and R.C. Hughes, 2001. Review of Chemical Sensors for *In-Situ* Monitoring of Volatile Contaminants.

Krabach, T., 2000. Breakthrough sensor technology for space exploration in the 21st century. Aerospace Conference Proceedings, *IEEE,* 6:565-569.

Nagel, D. J., 2000. Pervasive Sensing, *SPIE,* 4126:71-82.

Nagel, D. J., in press, Micro-Sensor Clusters, *Microelectronics Journal.*

Pister, K. S. J., J.M. Kahn, and B.E. Boser, 1999. Smart Dust: Wireless Networks of Millimeter-Scale Sensor Nodes, Highlight Article in 1999 Electronics Research Laboratory Research Summary.

Price, K.P, T.J. Crooks, and E.J. Martinko, 2001. Grasslands Across Time and Scale: A Remote Sensing Perspective, *Photogrammetric Engineering and Remote Sensing,* 67(4).

Smith, D., S. Carbotte, S. Cande, W. Ryan, S. Miller, and D. Wright, 2001. Data Management for Marine Geology and Geophysics Workshop Report: Tools for Archiving, Analysis and Visualization, Workshop Report, La Jolla, CA, May 14-16, 2001, www.geo-prose.com/projects/pdfs/data_mgt_report.low.pdf.

Staples, E.J., 2000. Field Analysis Using a Novel Electronic Nose as an Environmental Tool. Paper presented at the American Chemical Society, San Francisco, California.

DATA TECHNOLOGIES

Baker, K.S., B.J. Benson, D.L. Henshaw, D. Blodgett, J.H. Porter, and S.G. Stafford, 2000. Evolution of a multisite network information system: the LTER information management paradigm, *BioScience*, 50(11):963-978.

Berners-Lee, T., J. Hendler, and O. Lassila, 2001. The Semantic Web. *Scientific American,* http://www.sciam.com/2001/0501issue/0501berners-lee.html.

Foster, I., and C. Kesselman (eds.), 1998. *The Grid: Blueprint for a New Computing Infrastructure,* Morgan Kaufmann Publishers, ISBN 1-55860-475-8.

GBIF, Global Biodiversity Information Facility, http://www.gbif.org/.

Global Grid Forum, 2001. http://www.gridforum.org/.

HPWREN, High Performance Wireless Research and Education Network (see also ROADNet), http://hpwren.ucsd.edu/.

KNB, Knowledge Network for Biocomplexity, http://knb.ecoinformatics.org/.

LTER, NSF Long-Term Ecological Research, http://lternet.edu/.

Ludaescher, B., A. Gupta, and M.E. Martone, 2001. Model-Based Mediation with Domain Maps, *Proc. Intl. Conference on Data Engineering (ICDE)*, IEEE Computer Society, Heidelberg, Germany, April, 2001, 81-90.

NCEAS, National Center for Environmental Analysis and Synthesis, http://www.nceas.ucsb.edu/fmt/doc?/frames.html.

NEON, National Ecological Observatory Network, http://www.sdsc.edu/NEON/mar2000/neon2_report.html.

NetCDF, network Common Data Form, http://www.unidata.ucar.edu/packages/netcdf/

NPACI, National Partnership for Advanced Computational Infrastructure, http://www.npaci.edu/.

OpenGIS, http://www.opengis.org/.

ROADNet, Real-time Observatories, Applications, and Data management Network (see also HPWREN), http://roadnet.ucsd.edu/index.html.

Species Analyst, http://tsadev.speciesanalyst.net/.

SRB, SDSC Storage Resource Broker, http://www.npaci.edu/DICE/SRB/.

SCALABLE INFORMATION NETWORKS FOR THE ENVIRONMENT

Bowker, G.C., 2000. Biodiversity Datadiversity, *Social Studies of Science*, 30(5):643-684.

GeoGrid, Enabling the Creation and Use of GeoGrids for Next Generation Spatial Information, http://www.npaci.edu/DAKS/GeoGrid.

IBIN, International Biodiversity Information Network, http://wwwibin.org/about.htm

IRIS, Incorporated Research Institutions for Seismology, http://www.iris.edu.

Klemm, C.D., International Union for Conservation of Nature and Natural Resources, et al., 1990. *Wild Plant Conservation and the Law,* IUCN-The World Conservation Union, Gland, Switzerland.

MIX, Mediation of Information using XML, http:www.npaci.edu/DAKS/MIX.

National Research Council, 1995. *On the Full and Open Exchange of Scientific Data,* National Academy Press, Washington, D.C.

National Research Council, 1997. *Bits of Power: Issues in Global Access to Scientific Data,* National Academy Press, Washington, D.C.

Reichman, J. H., and P.F. Uhlir, 1999. Database Protection at the Crossroads: Recent Developments and Their Impact on Science and Technology, *Berkeley Technology Law Journal,* 14(2): 799-821.

Reichman, J.H., and P. F. Uhlir, publication pending. *Promoting Public Good Uses of Scientific Data: A Contractually Reconstructed Commons for Science and Innovation.*

Stiglitz, J.E., P.R. Orszag, and J.M. Orszag, 2000. *The Role of Government in a Digital Age,* Computer and Communications Industry Association, Washington, D.C.

APPENDIX 1

WORKSHOP AGENDA

San Diego Supercomputer Center
October 29-31, 2001 (Writing Day = November 1)

PowerPoint versions of these presentations are available at:
http://www.sdsc.edu/pbi/sine_workshop_agenda.html.

October 29 **Keynote by M. Cavanaugh (NSF) and Sensor Networks**
7:00 - 8:00 Shuttles to SDSC from the Radisson Hotel
7:30 - 8:00 Breakfast Buffet at SDSC

8:00 - 8:30 **Alison Withey** (SDSC) - Workshop background and objectives.
Fran Berman (SDSC) - Welcoming address.

8:30 - 9:00 **Margaret Cavanaugh** (NSF) - Environmental Cyberinfrastructure: turning data into knowledge.

9:00 - 9:30 **Deborah Estrin** (UCLA) - Next century challenges: scalable coordination in sensor networks.

9:30 - 10:00 **Gregory Bonito** (LTER) - In situ environmental sensor technologies.

10:00 - 10:30 Coffee Break

10:30 - 11:00 **Doug Goodin** (Kansas State U.) - Environmental remote sensing technologies.

11:00 - 11:30 **Dave Hughes** (Old Colorado City Communications) - Wireless environmental science.

11:30 - 12:00 **Kenneth Johnson** (MBARI) - Marine/aquatic sensor arrays.

12:00 - 12:30 **Frank Vernon** (Scripps Institution of Oceanography) - Wireless networks and sensor connectivity: HPWREN.

12:30 - 1:15 Lunch (catered)

1:15 - 1:30 **William Michener** (LTER) - Announcements and charge for the breakout sessions.

1:30 - 3:30 **Concurrent Breakout Sessions**
Theme 1: Design and implementation of aquatic and marine sensor networks.
(Facilitator, Orcutt; Reporter, Helly)

Theme 2: Design and implementation of terrestrial sensor networks.
(Facilitator, Waide; Reporter, J. Porter)
Theme 3: Sensor technologies. (Facilitator, Goodin; Reporter, Bonito)

3:30 - 4:00 Coffee Break

4:00 - 5:30 **Reports from Breakout Sessions**
(20 minute presentation by breakout session facilitator/10 minute discussion each)

6:00 - 7:30 Reception at the Radisson Hotel [sponsored by Cal-(IT)^2]

October 30 Data Technologies

7:00 - 8:00 Shuttles to SDSC from the Radisson Hotel
7:30 - 8:15 Breakfast Buffet at SDSC

8:15 - 8:30 **Alison Withey** (SDSC) - Announcements.

8:30 - 9:00 **Cherri Pancake** (Oregon State) - Enabling technologies and user requirements for data and information management and delivery.

9:00 - 9:30 **John Porter** (UVA) - Information systems for ecological research.

9:30 - 10:00 **Robert Peet** (UNC) - Taxonomic plot and specimen databases.

10:00 - 10:30 Coffee Break

10:30 - 11:00 **Jim Beach** (KU-BRC) - Biodiversity data retrieval and integration.

11:00 - 11:30 **Matt Jones** (UC - Santa Barbara) - Data integration, analysis, and synthesis.

11:30 - 12:00 **Jim Quinn** (UC Davis) - Technologies for integration and discovery of geospatial data.

12:00 - 12:30 **Mike Bailey** (SDSC) - Scientific data visualization.

12:30 - 1:15 Lunch (catered)

1:15 - 1:30 **William Michener** (LTER) - Announcements and charge for the breakout sessions.

1:30 - 3:30 **Concurrent Breakout Sessions**
Theme 1: Geospatial data integration.
(Facilitator, Quinn; Reporter, Stocks)
Theme 2: Distributed data access and retrieval.
(Facilitator, Beach; Reporter, Ludaescher)
Theme 3: Interfaces, portals, and knowledge environments.
(Facilitator, Pancake; Reporter, Jones)

3:30 - 4:00 Coffee Break

4:00 - 5:30 **Reports from Breakout Sessions**
(20 minute presentation by breakout session facilitator/10 minute discussion each)

October 31 Scalable Information Networks for the Environment
7:00 - 8:00 Shuttles to SDSC from the Radisson Hotel
7:30 - 8:15 Breakfast Buffet at SDSC

8:15 - 8:30 **Alison Withey** (SDSC) - Announcements.

8:30 - 9:00 **William Michener** (LTER) - Environmental information networks: the field station reality.

9:00 - 9:30 **Geoff Bowker** (UCSD) - Scaling environmental information networks: the human dimension.

9:30 - 10:00 **Warren Cohen** (USDA Forest Service) - Integration across scales: the role of remote sensing and models.

10:00 - 10:30 Coffee Break

10:30 - 11:00 **Raymond McCord** (Oak Ridge National Laboratory) - Regional databases and archives.

11:00 - 11:30 **Terry Smith** (UCSB) - Digital Libraries: conceptual & technical framework.

11:30 - 12:00 **Phil Papadopoulos** (SDSC) - Scalable computational infrastructure: workstations, clusters, grid computing.

12:00 - 12:30 **Chaitan Baru** (SDSC) - Data and knowledge-based grids.

12:30 - 1:15 Lunch (catered)

1:15 - 1:30 **William Michener** (LTER) - Announcements and charge for the breakout sessions.

1:30 - 3:30 **Concurrent Breakout Sessions**
Theme 1: Environmental networks: site to regional scaling.
(Co-Facilitators, Gage & Gosz; Reporter, Michener)
Theme 2: Environmental networks: regional to continental scaling.
(Facilitator, Baru; Reporter, Papadopoulos)
Theme 3: Data sharing, IPR, and human dimension issues.
(Facilitator, Uhlir; Reporter, Vande Castle)

3:30 - 4:00 Coffee Break

4:00 - 5:30 **Reports from Breakout Sessions**
(20 minute presentation by breakout session facilitator/10 minute discussion each)

November 1 Writing day for breakout session Reporters and Facilitators

APPENDIX 2

SINE WORKSHOP PARTICIPANTS

Andelman, Sandy
NCEAS
andelman@nceas.ucsb.edu

Arzberger, Peter
SDSC/UCSD
parzberg@sdsc.edu

Bachman, Mark
University of California, Irvine
mbachman@uci.edu

Bailey, Mike
SDSC/UCSD
mjb@sdsc.edu

Baker, Polly
NCSA
baker@ncsa.uiuc.edu

Baru, Chaitan
SDSC/UCSD
baru@sdsc.edu

Beach, James
University of Kansas
beach@ku.edu

Berman, Fran
SDSC/UCSD
berman@sdsc.edu

Bonito, Gregory
LTER Network Office
gbonito@lternet.edu

Bowker, Geoffrey
Communication Dept., UCSD
bowker@ucsd.edu

Braun, Hans-Werner
NLANR/SDSC
hwb@nlanr.net

Cavanaugh, Marge
National Science Foundation
mcavanau@nsf.gov

Cheeseman, John
Plant Biology/University of Illinois
j-cheese@uiuc.edu

Cohen, Warren
USDA Forest Service
warren.cohen@orst.edu

Cushing, Judith Bayard
The Evergreen State College
judyc@evergreen.edu

Estrin, Deborah
UCLA/Computer Science Department
destrin@cs.ucla.edu

Flikkema, Paul
Northern Arizona University
paul.flikkema@nau.edu

Frost, Eric
San Diego State University
frost@imagine.sdsu.edu

Gage, Stuart
Michigan State University
gages@msu.edu

Goodin, Douglas
Kansas State University
dgoodin@ksu.edu

Gosz, James
Chairman, LTER Network
jgosz@sevilleta.unm.edu

Graybeal, John
MBARI
graybeal@mbari.org

Greene, Thomas
National Science
tgreene@nsf.gov

Helly, John
SDSC
hellyj@ucsd.edu

Hughes, David
Old Colorado City Communications
dave@oldcolo.com

Itsweire, Eric
National Science Foundation
eitsweir@nsf.gov

Johnson, Kenneth
MBARI
johnson@mbari.org

Jones, Matthew
NCEAS
jones@nceas.ucsb.edu

Kloeppel, Brian
Coweeta LTER
kloeppel@sparc.ecology.uga.edu

Ludaescher, Bertram
SDSC/UCSD
ludaesch@sdsc.edu

Mantey, Patrick
University of California, Santa Cruz
mantey@soe.ucsc.edu

McCord, Raymond
Oak Ridge National Laboratory
mccordra@ornl.gov

Michener, Bill
LTER Network Office
wmichene@lternet.edu

Miller, Stephen
Scripps Institution of Oceanography
spmiller@ucsd.edu

Morris, Robert
UMASS-Boston
ram@cs.umb.edu

Ogle, Simeon
USC
sogle@rcf.usc.edu

Orcutt, John
Scripps Institution of Oceanography
jorcutt@igpp.ucsd.edu

Pancake, Cherri
Oregon State University/NACSE
pancake@nacse.org

Papadopoulos, Philip
SDSC
phil@sdsc.edu

Peet, Robert
University of North Carolina/NCEAS
peet@unc.edu

Keith Pezzoli
Urban Studies/UCSD
kpezzoli@ucsd.edu

Porter, Dwayne E.
University of South Carolina
porter@sc.edu

Porter, John
University of Virginia
jporter@lternet.edu

Quinn, Jim
University of California, Davis
jfquinn@ucdavis.edu

Rajasekar, Arcot
SDSC/UCSD
sekar@sdsc.edu

Reichman, Jim
NCEAS
reichman@nceas.ucsb.edu

Roy, Donna
USGS Center of Biological Informatics
droy@usgs.gov

Shapiro, Sedra
SDSU Field Station Programs
sshapiro@sciences.sdsu.edu

Skog, Judith
National Science Foundation
jskog@nsf.gov

Smarr, Larry
Cal(IT)2/UCSD
lsmarr@ucsd.edu

Smith, Terry
UCSB
smithtr@cs.ucsb.edu

Stevenson, Robert
U Mass-Boston
Robert.Stevenson@umb.edu

Stocks, Karen
SDSC/UCSD
stocks@sdsc.edu

Uhlir, Paul
The National Academies
puhlir@nas.edu

Alex Ushakov
UCSB
aushako@alexandria.ucsb.edu

Vande Castle, John
LTER Network Office
jvc@lternet.edu

Vernon, Frank
Scripps Institution of Oceanography
flvernon@ucsd.edu

Waide, Robert
LTER Network Office
rwaide@lternet.edu

Williams, Tom
NSF Wireless Field Tests
tomw@oldcolo.com

Willig, Michael
National Science Foundation
mwillig@nsf.gov

Withey, Alison
SDSC/UCSD
awithey@sdsc.edu

Zhang, Phoebe Y.
Rutgers University
phoebe@imcs.rutgers.ed