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.

S. Gage - Michigan State Univ.

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.*

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

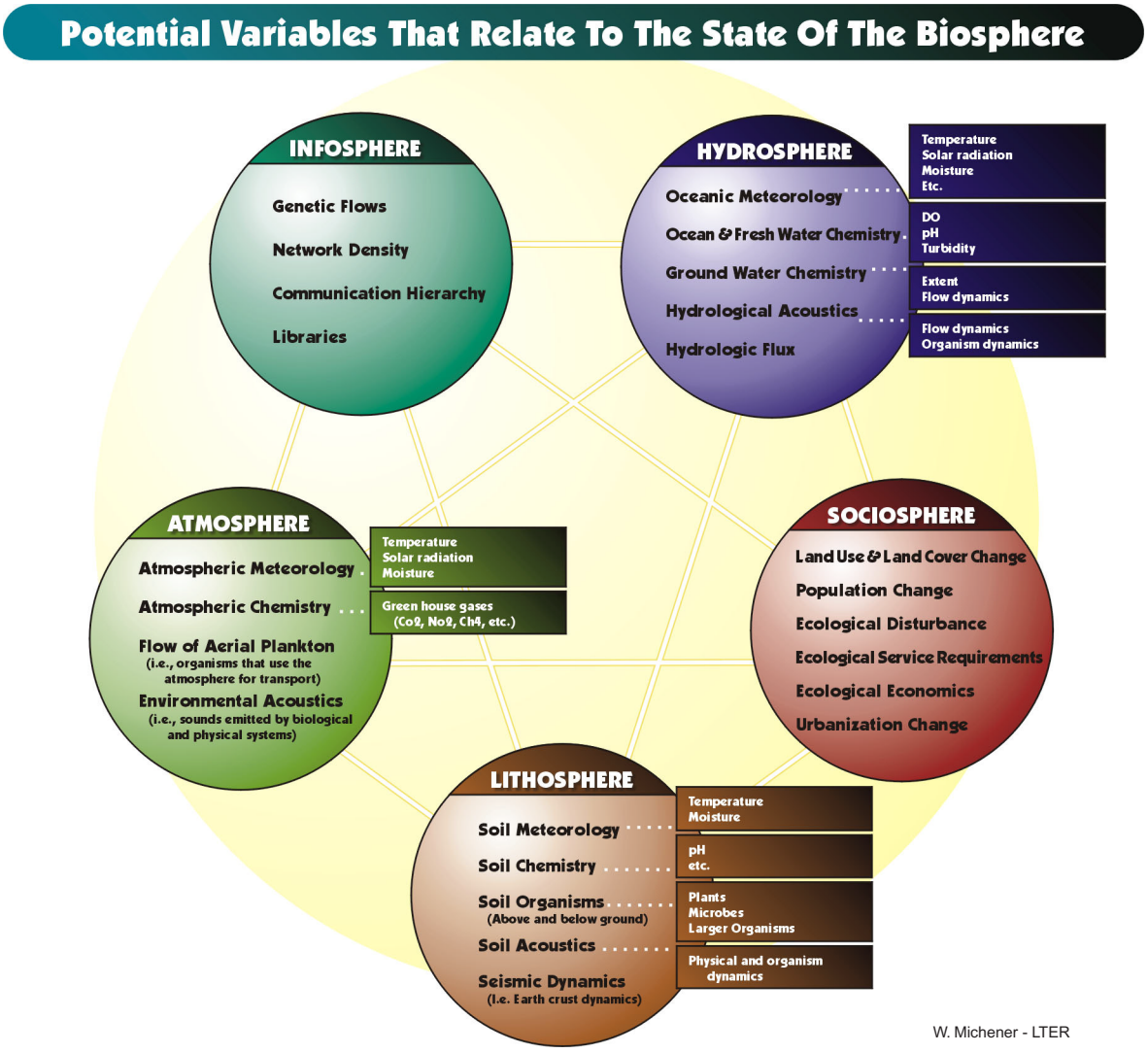
field sensors to the point(s) of analysis remains a significant problem.

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.



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 interna-

tional 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.

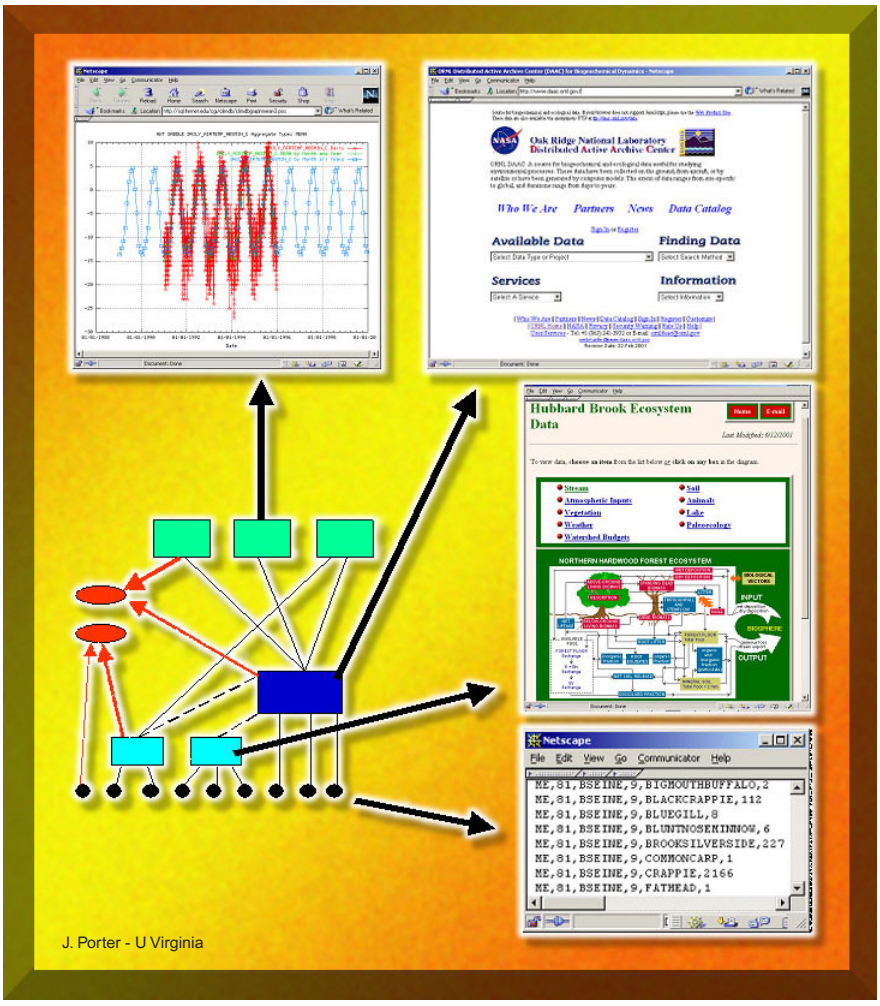
In such a large enterprise, the first step for the various participating parties is to "agree to agree." In terms of data and metadata standards, this means that there should be common agreement on the meta-standards that will be used. For example, the Extensible Markup Language (XML) is an example of 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.



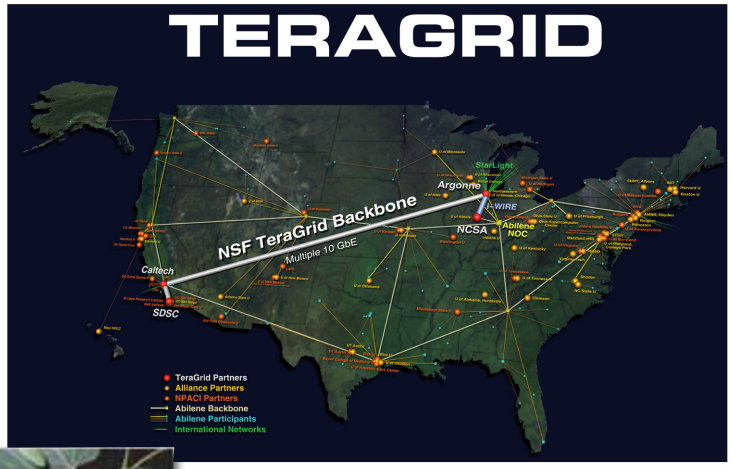
An idealized information system *would allow ready access by scientists (as depicted by the red ovals) to individual data sets and accompanying metadata (black circles: e.g. fish data in lower panel), project databases (aqua rectangles: e.g. Hubbard Brook Ecosystem Database - http://www.hubbardbrook.org/), 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 sub-disciplinary level, if not at the highest level of integration.

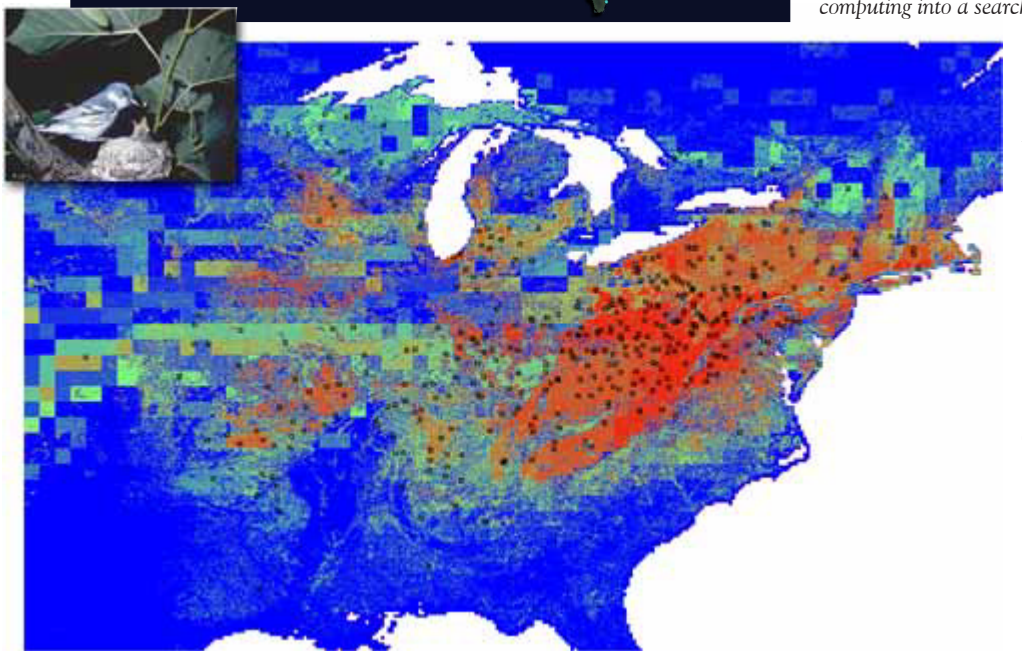
Data sharing and archiving

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

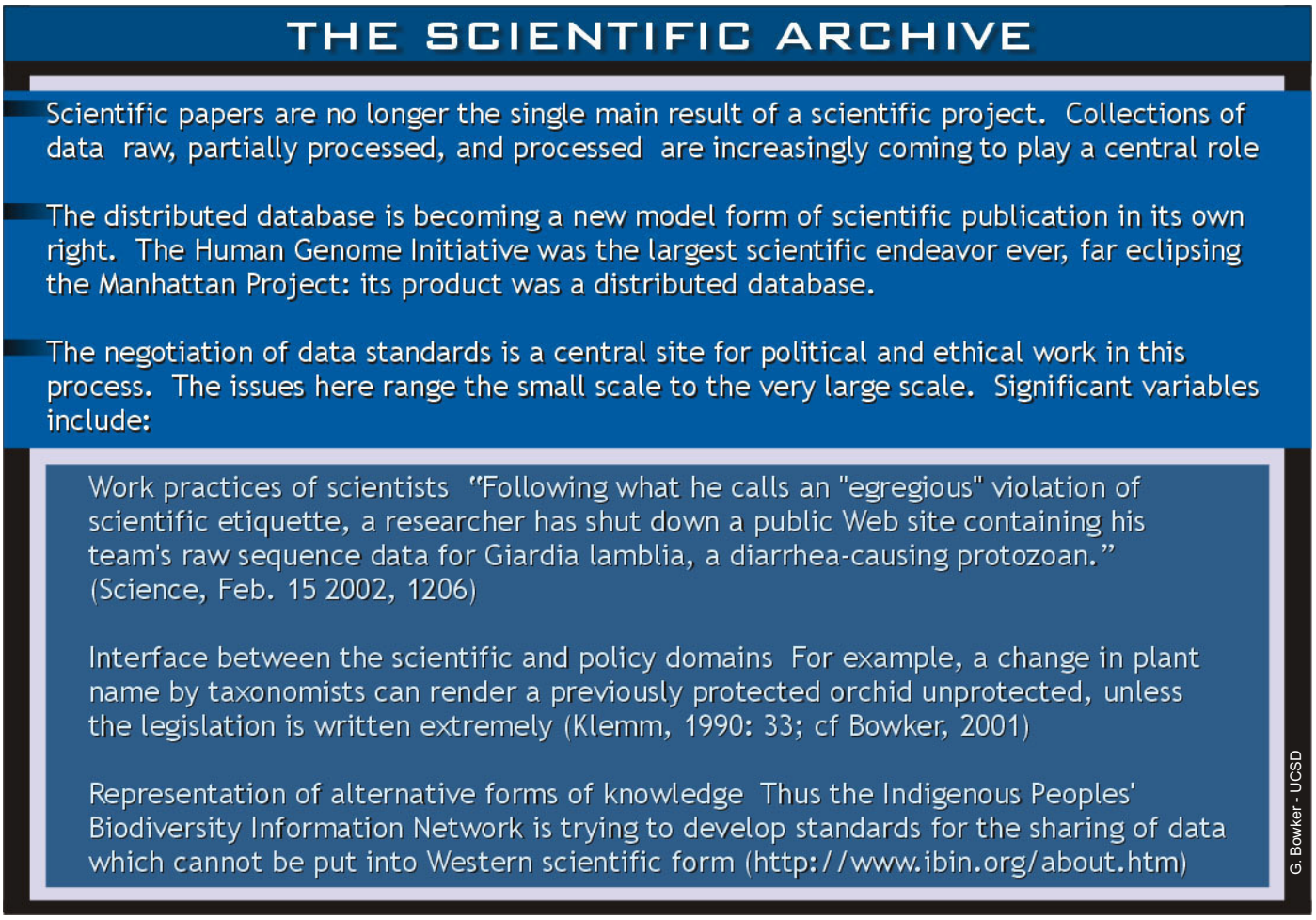
with each substantial update); and (c) ineligible or unprotectable components of otherwise protectable subject matter (e.g., factual data in databases, or ideas in copyrightable works).

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].



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 nondisclosure 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.