RESEARCH DIRECTIONS IN BIODIVERSITY AND ECOSYSTEM INFORMATICS

Report of an NSF, USGS, NASA Workshop on Biodiversity and Ecosystem Informatics held at NASA Goddard Space Fight Center, June 22-23, 2000.

Dave Maier, Eric Landis, Judy Cushing, Anne Frondorf,
Avi Silberschatz, and John L. Schnase (Editors)

April 1, 2001

This workshop was supported by the National Science Foundation Digital Government Program under grant EIA-0084541, the US Geological Survey, and the National Aeronautics and Space Administration. All opinions, findings, conclusions, and recommendations in any material resulting from this workshop are those of the workshop participants, and do not necessarily reflect the view of the sponsoring agencies.

See http://www.gsfc.nasa.gov/BDEI for additional information.

Suggested Citation:

Maier, D., E. Landis, J. Cushing, A. Frondorf, A. Silberschatz, and J. L. Schnase (Editors). 2001. *Research Directions in Biodiversity and Ecosystem Informatics*. Report of an NSF, USGS, NASA Workshop on Biodiversity and Ecosystem Informatics (NASA Goddard Space Fight Center, June 22-23, 2000, Greenbelt, Maryland). 30 pp.

RESEARCH DIRECTIONS IN BIODIVERSITY AND ECOSYSTEM INFORMATICS

Dave Maier, Eric Landis, Judy Cushing, Anne Frondorf,
Avi Silberschatz, and John L. Schnase (Editors)

Report of an NSF, USGS, NASA Workshop on Biodiversity and Ecosystem Informatics held at NASA Goddard Space Fight Center, June 22-23, 2000.[1]

EXECUTIVE SUMMARY

In June 2000, a group of computer scientists, biologists, and natural resource managers met to examine the prospects for advancing computer science and information technology (CS/IT) research by focusing on the complex and often unique challenges found in the biodiversity and ecosystem domain. We refer to this emerging, interdisciplinary field of study as Biodiversity and Ecosystem Informatics (BDEI). This report synthesizes the discussions and recommendations made at the workshop. It itemizes current BDEI challenges, lays out a national BDEI research agenda, and recommends actions to be taken within the national research agenda. It also proposes specific mechanisms to communicate and implement those actions. The following points summarize the conclusions of this forum:

- The CS/IT research community plays a foundational role in creating the technological infrastructure from which advances in the environmental sciences evolve;

- The next-generation CS/IT applications required by our expanding need to understand complex, ecosystem-scale processes will require the solution to significant, ground-breaking CS/IT research problems;

- Important new research opportunities for the CS/IT community are provided by the urgency, complexity, scale, and uniqueness of the data, processes, and problems presented by work in the biodiversity and ecosystem domain; and

- There is an increased need for governmental and industrial support of basic CS/IT research in order to respond to these challenges. Both the national CS/IT and environmental research agendas would derive significant, synergistic benefit from such investment.

In the remainder of this section, we introduce two major themes that weave throughout this report. First, the CS/IT demands of biodiversity and ecosystem research are drastically changing, thereby requiring new solutions to fit the altered landscape. Second, the CS/IT research community has a long and successful record of creating new solutions and enabling the technology transfer needed to put these ideas into practical use. It is therefore a wise investment of public monies to ensure that the emerging, interdisciplinary field of biodiversity and ecosystem informatics becomes a healthy and viable discipline.

[1] This workshop was supported by the National Science Foundation Digital Government Program under grant EIA-0084541, the US Geological Survey, and the National Aeronautics and Space Administration. All opinions, findings, conclusions, and recommendations in any material resulting from this workshop are those of the workshop participants, and do not necessarily reflect the view of the sponsoring agencies. See http://www.gsfc.nasa.gov/BDEI for additional information.

Biodiversity and Ecosystem Sciences

The most striking feature of Earth is the existence of life, and the most striking feature of life is its diversity. This biological diversity — or *biodiversity* — provides us with clean air, clean water, food, clothing, shelter, medicines, and aesthetic enjoyment. Biodiversity, and the *ecosystems* that support it, contribute trillions of dollars to national and global economies, directly through industries such as agriculture, forestry, fishing, and ecotourism and indirectly through biologically-mediated services such as plant pollination, seed dispersal, grazing land, carbon dioxide removal, nitrogen fixation, flood control, waste breakdown, and the biocontrol of crop pests. And biodiversity — the biological richness of ecosystems *per se* — is perhaps the single most important factor influencing the stability and health of our environment. Clearly, this is one of our most important knowledge domains, vital to a wide range of scientific, educational, commercial, and government activities.

There is an increasing need to understand and respond to complex environmental problems. Just as we are developing a capacity to predict long-term climate events, we would now like to predict public health and ecological outcomes far into the future. Unfortunately, we currently lack the technologies to do this. The environmental sciences are "resource limited" by fundamental inadequacies in the CS/IT tools that can be applied to problems of this scale. If we are to keep pace with our need for quality information about the living systems of our planet, we must produce mechanisms that can efficiently manage petabytes of a new generation of high-resolution, Earth-observing satellite data. We must understand how to integrate these new datasets with traditional biodiversity data, such as specimen data held in natural history collections, and genomic data from cellular- and molecular-level work. We must be able to make correlations among data from these and even more disparate sources, such as ecosystem-scale global change and carbon cycle data, compile those data in new ways, analyze them, and present the results in an understandable and usable way.

Despite encouraging advances in computation and communication performance in recent years, we are still unable to perform these activities on a large scale. It is only recently, for example, that IBM announced plans to build the world's fastest supercomputer — *Blue Gene* — which will attempt to compute the three-dimensional folding of human protein molecules. Given the thousands of proteins that are produced by the unknown millions of species on this planet, and given too that many of these molecules may have potentially significant economic value or environmental importance, we are clearly entering a new world of computer-mediated exploration.

Biodiversity and Ecosystem Informatics

Until recently, little attention has been paid to computer and information science and technology research in the biodiversity and ecosystem domain. The interdisciplinary field of biodiversity and ecosystem informatics (BDEI) is attempting to change that. We are pushing the boundaries in two directions by identifying research challenges that can simultaneously advance the environmental sciences and the computer and information sciences. The potential for such synergies is high because of the nature of work in the biodiversity and ecosystem domain.

The single most important factor influencing work in this field is the problem of complexity. This complexity arises from several sources. First is the underlying biological complexity of the organisms themselves. There are millions of species, each of which is highly variable across individual organisms, populations, and time. Species have complex chemistries, physiologies, developmental cycles, and behaviors resulting from more than three billion years of evolution. There are hundreds, if not thousands, of ecosystems, each comprising complex interactions among large numbers of species and between those species and multiple abiotic factors.

The second source of complexity is sociologically generated and includes problems of communication and coordination — among agencies, divergent interests, and groups of people from different regions, from different backgrounds, and with different points of view. Biodiversity and ecosystem data can be politically and commercially sensitive and entail conflicts of interest. The kinds of data scientists have collected about organisms and their relationships vary greatly in precision and accuracy, and the mechanisms used to collect and store these data are almost as diverse as the natural world they document. Many important observations are made by non-scientists, such as amateur birders and natural history enthusiasts. And the range of datasets with which these datasets must interact is unusually broad, including geographical, meteorological, geological, chemical, physical, and genomic sources. There is thus an unusual need to accommodate differences in data quality within a democratized community information infrastructure that is both formal and informal.

As in most biological and earth sciences, location is central. Much biodiversity and ecosystem data is *georeferenced* — it is tied to some place on the globe. Sometimes the designation of a location can be ambiguous or imprecise, especially with observations and samples taken in previous centuries. As a result, something as central to the science as a means for spatial referencing becomes a complex issue. Biodiversity and ecosystem data are also distinctive for being *species-referenced*. Genetic data is frequently associated with a species or sub-species, invasions and extinctions are tracked at the species level, and much of the characterization of an ecosystem is described through the number and distribution of its constituent species. However, the naming of species is an abstract process, deeply embedded in long-standing scientific cultural processes — incomplete, subject to local variation, and changing with time. In the ongoing process of species discovery, different scientists may assign two or more names to the same species, and a single species name may be applied to what turns out to be distinct species. To make matters worse, most species on the planet have not yet been named and classified, and there is no authoritative listing of all the species we do know. In this field, ontological complexities abound!

Many key biodiversity and ecosystem questions involve flux — changes in range, numbers, distribution, genetics, and proportions over time. Extinctions, migrations, incursions, restorations, predicted environment impacts are all issues of flux. However, seldom does one dataset span enough time, area, or include enough species to answer a specific question by itself. Scientists often require that biodiversity and ecosystem data be assembled from different sources into time sequences of comparable datasets, realizing that the component datasets may have been compiled for quite different purposes. Scientists also often deal with data at small scales over a large area or extended periods of time. Many significant situations will be lost if standard methods for moving to larger scales are used.

Finally, historical information serves prominently in the work of biodiversity and ecosystem scientists. Examples include plant and animal specimens and their labels, publications (some dating back 250 years), maps, and personal field notebooks. The study of biodiversity and ecosystems requires the analysis of trends, adaptations, and long-term relationships. These historical sources are thus often as pertinent as contemporary data. An additional and significant problem is that many of the historical information sources are not yet in digital form. For example, over 750 million natural history specimens and their accompanying metadata remain to be digitized in the US alone.

Because of these complexities, humans still play a crucial role in the processing of biodiversity and ecosystem data. This information is simply not as amenable to automatic correlation, analysis, synthesis, and presentation as many other types of information. People act as sophisticated filters and query processors — locating resources on the Internet, downloading datasets, reformatting and organizing data for input to analysis tools, then reformatting again to visualize results. This process of creating higher-order understanding from dispersed datasets is a fundamental intellectual process in the biodiversity and ecosystem sciences, but it breaks down quickly as the volume and dimensionality of the data increase. Who could be expected to understand millions of cases, each having hundreds of attributes? Yet problems on this scale are common in biodiversity and ecosystem research.

Biodiversity and Ecosystem Informatics Research Agenda

Given this context, there are clearly areas where computer science and information technology research could be advanced — with great social and scientific benefit — by focusing on challenges in the biodiversity and ecosystem domain. These synergistic opportunities fall into three major categories: acquisition and conversion of data and metadata, analysis and synthesis of data and metadata, and dissemination of data and metadata. Some specific opportunities include the following:

Acquisition and Conversion of Data and Metadata

- Modernizing the Biological Library – The accumulated volume of biological information and data collected over the past 250 years is massive. Improving methods for organizing, storing and retrieving these records is extremely critical. New techniques and tools must be developed for information extraction, text understanding, and cross-lingual information retrieval, making this an important non-business application domain for research on data integration, data cleansing, data warehousing, and archiving.

- Digitizing the Biological Legacy – America's museums and laboratories maintain at least 750 million biological specimens. There is an urgent need to convert them, their documentation, and new specimens into metric-quality digital formats. This provides an excellent opportunity to advance research on lossless image compression, 3D image understanding, robotics, and the problem of integrating physical artifacts into digital libraries.

- Multi-dimensional Observation and Recording – Efforts are needed to enable the collection of detailed information about the Earth in multiple dimensions and at multiple scales. This provides rich opportunities for research on scaling sensor-fusion techniques to large fields and developing and testing temporal-spatial data access methods.

- Mobile Computing – New instrumentation is needed to bring knowledge to the field and to collect, store, and transmit data from the field. Specific opportunities here include applications of human-computer interaction research to multi-model interfaces, hands-free systems, wearable computers, remote presence, robotics, and human augmentation.

- Taxonomic Freedom – Changes in biological names and classification schemes over time and discipline presents enormous challenges. There is a need to integrate various interpretations, views, and versions of taxonomic data and make it available in a simple, easy-to-understand formats. This provides an unusually challenging context in which to examine the flexibility and robustness of knowledge representation systems, particularly their temporal and versioning aspects and their support for cross-ontology linking and translation.

Analysis and Synthesis of Data and Metadata

- Comparing Across Scales – Biological data from different sources and times are frequently collected and presented in different scales and resolutions resulting in a loss of detail when multiple datasets are required for data synthesis and analysis. Tools and procedures to facilitate analysis across scales are needed, which provides an important opportunity for research on adaptive- and multi-resolution techniques for computation and modeling.

- Modern Modeling – Researchers, managers, and policy-makers require *models* for biological decision-making rather than disaggregated collections of data and facts. Improved spatio-temporal modeling of biological, ecological, and social processes are required, providing a fertile area for multi-modal data assimilation research and high-performance computing.

- Taxonomic Retooling – Taxonomists need new and improved tools for naming and defining species, changing and manipulating taxonomic organization, and performing other tasks regarding taxonomic content and structure. This provides a rich domain for research in knowledge acquisition and hierarchical display techniques.

- Making Data Usable – Too frequently, decision-makers underutilize research results. To enhance the use of biological data, decision-makers require systems that will facilitate the synthesis and analysis of scientific data and research results. This is an excellent application area for research on uncertainty analysis, reasoning with incomplete information, and automatic summarization.

- Machine Processable Metadata – Current metadata is largely for human consumption. It is used to document and help interpret datasets. However, much more value will be gained from it when it is complete, correct, and descriptive enough to help automate data manipulation tasks, such as summarization, combination of datasets, and conversion of data to appropriate forms for use in models and statistical tools. There are important opportunities here for testing of data-based inferencing technology and metadata-based information integration research.

- Need for Speed and Accuracy – Many tasks in data management are iterative and require considerable time. Researchers are frequently challenged with data entry and pattern discovery procedures and are required to estimate the quality of utilized data. Meanwhile species are disappearing at a rate greater than they can be recorded. This is a challenging domain for research on data reduction and data mining algorithms, including parallel implementations, modeling and analytic techniques with tunable accuracy, and data quality metrics.

Dissemination of Data and Metadata

- Visualization – Users of biodiversity and ecosystem data and information, including land managers, policymakers, educators, non-governmental organizations, industry, and others outside biological research, need visualization techniques to better understand data, relationships among data, natural processes, and management actions over time. This leads to opportunities for research on advanced display and visualization techniques, including display of uncertainty, user-adaptive display, and multi-dimensional data visualization.

- Interdisciplinary Collaboration and Communication – Stakeholders of biodiversity and ecosystem data are growing in numbers and breadth. No longer are management decisions made solely by individuals or single agencies, but involve communities of individuals. This calls for the development of computer-supported cooperative work and remote collaboration research suited for participants with widely varying roles, specialties, and training. It also provides opportunities to study cross-domain mapping, data integration, data quality management, ontologies, and other knowledge representations.

- Data Management Guidelines – Biodiversity and ecosystem information is frequently used in complex and potentially controversial political, economic, and environmental discussions and decision-making. Informatics issues arising from this context include issues of data security, data sharing policies, intellectual property rights, quality assurance, and reuse of data. This provides important opportunities for research on data models for representing annotation and provenance, explicit modeling of data product generation, and policy development and dissemination techniques.

Biodiversity and Ecosystem Informatics Research Agenda Implementation Plan

A concerted effort should be made to build a sustainable biological information infrastructure that proactively engages the broader CS/IT community in BDEI research. The following are among the specific actions that should be taken:

- Interdisciplinary Planning Groups – Interdisciplinary planning groups, comprising members of the biodiversity and ecosystem and CS/IT research communities, should be established to serve as a mechanism for articulating and communicating the special informatics challenges from one community to research actions in the other community. These planning groups should identify existing CS/IT technologies that could be transferred from other domains, long-term basic CS/IT research questions, short-term CS/IT research that needs to be pursued, and infrastructure needs including, equipment, facilities, networks, and personnel.

- Matching Research Needs with Available and Appropriate Mechanisms – Efforts to implement any research agenda item need preliminary study to determine how these research actions could benefit from existing programs. These programs include mechanisms for funding, partnerships, interdisciplinary training and teaming, and resource sharing.

- Communicating the Research Agenda – Every effort should be made to communicate the BDEI research agenda to an audience that includes researchers in computer science and the biodiversity and ecosystem sciences, as well as the many agencies and foundations that support their efforts. Recommended actions include developing extended workshops or seminars; building a multi-sector, multi-disciplinary community; developing interdisciplinary "matchmaking" mechanisms; adding a CS/IT component to existing biodiversity and ecosystem projects; developing venues for multi-disciplinary activities; and promoting biodiversity and ecosystem informatics through the dissemination of reports, publications, email distribution lists, and websites.

- Short-term Critical Actions that Require Immediate Attention – Several activities should be acted upon immediately to "jump start" BDEI research. Paramount among these is an urgent plea from the scientific community for the formation of an NSF, USGS, NASA interagency strategic partnership to promote BDEI research. Within the next fiscal year, every effort should be made to launch a high-profile solicitation for cutting-edge research in this area. This activity should highlight the urgency and importance of biodiversity and ecosystem informatics problems and opportunities, and provide a form for organizing problem-specific, interdisciplinary consultative and investigative teams and pursuing problems relevant to the core missions of the sponsoring agencies.

Conclusion

A more complete consideration of these issues is presented in the report that follows. The workshop and report basically attempt to make the point that the biodiversity and ecosystem sciences are fundamentally information sciences, and worthy of special attention from the computer science and information technology community because of their distinctive attributes of scale and socio-technical complexity. At almost every turn, scale, complexity, and urgency conspire to create a particularly wicked set of problems. Working on these problems will undoubtedly advance our understanding and use of information technologies, and, even more important, give us the tools to protect and manage our natural world so as to provide a stable and prosperous future.

TABLE OF CONTENTS

RESEARCH DIRECTIONS IN BIODIVERSITY AND ECOSYSTEM INFORMATICS

Dave Maier, Eric Landis, Judy Cushing, Anne Frondorf,
Avi Silberschatz, and John L. Schnase (Editors)

Report of an NSF, USGS, NASA Workshop on Biodiversity and Ecosystem Informatics held at NASA Goddard Space Fight Center, June 22-23, 2000

INTRODUCTION

> *It's Tuesday afternoon. Karen Culver has just been asked by her boss at the state fish and game agency to address a meeting of the Eagle River Watershed Council. The council is presenting a restoration plan for Silver Creek, with a particular goal of improving bull trout habitat. One aspect of the plan involves the removal of a small diversion channel that feeds an irrigation pond, with the expectation that this would improve stream flows and lower water temperatures in the summer. The proposal is to replace the irrigation water drawn from the creek using a pipeline or culvert from another nearby water source to the pond. Local landowners and members of the public have been at odds about whether closing the existing channel would have much benefit, and what the adverse effects and costs might be for installing a culvert or a pipeline.*
>
> *Karen thinks she could compile the scientific information relating to these issues in about six weeks. To do this, she needs topographic maps of the Silver Creek drainage and surrounding area to identify possible sources for replacement irrigation water and likely routings of a culvert or pipeline. She also needs to locate any recent hydrology studies and fish counts of Silver Creek. Karen must also check what the ownership and land use is of areas that a culvert may cross. If she could retrieve the appropriate meteorological data, along with the topographic and hydrologic data, maybe her co-worker, Tom Hamilton, could run a simple model to project Silver Creek's stream flow, water temperature, and downstream sedimentation after closing the channel. With that information, she might be able to look for streams with similar characteristics and what bull trout populations they support. Then she would need to get those tables of numbers into a form understandable by everyone at the upcoming meeting. Maybe she can go through agency archives to see if there are any historical surveys of the area before the channel was dug in 1932. She also wonders if there are any sensitive populations of other plant or animal species currently dwelling in the creek or channel. It would be helpful to go into the field and examine the Silver Creek area. But it's 220 miles away in the southwestern corner of the state. Besides, the watershed council meeting is Friday morning. Karen has three days, not six weeks.*

Karen's situation typifies problems that arise in trying to answer management and scientific questions about species diversity and ecosystem health using the current information infrastructure in this domain. Relevant information is difficult to locate (if it exists), and may be in a variety of digital and non-digital forms. Integrating the information and putting it into a form suitable for use with a specific analytic tool usually involves intensive and time consuming human interactions with the data. Visualizations of datasets are not easy to construct quickly. And the questions are posed in a climate of increasing public scrutiny of agency decisions and concern about species and ecosystem preservation. We will return to Karen later in this report to see how this situation could be improved with some of the advances in biodiversity and ecosystem informatics proposed here.

BACKGROUND AND CURRENT SITUATION

Many agencies and organizations are involved in research on and the monitoring and management of our biodiversity and ecosystems. Together, these partners and collaborators in federal, state, and local government agencies, academia, non-profit organizations, and private industry are concerned with developing a national data and information infrastructure to better support the collection, management, dissemination, and application of biodiversity and ecosystem data and information. The vision of an enhanced 21st-century biodiversity and ecosystem infrastructure is described in the 1998 *Teaming with Life* report of the President's Committee of Advisors on Science and Technology (PCAST, 1998).

The PCAST report called for developing the "next generation" of the National Biological Information Infrastructure. NBII-2 will not only support more effective biodiversity and ecosystem sciences and resource management, it will also provide a rich set of new challenges for the computer science and information technology (CS/IT) research and development community. The challenge to *both* these communities now is to work together to articulate and pursue an *interdisciplinary* research agenda that will advance CS/IT research into new and exciting directions, while helping to advance biodiversity and ecosystem sciences and resource conservation. Towards this end, the National Science Foundation, along with the US Geological Survey and the National Aeronautics and Space Administration sponsored a workshop entitled "Research Directions in Biodiversity and Ecosystem Informatics" in June 2000. This workshop brought together scientists representing a broad cross-section of CS/IT disciplines and scientists and resource managers from the biodiversity and ecosystem community. (See Appendix A-IV for a list of workshop participants.)

The workshop's main goal was to examine the prospects for advancing computer science and information technology (CS/IT) research by focusing on the complex and often unique challenges found in the biodiversity and ecosystem domain. We refer to this emerging, interdisciplinary field of study as *Biodiversity and Ecosystem Informatics* (BDEI). This report synthesizes the discussions and recommendations made at the workshop. It itemizes current BDEI challenges, lays out a national BDEI research agenda, and recommends actions to be taken within the national research agenda. It also proposes specific mechanisms to communicate and implement those actions.

Workshop on Biodiversity and Ecosystem Informatics

The workshop was held June 22 and 23, 2000. After initial, overview presentations on pertinent national and global-level biodiversity and ecosystem information networks (see Appendices I-III), the group heard two case study presentations that helped provide a representative introduction to the requirements, issues, and challenges inherent in the biodiversity and ecosystem sciences and resource management arena. Following the case study presentations, the group began discussions to identify and articulate the informatics challenges of the biodiversity and ecosystem domain. Full group discussions on the domain's informatics challenges were then followed by breakout group discussions on the specific types of CS/IT research questions that flow from the informatics challenges. On the second day, the full group discussed, further refined, and synthesized the findings of the previous day's breakout groups. The afternoon sessions focused on breakout group discussions of key aspects of implementing the BDEI research agenda, including communication, funding opportunities, and facilitating interdisciplinary research. The concluding session focused on group discussion of the most significant issues and ideas raised during the workshop and recommendations for short- and medium-term "next steps."

The Nature of Biodiversity and Ecosystem Data – What Makes it Different?

Biodiversity and ecosystem data are distinctive, and information management challenges in this domain are, in many ways, unique when compared to those found in other disciplines. Not only is there a tremendous volume of stored data, it is collected, archived and disseminated in every conceivable scale, format, and location, and represents multiple points or ranges of time and space. The requirements of those accessing biodiversity and ecosystem data are extraordinarily diverse. Users demand access to biodiversity and ecosystem information through place names, latitude and longitude, species names, author, time, and content. Not only is it presented in a variety of formats, scales, etc., there are also non-biological data important to biodiversity and ecosystem questions, such as data about climate, geology, topography, land use and ownership, political boundaries, roads and other cultural entities, and population densities, to name a few.

As in most biological and earth science domains, location is central. Questions of diversity are almost always couched in terms of some area or place name. Thus much biodiversity and ecosystem data is *georeferenced*—it is tied to some place on the globe. Sometimes the designation of a location can be ambiguous or imprecise, especially with observations and samples taken in previous centuries. A designation of location at a country or even a continent level is not unheard of in natural history collections. Current methods rely on better maps, remote imagery, and GPS systems and produce datasets having high accuracy. As a result, something as central to the science as a means for spatial referencing is thus a complex issue.

Biodiversity and ecosystem data are also set apart because much of it is *species-referenced*. Genetic data is frequently associated with a species or sub-species, invasions and extinctions are tracked at the species level, and much of the characterization of an ecosystem is described through the number and distribution of its constituent species. However, unlike coordinate systems for geographic locations, which are universal, consistent, and relatively unchanging, the naming of species is incomplete, subject to local variation, and will change with time. Most species on the planet have not yet been named and classified, and there is no authoritative listing of all the known species names (though efforts in this direction are underway [See Appendix A-II: Global Biodiversity Information Facility]). Further, the mapping of names to organisms is an abstract process and evolves over time. It is possible that two or more names for the same species have been assigned by different scientists, or that the same name has been applied to what turns out to be two distinct species. (Physical *type specimens* are key to resolving such naming ambiguities.) Even if a complete and consistent lexicon of species names is produced, there are centuries of observations, specimens, and publications using the taxonomic schemes of their historic places and times.

To appreciate the problem with species-referenced data, consider the following analogy. Imagine dealing with observations that are located relative to maps by 15th century cartographers. These maps are incomplete, and they disagree on names of particular places. Or, where they agree on place names, they might disagree on the location of the place. They might even contain places that do not actually exist. Looking at these maps, it might seem that the Earth's features are changing over time, though generally these changes reflect improved knowledge of what is actually there.

The naming of species is not the only aspect of biological data where classification schemes and terminology are ambiguous. Other characteristics, such as ground cover, vegetation type, and soil type, have classifications and terms that may vary over time, place, disciplines, institutions, and individuals. Any effective system for managing and querying biodiversity and ecosystem data must deal directly with changing classification schemes and varying terms and definitions rather than ignoring the problem. The ambiguities in terms and changes in classification must be explicitly modeled, so, for example, it is possible to access old data using current terminology, or trace information about a given category across time and place.

Another important aspect is that many key biodiversity and ecosystem questions involve flux—changes in range, numbers, distribution, genetics, and proportions over time. Extinctions, migrations, incursions, restorations, predicted environment impacts are all issues of flux. However, seldom does one dataset span enough time, area, or include enough species to answer a specific question by itself. When did exogenous species of shellfish first appear in the Great Lakes? Is the population of salmon in this river system becoming more or less genetically diverse? What are the long-term ecological consequences of exotic species invasions? The information to answer such questions must often be pieced together from many sources. Scientists require that biodiversity and ecosystem data be assembled into time sequences of comparable datasets, realizing that the component datasets may have been compiled for quite different purposes.

Unlike many disciplines, such as physics or medicine, that primarily rely on current or recent data, historical information serves prominently in the work of biodiversity and ecosystem scientists. Examples include plant and animal specimens and their labels, publications (some dating back 250 years), field notebooks and observation files (often kept by hand), and maps. The study of biodiversity and ecosystems requires the analysis of trends, adaptations, and long-term relationships. Therefore these historical data are often as pertinent as contemporary data. Unfortunately, many of the historical information sources are not yet in digital form. For example, over 750 million natural history specimens and their accompanying metadata remain to be digitized in the US alone.

Another strong requirement for storing and manipulating biodiversity and ecosystem data not found in other disciplines is the need to deal at small scales over a large area or extended period of time. Many significant situations will be lost if standard methods for moving to larger scales are used, such as elision of features below a certain size. For example, the known infestation of kudzu in the western United States consists of two plots of 1/4 and 1/8 acre, out of over a million square miles of land area. Riparian zones are often pathways for invasion of species, but are essentially one-dimensional features in a two-dimensional terrain. Small areas with a differing ecosystem, such as aspen groves in meadows and grasslands, may sustain a disproportionate number of the distinct species in a larger region. Users of biodiversity and ecosystem data need to be able to preserve selective portions of a dataset or map at higher detail than that required for the dataset or map as a whole, particularly when moving to a larger scale.

It is important to realize that the historical and current record for biodiversity and ecosystem data will never be as complete as desired. Some research question or management decision will always benefit from a dataset over a larger area, or a sampling regimen at a smaller granularity or at shorter or longer time intervals, or a complete census of a species in an area versus a sample. The challenge to meet these new requirements for biodiversity and ecosystem information providers is grand, but with the right tools and processes, achievable.

Biodiversity and Ecosystem Informatics – Current Needs and Unique Challenges

We are confronted by both opportunities and challenges in the development of biodiversity and ecosystem informatics. Our greatest opportunity is that enhancement of the information infrastructure for biodiversity and ecosystems will allow us to address fundamentally new types of questions that span levels of biological organization. How do ecological factors affect diversity at different levels of phylogenetic relatedness? Which genes are held in common across taxa inhabiting streams, and which are unique? What are the biotic influences on climate change or the carbon cycle? To answer these (and other) questions, information systems need to make the connections between taxa, specimens, phylogenies, geographic locations, and environmental factors. Specific challenges include the following:

Varying methodologies, measurements, taxonomies, and questions – Traditional information systems (typically designed for business) are based on assumptions that may not apply to biodiversity and ecosystem data. Traditional applications emphasize data integrity and internal consistency, while biodiversity and ecosystem data, by nature, often contain inconsistent observations driven by differences in methodologies, measurement precision, or even different taxonomic information collected in the past relative to current (and future) taxonomic usages. Traditional applications also emphasize the creation of "standard" reports, while biodiversity and ecosystem applications need to be adaptable to new types of questions. Adapting or altering traditional software designs to meet the needs of scientists is a major challenge for biodiversity and ecosystem informatics.

Multiple spatial and temporal scales in a vast volume – Biodiversity and ecosystem applications need to deal with a variety of temporal and spatial scales, from the microsecond to the millennium and from the gene to the biosphere. For instance, managers and researchers often require historical data to survey relationships and trends, and to estimate the costs and benefits associated with planned management or policy actions. These systems also need to deal effectively with immense quantities of data, much of it (such as data on museum specimens or species descriptions) not currently in digital form. Where digital information exists, it is often distributed among systems that may not be interoperable.

Breadth of participation – Many important biodiversity and ecosystem observations are made by non-scientists. Amateur birders, community planners, students, gardeners, hikers, natural history enthusiasts, and others often create potentially useful datasets. There is a need to develop tools for the community construction of biodiversity and ecosystem information, and create the mechanisms that allow us to establish relationships across all levels of biological data and organization within a distributed information infrastructure. The resulting applications will need to accommodate and evaluate differences in data quality in this formal and informal infrastructure, and be flexible enough to provide multiple views of the underlying data — including alternative phylogenies and taxonomies, georeferencing systems, and biodiversity and ecosystem models.

CASE STUDIES

The following two case studies were presented at the workshop and illustrate typical informatics problems facing biodiversity and ecosystem researchers and resource managers.

CASE STUDY 1: Forecasting Exotic Plant Invasions In Colorado and Utah

Introduction

Invasive species cost the United States of America an estimated $137 billion per year—with $29 billion per year in crop losses alone (Pimentel, 1999). Invasive plant species not only compete with crops, they compete with native plant species, attract pollinators, alter nutrient cycling and fire frequency, and can cause the extirpation or extinction of native plant species (Mack, 2000). Examples of invasive species include Chestnut blight and Dutch elm disease in northeastern forests and parklands; yellow starthistle, European wild oats, water hyacinth, and white pine blister rust in California; tamarisk, and African lovegrass in the southwest; cheatgrass, smooth cordgrass, hydrilla, and white pine blister rust in the northwest; purple loosestrife and Kentucky bluegrass in the Midwest; and kudzu, water hyacinth, and Brazilian pepper in the southeast. Hundreds of invasive species affect Hawaii and Pacific territories. Land management agencies recognize these problems and have listed invasive plant species as top priorities for research and resource management activities.

Invasion of exotic plant species creates a formidable stress on natural ecosystems (Stohlgren *et al.*, 1999a,b). Degradation of habitats resulting from exotic invasion can be assessed by quantifying native and exotic plant diversity at multiple scales. Experience from research projects on Federal and non-Federal lands has emphasized multi-scale, statistically robust sampling, both in the field and in the laboratory using remote sensing, GIS analysis, statistics, and modeling (Stohlgren *et al.*, 1998, 1999a,b). Because of the widespread use of these methods, comparability of field data among agencies allows for local, regional, and national syntheses and improved predictive modeling capabilities. We are equally concerned that the past and future spread of invasive species affects carbon storage, nitrogen cycling, and fire dynamics in many ecosystems.

The predictive modeling challenge in assessing exotic plant invasions is still immense and forecasting is difficult. Various species are intentionally and unintentionally introduced into the United States each year. Some are met with inhospitable sites, strong competitors, or other biotic constraints, but other species find a new home, and we have an incomplete understanding of what makes a habitat vulnerable to invasion. Likewise, the attributes of highly invasive species are not well documented or understood. However, we can look to the landscape for clear direction. Several studies have shown that exotic plant species have invaded hot spots of native plant species richness (Stohlgren *et al.*, 1997, 1998, 1999). Our goals should be to: (1) accurately describe current locations and physical environment for many invasive plant species; (2) predict potential habitats of invasive species; (3) forecast rates of spread for selected exotic plant species; and (4) evaluate the effects of invasive alien species on ecosystem structure and function to maintain native biodiversity and natural ecological processes.

Current Capabilities

Lots of vegetation plot data exist on native and exotic plant species richness and foliar cover throughout Colorado and Utah. Colorado datasets include over 400 plots from The Colorado Natural Heritage Program, over 200 plots from the US Geological Survey, and over 100 plots from the USDA Forest Service's Forest Health Monitoring Program and other sites. Many of the datasets contain information on soils (texture, nitrogen, carbon), and can be linked to specific vegetation types. In Utah, over 150 plots have been established. Remote sensing data are available (Landsat TM imagery) and are georeferenced for some portions of the areas (e.g., Rocky Mountain National Park in Colorado and Grand Staircase-Escalante National Monument in Utah). Spatial analysis algorithms have been very successful in early tests (Kalkhan *et al.,* 1999).

We have successfully used this landscape analysis approach to address resource issues in Rocky Mountain National Park, Colorado, and to assess weed invasions in the Central US (Stohlgren *et al.*, 1998b, 1999a). Specifically, we are quantifying the effects of elk grazing on plant diversity, identifying areas of high or unique plant diversity needing increased protection, and evaluating the patterns of non-native plant species on the landscape. This same approach is ideally suited for other areas. It relies on new multi-scale sampling methods that have been extensively peer-reviewed in the scientific literature (Stohlgren *et al.*, 1995, 1997a,b,c, Stohlgren *et al.*, 1998a,b, 1999a,b, Kalkhan *et al.*, 1998).

Current Constraints and Hurdles to Overcome

There are however, several constraints and limitations to forecasting exotic species invasions. First, existing computer capabilities are woefully inadequate. Moderate-resolution spatial model simulations take a week to 10 days on a Sun Workstation for each dependent variable. Coarse-resolution models (degrading the resolution of the remotely sensed data by increasing the grid size take 3 days, but are far less informative. Second, high-resolution remotely sensed data (side-looking radar, high-resolution digital elevation models, etc.) are too cumbersome for our existing computer capabilities for regional, much less statewide, scales.

Species richness and cover data must be linked to productivity, carbon storage (vegetation and soils), and rapid vegetation change at local, landscape, and regional scales. This capability will require multi-scale and multi-phase sampling of several species at many sites along environmental gradients. High-performance computer capabilities are needed to integrate high-resolution remotely sensed data with detailed field data on vegetation, soils, and topography to develop "real time" simulations and forecasts of invasive plant species in Colorado or Utah as new data arrive (as new vegetation plots are established).

Key Informatics Issues:

1. Need for faster algorithms and more effective geospatial models.
2. New modeling techniques that incorporate spatial analyses.
3. Uncertainty maps of invasive species distributions must be constructed.
4. Real-time field data updates of existing models needed.
5. Better use of remote sensing, ground truthing, and stratification.
6. Augmentation of current sampling methods for multi-scale plots at multiple slopes, aspects and elevations.

References

Mack, R.N., D. Simberloff, W.M. Lonsdale, H. Evans, M. Clout, F. Bazzaz. 2000. Biotic invasions: causes, epidemiology, global consequences, and control. *Ecological Society of America, Issues in Ecology* 5.

Pimentel, D., L. Lach, R. Zuniga, and D. Morrison. 1999. Environmental and economic costs associated with non-indigenous species in the United States. (http://www.news.cornell.edu/releases/Jan99/species_costs.html)

Stohlgren, T.J., M.B. Falkner, and L.D. Schell. 1995. A modified-Whittaker nested vegetation sampling method. *Vegetation* 117: 113-121.

Stohlgren, T.J., G.W. Chong, M.A. Kalkhan, and L.D. Schell. 1997a. Rapid assessment of plant diversity patterns: A methodology for landscapes. *Environmental Monitoring and Assessment* 48: 25-43.

Stohlgren, T.J., M.B. Coughenour, G.W. Chong, D. Binkley, M.A. Kalkhan, L.D. Schell, D. Buckley, and J. Berry. 1997b. Landscape analysis of plant diversity. *Landscape Ecology* 12: 155-170.

Stohlgren, T.J., G.W. Chong, M.A. Kalkhan, and L.D. Schell. 1997c. Multi-scale sampling of plant diversity: Effects of the minimum mapping unit. *Ecological Applications* 7: 1064-1074.

Stohlgren, T.J., K.A. Bull, and Y. Otsuki Y. 1998a. Comparison of rangeland sampling techniques in the central grasslands. *Journal of Range Management* 51:164-172.

Stohlgren, T.J., K.A. Bull, Y. Otsuki, C. Villa, and M. Lee. 1998b. Riparian zones as havens for exotic plant species. *Plant Ecology* 138: 113-125.

Stohlgren, T.J., D. Binkley, G.W. Chong, M.A. Kalkhan, L.D. Schell, K.A. Bull, Y. Otsuki, G. Newman, M. Bashkin, and Y. Son. 1999a. Exotic plant species invade hot spots of native plant diversity. *Ecological Monographs* 69: 25-46.

Stohlgren, T.J., L.D. Schell, and B. Vanden Heuvel. 1999b. Effects of grazing and soil quality on native and exotic plant diversity in Rocky Mountain grasslands. *Ecological Applications* 9: 45-64.

Westbrooks, R. 1998. Invasive plants, changing the landscape of America: Fact book, Federal Interagency Committee for the Management of Noxious and Exotic Weeds (FICNMEW), Washington, D.C. 109 pp.

CASE STUDY 2: Meeting Information Needs of Listed Columbia River Salmon for Biological Decision Analysis

Introduction

Twelve species of Columbia River Basin salmon have been listed under the Endangered Species Act. As a result, numerous management decisions, on many temporal and spatial scales, are made with the purpose of protecting and restoring salmon populations in the Columbia River and its tributaries. Managing the data, information, and analyses associated with these complex and inter-related decisions has become a monumental challenge. A key component of this challenge is to provide decision-makers with a *decision support system* that is grounded in a probabilistic decision analysis approach.

The Columbia River Basin includes four states and two Canadian provinces. The basin has a complex network of mainstream hydroelectric dams, numerous smaller diversion dams, and many water withdrawal systems for agricultural irrigation. Management of the Columbia River relies on many sources of data and forecasting models, as does the management of salmon migration through the river. Water management decisions must balance the needs for salmon recovery, sturgeon recovery, and bull trout recovery while maximizing power production, flood control, and irrigation needs. The life cycle and ocean migration patterns of salmon dictate that data need to be collected on a coastwide and basinwide scale. Anthropogenic factors influencing salmon include: agriculture and urbanization affecting the spawning and freshwater life stages, hydroelectric dams and water storage dams affecting the juvenile and adult migration corridor, and, for chinook and coho, harvesting activities that take place in the ocean from California to Alaska. In addition, there are numerous river fisheries within the Columbia River system whose activities affect all salmon species.

Types of Data Collected

A considerable amount of data has been collected and archived since the listing of various salmon species under the Endangered Species Act. For juvenile salmon, the data collected include:

- tributary trap counts of smolts,
- release numbers of tagged smolts,
- release numbers of hatchery smolts (tagged and untagged) by species,
- dam counts, and
- recovery of tagged juvenile salmon at dams.

For adult salmon the types of data collected include:

- counts by species at dams and hatcheries,
- tagged salmon recoveries at hatcheries,
- catch estimates from river fisheries (tagged and untagged),
- fishing "effort" data for river and ocean fisheries,
- tributary spawning escapement counts from nests, live fish, and carcasses, and
- recovery of tagged salmon from spawning grounds.

Given the salmon's migratory pattern (see Figure 1), these data must be integrated into a "life cycle recovery analysis" to provide decision makers with sufficient understanding of the possible affects of any management action, be it on a local, regional, or international scale.

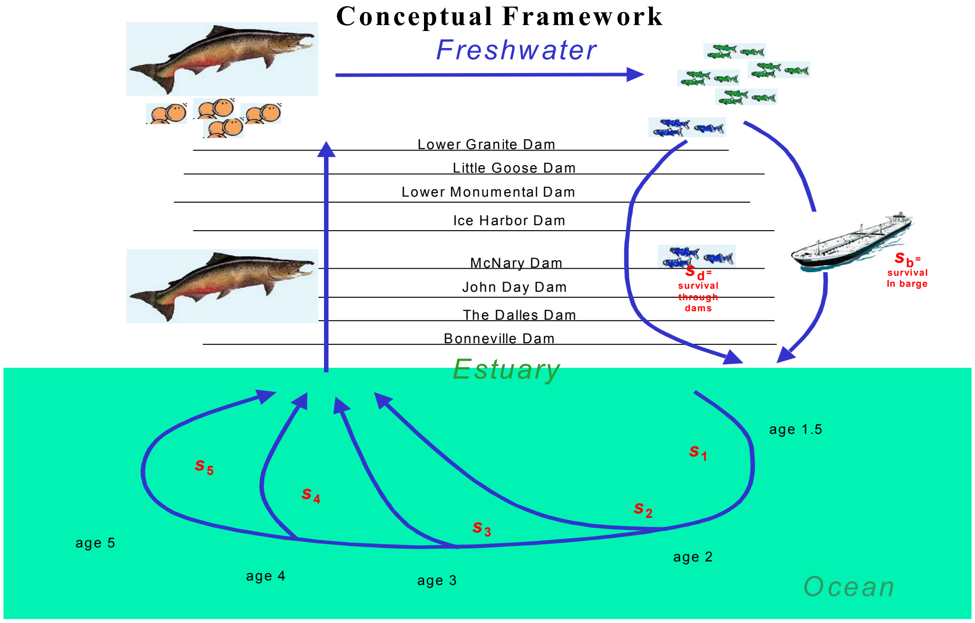

Figure 1. A time series of juvenile and adult data integrated into life-cycle recovery analysis demonstrates various management actions across habitat, hydro, harvest, and hatcheries.

The Need – Biological Decision Analysis

Decision makers require a process to evaluate the various combinations of management actions that will identify the likelihood of salmon recovery and the risks associated with those actions, and this complex process should be facilitated with a state-of-the-art, information-rich decision support system. Such a biological decision analysis system would assist in evaluating management actions for salmon harvests, hatchery obligations, hydro-electric decisions, including dam breaching, and large-scale land management agreements that affect fresh water habitat. Specifically, the objectives of a biological decision analysis system would include the ability to:

- evaluate the evidence for different hypotheses about how environmental factors influence survival,
- forecast the likelihood of survival and recovery resulting from different management actions, and
- identify management options with the least risk.

Key Informatics Issues:

1. Multiple, disparate data sources.
2. Requirement for interagency collaboration.
3. Massive volume of data representing wide spatio-temporal variation.
4. Need for computational modeling.

INFORMATICS CHALLENGES

These case studies illustrate some of the informatics challenges in the biodiversity and ecosystem domain. Here we elaborate on some of these issues, based on discussions at the workshop.

Data Integration

Biodiversity and ecosystem datasets are commonly maintained in monolithic, stand-alone systems that are developed in response to specific issues, projects, or mandates. These datasets and systems were designed independent of one another and frequently maintain their own standards for vocabulary, metadata, formats, scale, syntax, and access. As a result, biodiversity and ecosystem datasets are seldom effectively integrated with one another for purposes of access, synthesis, or distribution. True and meaningful integration of data is of particular significance in this domain, since the very nature of the ecological sciences is integrative, and most of the systems and processes we seek to understand comprise profoundly complex biotic and abiotic interactions. The science community thus faces an enormous challenge in locating, analyzing, and synthesizing existing data into useful knowledge for research, management, or policy-making. To compound the problem, the number of biodiversity- and ecosystem-related datasets is increasing at a rapid pace as resource agencies, universities, and conservation organizations respond to society's legal and social needs for biodiversity and ecosystem research and information. Numerous impediments to linking datasets exist. These impediments include data structure, semantic, and organizational challenges.

Data structure challenges – Several dataset-level metadata standards and numerous data formats are in use by data collectors. The use of multiple standards and formats (e.g. text, GIS, spreadsheet, database, video, audio, etc.) is problematic for those searching and correlating or integrating multiple datasets. While some improvements in the use of data and metadata standards have occurred (e.g. expanded use of the Federal Geographic Data Committee [FGDC] standard), many biologists continue to use their own (or no) data and metadata formats.

Semantic challenges – Successful data connectivity requires more than physical connectivity via technological solutions. Biologists are faced with a wide array of semantic options for identifying species, actions, land features, etc. Many biologists develop their own controlled vocabularies for project or organizational use and others may use recognized thesauri. Attempts to compare data or research results across datasets with differing controlled vocabularies can lead to missed or misinterpreted data.

Organizational challenges – Multiple-agency information and data sharing seldom occur, in part due to differing mandates, locations, funding, and personnel. Interagency efforts to develop common standards or protocols for collecting, archiving, and maintaining biodiversity and ecosystem data on common issues or questions are relatively uncommon. Often, data are collected and archived using the resources at hand in a given office or organization. For example, biologists may store collected data in Microsoft Excel as it is available to them and they know it, rather than in a more powerful environment, such as a database management system.

Information Intensity

The complexity of the Earth's biodiversity and its ecosystems is reflected in the massive amount of biological data that has been generated and archived for public use. These data represent the nation's biological legacy and include every level of detail from genes to ecosystems. New technologies in remote sensing further expand the "ecosystem catalogue" with terabytes of satellite imagery. With the advent of the Internet and World Wide Web, scientists, resource managers, teachers, and students, conservationists, lawmakers, and the general public have all begun to access and utilize this information at an increasing rate for an incalculable number of applications, 24 hours a day, 365 days a year.

Unlike with many other disciplines, biodiversity and ecosystem data are seldom outdated — information on species and their interactions with environmental factors collected in the eighteenth century continue to be relevant in the twenty-first century. The collection and archiving of new biological data—to add to existing knowledge—will continue at an increasing rate as more efficient data collection tools are developed and new demands for information are expressed.

New Instrumentation

New instrumentation is needed to improve biodiversity and ecosystem informatics from the data collection process to analysis and dissemination of the resulting information and knowledge. Highlighted throughout the workshop was the need for easy-to-use handheld field instruments with remote data transfer capabilities to enhance *in situ* discovery. Determining the best means of applying new space-based remote sensing instruments, such as laser- and radar-based sensors, hyperspectral sensors, etc., will be required. A need was also noted for tools to analyze and synthesize collected data and deliver the results of that data in presentable and understandable formats for decision-makers.

Finally, high priority was placed on developing improved scanning tools for metrically and chromatically accurate images of physical specimens. It was noted that millions of physical specimens are currently residing in an uncertain state at numerous labs and museums throughout the United States. These specimens, some of endangered or extinct species, are subject to natural degradation, and physical access to them by researchers is limited. Without digital archiving of these specimens the opportunity for advanced research and biological understanding will, at the least, be limited to the few who have access, and in the worst scenario, be lost. A sense of urgency was expressed for research into the development of these scanning tools.

Need to Identify and Record Species

Biologists have identified approximately 1.8 million living species of organisms, but vast numbers of remain to be discovered. The grand total for all life is currently estimated to be between 10 and 100 millions species. However, less than one-third of species that occur in the US have been discovered and the percentage is much lower for other parts of the world. We are in the midst of the sixth major extinction event of the planet's history, this one primarily the result of human modifications to the environment (PCAST, 1998). Given this context, it is disturbing to realize that species discovery is still largely a manual process requiring fieldwork and months or years of laboratory analysis and publishing activities. Current work practices — largely unchanged over the past two centuries — are unable to keep pace with rate of habitat destruction and species loss. If we ever hope to fathom the Earth's biodiversity, the biodiversity and ecosystem enterprise must reinvent itself — develop wholly new approaches to dealing with global-scale problems in a rapidly changing, information age.

Herein lie some of the most interesting informatics research challenges. These challenges cut across some of the most important research themes in computer science today, such as collaboration-in-the-large, instrumented (species) discovery, and computational exploration. The dominant mode of discovery for the biodiversity and ecosystem enterprise will increasingly become collaborative and model- and computation-driven. The research opportunities here are unlimited.

Multiple Data Formats

Biodiversity and ecosystem information exists in thousands of independent, individual and institutional databases, and in laboratory and personal field journals scattered throughout the country. The determination of which format any particular dataset is stored in is a function of several variables, including the availability and understanding of data collection and archiving software, the purpose for which the data are used, time constraints of the data collecting organization, and the amount of funding available for purchasing software and conducting staff training. The availability and use of multiple formats for storing biodiversity and ecosystem data will, in all likelihood, continue to occur. The CS/IT challenge is to ensure maximum compatibility and comparability of data across formats, while not constraining data collection, archiving, conversion, and distribution efforts.

Complexity of Biological Process

Knowledge about biodiversity and ecosystems is vast and complex. The complexity arises primarily from two sources. The first is the underlying biological complexity of the organisms themselves. There are millions of species, each of which is highly variable across individual organisms, populations, and time. Species have complex chemistries, physiologies, developmental cycles, and behaviors resulting from more than three billion years of evolution. There are hundreds, if not thousands, of ecosystems, each comprising complex interactions among large numbers of species and multiple abiotic factors. There is an immediate need to describe and model ecosystem processes and biological systems in a machine-processable form through computational modeling. The second source of complexity is sociologically generated. The sociological complexity includes problems of communication and coordination—among agencies, among divergent interests, and among groups of people from different regions and different backgrounds, such as academia, industry, and government—all having different views and requirements.

Evolving Nature of Biological Sciences

Early biologists were principally concerned with species identification, descriptions, and range. Today's biologists deal with the complexity of bio-life cycles. For example, the salmon case study presented above identified the broad spatial and temporal scale that fisheries biologists must deal with in developing management plans for the Columbia River Basin. Within any given spatial-temporal scale, factors influencing salmon life cycles may include numerous biological and non-biological events and processes for which the biologist must find information in order to help support those making management decisions. These events and processes may include climate, geology, fire, hydropower development, land use, vegetation, water quality, exotic species, harvests, and much more. Locating, retrieving, and analyzing such diverse data requires interdisciplinary collaboration among all data providers and data users. A single project might require participation of scientists representing several agencies or offices. Adding to the complexity of data needs, biologists and decision-makers often seek these data for multiple years to better understand and describe natural and anthropogenic trends.

Ecological Prediction

Biodiversity and ecosystem sciences require extensive knowledge and understanding of past events, interrelationships, and outcomes in order to gain new scientific understanding and predict future outcomes. The biologist is constantly challenged to predict results, risks, and timelines for proposed actions. As new discoveries and information accrue, adjustments in research or management strategy are required—in essence, an ongoing adaptive process. Computational models that analyze known information, incorporate new discoveries, and predict interactions and change under various conditions are needed.

These new models must take into account the complex and adaptive qualities of ecosystems as well as the spatial and temporal aspects of biological data. The models must include tools for automated data and information updating, building "what if" scenarios, developing visualization of predicted outcomes over time, and computing confidence or uncertainty characteristics.

Political Mandates

Public natural resource agencies must observe numerous mandates and directives that, among other things, dictate the types of data they collect and how they distribute those data. For example, in some cases, the Endangered Species Act influences what data are collected and how they are collected. The Freedom of Information Act outlines how data are to be distributed, including who has access to what data. Other directives influence monitoring requirements as well as quality assurance activities. Some directives require documents to be developed for, and open to, public comment. This requirement has a bearing on content, format, scale, scope, and other factors affecting the data management process.

Multiple Scales and Purposes

Multiple users access data for multiple purposes. The users of a particular dataset can conceivably include everyone from elementary school pupils to Nobel Laureates. Their use of the data can vary from showing the range of a particular species to investigating the effects an invasive species has on the forest canopy or the economy in a given region. This wide range of users of, and uses for, data presents a significant challenge for biodiversity and ecosystem informatics. How can a single dataset serve such a wide range of interests and needs? To avoid recollecting data for different users, a single dataset must meet the need of the most detailed users while not losing information or accuracy at smaller scales for the more general user.

Limits on Resources

As new demands are placed on public resource agencies to collect, archive, maintain, and provide biological data, additional burdens are placed on the agencies' physical, financial, and human resources. This situation can create a bottleneck in the flow of data and threatens the nation's capacity to record, maintain, and make accessible its biological records. Biodiversity and ecosystem informatics research can assist by developing new technologies and standard processes that meet resource agencies budgets of time, expertise, and funding.

BIODIVERSITY AND ECOSYSTEM INFORMATICS RESEARCH AGENDA

Based on the informatics challenges as identified through the case studies and discussions between biologists and computer scientists at the workshop, participants recommended 13 activities that should be included in a national BDEI research agenda. These activities can be grouped under three topics: acquisition and conversion of data and metadata, analysis and synthesis of data and metadata, and dissemination of data. Identification of specific CS/IT research opportunities follows each recommendation.

FOCUS AREA 1: Acquisition and Conversion of Data and Metadata

Modernizing the Biological Library – The accumulated volume of biological information and data collected over the past 250 years is massive. Unlike some disciplines, biological information is never out of date and will be required into the foreseeable future. Improving methods for organizing, storing and retrieving these records is extremely critical. New techniques and tools must be researched and developed for:

- digitizing the existing corpus of scholarly work related to biodiversity and ecosystems on a large scale,
- simultaneous acquisition of biological data from multiple sources including agencies, research stations, and individual scientists;
- correlating data from different formats including GIS, tabular, and text;
- storing massive amounts of complex biological data and information, including petabytes of high-resolution satellite imagery;
- handling variation in spatial and temporal scale across datasets;
- providing centralized and high-speed access to data and metadata;
- automating the capture of metadata from archived data and information;
- managing volume, quality, and versioning of data; and
- integrating data through essential attributes such as georeferences, URLs, and taxonomy.

CS/IT opportunities: Testing techniques for information extraction, text understanding, and cross-lingual information retrieval. Non-business application domain for data integration, data cleansing, data warehousing, and archiving research. Evaluation of temporal data models, cross-media linkage methods, and multi-scale representations.

Digitizing the Biological Legacy – America's museums and laboratories maintain at least 750 million biological specimens. There is an urgent need to convert them, their documentation, and new specimens into metric-quality digital formats. Many scientists would benefit from gaining network access to accurate digital representations of these specimens. Research needs to be conducted to:

- develop new holographic scanners of high-level accuracy,
- refine virtual reality and three-dimensional modeling, and
- improve the processes and techniques for labeling specimens with annotations, collection methods, taxonomic data, and other relevant metadata.

CS/IT opportunities: Test domain for image processing, video sensing, lossless image compression, and 3-D image understanding technology. A possible application for advanced robotic manipulation capabilities.

Multi-dimensional Observation and Recording – Efforts are needed to enable the collection of detailed information about the Earth's surface in multiple dimensions and at multiple scales. To accomplish this, workshop participants recommended research to:

- develop new remote sensing methods, particularly for species identification, habitat characterization, and biodiversity "hot spot" identification;
- enable the integration of data from disparate sensors; and
- store map information in four dimensions for complete spatial and temporal analysis.

CS/IT opportunities: Scaling sensor-fusion techniques to large fields. Testing temporal-spatial data access methods.

Mobile Computing – New instrumentation is needed to bring knowledge to the field and to collect, store, and transmit data from the field. Specific research activities include:

- developing techniques and tools for *in situ* access of remote data storage of species and ecosystem characteristics,
- methodologies for extracting subsets of data for regionalization of species and environmental attributes, and
- designing portable multi-media data recorders.

CS/IT opportunities: Application of human-computer interaction research in augmented reality, multi-model interfaces, hands-free systems, wearable computers, and remote presence. Possible area for using adaptive communication techniques for delivery of multi-media information and robotics and human augmentation.

Taxonomic Freedom – Changes in, and adaptations of, taxonomic names and classification schemes over time and discipline present challenges to biologists in recording and locating relevant data and information. The creation of new, and sometimes independent and concurrent, taxonomic classifications will continue. There is a need to integrate various interpretations, views, and versions of taxonomic data and make it available in a simple, easy-to-understand format. Tools are needed to:

- perform automated associations across different taxonomic schemes,
- display and browse multiple taxonomic classification systems simultaneously, and
- support the option of reorganizing taxonomic classifications to meet new findings and systematic knowledge.

CS/IT opportunities: Examining the flexibility and robustness for knowledge representation systems, particularly their temporal and versioning aspects and their support for cross-ontology linking and translation.

FOCUS AREA 2: Analysis and Synthesis of Data and Metadata

Comparing Across Scales – Biological data from different sources and times are frequently collected and presented in different scales and resolutions resulting in a loss of detail when multiple datasets are required for data synthesis and analysis. Tools and procedures to facilitate comparison and analysis across scales while retaining important details are needed. Recommended research activities include developing:

- methods of specifying what data to present at various levels of scale,
- techniques and policies for data "averaging" at various levels of scale, and

- processes for extrapolating between specific ecosystem attributes and whole ecosystems without compromising conclusions.

CS/IT opportunities: Testing adaptive- and multi-resolution techniques for computation and data modeling.

Modern Modeling – Researchers, managers, and policy-makers require models for biological decision-making rather than simply studying disaggregated collections of data or facts. Further, because of the complexity of biological processes and interactions, questions are frequently "fuzzy" in nature. Improved spatio-temporal modeling of biological resources, and biological and social processes, is required. Identified computer science research activities for modeling include investigating:

- high-performance modeling and simulation development,
- methods and techniques for incorporating data into models,
- how to define model parameters and develop automated parameter induction within models, and
- standards and methods for validating bioscience and social models.

CS/IT opportunities: A fertile area for testing developments in data assimilation research, including 4-D assimilation, continuous assimilation, variational assimilation, and multi-model assimilation.

Taxonomic Retooling – Taxonomists need new and improved tools for naming and defining new species, changing and manipulating taxonomic organization, and performing other tasks regarding taxonomic content and structure. Useful tools for taxonomists would facilitate:

- simultaneous display and browsing of taxonomic content of different schemes dating back several hundred years,
- editing taxonomic content and structure, and
- validating content and structure of existing classification schemes.

CS/IT opportunities: Test domain for research in knowledge acquisition and hierarchical display techniques.

Making Data Usable – Too frequently, decision-makers underutilize research results. To enhance the use of biological data, decision-makers require systems that will facilitate the synthesis and analysis of scientific data and research results. Such decision-support systems must include new techniques to:

- detect and represent biological and social changes over time and space, and
- identify gaps in, and duplication of, relevant data and information.

CS/IT opportunities: Application area for uncertainty analysis techniques, reasoning with incomplete information, and automatic summarization.

Machine Processable Metadata – Current metadata is largely for human consumption. It is used to document and help interpret datasets. It is sometimes collected and made available for searching in order to locate particular data. However, much more value will be gained from it when it is complete, correct, and descriptive enough to help automate data manipulation tasks, such as summarization, combination of datasets, and conversion of data to appropriate forms for use in models and statistical tools. Reaching this level of sophistication requires:

- understanding the requirements on metadata to allow machine interpretation of it,
- producing robust systems for the collection and checking of metadata, and
- developing metadata-driven tools for data manipulation.

CS/IT opportunity: Testing of data-based inferencing technology and metadata-based information integration research.

Need for Speed and Accuracy – Many tasks in data management are iterative and require considerable time. Researchers are frequently challenged with data entry and pattern discovery procedures and are required to estimate the quality of utilized data. Meanwhile species are disappearing at a rate greater than they can be recorded. Automating many of these tasks can assist bio-scientists in time and accuracy. Workshop participants recommended that CS/IT research be conducted to:

- facilitate the extrapolation of data sampled from an ecosystem to the totality of that ecosystem,
- develop techniques that indicate or rank quality and indicate the reliability of accessed biological data,
- make improvements in data mining processes, and
- develop processes for automating data and metadata entry.

CS/IT opportunity: Challenging domain for data reduction and data mining algorithms, including parallel implementations; modeling and analytic techniques with tunable accuracy; and data quality metrics.

FOCUS AREA 3: Dissemination of Data and Metadata

Visualization – Users of biodiversity and ecosystem data and information include individuals that require visualization of natural processes and management actions over time. Use of such visualization techniques can be for public discourse, adaptive management process (e.g. evaluating management options) or educational purposes, as well as scientific inquiry. In particular, land managers, policymakers, educators, non-governmental organizations, industry, and others outside biological research need visualization techniques to better understand the data and relationships among data. Research is required to:

- better characterize the needs and requirements of data users,
- gain insight into how to adapt systems to specific user needs and behaviors, and,
- find techniques to indicate gaps, inconsistencies, and uncertainties when viewing data.

CS/IT opportunity: Application of advanced display and visualization techniques, including display of uncertainty, user-adaptive display, and multi-dimensional data visualization.

Interdisciplinary Collaboration and Communication – Stakeholders of biodiversity and ecosystem data are growing in numbers and breadth. No longer are management decisions made solely by individuals or single agencies, but involve communities of individuals. Additional research needs to be conducted in methods to facilitate community involvement in biodiversity and ecosystem research and decision-making. These collaborative efforts must occur throughout a project's life cycle from its conception and, in fact, need to be second nature to the field of biodiversity and ecosystem informatics. Specific areas for research include:

- improving avenues of interdisciplinary communication between scientists, resource managers, decision-makers, and the public;
- investigating ways of sharing resources, including facilities, personnel, and funding, among all stakeholders;
- developing and testing innovative models for enhancing collaborative computer and biological science research; and
- developing techniques for bridging the use of different standards for temporal, semantic, and spatial references by different disciplines and communities.

CS/IT opportunity: Calls for the development of computer-supported cooperative work and remote collaboration research suited for participants with widely varying roles, specialties and training. Also test case for cross-domain mapping and integration of data, ontologies, and other knowledge representations.

Data Management Guidelines – Biodiversity and ecosystem information is frequently used in complex and potentially controversial political, economic, and environmental discussions and decision-making. Informatics issues arising from this context include issues of data security, data sharing policies, intellectual property rights, quality assurance, and reuse of data. Further development of technological tools to ease the burden of meeting the demands that these issues bring is needed. CS/IT research is needed in:

- the development of controlled read/write access programs;
- developing policy templates for IPR, copyright, access, and distribution, acknowledgements, and quality assurance and control;
- tracking use and property rights to support conformance with policies; and
- keeping the pedigree or provenance associated with data to better understand its sources and history. (For example, Is it primary or secondary data? Who collected it? What methodology was used? and What programs were used in subsequent processing?).

CS/IT opportunities: Application of data models for representing annotation and provenance, explicit modeling of data product generation. Possible area for investigation of data dissemination techniques and policies.

IMPLEMENTING THE RESEARCH AGENDA

Implementation of the biodiversity and ecosystem informatics research agenda described here will require a number of actions. Workshop participants recommended three areas where concerted effort should be made to proactively engage the broader CS/IT community in BDEI research: form implementation planning groups; utilize appropriate and available mechanisms for funding, partnering, collaborating, and resource sharing; and identify short-term “critical actions” that require immediate attention.

Interdisciplinary Planning Groups

Interdisciplinary planning groups, comprising members of the biodiversity and ecosystem and CS/IT research communities, can serve as a mechanism to articulate and communicate the special informatics challenges from one community to research actions in the other community. Such planning groups will be more effective if they are formed to address specific biodiversity and ecosystem “problem areas” (e.g. salmon conservation, invasive species, carbon cycle, etc.) rather than general research issues. The planning groups could identify the following:

- existing CS/IT technologies that could be transferred from other domains,
- the long-term basic CS/IT research questions,
- the short-term CS/IT research that needs to be applied, and
- the infrastructure needs including, equipment, facilities, networks, and personnel.

Matching Research Needs with Available and Appropriate Mechanisms

Efforts to implement any research agenda item need preliminary study to determine how these research actions could benefit from existing programs. These programs could include mechanisms for funding, partnerships, interdisciplinary research, and sharing resources with compatible programs.

Funding options – Funding sources for BDEI research activities are numerous. NSF's Information Technology Research (ITR), Biocomplexity, and Digital Government programs provide potential venues as does NASA's High Performance Computing and Communications (HPCC), Global Change, Carbon, and Terrestrial Ecology programs. In addition, the Department of Defense provides funding opportunities that may be appropriate, and new activities, such as the proposed National Ecological Observation Network (NEON) (see Appendix A-III) and the Global Biodiversity Information Facility (GBIF) (see Appendix A-II), may create additional opportunities in the future.

Partnerships – The tasks required to accomplish many research agenda items is monumental. The development of partnerships and the use of existing partnerships such as the National Partnership for Advanced Computational Infrastructure (NPACI) are highly encouraged.

Interdisciplinary training and teaming – Biological scientists need tools and computer scientists seek challenging problems that will result in applied technologies. Too frequently, meaningful interchange between the biological and computer science disciplines does not occur. Efforts must be made to link these two disciplines from the project planning process through testing and evaluation of resulting technologies. Cross-training programs through academic institutions, workshops, and other avenues need to be designed, funded, and implemented. Cross-teaming of biological and computer scientists can be encouraged through innovative funding mechanisms that support an initial "spin-up" phase to develop interdisciplinary teams, basic infrastructure, and detailed research plans before launching into the main research activities. NSF's Digital Government program is responding to this need and can serve as example for facilitating interdisciplinary processes in research and application of information technologies.

Sharing resources – Many institutions have developed technologies and infrastructure that could be leveraged to strengthen the implementation of selected activities within the proposed research agenda. Examples of such institutions include the National Biological Information Infrastructure (NBII), Global Biodiversity Information Facility (GBIF), Long-Term Ecological Research (LTER), and National Ecological Observation Network (NEON).

Short-term Critical Actions that Require Immediate Attention

Workshop participants noted a sense of urgency to begin implementation of the new BDEI research agenda. Several activities were noted that could be acted upon immediately to "jump start" BDEI research: developing an interagency strategic partnership to begin preliminary and requisite work on the research agenda, recognizing the time critical nature of specific biodiversity and ecosystem problems, and organizing problem-specific consultant teams to develop appropriate implementation plans.

Interagency strategic partnership – Workshop participants endorsed a strategic partnership between NSF, NASA, and USGS to advance cutting-edge computer and information science research and research in biodiversity and ecosystem sciences. By pooling existing resources, this interagency BDEI R&D effort would immediately launch interdisciplinary "seed" projects and prepare for the start of a major joint program in FY2003 that would address long-term research agenda recommendations. These partner agencies would also jointly support low-cost "community building" activities for the community. Such activities could include using Web-based information services to announce the availability of BDEI "challenge problems" and data testbeds for CS/IT research activities.

Recognizing the urgency and importance of biodiversity and ecosystem problems – The time-critical nature of many biodiversity and ecosystem science and conservation issues could be more effectively communicated to prospective CS/IT researchers. The urgency of these problems is an important factor in the overall challenge and excitement of working in this domain, and can provide researchers with a sense of accomplishment, knowing that they are making important contributions to problems of global concern. It was recommended that every effort in this direction be made.

Organizing problem-specific consultant teams – The use of specialized, issue-specific, teams of specialists was endorsed to develop appropriate implementation plans. It was envisioned that these plans would be written for a specific biodiversity and ecosystem problem area, e.g. invasive species, and included:

- long-term vision of biodiversity and ecosystem applications and information management,
- basic research (e.g., for predictive models) needed from the computer science community,
- applied research (e.g., for data acquisition) required, and
- opportunities for meaningful interdisciplinary CS/IT and biology collaboration.

COMMUNICATING THE RESEARCH AGENDA

Workshop participants were asked to develop recommendations for how best to communicate the BDEI research agenda to an audience that includes researchers in computer science and the biodiversity and ecosystem sciences, as well as the many agencies and foundations that support their efforts.

Develop extended workshops or seminars – The duration of these events would be, at a minimum, one month and would focus on specific biodiversity and ecosystem informatics problems using case studies as testbeds. Biologists and CS/IT scientists would work together to learn and share, with a goal of developing a product that is applied and tested in the field. In the very near term, a request for proposals to conduct the workshops needs to be developed and advertised.

Build a multi-sector, multi-disciplinary community – Considerable discussion during the workshop centered on developing mechanisms for linking CS/IT researchers with biodiversity and ecosystem researchers. Some progress, such as NSF's Digital Government Program, has been made in this direction and may be used as models. Collaboration between the two communities would achieve a greater degree of success if the projects were conducted on specific testbeds. (See examples of Bio-CS/IT collaborations in the Appendix).

Develop "matchmaking" mechanisms – CS/IT researchers attending the workshop noted their lack of awareness of opportunities to collaborate with biodiversity and ecosystem scientists on challenging informatics questions and problems. Biodiversity and ecosystem scientists were likewise unaware of the extent to which the CS/IT community might benefit from collaborations on the many unique and challenging research questions in this field. Professional societies could be used to launch "matchmaking" activities such as posting contact lists and publishing white papers. A specific suggestion along these lines was to establish a corpus of BEDI "challenge problems" contributed by biodiversity and ecosystem scientists. Such a challenge problem would describe the desired capability, the limitations of current solutions, and the characteristics of a good solution (e.g. efficiency, scale, precision). Each problem would be accompanied by one or more datasets on which to test new approaches. The ready availability of test data, along with the presumed importance of the problems, could entice computer researchers to try out their new ideas and latest developments in the biodiversity and ecosystem domain.

Add CS/IT component to existing biodiversity and ecosystem projects – Biodiversity and ecosystem projects are staffed by scientists trained (and interested) in the biological sciences. The new emphasis on informatics within the biological sciences presents special challenges that are best addressed by members of the CS/IT community. Incorporating or adding a CS/IT research emphasis into existing biodiversity and ecosystem initiatives will provide cost- and time-effective opportunities for CS/IT researchers and bolster the informatics aspects of existing biodiversity and ecosystem research projects.

Develop venues for multi-disciplinary activities – Currently, computer science and the ecological sciences are "vertically integrated" and seldom present opportunities for members across disciplines to meet or work together for a common objective. Workshop participants recommended that an effort to establish university departments or research centers and inter-disciplinary curricula in biodiversity and ecosystem informatics be sponsored and developed. Participants also felt that publishing a new journal (or special issues of existing journals) on biodiversity and ecosystem informatics would provide a high-profile forum for the exchange of challenges and ideas across disciplines.

Promoting biodiversity and ecosystem informatics through dissemination of reports – Much still needs to be accomplished with regards to articulating and promoting the biodiversity and ecosystem informatics mission. Continuing workshops and disseminating the findings and recommendations in this report are critical to broadcasting the biodiversity and ecosystem informatics research agenda. These reports should be made available to both biodiversity and ecosystem and CS/IT researchers and should provide background and the current status of both communities. Reports should also include visionary case studies that motivate interdisciplinary research initiatives in biodiversity and ecosystem informatics. It is also important to provide "experience papers" describing successful (and maybe unsuccessful) collaborations, past and present.

Promoting the biodiversity and ecosystem informatics agenda through publications, email distribution lists, and websites – A wealth of publications within both communities is available to engage scientists in the biodiversity and ecosystem informatics research agenda. Examples of publications include SIGMod and the Bulletin of the Ecological Society of America. Professional societies can assist in this effort by offering links on their websites and emails to their members.

CONCLUSION

It's 2010, and Karen Culver is once again evaluating the proposed closure of the Silver Creek diversion channel. Back in 2001, the channel wasn't closed. While she felt closing the diversion channel would help the bull trout population, in the limited time available, she wasn't able to get all the information she needed to make an effective presentation to the local council.

The channel is now in need of repair, and closure is being considered again. Karen is at the site where the channel leaves the stream, to get a feel for the situation. She dons a pair of visualization goggles that interface with her portable computer. Using voice commands, Karen can overlay her view of the terrain with different maps and datasets. She quickly superimposes land ownership, topographic lines, and locations of previous biological studies. She is also able to view the creek in false color, to see seasonal temperature variations, and flow rates. She focuses her gaze on the channel and brings up counts of species that have been surveyed there. She notes there has been an observation of a species of tiger salamander that is listed as threatened.

She switches to the screen of her portable to look at the area in plan view. She examines some aerial imagery of the drainage and adds a map showing the location of the farms that draw water from the irrigation pond that the channel supplies, plus another map showing land ownership and use in the area. Using this information she starts sketching a route to nearby Crabb Lake that could supply replacement water.

Karen now turns to the effects of channel closure. She gets her co-worker, Tom Hamilton, online, and he helps her select a model to use for predictions. He shows her how to work a wizard that can help select and convert appropriate datasets for use with the model. Within about fifteen minutes Tom and Karen have located suitable topographic and meteorological data, and the wizard has suggested two possible hydrologic datasets. Tom recommends using the second one, as it has more complete historical coverage. The model is then dispatched to run remotely on a compute server, to work through the range of expected stream temperatures and flows if the channel were closed. Although Karen isn't explicitly aware of it, the computation is actually split into three parts that take place on three different high-performance cluster computers.

Karen is also wondering about sedimentation of downstream gravel beds where bull trout currently lay eggs. She does a similarity search for documents about other stream modifications in areas with comparable soil types and hydrology. She finds six and examines them to find which most closely match the current situation.

The model calculations on predicted temperatures and flows after closing the channel are done and have been transferred to Karen's portable, as well as sent to Tom back at the main office. She gets him back online to help interpret them in terms of effect on fish. He helps her construct a plot comparing the periods each year when water temperature or oxygen levels are likely to adversely affect the fish. They compare that plot to one based on records from a recent year. They see that the closure would likely yield a great improvement, with periods of adverse conditions being both less frequent and of shorter duration. With a little help from Tom, she launches a task to render an animation of water conditions before and after closure, with a color spectrum representing favorable to adverse conditions. That task is routed to a remote server; all Karen cares about is that an MPEG-9 file for the animation is downloaded onto her portable when she gives her presentation to the watershed council in the afternoon.

The one nagging issue in Karen's mind still is the tiger salamanders in the creek. She'd really like to know if that species of salamander was present before the channel was dug (and thus can be expected to survive if the creek returns to a similar state). Unfortunately, amphibian survey data on Silver Creek only go back about 15 years. Karen has an idea, however. She dispatches a query through the National Biological Information Infrastructure to search holdings of natural history collections throughout the country. In about four minutes, she gets back two records of tiger salamanders collected at Silver Creek, in 1914 and 1933. She is quite impressed by the results, as the query system knew that Silver Creek was called Sinners Creek before 1920, and that the scientific name of that particular species had been modified in the 1950s. She is able to view the digitized label information for the 1933 specimen, which contains an annotation that tiger salamanders were abundant at several places in the stream, including one site near the channel junction. She is reassured that there likely will be suitable habitat for the salamanders if the channel is closed, though there will still need to be some further study.

Karen sets off to her afternoon meeting with the council feeling much more confident about the presentation she's going to make than she did ten years earlier. She did in three hours what she was unable to do in three days in 2001.

REFERENCES

Invasive species case study powerpoint presentation (6.2 mb).
http://www.nrel.colostate.edu/projects/stohlgren/images/Inventory-monitoring.ppt

President's Committee of Advisers on Science and Technology (PCAST). March 1998. *Teaming with Life: Investing in Science to Understand and Use America's Living Capital*.
http://www.whitehouse.gov/WH/EOP/OSTP/Environment/html/teamingcover.html

RELATED READING

Bowker, Geoffrey. *Work and Information Practices in the Sciences of Biodiversity*. Presented at "Marking the Millennium" – 26th International Conference on Very Large Databases Cairo, Egypt, 10-14 September 2000.
http://www2.aucegypt.edu/vldb2000/bowker.pdf

Nielsen, Ebbe, James Edwards and Meredith Lane. *Biodiversity Informatics: The Challenge of Rapid Development, Large Databases, and complex Data*. Keynote presentation at "Marking the Millennium" – 26th International Conference on Very Large Databases Cairo, Egypt, 10-14 September 2000.
http://www2.aucegypt.edu/vldb2000/NeilsenE.pdf

Schnase, John. *Research Directions in Biodiversity Informatics*. Presented at "Marking the Millennium" – 26th International Conference on Very Large Databases Cairo, Egypt, 10-14 September 2000.
http://www2.aucegypt.edu/vldb2000/schnase.pdf

Website for "Marking the Millennium" – 26th International Conference on Very Large Databases Cairo, Egypt, 10-14 September 2000.
http://www2.aucegypt.edu/vldb2000/

APPENDICES

A-I. National Biological Information Infrastructure (NBII)
http://www.nbii.gov/

The National Biological Information Infrastructure (NBII) was established in 1993 to help create a national partnership for sharing information on the nation's biodiversity and ecosystems. The philosophy of the NBII is to build a distributed electronic federation of biological data and information sources and, as part of this effort, provide an operational infrastructure that includes policies, protocols, and standards for participation. These "guidelines" support discovery, retrieval, integration, and application of biological data across NBII's distributed network. The NBII is a broad collaborative activity involving many agencies and organizations. The US Geological Survey provides overall coordination and leadership. The range of NBII's content and the development of operational guidelines are two areas of focus.

Natural history collections and museums represent one example of a very important and rich biological data "source" for the NBII. These institutions produce a tremendous amount of biodiversity and ecosystem information (much of which is not even digital). NBII staff is working directly with museums, as well as strategically on a national basis, to facilitate future access to these important biological collections.

The goal for NBII is to have diverse biodiversity and ecosystem content from many types of sources linked together and interoperable so users can locate all relevant information for a specific question in a single search. As an example, a single query may result in retrieving museum specimen data, satellite imagery data, an ecological model, and a technical report. NBII partners and collaborators are working together to develop the "underpinnings" or infrastructure of the NBII federation, including:

- providing a standardized way for scientists to describe and document their biological data and information;
- providing an online, consistent reference for biological nomenclature; and
- providing a comprehensive biological sciences thesaurus or controlled vocabulary.

The first element is being accomplished through the adoption of accepted dataset-level metadata standards (i.e., the biological data profile of the FGDC metadata content standard). All metadata produced following this standard is made accessible for online searching through NBII's metadata clearinghouse.

While spatial coordinates locate objects in the spatial data world, species names locate objects in the biological data world. Providing access to a scientifically credible, consistent biological species nomenclature reference system is critical to success of the NBII infrastructure.

The third important infrastructure component is the development of a biological controlled vocabulary or thesaurus that would be available as a consistent reference for use by content providers in documenting contributions and by NBII's customers in locating relevant content. This will be accomplished by building on existing vocabularies, "knitting" these vocabularies together, and making the resulting product available for use online.

A-II. Global Biodiversity Information Facility (GBIF)

http://www.gbif.org/

Working in close cooperation with established programs and organizations that compile, maintain, and use biological information resources, the Global Biodiversity Information Facility (GBIF) will be an interoperable network of biodiversity and ecosystem databases and information technology tools. GBIF will enable users to navigate and put to use the world's vast quantities of biodiversity and ecosystem information to produce national economic, environmental, and social benefits.

The central purpose of establishing GBIF is to design, implement, co-ordinate, and promote the compilation, linking, standardization, digitization, and global dissemination of the world's biodiversity and ecosystem data, within an appropriate framework for property rights and due attribution. It will have the characteristics of a large, distributed public domain databases with a number of interlinked and interoperable modules (databases, software and networking tools, search engines, analytical algorithms, etc.). GBIF will:

- Be a distributed facility, while encouraging co-operation and coherence;
- Be global in scale, though implemented nationally and regionally;
- Be open to participation by individuals from all countries, and offering potential benefits to all countries, while being funded primarily by those countries that have the greatest financial capabilities;
- Help bridge human language barriers by promoting standards and software tools designed to facilitate their adaptation into multiple languages, character sets and computer encodings;
- Serve to disseminate technological capacity by drawing on and making widely available scientific and technical information; and
- While aiming to make biodiversity and ecosystem information universally available, facilitate respect for the contribution made by those gathering and furnishing this information.

Operationally, GBIF will be established as a freestanding international organization. It will be supported by a small secretariat that will work internationally to co-ordinate national and regional efforts and bring focus to the organization and its activities. In addition, it will manage a small amount of seed money to be used to support activities being conducted by other agencies.

A-III. National Ecological Observatory Network (NEON)
http://www.archbold-station.org/abs/neon/index.html

The National Science Foundation's proposed National Ecological Observatory Network (NEON) will establish 10 observatories located around the country that will serve as national research platforms for integrated studies in field biology. Each observatory will provide state-of-the-art infrastructure to support interdisciplinary, integrated research. Collectively, the network of 10 observatories will allow scientists to conduct comprehensive, continental-scale experiments on ecological systems. Each NEON observatory will include site-based experimental infrastructure; natural history archive facilities; and facilities for biological, physical and data analyses. In addition, the 10 NEON observatories will be linked via a cutting-edge communications network.

The objectives of NEON are:

- To provide a state-of-the-art national facility for field biologists to conduct cutting edge research spanning all levels of biological organization from molecular genetics to whole ecosystem studies and across scales ranging from seconds to geological time and from microns to kilometers;
- To interconnect the geographically distributed parts of the facility into one virtual installation via communication networks so that members of the field biology research community can access the facility remotely; and
- To facilitate predictive modeling of biological systems via data sharing and synthesis efforts by users of the facility.

The diversity of ecosystems that comprise our national landscape, from forests to grasslands and deserts to tundra, precludes using only a single observatory for biocomplexity research. NEON will constitute a distributed and virtual national laboratory for field biology and will foster integrated, interdisciplinary research through long-term collaborations and data sharing. NEON is needed to understand how our nation's ecosystems function and to predict their response to natural and anthropogenic events. NEON will also systematically and directly support application of new technologies (e.g., functional genomics, molecule-specific stable isotopes, etc.) to advance ecological research.

Each NEON observatory will contain a suite of common instruments for continental-scale measurement and analysis. In addition, each observatory will have unique infrastructure to address site-specific research questions. Intensive studies at each observatory will be facilitated by standardized equipment for integrated field research (e.g., high resolution global positioning grid arrays, mesonet scale meteorological equipment, eddy flux correlation towers, hydrological facilities, etc.) and laboratory analyses (e.g. confocal microscopes, DNA sequencers, stable isotope mass spectrophotometers, CHN Analyzers, ultracold tissue archives, and digital museum technology). The observatories will have scalable computation capabilities and will be networked via satellite and landlines to the vBNS, to each other, and to specialized facilities, such as supercomputer centers.

NEON will provide a superb platform for educational uses and outreach. K-12 students, undergraduate students, and the general public heavily use biological field stations, potential affiliates of NEON observatories. NEON communications and research facilities will introduce students at all levels to cutting-edge ecological research. Many experimental research sites managed by potential NEON members are located close to community colleges, land-grant colleges, and HBCUs. In addition to its value for scientific and education purposes, NEON activities will develop a wide range of data that will be of value to a broad array of users. The general public will be able to access NEON databases, as will decision-makers.

A-IV. List of Workshop Participants and Observers

PEGGY AGOURIS
University of Maine
5711 Boardmann Hall, Room 342
Orono, ME 04469-5711

Phone: 207-581-2180
Fax: 207-581-2206
Email: peggy@spatial.maine.edu

GEOF BOWKER
Department of Communication
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0503

Phone: 858-534-7192
Fax: 858-534-7315
Email: bowker@ucsd.edu

LARRY BRANDT
National Science Foundation
4201 Wilson Blvd, Room 1160
Arlington, VA 22230

Phone: 703-306-1981
Fax: 703-706-0589
Email: lbrandt@nsf.gov

MICHAEL L. BRODIE
GTE Laboratories
40 Sylvan Road
Waltham, MA 02451-1128

Phone: 781-466-2256
Fax: 781-466-2439
Email: brodie@gte.com

ROZ COHEN
NOAA National Oceanographic Data Center
SSMC3, #4822
1315 East-West Highway
Silver Spring, MD 20910-3282

Phone: 301-713-3267 x146
Fax: 301-713-3300
Email: Rosalind.E.Cohen@noaa.gov

GLADYS COTTER
US Geological Survey
MS 300
12201 Sunrise Valley Drive
Reston, VA 20192

Phone: 703-648-4182
Fax: 703-648-4042
Email: gladys_cotter@usgs.gov

JUDY CUSHING
The Evergreen State College
2700 Evergreen Parkway NW
Olympia, WA 98505

Phone: 360-866-6000
Email: judyc@evergreen.edu

JIM EDWARDS
National Science Foundation
4201 Wilson Blvd, Room 605
Arlington, VA 22230

Phone: 703-306-1400
Fax: 703-306-0343
Email: jledward@nsf.gov

MIKE FRAME
US Geological Survey
MS 300
12201 Sunrise Valley Drive
Reston, VA 20192

Phone: 703-648-4164
Fax: 703-648-4224
Email: mike_frame@usgs.gov

MICHAEL FREESTON
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106

Phone: 805-893-8589
Email: freeston@Alexandria.ucsb.edu

ANNE FRONDORF
US Geological Survey
MS 300
12201 Sunrise Valley Drive
Reston, Virginia 20192

Phone: 703-648-4205
Fax: 703-648-4224
Email: anne_frondorf@usgs.gov

VALERIE GREGG
National Science Foundation
4201 Wilson Blvd.,
Arlington, VA 22230

Phone: 703-306-1981
Fax: 703-306-0610
Email: vgregg@nsf.gov

MILT HALEM
Chief Information Officer
NASA Goddard Space Flight Center
Greenbelt, MD 20771

Phone: 301-286-5177
Fax: 301-286-1777
Email: halem@gsfc.nasa.gov

CHRISTINA HARGIS
US Forest Service
Forestry Sciences Lab
2500 S. Pine Knoll
Flagstaff, AZ 86001

Phone: 520-556-2182
Fax: 520-556-2130
Email: chargis@fs.fed.us

JOHN HELLY
San Diego Supercomputer Center
9500 Gilman Drive
La Jolla, CA 92024-0527

Phone: 858-534-5060
Fax: 858-822-3631
Email: hellyj@ucsd.edu

EDUARD HOVY
Information Science Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695

Phone: 310-448-8731
Fax: 310-823-6714
Email: hovy@isi.edu

ERIC LANDIS
Oregon Graduate Institute
13875 Tangen Road
Newberg, OR 97132

Phone: 503-538-6683
Email: elandis@ix.netcom.com

MEREDITH LANE
The Academy of Natural Sciences
1900 Benjamin Franklin Parkway
Philadelphia, PA 19103

Phone: 215-405-5060
Fax: 215-299-1028
Email: lane@acnatsci.org

MIRON LIVNY
University of Wisconsin – Madison
1210 West Dayton Street
Madison, WI 53706-1685

Phone: 608-262-0856
Fax: 608-262-9777
Email: miron@cs.wisc.edu

DAVID MAIER
Oregon Graduate Institute
20000 NW Walker Road
Beaverton, OR 97006

Phone: 503-748-1154
Email: maier@cse.ogi.edu

REAGAN MOORE
San Diego Supercomputer Center
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0505

Phone: 858-534-5073
Fax: 858-534-5152
Email: moore@sdsc.edu

JOHN PORTER
VCR Long-Term Environmental Research
Department of Environmental Sciences
University of Virginia
PO Box 400123
291 McCormick Road
Charlottesville, VA 22904-4123

Phone: 804-924-8999
Fax: 804-982-2137
Email: jhp7e@virginia.edu

JIM QUINN
Dept. of Environmental Science and Policy
University of California, Davis
Davis, CA 95616-8576

Phone: 530-752-8027
FAX: 530-752-3350
Email: jfquinn@ucdavis.edu

JOHN L. SCHNASE
Earth and Space Data Computing Division
NASA Goddard Space Flight Center
Greenbelt, MD 20771

Phone: 301-286-2451
Fax: 301-286-1777
Email: schnase@gsfc.nasa.gov

BERNARD SHANKS
US Geological Survey
MS 300
12201 Sunrise Valley Drive
Reston, VA 20192

Phone: 703-648-4084
Fax: 703-648-4039
Email: bernard_shanks@usgs.gov

AVI SILBERSCHATZ
Lucent Technologies Bell Labs
600 Mountain Avenue
Murray Hill, NJ 07974-0636

Phone: 908-582-4623
Fax: 908-582-6923
Email: avi@bell-labs.com

JIM SMITH
Laboratory for Terrestrial Physics
NASA Goddard Space Flight Center
Bldg. 33, Rm. G125D, Code 920
Greenbelt, MD 20771

Phone: 301-614-6020
Fax: 301-614-6015
Email: jasmith@hemlock.gsfc.nasa.gov

ANTHONY STEFANIDIS
University of Maine
5711 Boardman Hall, Room 348
Orono, ME 04469-5711

Phone: 207-581-2127
Fax: 207-581-2206
Email: tony@spatial.maine.edu

BRUCE STEIN
Association for Biodiversity Information
4245 North Fairfax Drive
Suite 100
Arlington, VA 22203-1606

Phone: 703-841-2711
Fax: 703-525-8024
Email: bstein@tnc.org

TOM STOHLGREN
Midcontinent Ecological Science Center
US Geological Survey
Natural Resource Ecology Laboratory
Colorado State University
Fort Collins, CO 80523

Phone: 970-491-1980
Fax: 970-491-1965
Email: tom_stohlgren@usgs.gov

WOODY TURNER
Mail Code YS
NASA Headquarters
Washington, DC 20546-0001

Phone: 202-358-1662
Fax: 202-358-2771
Email: wturner@hq.nasa.gov

MATT WARD
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609

Phone: 508-831-5671
Fax: 508-831-5776
Email: matt@cs.wpi.edu

MARIA ZEMANKOVA
National Science Foundation
4201 Wilson Blvd, Room 1115
Arlington, VA 22230

Phone: 703-306-1926
Fax: 703-306-0599
Email: mzemanko@nsf.gov