Korean J. Ecol., 23(2): 101~106, 2000

An Interface between Computing, Ecology and Biodiversity: Environmental Informatics

Stockwell, David, Peter Arzberger, Tony Fountain and John Helly

San Diego Supercomputer Center and National Partnership for Advanced Computational Infrastructure, University of California, San Diego, La Jolla, CA 92093, U.S.A.

ABSTRACT: The grand challenge for the 21st century is to harness knowledge of the earth's biological and ecological diversity to understand how they shape global environmental systems. This insight benefits both science and society. Biological and ecological data are among the most diverse and complex in the scientific realm, spanning vast temporal and spatial scales, distant localities, and multiple disciplines. Environmental informatics is an emerging discipline applying information science, ecology, and biodiversity to the understanding and solution of environmental problems. In this paper we give an overview of the experiences of the San Diego Supercomputer Center (SDSC) with this new multidisciplinary science, discuss the application of computing resources to the study of environmental systems, and outline strategic partnership activities in environmental informatics that are underway. We hope to foster interactions between ecology, biodiversity, and conservation researchers in East Asia-Pacific Rim and those at SDSC and the Partnership for Biodiversity Informatics.

Key Words: Biodiversity, Computing, Ecology, Environmental informatics, PBI, SDSC, Systematics.

THE MOTIVATION

The grand challenge for the 21st century is to harness knowledge of the earth's biological and ecological diversity to understand how they shape global environmental systems. This insight is critical for managing natural resources, sustaining human health, ensuring economic stability, and improving the quality of human life. As the conversion of natural systems to agriculture and urban systems decreases biological and ecological diversity, the need for this information becomes more urgent. In fact, at the current rate of environmental degradation and species extinction, the worldwide biological science community, working with society-at-large, has approximately 50 years or so to answer the challenge of declining diversity and its ramifications (Systematics Agenda 2000 1994, Wilson 1998, Raven and Wilson 1992), leading Wilson to predict that the coming century will be the century of the environment (Wilson 1998).

Many universities now offer courses in environmental informatics, an action arising from stakeholders in the academic and existing environmental regulation and policy management structure awakening to the need for sophisticated information systems if they are to quickly respond to current conditions. This is informed by the experiences of several groups in Canada, the U.S., Australia, and Europe, where the assumption that environmental science is orchestrated by planners and conciliators on one hand, and civil and environmental engineers and scientists on the other, has been challenged. Environmental informatics addresses problems at the forefront of computing and statistical science, and thus encounters obstacles in storage: management and retrieval for large scientific databases (such as distribution ownership, heterogeneity, access, security, legacy, and quality assurance): noisy and uncertain task environments: irregular spatial and temporal data distributions: knowledge representation: statistical computing: expert systems: visualization: decision support, and Web-based collaboration.

IN PURSUIT

To progress toward the grand challenge, existing knowledge must first be successfully inventoried. There are several key interrelated strategies for doing so: (a) Developing a research framework to understand the underlying processes that will allow the field to move from a qualitative to a predictive science: (b) establishing collaborations to tackle the complex, multidisciplinary nature of the problem: (c) harnessing technology to develop new tools for conducting science, and to create and then store, query, and understand the data, and: (d) aggressively creating educational opportunities at all levels and across all strata of society so

that political and social practice may be informed by scientific knowledge. We would like to elaborate a bit on each of these strategies.

The role of fundamental research (a) is to elucidate the states, processes, and interactions of the natural environment under anthropogenic influences. The study of environmental pattern and process requires both analysis and synthesis to provide a foundation of information (NSB 1999) across all scientific, engineering, and mathematics disciplines.

The dynamics of Earth's biodiversity are vastly complex, with multiple, diverse, and dispersed causes (Jasanoff *et al.* 1998). Therefore, it cannot be understood, managed, or controlled through scientific activity organized on single or traditional disciplinary lines. Instead, biodiversity requires cross-domain, interdisciplinary approaches (b). The data and tools (conceptual, physical, computational, etc.) - many of which are still under development - required to investigate its dynamics are beyond the scope of any single investigator and often beyond the mission, infrastructure, or expertise of any single institution. Therefore research in this field requires interdisciplinary, collaborative teams working within and across institutions.

Ecological and systematics data are among the most heterogeneous and complex in the scientific realm. By their very nature they are multidisciplinary and multiscalar, with origins ranging from geochemistry to climate patterns with all levels of life sciences in between. In addition, the data are from many locales, are highly disparate in format, and span enormous scales of space and time. Nevertheless, integrated information is essential to a comprehensive understanding of the patterns and processes of both ecology and systematics as well as their area of overlap (e.g., biodiversity). The move from data to information to knowledge is only possible using computational and information technologies (c).

Education (d) is necessary at all levels, to train and inform both researchers and the public. Within research, individuals must understand the underlying theories of the processes, be keen observers, and be able to develop and apply relevant technologies to address the challenges of integration, synthesis and analysis. At a societal level, environmental biological diversity affects and is affected by all humans on Earth, and the Earth's citizens must be informed and involved so that reasonable policy and social practices, based on knowledge, can be established.

For the purposes of this paper, we would like to present the steps and experiences that SDSC and its partners have undertaken and met with while employing the above strategies. For background, though, it is first useful to answer the question, Why is a supercomputing center interested in environmental issues?

The original mission of SDSC - which is funded by the National Science Foundation (NSF) - was to provide service (i.e., access to cycles and corresponding support) to the academic community. But during the first few years of our existence we learned that the original service mission should not and could not be separated from development and research. This realization changed our mission to "providing world leadership in advancing knowledge through the application and development of advanced computational technologies." In our ongoing efforts to address this mission we actively engage in outreach to new communities of users.

At the time of SDSC's inception in 1985, the primary new community targeted was the physical sciences, who had sophisticated computational models that demanded high-end computing. Since then there have been several major technical and sociological changes that have changed the nature of supercomputing and thus the communities with which SDSC is involved. The Internet and Web applications have impacted how scientists interact with computers and each other, the types of services that a center like SDSC can provide, and even how society-at-large interacts and accesses information. Data storage and retrieval technologies have expanded tremendously concurrent with an explosion in the amount of data generated within many disciplines (biology in particular). The speed of computation has also increased: personal computers are now as powerful as the machines once only available at a handful of national laboratories. When SDSC began its work, our "revolutionary" CRAY X-MP/48 ran at 840 million calculations per second, comparable to the top speeds of a top-of-the-line PC available today. Finally, research has become increasingly multidisciplinary (Jasanoff *et al.* 1998).

Today, environmental informatics offers SDSC a compelling imperative and opportunity to apply computing and information technology within yet another new community, and through key hires and strategic partnerships we have acquired the expertise to make a contribution. The following examples, which build upon each other, illustrate how SDSC has strategically pursued its role within the grand challenge.

Computational ecology

In 1995, as a first step toward defining opportunities at the interface between computing and ecology, John Helly organized a workshop at SDSC to "gather together ecologists and computer scientists for the purpose of identifying those technology issues which impede the progress of ecological research" (Helly *et al.* 1996). Workshop participants identified the three key areas of technology need: data management, modeling, and visualization.

Common to these areas is the need to develop standards to facilitate the sharing of data, the comparison of results, and the integration of model components. Unique to the problem of data sharing is the issue of the proprietary nature of research data and the lack of institutional incentives for data sharing along with the usual issues of intellectual property. The use of visualization continues to be inhibited by the highly specialized knowledge still required to effectively utilize current visualization packages along with fundamental questions such as how should statistical error be represented in visual presentations. The modeling community is challenged by questions such as how multi-scale, multi-resolution models may be integration across disciplines, as well as the desire for collections of standard realizations of model components to minimize the redundant software development and facilitate the comparison of modeling analysis (Helly *et al.* 1996).

Data, visualization, and modeling have since formed a framework for our other activities within environmental informatics, including:

Biodiversity informatics

In 1997, the NSF awarded SDSC with funds to initiate activities in computational biodiversity and we hired David Stockwell, who has worked in Australia on a number of environmental issues and who had participated in part of the Environmental Resource Information Network (ERIN). David Stockwell developed the Biodiversity Species Workshop (http://biodi.sdsc.edu), a Web-based application for experimental analysis, mapping, and prediction of worldwide species distribution data (Stockwell and Peters 1999). Researchers can submit their own latitude and longitude data through a Web browser, conduct preliminary mapping and manipulation of that data, track changing distributions with climate change, and visualize the outputs using GIS, animation, and/or virtual reality. The central function of the Workshop is to develop predictions of spatial distribution of species using algorithms that range from simple to bioclimatic limit fitting to an artificial intelligence algorithm (Stockwell 1999).

The Biodiversity Species Workshop, which uses CGI Perl scripts in a frame-based web application, has already produced interesting results and holds the potential for other applications. Recently, a study that made use of the Workshop was conducted to address issues of conversatism of ecological niches in evolutionary time. The study looked at 37 pairs of closely related forest species, separated by an arid barrier, and concluded that the "geographic barriers are the major force driving the formation of new species" (Peterson 1999) (i.e., geographic barrier rather than ecology was the critical factor in speciation).

The Peterson study was based on data from over 5 years of painstaking negotiation and database development with museums around the world. A new tool, the Species Analyst (http://chipotle.nhm.ukans.edu), developed by David Vieglais, a research associate at the University of Kansas Natural History Museum, promises to these data at a scientist's fingertips. Species Analyst pulls up records from disparate biodiversity databases, many of which are located at natural history museums across the country, and can even access databases compiled using incompatible software. "It taps about 1.5 million records so far, using a computer program written according to a standard protocol (Z39.50) that libraries have long used to share bibliographic databases. The Kansas team's tool sends a query to each database, then pools the data it gets back. Putting the Web site to work is easy: Just tick any of the boxes next to each of the nine collections now on the Web, type a search term such as species name, and hit the query button. In seconds the site will produce a world map showing where the plant or animal has been found, along with a table--available as an Excel spreadsheet--that lists each specimen in the museums' collections and the date, collector, geographic coordinates, and so on" (Kaiser 1999).

Both the Species Analyst and the Biodiversity Species Workshop can be used alone, or linked together to produce a very powerful tool for integrating data and predicting spatial distribution of species. Other potential uses for this software suite include prediction of location of species (for determining where to send an expedition), endangered species preservation design, and managing deleterious invading species such as the hardwood-chomping Asian longhorn beetle or the house finch.

Maintaining and Providing Access to Data

John Helly, who had been involved in environmental modeling work, was recruited to SDSC in 1993. Helly has a background in computer science and has played a key role in data access and publication issues. He has worked with the Ecological Society of America (ESA) on their Future of Long Term Ecological Data (FLED) project, been involved in some initial International Long Term Ecological Research site visits, and plays a key role in the National Partnership for Advanced Computational Infrastructure.

In 1995 Helly was the key SDSC participant in an interagency project involving 31 institutions or agencies that have interest in or responsibility for monitoring data about the San Diego Bay (http://sdbay.sdsc.edu). The project provided a forum for discussion between agencies that had not before shared data, produced a platform and common grid for sharing data, and produced several hydrological models of the bay within which one can visually navigate the bay and simulate the impact of potential oil spills.

The San Diego Bay project, combined with insights gained from the FLED committee report, led to a collaboration between ESA, the National Center for Ecological Analysis and Synthesis (NCEAS), the Bishop Museum, and the California State Resources Agency, to develop an ecological data repository to support the acquisition, management, archival, and sharing of valuable ecological research data. The group created the Caveat Emptor Ecological Data (CEED) Web site (http://ceed.sdsc.edu), which provides a complete methodology for and implementation of a controlled method for publishing and sharing scientific data of any attribution type. Currently, CEED contains collections data from the Bishop Museum in Hawaii and data acquired through the San Diego Bay project.

PARTNERSHIPS

To have an impact on the grand challenge and address the funding realities of research in the U.S., SDSC has also pursued multidisciplinary partnerships with other likeminded major players - individual investigators and their institutions - within or related to environmental informatics. Collaborations of these sorts provide an opportunity to overcome cross-disciplinary barriers, such as those between ecology and systematics - where the overlap has not been great - and between these two and computer science. Four of the partnership activities SDSC is involved with are:

LTER Network Office

In 1996, the University of New Mexico (UNM) with SDSC teamed on a successful bid to become the Long Term Ecological Research (LTER) Network Office (http://www.sdsc.edu/sdsc-lter). Our two institutions shared a vision that the ecology community - in particular, the LTER sites - would take advantage of SDSC's high-performance computational resources to confront the scientific challenges underlying LTER. Our goals include promotion and dissemination of new computational technologies relevant to research throughout the LTER network, providing leadership in data management and dissemination to the LTER community and other scientific communities, and enhanced interaction between the LTER network and the larger environmental and scientific communities. Toward this end, in 1999 SDSC hired Tony Fountain to be its key liaison with the LTER community.

National Partnership for Advanced Computational Infrastructure (NPACI)

NPACI (http://www.npaci.edu), led by UC San Diego and SDSC, was formed in 1997 when the NSF supplanted the original Supercomputer Centers Program - under which SDSC had been created in 1985 - with the Partnerships for Advanced Computational Infrastructure (PACI) program. PACI combines computer and computational scientists to develop and deploy an advanced computational infrastructure for academic research. NPACI, to pair application "push" and technology "pull," established "thrust areas" in Technologies (Programming Tools and Environments, Data- Intensive Computing, Metasystems, and Interaction Environments) and Applications (Earth Systems Science (ESS), Molecular Science, Neuroscience, and Engineering). In each NPACI project, application need is linked to a critical computational technology under development. NPACI ESS projects include:

- Multiscale, Multiresolution Modeling - attempting to link together global (e.g. atmospheric and ocean), regional (e.g. mesoscale climate), and local (e.g. bay and estuary) models together;
- Quantitative Geography for Ground Truth - allowing environmental studies of global modeling for climate, the biosphere, and the carbon cycle;
- Surface Water Transport and Flow - to model the flow of bays and estuaries;
- Real-Time Coastal Data Acquisition Systems

(REINAS) - supporting regional-scale environmental science, monitoring and forecasting;
- Biological Scale Process Modeling - quantifying biodiveristy and its ecosystem function.

These projects are working closely with the Data-Intensive Computing thrust area (http://www.npaci.edu/DICE), which is developing tools (such as the Storage Resource Broker) to share, manage, and query distributed data, and working on several digital library technologies

LTER biological scale process modeling

Since the landscapes and ecosystems of the LTER sites are representative of larger physiographic, climatic, and ecological provinces, the findings at LTER sites are representative of larger geographic domains. NPACI, the LTER Network Office, and the University of Kansas have been involved in an effort to help LTER sites develop a regional perspective in their studies. Towards that goal, John Helly, Stuart Gage, and Bob Waide organized a series of regional modeling workshops at SDSC and invited members of the LTER, biodiversity, and computing communities to attend, explore issues of common interest, and forge new scientific collaborations.

The first workshop, a joint activity of LTER and NPACI, held in December 1998, had a primary goal of identifying ecological models that could make use of high-performance networks, computers, and data management facilities. A secondary goal was to define a series of regionalization experiments making use of these models. Twelve LTER models were identified, covering a variety of ecosystems, from coastal regions to deserts to urban centers, and including such key ecosystem processes as hydrology, atmospheric processes, and biogeochemistry. Twelve regionalization experiments making use of these models were defined. These included plans to link models across domains (e.g., linking bio-geochemistry and atmosphere models to capture terrestrial and atmospheric feedback processes; ecology and museum collections to understand the relationship between biodiversity and ecosystems function). The experiments also called for simulations at broader temporal and spatial scales and at finer resolutions than previously performed. Both these factors increase the computational burden of the simulations, necessitating the high-performance computing resources of SDSC. The center's visualization resources are also called on, as they provide an opportunity to illustrate the impact of LTER regional research to the public.

A second regional modeling workshop was held in November 10-17, 1999, and was organized by Tony Fountain, Stuart Gage and Bob Waide. The workshop's goals were to report results in the development of regionalized versions of models selected in the first workshop, and to define procedures for the validation of the regional models, including data and computation. A third regional modeling workshop is planned in conjunction with the LTER All-Scientists Meeting in Snowbird, Utah, in August 2000. At that point the results of the modeling experiments will be presented to the larger LTER community. Through this process we hope to encourage members of the ecology community to take full advantage of the computational resources available to researchers, to demonstrate the value of looking at regionalization, and to strengthen the ties between the ecological and museum collections community.

Knowledge and Distributed Intelligence (KDI)

SDSC is involved with two KDI projects, both of which are aimed at producing an infrastructure for distributed data handling at the interface of ecology and systematics. The first, Knowledge Networking of Biodiversity Information (NWBI), led by the University of Kansas, will develop high-performance network access to biological collection and classification databases through the implementation of international information retrieval protocols and metadata profiles, partly in collaboration with the U.S. Integrated Taxonomic Information System (ITIS). The project will also integrate museum databases with other earth systems sciences data, such as terrain and climate information, for predictive modeling and visualization of species distributions and related biodiversity phenomena. Finally, NWBI will initiate testbed research projects to assess the ecosystem function of biodiversity and the pattern and processes governing decline in amphibian populations worldwide.

The other KDI-A Knowledge Network for Biocomplexity: Building and Evaluating a Metadata-based Framework for Integrating Heterogeneous Scientific Data (led by the National Center for Ecological Analysis and Synthesis (NCEAS, http://www.nceas.ucsb.edu) and the LTER Network Office)-will integrate the distributed and heterogeneous information sources necessary to develop and test theories in ecology and its sister fields, into a standards-based, open architecture, knowledge network. Drawing on recent advances in metadata representation, the network will provide conceptually-sophisticated access to integrated data products

acquired from distributed, autonomous data repositories. It will also include advanced tools for exploring complex data sets from which hypotheses can be tested.

PARTNERSHIP FOR BIODIVERSITY INFORMATICS

Given shared vision of the grand challenge of environmental informatics, NCEAS, the LTER Network Office, the University of Kansas Natural History Museum and Biodiversity Research Center (KUNHM), and SDSC have formed a Partnership for Biodiversity Informatics (PBI). The mission of PBI is to conduct informatics research (i.e., the study and application of information technology) to advance knowledge discovery in systematics, ecology, and biodiversity. To accomplish this mission, the PBI will

1) enable research at the interface between ecology and systematics;
2) enhance knowledge discovery through the management, sharing, and integration of data and information;
3) demonstrate the value of collaborative research in a shared data environment; and
4) Expand access to information and knowledge for research, resource management, policy decision making, and education.

While this Partnership has just been formed, we see limitless potential to bring together the four founding organization's expertise and begin taking positive steps toward addressing the grand challenge of the 21st century.

SUMMARY

Harnessing knowledge of earth's biological and ecological diversity and how it shapes global environmental systems is a compelling problem. To address it, SDSC has developed tools that are now in use and has established key partnerships to conduct research, develop infrastructure, and train individuals through a multidisciplinary approach. We invite comments on our method, and would be interested in pursuing additional collaborations on tool and infrastructure development, and research using our tools. Possible pursuits involving environmental informatics include mapping and monitoring the environment, products and productivity from the environment, and avoidance and mitigation of natural disasters.

ACKNOWLEDGEMENTS

We wish to thank members of the Partnership for Biodiversity Informatics for their insight, in particular Jim Reichman, Leonard Krishtalka, Jim Beach, and Bob Waide. In addition, we wish to thank Amy Finley for her editing which led to an improved paper.

We have been supported in part by NSF ACI 9619020 (PA and JH), NSF DBI 9873021 and NSF DBI 9808739 (DS), and NSF DEB 9634135.

LITERRATURE CITED

Finley, A. 1999. Biocomplexity and the Art of Planetary Management, Envision, Jan - March, 1999.

Interim Report. 1999. Environmental Science and Engineering for the 21st Century: The Role of the National Science Foundation (NSB 99-133). July 29, 1999.

Helly, J., T. Elvins, D. Sutton, D. Martinez, S. Miller, S. Pickett, and A.M. Ellison. 1999. Controlled Publication of Digital Scientific Data.

Helly, J., S. Levin, W. Michener, F. Davis and T. Case. 1996. The State of Computational Ecology, SDSC.

Jasanoff *et al.* 1998. Conversations with the Community: AAAS at the Millenium. Science 278: 2066.

Kaiser, J. 1999. Searching Museums from Your Desktop. Science 284: 888.

Gross, K., E. Allen, C. Bledsoe, R. Colwell, P. Dayton, M. Dethier, J. Helly, R. Holt, N. Morin, W. Michener, S.T. Picket and S. Stafford. Report of the Committee on the Future of Long-term Ecological Data (FLED), Ecological Society of America.

Michener, W.K. 1996. Environmental Informatics: Methodology and Applications of Environmental Information Processing, Ecology (in book reviews) 77: 1309.

Peterson, A.T., J. Soberon and V. Sanchez-Cordero. 1999. Conservatism of Ecological Niches in Evolutionary Time. Science 285: 1265-1267.

President's Committee of Advisors on Science and Technology (PCAST). 1998. Teaming with Life: Investing in Science to Understand and Use America's Living Capital, March, 1998.

Raven, P. and E.O. Wilson. 1992. A Fifty-year Plan for Biodiversity Surveys. Science 258: 1099.

Stockwell, D.R.B. 1999 Genetic Algorithms II. *In* A.H. Fielding. (ed.), Machine Learning Methods for Ecological Applications. Kluwer Academic Publishers, Boston. pp. 123-144.

Stockwell, D.R.B. and D. Peters. 1999. The GARP Modeling System: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science 13: 143-158.

Systematics Agenda 2000. 1994. Charting the Biosphere. ("Technical Report"). Published by SA2000, Department of Ornithology, American Museum of Natural History, New York, NY 10024. 33 p.

Wilson, E.O. 1998. Integrated Science and the Coming Center of the Environment. Science 279: 2048.

(*Received October 11, 2000*)