

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 Flight
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 solutions 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 requirements 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://bio.gsfc.nasa.gov> 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 systems 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 methods 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 important questions 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 nearly one billion 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 present 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 knowl-

edge 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 scientific metadata is largely for human consumption. It is used to document and 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 time consuming. 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 articulate and communicate 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. There also is a need to insure that new CS/IT research is effectively applied to real biodiversity and ecosystem test cases.
- 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 immediately initiated 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 forum 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 emphasize 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.