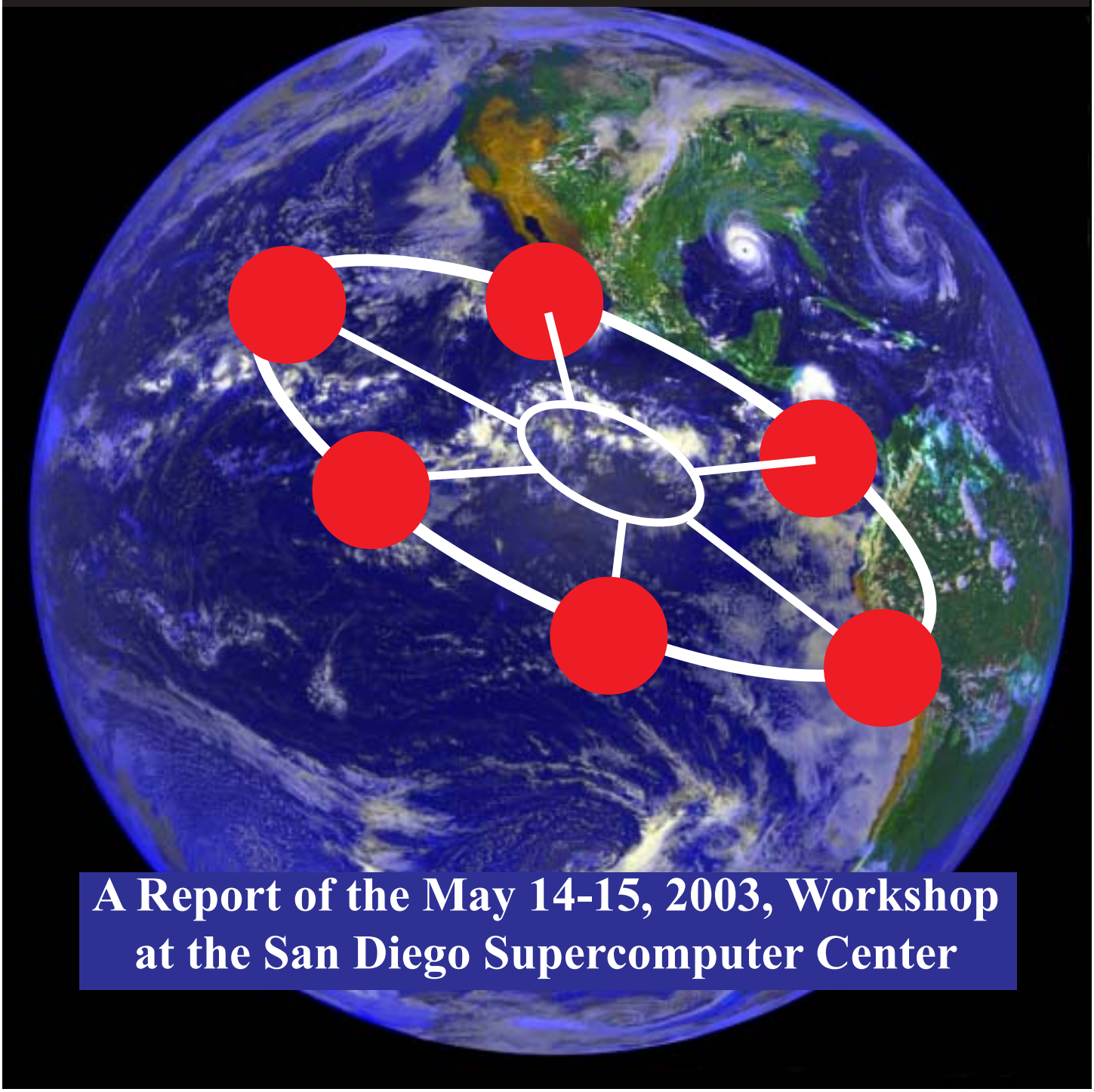
Kansas Geological Survey Open-file Report 2004-48

GEOINFORMATICS:

BUILDING CYBERINFRASTRUCTURE FOR THE EARTH SCIENCES



A Report of the May 14-15, 2003, Workshop at the San Diego Supercomputer Center

Workshop Organizers

Lee Allison, Kansas Geological Survey
Chaitan Baru, San Diego Supercomputer Center
Thomas Jordan, Southern California Earthquake Center
Workshop Facilitator: Robert Crangle, Rose and Crangle, Ltd

GEOINFORMATICS:

BUILDING CYBERINFRASTRUCTURE FOR THE EARTH SCIENCES

A Report of the May 14-15, 2003, Workshop at the San Diego Supercomputer Center

Workshop Organizers

Lee Allison, Kansas Geological Survey

Chaitan Baru, San Diego Supercomputer Center

Thomas Jordan, Southern California Earthquake Center

Workshop Facilitator: Robert Crangle, Rose & Crangle, Ltd

This Workshop was funded in part by the United States National Science Foundation under Grant EAR-0327961

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This is the report of a workshop held at the San Diego Supercomputer Center on May 14-15, 2003. The workshop was designed as a focused session to follow up suggestions from earlier reports and meetings that the many subdivisions of the earth sciences should make a significant commitment to information technology infrastructure, or cyberinfrastructure, to serve the needs of their entire community. As a matter of shorthand, this had become known as "geoinformatics."

In response to a proposal by the organizers, the Division of Earth Sciences (EAR) in the National Science Foundation (NSF) funded this workshop to consider the next steps in geoinformatics. The workshop was entitled "Building the Geoinformatics System."

Community interest was such that the workshop was oversubscribed. It attracted a wide spectrum of earth science researchers, knowledgeable information technologists, and education and outreach specialists. It also drew concerned persons from allied environmental sciences (atmospheric, oceanic, and so forth), the U.S. Geological Survey (USGS), and the NSF.

Participants worked hard, both as a committee of the whole and in small groups, to reach a common understanding of the issues, the challenges, and the path to be taken. A fundamental objective of the workshop was to assess not only community desires, but also the resources necessary to fulfill them.

Nine primary findings emerged:

- **Both interest and momentum towards development of cyberinfrastructure (geoinformatics) are found throughout all the domains of earth sciences.**
- **The relevant federal agencies are supporting significant geoinformatics activities.**
- **"GeoGoogle" is the symbolic goal: providing access to as many interested persons from as many disciplines and levels of sophistication as possible.**
- **Hundreds of geoinformatics-enabling projects are underway, largely independent of each other and not using common standards or metadata.**
- **Various organizational models exist to foster geoinformatics coordination.**
- **Issues of attribution of authorship and ownership of various datasets and databases are of great importance, but should not be 'deal-breakers' to moving forward.**
- **The earth scientists, rather than the information technologists, must lead the way.**

- **Successful coordination of geoinformatics will create transformational opportunities throughout all the geosciences.**

- **It is possible and practical for an interim steering group to begin to address these issues and to encourage organization around the grand earth science challenges, shared instrumentation, facilities, and other resources.**

Participants agreed at the end on four significant recommendations for action:

- **To understand the cultural and technical requirements among the various subdisciplines of the earth sciences (and allied fields) for enabling geoinformatics.**

- **To determine what form of organization is most acceptable to the community to enable geoinformatics.**

- **To define implementation tactics for geoinformatics.**

- **To develop and continually refine processes based on community participation.**

Many of the questions posed at the workshop were addressed and resolved by consensus. However, some were not. They primarily surround the form of an organization that must emerge to enable geoinformatics implementation. Who should run such an organization, how should its decisions be made, and where should the administrative components be housed?

The workshop endorsed the creation of a group (variously referred to as a Geoinformatics Task Force or an Interim Steering Committee) to investigate these matters. They suggested that this group report to the entire community on its conclusions.

BACKGROUND

Motivation - The Revolution is Underway

"We need a common way for everybody to communicate with information and knowledge."

John Scully, formerly of Apple Computer, *Baseline* magazine, December 2001.

"A unique challenge…is the integration of four-dimensional…digital data… to obtain a holistic view of complex Earth and human systems."

Complex Environmental Systems – Synthesis for Earth, Life, and Society in the 21st Century, A 10-Year Outlook for the National Science Foundation, NSF Advisory Committee for Environmental Research and Education, January 2003.

"Enabled by information technology, a qualitatively different and new science and technology infrastructure has evolved."

National Science Board, Science and Engineering Infrastructure for the 21st Century, February 2003.

"An ongoing series of cyberinfrastructure workshops would facilitate discussion, cooperation, and community building."

Cyberinfrastructure for Environmental Research and Education (Report from a workshop held at the National Center for Atmospheric Research, October 30-November 1, 2002), Environmental Cyberinfrastructure Steering Committee, September 2003.

Early in the workshop Walter Snyder of EAR reported to the participants that the number of groups within the earth science community that are presently involved in or developing geoinformatics efforts has reached a size where the management and coordination has become critical. He compiled a list of more than 400 earth scientists from 125 universities and colleges involved in various NSF-supported geoinformatics programs and projects. This excludes their state and federal agency and private industry collaborators. He emphasized that there are not only many major projects, but also a lot of smaller ones, underway.

The report of the NSF Blue Ribbon Advisory Panel on Cyberinfrastructure (NSF BRAPC, 2003) outlines a vision for the infrastructure necessary to support the revolution in how science is done. Like most revolutions in their early stages, the geoinformatics revolution is scattered and disorganized**.** In early 2003 significant activity was underway in both the environmental spatially based disciplines and in computer science, but they are proceeding in a largely uncoordinated fashion.

A generation of access to the internet/worldwide web has combined with low cost but very powerful computing power to foster hundreds, and perhaps thousands, of innovative

projects. The revolution can continue, unfocused and uncoordinated, in a slow, inefficient process. Or, the community can facilitate coordination and incentives for cooperation and collaboration, with the intent of enabling as quickly as possible a comprehensive integrated process that enhances scientific inquiry.

The active geoinformatics community is growing rapidly. For example, at the "Solid Earth Science" workshop at the Geological Society of America Annual Meeting in Denver, Fall 2002, the 60+ attendees immediately stated the need for a "digital information system." This is clearly a community-based effort with the community ahead of NSF's ability to provide enough funding.

The needs of these growing geoinformatics efforts cannot be met with disjointed, ad hoc approaches. Through this and other workshops, and discussion with the existing organizations, it is clear that the community: 1) are looking for a larger effort to help coordinate and facilitate their ongoing and planned programs, and 2) are pushing for significant sources of funding to foster their growing cyberinfrastructure needs.

Goals of the Workshop

Both the purpose and context of the workshop were succinctly stated by NSF's Snyder:

> "We realize that natural systems are inherently complex, heterogeneous and nonlinear, and span huge temporal and spatial scales – this encompasses the science challenge. The science thus defined dictates that cyberinfrastructure must be flexible, be able to grow and morph to meet the domain science needs, and must encompass needs of very large to the small, focused studies. In short, it must be a true system that meets the general needs, but also works with the domain scientists to help forge the basis for tomorrow's science.
>
> "Within the Earth Sciences, geoinformatics, and more generally, cyberinfrastructure, is also an integrative tool. New science will be developed as the traditional walls between geoscience subdisciplines are breached - now that is done rarely by insightful individuals, occasionally by innovative teams. In short, CI binds together traditionally disparate efforts. What we have seen already during workshops is an enthusiasm for cross-boundary collaboration driven by common cyberinfrastructure needs and goals. We need to encourage and facilitate this - but such collaboration cannot be dictated. It is a question of management, because you cannot set up a one-time solution, rather it must be nurtured for a longer time."

His comments immediately led to a discussion of organization: whether what is needed is a single, sanctioned system, an environment in which multiple approaches to geoinformatics would vie for acceptance by the scientific community, or some third way. It was agreed that advances in technology were quickly allowing distributed, disparate geoinformatics components

to be easily linked and made interoperable through middleware. Accordingly, the group dropped the use of the word 'system' as either pejorative or misleading.

A fundamental objective of the workshop was to assess not only community desires, but also the resources necessary to fulfill them. Many questions were raised:

- What is the full list of the relevant subdisciplines, elements, allies, funders, and information consumers of the earth science community?
- What are the information needs of the community?
- There is a spectrum of organization, from distributed coordination to centralized management – what are the models that might work for organizing geoinformatics in the earth sciences community?
- If a geoinformatics organization is put in place, what functions (including coordination, collaboration, and interaction with federal agencies and interagency initiatives) should it either support or have responsibility for?
- If an organization is put in place, how should it make decisions and who should run it?
- If an organization is put in place, where should its administrative components be housed?

Rationale for Infrastructure

It is important to promote informal and meaningful long-term coordination and collaboration among the elements of the earth sciences community, the information technology community, allied fields, and interested governmental units. Many representatives of most of the known groups in these categories, including ocean sciences, biodiversity, and atmospheric sciences, attended the workshop.

They not only participated in the needs examination, they established a preliminary list of primary end-user needs and the major issues facing each separate group. There appeared to be consensus on many of the elements of an emerging geoinformatics infrastructure and a willingness to consider organizational models.

Information Structure

Multiple datasets and databases, characterized by various formats and metadata, exist in the geosciences. Many have been digitized; many have not. Some of the databases have huge digital files; some are hand-written notes from a series of single-point field observations.

The development of the World Wide Web in the 1990's created recognition that many of these datasets could be viewed from afar. Web services came into being, permitting development of and access to documents and data from remote locations.

Growth of computing power, storage capacity, metadata, and networking are transformational enabling technologies. The development of web services in recent years makes it conceptually possible for anybody – researcher, schoolteacher, legislative assistant,

petroleum industry engineer, chamber of commerce clerk, librarian – anybody – to quickly conduct a global search of existing digital information to see if what they need exists and can be viewed, downloaded, and worked with at their own location.

This vision can only happen if the hardware, software, metadata, and digital datasets can be linked seamlessly and easily. The workshop participants fully understood and agreed to this principle.

Cyberinfrastructure in the Geosciences

Organization of earth science disciplines has evolved over many years. Each domain has, to some extent, its own culture and structure as well as its own professional societies, journals, and history. Some domains have comprehensive digital datasets and databases; others less so, or even not at all. Some have massive collections of physical samples and cores; others do not.

Cyberinfrastructure in the geosciences, or geoinformatics, defines the integration of computer technology, scientific activities, and geospatial coordinates. **Geoinformatics has come to mean the integrated use of computer technology in the service of scientific activities that require, generate, or use geospatial coordinates.** This unifying theme is found in a growing set of collaborating earth sciences activities. Many of the fields are organizing for improved geoinformatics.

The chronostratigraphic community has built a consensus-forming coalition known as Chronos. The hydrogeologic community through CUAHSI (Consortium of Universities for the Advancement of the Hydrologic Sciences, Inc.) and the HIS (Hydrologic Information Systems) committee has an aggressive geoinformatics plan. The Community Modeling Environment project managed by SCEC (Southern California Earthquake Consortium) is one of only two large-ITR (Information Technology Research) projects funded by the NSF. At the fall 2002 annual meeting of the Geological Society of America, leaders in tectonics and geochemistry formed a steering committee to start similar efforts in their areas. The GEON (Geologic Network) and DLESE (Digital Library for Earth Science Education) projects are recognized as influential coordinating forces.

The anticipated Earthscope initiative requires an extensive Earth Science Information System in order to take best advantage of the opportunities. Geophysicists want access to integrated earth science data to help configure their geophysical arrays. Both geologists and geophysicists will want to integrate newly acquired geophysical results from Earthscope with existing geological data to better characterize and understand the four-dimensional (4-D) nature of the earth.

GEON is collaborating with the SCEC on earthquake database architecture. This synergy alone between the two large NSF ITR projects in the NSF EAR will have profound impacts on instigating further coordination among the earth sciences.

The merging of the NAVDAT (North American Volcanic [& Igneous] Database), Pet DB (marine petrology database), and GEOROC (German-based igneous database) into Earthchem is providing improved data access and opportunities to researchers. This integration of distributed databases is important because it not only crosses disciplinary (and NSF divisional) lines, it is international by organization and scope.

Leading geoscience societies (including the Geological Society of America, American Association of Petroleum Geologists, Society of Exploration Geophysicists, Mineralogical Society of America, SEPM – Society for Sedimentary Geology, Geological Society of London, and the American Geological Institute) are collaborating on creating an electronic journal aggregate, called "GeoScience World" (GSW), involving 35 of the leading geoscience journals.

The Geospatial One-Stop, via the FGDC (Federal Geographic Data Committee) and the NSDI (National Spatial Data Infrastructure), is creating a top-down system promoting standards across governmental agencies. The FGDC recognizes that 85% of government data have spatial components. FGDC is enabling NSDI by establishing system-wide standards to disseminate government data uniformly. State geographic information system centers want input to protect their needs and interests. An early emphasis of the FGDC has been to promote metadata standards.

Private corporations are preceding with proprietary cyberinfrastructure systems largely outside of the view of government and academic persons. Many petroleum service companies are putting together large databases, software tools, and support services for marketing to their industry. Participants believed that some of these efforts are far-reaching and useful. One goal of geoinformatics would be to include interoperability with these systems.

The NGDRS (National Geoscience Data Repository System) has geoscience data repositories feeding a centralized online supercatalogue about their respective holdings. The supercatalog is accessed through a web-based graphical geographical browser. Users may select data visually or through searches on specific criteria. The catalog lists the holdings of participating public and private geoscience data repositories. While much of the catalog data are metadata, direct online data access is possible in some instances. The goal is to give users information about the vast holdings of geoscience data. It is jointly sponsored by the U.S. Department of Energy (DOE) and the petroleum industry; it is operated by the American Geological Institute.

A related community (oceanography) is pursuing coordination efforts similar to those in the earth sciences (OITI, 2002). Similar plans are underway in biodiversity sciences, and in atmospheric and environmental sciences. While each community has its own needs and unique attributes, early coordination among all the sciences is crucial to the long-term successful implementation of geoinformatics.

Each of these initiatives adds one or more critical components to cyberinfrastructure. However, each is moving independently, with only limited and anecdotal collaboration. The history of the development of the World Wide Web (Gillies and Cailliau, 2000) is instructive. It

was initially designed and built by combining existing components. Today a managing organization maintains open standards while encouraging development of new capabilities. The participants agreed that geoinformatics should emulate this balanced interaction of bottoms-up creativity with the need for broadly adopted standards.

The NSF role is critical in moving to the next level. NSF has sponsored much of the work done in the development of cyberinfrastructure, often through its CISE (Computer Information Sciences and Engineering) directorate. CISE-funded projects for the most part are oriented to understanding form, procedure, process, standards, communications and logic – not to substantive content of an established domain of knowledge. The NSF ITR funded projects are cross-disciplinary, encouraging ties between the domain sciences and information technology.

NSF sponsored a workshop on environmental cyberinfrastructure in Boulder, Colorado, in October-November 2002, bringing together 90 researchers from the solid earth sciences, biodiversity, oceanography, atmospherics, environmental engineering, and information technology communities. That workshop recommended that "NSF should establish a cross-cutting program to support the development and deployment of cyberinfrastructure for environmental research and education," to include support for establishing an "Environmental Research and Education Cyberinfrastructure Federation" (ECI Steering Committee, 2003).

The 2003 Atkins report (*Blue Ribbon Panel on Cyberinfrastructure* [NSF BRAPC, 2003]) laid out an agenda for ECI from the technology side, in a compelling explanation of what can be achieved in science and engineering as a result. This report prepared the foundation for the domain sciences to build upon.

NSF promoted cross-disciplinary collaboration through its NSF-wide Advisory Committee on Environmental Research and Education (AC-ERE) panel (Futrell and AC-ERE, 2003), which serves as the parent to the ECI (environmental cyberinfrastructure) effort. The full AC-ERE (2003) report further stated that "Environmental cyberinfrastructure and observing systems are long-term and expensive endeavors requiring the efforts of multiple agencies." ECI will be a core component of the developing ERE initiative.

In February 2003, the National Science Board issued an NSF policy statement that for domain sciences to take full advantage of the capabilities of cyberinfrastructure, significant funding must be provided to the sciences (NSB, 2003). The NSB observed that, "Recent concepts of infrastructure are expanding to include distributed systems of hardware, software, information bases, and automated aids for data analysis and interpretation. Enabled by information technology, a qualitatively different and new S&E [Science & Engineering] infrastructure has evolved."

The NSB recognizes that the NSF budget dedicated to science and technology infrastructure is insufficient and urges additional resources. It also recommends that special emphasis be given to develop and deploy an advanced cyberinfrastructure to enable new science and technology in the 21st century (NSB, 2003).

Challenges

The biggest challenge may be cultural: getting scientists who have long worked only in their own fields to coordinate with others. They are understandably concerned about having to convert existing datasets and databases into something else, which they see as both potentially expensive and clearly disruptive. This is why workshop participants believed the future of geoinformatics is in the creation of a truly interoperable environment though the development of appropriate middleware.

Coordination needs to be achieved. However, there are also challenges in integrating so-called 'flat' databases into the more robust and flexible relational database environment.

ORGANIZATION AND CONDUCT OF THE WORKSHOP

Organizers

Workshop organizers are Lee Allison, Chaitan Baru, and Thomas Jordan. Each has been centrally involved in geoinformatics through individual projects and broader policy roles.

Allison directs the Kansas Geological Survey (KGS), which has a $1-2 million per year geoinformatics initiative to digitize and dynamically serve online via web services, all the data related to petroleum and ground water and all GIS data for the state. Funding is from a mixture of state, industry, and DOE sources. Current KGS projects include creating seamless interoperability among databases from more than 140 organizations in 47 states and three Canadian provinces in regards to CO_2 sequestration (Natcarb project), integrating hundreds of global environmental data layers with the Ocean Biogeographical Information System (OBIS) databases, and implementing a prototype geoinformatics system tying together databases, online software toolboxes, tutorials, and workspace (GEMINI project). Allison served on the NSF Panel on Environmental Cyberinfrastructure. He was a member of the organizing committee for the fall 2002 NSF-sponsored Boulder workshop on environmental cyberinfrastructure.

Baru is the GEON principal investigator. GEON has a focus on the 4-D structure and evolution of continents and is addressing the need to craft the many relatively raw datasets into mature databases to grow and evolve. Baru has worked for many years on the problems of interlinking and sharing interdisciplinary databases by creation of a robust toolbox of open source software (including middleware) for analysis, modeling, and visualization. He is recognized for his cyberinfrastructure work in managing a highly distributed and diverse network. Baru is the Program Director for Data and Knowledge Systems at the San Diego Supercomputer where he leads a variety of projects involving Grid computing, geospatial referenced data systems, and bioinformatics.

Jordan is the director of the Southern California Earthquake Center (SCEC) and Principal Investigator for an NSF ITR grant to develop an information infrastructure for system-level earthquake research. He served on the organizing committee for the NSF-sponsored fall 2002 workshop on environmental cyberinfrastructure. He is a member of the NSF Advisory Committee for the GEO directorate (AC-GEO) and is leading the committee's effort on environmental cyberinfrastructure.

Invitation process and attendees

The original target was to invite 10 to 15 participants who were recognized as leaders in developing components and aspects of geoinformatics. A small number of observers were invited from the oceans, atmospheres, and biology communities, federal initiatives, and industry.

Word quickly 'got out' about the workshop. Many persons not originally identified, some of them not previously known to the organizers, asked to attend as representatives of their

communities. When the workshop opened on May 14, 2003, the conference room was crowded with three times the originally planned number. A complete participant list is provided in Appendix B.

Although the number of attendees was unexpected, the result was a breadth and diversity that greatly contributed to the discussion and outcomes. It demonstrated the level of activity and enthusiasm in the community for developing a common understanding and to be part of the solution.

Web site

Materials were placed on a webpage hosted by the KGS at www.kgs.ku.edu/Geoinfo2/Cyber/index.html. Later they were moved to www.geoinformatics.info. This is intended to serve as a dedicated site for ongoing geoinformatics activities and information. Links are provided to websites of many of the participants' projects and to other informatics activities.

Facilitation

With the attendee list growing in size and diversity, the organizers retained a professional facilitator. This permitted them to concentrate on the substance of the workshop, freely entering discussions without having to also moderate them. The facilitator was Bob Crangle, a management consultant and attorney with a substantial background in facilitation, planning and evaluation activities with NSF, other agencies, the sciences, universities, and other sectors. In advance of the workshop itself, Crangle conducted extensive telephone and email exchanges with most of the participants.

Materials

Appendix A has the preliminary workshop agenda. The actual workshop schedule deviated from this agenda; the modifications reflected the progress of the discussions. There was real-time feedback from what had just happened into what was then scheduled next.

Opening Discussion

After the opening charge to the participants by NSF's Snyder, Crangle reported on the results of his telephone and email interactions with each participant prior to the workshop. This was done interactively, and the resulting responses and comments immediately plunged the workshop into the heart of the substantive issues.

Crangle reported that the participants seemed to have a common and broad consensus that work must accelerate on geoinformatics, with emphasis needed on standards and metadata. The pre-workshop goal of geoinformatics was that useful, seamless connections of dispersed, divergent geologic datasets and databases were needed by scientists, other professionals, and the public.

There were many examples given to Crangle about legacy datasets and databases not yet digitized, as well as many which include field notes and characterizations by individual geologists and field workers. Points were also made that a dataset ranges from a single observation at a spatial location to high-volume streams of data from automated sensors.

Many of the pre-workshop comments were directed at cultural issues that sometimes separate different fields. Terminology can be an issue, especially so between information technology experts and the earth sciences, but also between some of the subdisciplines and with the allied disciplines. There is also a chronic need for more funding to advance initiatives already underway and to provide for more collaboration and connection among the various fields and disciplines.

A repeated pre-workshop theme appeared to be the need to be aware of the users. Data users were said to run the gamut from university researchers in allied or other professions to industrial users to large numbers of interested teachers, students, and lay persons. The population of concern is a rich mixture of providers, reporters, and users who often need to have information from a single source either processed or reported at different levels of detail.

As this discussion proceeded, the concept of a "geoGoogle" framework emerged: creating a data universe where users at all levels could search for information relevant to their very different concerns. Creation of such a framework appears to require development of middleware not presently available.

Another pre-workshop theme, not quite as prevalent, was an expressed concern about data ownership. Ownership issues include authorship and professional recognition, the possibility of economically important data, and the issues that arise when databases (or even datasets) have mixtures of public and private data from multiple sources.

PRINCIPAL FINDINGS

Discussion on some of the findings is found above. As appropriate, however, additional comments are provided below.

- **Both interest and momentum towards development of cyberinfrastructure (geoinformatics) are found throughout all the domains of earth sciences.**

There are differences in the nature of the scientific questions and data types within each of the domain sciences, and cultural differences about how they use data and models. These differences are not fundamental barriers and they seem to be narrowing. It is illustrative that other environmental sciences are moving ahead with similar efforts to build cyberinfrastructure (e.g., OITI Steering Committee, 2002).

- **The relevant federal agencies are supporting significant geoinformatics activities.**

Many federal agencies are supporting significant geoinformatics activities. The NSF and USGS participants concurred that it is in the national interest to move forward with geoinformatics, and that their organizations and other agencies were prepared to support it.

Geospatial One-stop is a priority initiative reaching all the way to the White House. However, its scope is so large that some agencies are skeptical of meeting the targeted deadlines; they are continuing their own geoinformatics initiatives. An effective community-based geoinformatics initiative should be inclusive of the federal agencies. The community-based effort offers a realistic path to achieving their goals in a possibly much shorter time frame.

- **"GeoGoogle" is the symbolic goal: providing access to as many interested persons from as many disciplines and levels of sophistication as possible.**

Participants want a web-based capability and interoperability that is robust, dynamic, easily usable, universally accessible, and free. A key component in the concept of "geoGoogle" is a 4-D search engine that underlies the portal or browser. This would facilitate access and interoperability among the user as well as provider community.

The previously accepted wisdom of topical centralized databases is rapidly evolving to take advantage of improved networking and bandwidth. The key concept today is interoperability. Many groups and researchers are developing online "geoportals" as central clearinghouses for their topical datasets and databases, and to link to related web sites.

The emerging vision is for a geoinformatics portal (or multiple, linked portals) with strong search capabilities. News, discussion forums, reports, and links will be showcased, but the emphasis will be on a geoinformatics browser with searchable access to specific data or databases, software tools, computational capability, and sensor/observation feeds. There are

unresolved complexities in searching for data based on both three-dimensional geography and geologic time.

- **Hundreds of geoinformatics-enabling projects are underway, largely independent of each other and not using common standards or metadata.**

Each of the individual earth science domains has developed its own journals, culture, and societies, to meet its particular needs, procedures, and goals. This has generally been done without consideration of the needs of users (or providers) in other domains.

- **Various organizational models exist to foster geoinformatics coordination.**

There are membership consortia such as CUAHSI, IRIS, the World Wide Web and SCEC; federations such as ESIP; university and corporate cooperatives linked with memoranda of understanding or other agreements; university-hosted research centers; and stand-alone non-profits such as UCAR/NCAR. When the community is clear about what it wants to do and how it wants to function, the legal and administrative issues will readily fall into place.

- **Issues of attribution of authorship and ownership of various datasets and databases are of great importance, but should not be 'deal-breakers' to moving forward.**

Owners of datasets and databases have from time to time exhibited highly parochial attitudes, but even in university research settings there are examples where data owners respond very well to factors such as prestige and convenience in having one's data open to peers. They are very often also interested in collaborative opportunities that might lead to future funding. Peer review and professional recognition are enhanced through making full availability of datasets and databases. Scientists will respond far better to this model than to any attempts to coerce or mandate open files.

- **The earth scientists, rather than the information technologists, must lead the way.**

Content must drive process. Geoinformatics will be simplified if information technology specialists create the middleware to let users and providers seamlessly access and place data without having to conform to restrictive universal standards and metadata.

Leadership and associated funding must come from within the earth sciences without waiting for the information technologists. There will be continued information handling breakthroughs, but the basic philosophy and policy must be based on scientific community leadership.

- **Successful coordination of geoinformatics will create transformational opportunities throughout all the geosciences.**

The developments anticipated by the participants will foster unprecedented levels and scopes of data coordination and human collaboration. It is not an exaggeration to refer to these opportunities as transformational to the study of complex global questions involving the earth sciences.

- **It is possible and practical for an interim steering group to begin to address these issues and to encourage organization around the grand earth science challenges, shared instrumentation, facilities, and other resources.**

An interim steering group or task force should develop options for consideration by the communities. Options should address organizational models, funding, research issues, decision making, administrative support, coordination, and so forth. This group should include all who wish to participate, be accessible to all parts of the community, and fairly represent the needs of each domain.

RECOMMENDATIONS

An Interim Steering Group

The most important recommendation relates to the last finding, immediately above. A small group of the ready, willing, and able should become an interim coordinating body to do something about the issues identified in the workshop.

Several participants volunteered on the spot. They will provide a core group to "boot-strap" into a larger team. The entire group of workshop participants reacted well on the second day to the notion that the interim steering group should organize around grand challenges or shared resources to address the issues of infrastructure, communications, and standards.

The grand challenges are a matter for debate. The popularly referred to "Keck report" (NRC, 1993) of a decade ago laid out a comprehensive overview of research needs for the earth sciences that is still valid. Many challenges have strong societal relevance. They would seem to include understanding global climate change, protecting our environment, maintaining national security, enhancing the availability into the future of our natural resources, applying what we know toward a better understanding of extraterrestrial bodies, and predicting and protecting against natural disasters.

As the organization for implementation takes shape, the interim and expanded coordinating body should actively inform and listen in meetings of ITR recipients, other geoinformatics project leaders, and professional societies. Reports in society newsletters and professional journals are to be expected.

Getting Started: Domain Requirements

Fundamentally, each scientific domain must take responsibility for making its data accessible, as well as for its own science standards and methodologies (including data assessment, quality, archiving, access, and linkages to physical data samples). Each self-organized geoscience domain should inventory its existing knowledge base and present it in a common format to the rest of the geosciences community. During the workshop it was confirmed that many of the groups are already doing this.

It would be useful for each domain's report to describe its existing datasets, a description of applicable metadata, and a description of its knowledge 'format' (such as digital, field notes, physical samples, automated sensor read-outs, and so forth). Putting such information into a three-dimensional matrix or some similar visibility tool will permit the other domains to better understand how existing knowledge is organized, and would facilitate the development of useful search techniques. The "geoGoogle" concept is intended to provide an intelligent interface that seeks out and identifies such data with minimal additional documentation required from the data providers.

It would also be useful if each domain would consider the 'transformational opportunities' that might emerge if it had total access to all the digital geo-knowledge: new questions, new structures, new challenges that cannot now be addressed.

Available Tools

A task force of cyberinfrastructure specialists with previous participation in geoinfrastructure projects should be chartered to assess the ability of existing information technology tools and methodologies. It should recognize the knowledge needs of all of the geosciences, address security and data ownership needs, and focus on developing middleware solutions which have the least impact on forcing users to alter existing database structures and metadata.

Because subdisciplines are at different stages of geoinformatics development, the interim steering group and the cyberinfrastructure task force should develop and maintain readily viewable formats to track the status of the community as a whole. This would help to understand each domain's time line and identify individuals in each domain who are taking on the tasks involved.

Electronic Reporting

Each domain seems to be already developing plans for digital reporting of its knowledge base. This includes fully retrievable datasets with associated metadata. These formats will provide the kind of professional recognition that will overcome many, perhaps most, of the issues of proprietary data. The electronic journal aggregate in the earth sciences, *GeoScience World*, is a useful model.

Electronic reporting will gradually expand into news about developments in geoinformatics, peer-reviewed articles, infrastructure and tool development, and so forth. An electronic journal, as is also true in print media, needs a publication statement, name, editorial board, funding model, set of operating procedures, and associated administrative support.

Participants noted that career enhancement and professional recognition depend heavily on the reward system. Because cyberinfrastructure is new, contributions to electronic reporting may not be seen as valid as to print reporting. Each domain should promote a citation system for electronic databases, sample repositories, and products so there are professional rewards for attributions.

Implementation of Geoinformatics

The information technology task force and the interim steering group should define the practical attributes of geoinformatics. It must include middleware, within the grid computing context, adequate to enable authorized users to freely access existing digital databases in an interoperable process. It must provide physical and data security and protection of proprietary and attributed information.

Geoinformatics must eventually serve projects ranging from poorly funded single-investigator fieldwork to massive real-time data acquisition efforts of large national centers. It must be flexible enough (some participants used the word "organic") to grow with new demands, capabilities, and transformative advances in the earth sciences.

Education and outreach aspects require careful planning. Above all it should be a service to the relevant communities. A distributed, hierarchical system of permanent libraries and data repositories, such as that of the American Geological Institute's National Geoscience Data Repository System, will be one of many elements.

Geoinformatics should let non-scientists such as public policy specialists, teachers, general librarians, industry staff, regulators, professional society members, and community leaders have fast and easy access to relevant, accurate, four-dimensional earth science information. It should be persistent, have options for portals and access through it to electronic publications, and include advanced access, workspace, and communications tools. It should be a reliable utility that serves community needs in as painless a way as possible.

Geoinformatics is intended to be open, free, persistent, user-driven, adaptive, flexible, innovative and organic (Allison et al., 2002). It will be designed to serve researchers, but it is likely and desirable that others will become users and perhaps contributors. Use of the system by government and industry would be an indication of its utility and influence. Eventually, participants expect that NSF and other funding agencies will require participation in geoinformatics (i.e., at least to the extent that data and tools are made readily accessible) as a condition of grants, contracts, and other agreements.

Organization, Administration, and Funding

Participants agreed that a geoinformatics organization is needed which has a competent and dedicated (rather than purely voluntary) administrative staff. Geoinformatics components and resources will be widely distributed and will evolve and grow rapidly in many complex ways.

The individual earth science domains have unique cultures, traditions, and needs, which will complicate coordination. The development, distribution, and effective utilization of new user tools will not be simple. Users will come from all sectors – public, private, academic, and international. This is not a simple job. However, it is also critical that the communities see geoinformatics as a net benefit to them. It cannot be, or seem to be, a competitor to them for NSF or other agency funding.

The administrative staff and required central servers and other equipment need a physical home. In all likelihood, that home should not routinely move around once established. Agreement on a location is needed. As with the other questions, these are matters which need to be addressed by the interim steering committee.

REFERENCES

Allison, M. Lee, Walter S. Snyder, and J. Douglas Walker, 2002, *Geoinformatics: A Nascent Revolution in the Earth Sciences*, GSA Today, v. 12, no. 7, p. 17.

Environmental Cyberinfrastructure Steering Committee, September 2003, *Cyberinfrastructure for Environmental Research and Education (Report from a workshop held at the National Center for Atmospheric Research, October 30-November 1, 2002):* Boulder, CO, National Center for Atmospheric Research, 15 p.

Futrell, J., and the AC-ERE, May 2003, *Environmental Cyberinfrastructure (ECI: Tools for the Study of Complex Environmental Systems),* AC-ERE 1, May 2003: Arlington, VA, National Science Foundation.

Gillies, James, and Robert Cailliau. 2000, *How the Web Was Born*: Oxford University Press, Oxford, Great Britain, 372 p.

National Research Council, Committee on Status and Research Objectives in the Solid-Earth Sciences, 1993, A Critical Assessment, Board on Earth Sciences and Resources, Commission on Geosciences, Environment, and Resources, Solid-Earth Sciences and Society: National Academy Press, Washington, D.C., 368 p.

National Science Board, February 2003, *Science and Engineering Infrastructure for the 21st Century: The Role of the National Science Foundation*, NSB-02-190: Arlington, VA, National Science Foundation.

NSF Advisory Committee for Environmental Research and Education, January 2003, *Complex Environmental Systems: Synthesis for Earth, Life, and Society in the 21st Century*, *A 10-Year Outlook for the National Science Foundation:* Arlington, VA, National Science Foundation, 68 p.; http://www.nsf.gov/geo/ere/ereweb/acere_synthesis_rpt.cfm

NSF Blue Ribbon Advisory Panel on Cyberinfrastructure, February 2003, *Revolutionizing Science and Engineering through Cyberinfrastructure* ["Atkins report"]: Arlington, VA, National Science Foundation; http://www.cise.nsf.gov/evnt/reports/toc.htm

Ocean Information Technology Infrastructure Steering Committee, 2002, *An Information Technology Infrastructure Plan to Advance Ocean Sciences*: Washington, D.C., National Oceanographic Partnership Program, 80 p.; http://www.geo-prose.com/oiti/report.html

APPENDICES

Appendix A: Preliminary Agenda

Building Geoinformatics—Workshop Agenda

The workshop is planned to be held at the San Diego Supercomputer Center in La Jolla, California, on May 14 and 15 (Wednesday-Thursday), 2003.

The workshop organizers will moderate the breakout sessions. An external facilitator will lead the group discussion and at least one organizer will serve as a recorder of the group sessions and results.

Wednesday, May 14, 2003, San Diego Supercomputer Center

7:30 Registration, continental breakfast

8:00 Welcome and introductions--Allison

8:10 Purpose and structure of workshop--Allison, Crangle

8:20 NSF perspective--Snyder

8:30 Roundtable review of earth science community organizations and plans (handouts and materials in workshop binder)

- Chronostratigraphy:
- Hydrogeology-surficial processes:
- Solid earth sciences:
- Geophysics:
- Education:
- Geoinformatics:

9:00 Roundtable review of other community activities (handouts and materials in workshop binder)

- Oceans: Steering Committee
- Atmospheres: Steering Committee
- Biology:
- NSF:

9:30 Breakout session #1 (3 sections, same topics)

1. Definition of this community
    a. Is this a community, or not?

b. Who might be missing?
c. What descriptive or problem-oriented characteristics can be used to define us as a community?

2. A triage on community priorities
   a. Identify topics everyone agrees are in, or out, and then concentrate on the topics in between.

10:30 Coffee Break

11:00 Breakout session #2 (3 sections, same topics)--participants mixed from Session 1

1. Opportunities

2. Emerging or blatant crises

12:00 Catered lunch at SDSC

1:00 Committee of the Whole--**Breakout reports and discussion**

2:00 Breakout session #3 (3 sections, same topics)

1. Community goals
   a. Identify the earth science community elements
   b. What does the community want to achieve?

2. Delivery of services to the community
   a. Identify goals for the community
   b. Coordination vs. management
   c. Distributed vs. centralized
   d. Develop a community-based system
   e. What are the alternative organizational models?

3:30 Cookie break

4:00 Committee of the Whole--**Discussion question**: IF an organization were to be set up, what would we expect it to do for all of us, and for each of us?

5:00 Recess

5:30 Social gathering at hotel bar on the beach

7:00 Group dinner (locations tbd)

Thursday, May 15, 2003, San Diego Supercomputer Center

8:00 Continental breakfast

8:30 Committee of the Whole (continuation of previous discussions)

1. Recap of services to be provided to the community
2. How can those services be provided?
    a. Participants/members
    b. Decision processes
    c. Needed resources
    d. Communication systems

10:30 Coffee Break

11:00 Committee of the Whole--**Discussion of organization issues**

12:00 Catered lunch at SDSC

1:00 Committee of the Whole

1. What do we want to do next?

2. Coordination and collaboration with extra-earth science groups

3. Interactions with NSDI/FGDC, ESIP, and other federal initiatives

4. Interactions with industry

3:00 Cookie break

3:30 Committee of the Whole

1. Have we missed anything?
2. Preparation of workshop final report
3. Set up steering committee and task forces as necessary
    a. Collaborations and coordinations
    b. Next meeting?
    c. Funding

5:00 Adjourn

6:00 Organizing Committee social and working dinner (open to all)

Appendix B: Workshop Participants

Organizers

M. Lee Allison
Kansas Geological Survey
University of Kansas

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

Thomas Jordan, Director
Southern California Earthquake Center
University of Southern California

Facilitator

Robert D. Crangle
Rose & Crange, Ltd

Confirmed Participants

Tim Ahern
IRIS Data Management Center

Stan Blum
California Academy of Sciences

Steven Bohlen
Joint Oceanographic Institutions
Ocean Drilling Program

Harold Bolton
USGS Albuquerque Seismological Lab

Boyan Brodaric
Natural Resources Canada
Earth Science Sector
Geological Survey of Canada

Suzanne Carbotte
Lamont-Doherty Earth Observatory
Columbia University

Richard W. Carlson
Carnegie Institution of Washington
Department of Terrestrial Magnetism

Cinzia Cervato
Geological & Atmospheric Sciences
Iowa State University
(CHRONOS)

Joe Futrelle
NCSA, University of Illinois
Champaign, Illinois
(Digital Library Technologies Group)

Arthur Goldstein
Colgate University

Michael C. Gurnis
California Institute of Technology

Linda Gundersen
U.S. Geological Survey

John Helly
University of California at San Diego
San Diego Supercomputer Center

Chris Jones
Partnership for Interdisciplinary Studies
of Coastal Oceans (PISCO)
University of California at Santa Barbara

Dave Jones
StormCenter Communications, Inc.
(ESIP Federation)

Anke Kamrath
San Diego Supercomputer Center
University of California, San Diego
(NEES)

Christopher M. Keane
American Geological Institute

G. Randy Keller
Department of Geological Sciences
University of Texas at El Paso
(GEON)

James L. Kinter III
Center for Ocean-Land-Atmosphere Studies

Rich Lane
National Science Foundation
Geology and Paleontology Program

David Maidment
Dept. of Civil Engineering--CRWR
University of Texas at Austin
(CUAHSI)

Mary Marlino
University Corporation for Atmospheric Research
(Digital Library for Earth Science Education)

Charles Marshall
Museum of Comparative Zoology
Harvard University
(Paleobiology Database)

Chuck Meertens
UNAVCO, Inc.
Boulder, Colorado

John A. Orcutt
Scripps Institution of Oceanography
(CORE)

A. Krishna Sinha
Department of Geological Sciences
Virginia Tech
(GEON)

Judy Scotchmoor
Museum of Paleontology
University of California, Berkeley

Walt Snyder
National Science Foundation
Division of Earth Sciences

J. Alexander Speer
Mineralogical Society of America

Hubert Staudigel
University of California, San Diego
Institute Geophysics & Planetary Physics

Lars Stixrude
Department of Geological Sciences
University of Michigan
(COMPRES)

Bruce R Wardlaw
U.S. Geological Survey

Herman Zimmerman
National Science Foundation
Division of Earth Sciences

Appendix C: Acronyms

AC-ERE	Advisory Council, Environmental Research and Education panel, NSF
AC-GEO	Advisory Council, Geosciences Directorate, NSF
CISE	Computer Information Sciences and Engineering Directorate of NSF
CME	Community Modeling Environment (NSF-ITR project based at SCEC)
CUAHSI-HIS	Coalition of Universities for the Advancement of the Hydrologic Sciences, Inc. – Hydrologic Information Systems Committee
DLESE	Digital Library for Earth Science Education
DOE	U.S. Department of Energy
EAR	Earth Sciences Division of the National Science Foundation's Geosciences Directorate
ECI	Environmental Cyber-Infrastructure
ERE	Environmental Research and Education Panel, NSF
ESIP	Earth Science Information Partnership
FGDC	Federal Geographic Data Committee (based with the USGS)
GEMINI	Geo-Engineering Model for Internet Informatics (DOE project at KGS)
GEO	Geosciences Directorate, NSF
GEON	Geologic Network (large ITR project funded by NSF)
Georoc	Geologic rock data base
GIS	Geographic information systems
GSW	GeoScience World, electronic publishing aggregate
HTML	Hyper text markup language
HTTP	Hyper text transfer protocol
IRIS	Incorporated Research Institutions for Seismology
IT	Information Technology
ITR	Information Technology Research funding initiative, NSF
KGS	Kansas Geological Survey, University of Kansas
Natcarb	National Carbon sequestration database project (DOE project at KGS)
NAVDAT	North American volcanic (and volcanic) database
NCAR	National Center for Atmospheric Research
NSB	National Science Board
NSDI	National Spatial Data Infrastructure (federal initiative)
NSF	National Science Foundation
OITI	Ocean Information Technology Infrastructure
OMB	Office of Management and Budget (in the White House)
PetDB	Petrologic database
SCEC	Southern California Earthquake Center (runs the CME ITR project), University of Southern California
SDSC	San Diego Supercomputer Center, University of California
SEPM	Society for Sedimentary Geology
UCAR	University Consortium for Atmospheric Research
URL	Universal resource locator
USGS	United States Geological Survey
WWW	World Wide Web
W3C	World Wide Web Consortium