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

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.