Building the Next Generation Biological Information Infrastructure

John L. Schnase
The Missouri Botanical Garden
Center for Botanical Informatics, LLC

Meredith A. Lane
Division of Botany, Natural History Museum
Department of Botany, Division of Biological Sciences
University of Kansas

Geoffrey C. Bowker
Graduate School of Library and Information Sciences
University of Illinois at Urbana-Champaign

Susan Leigh Star
Graduate School of Library and Information Sciences
University of Illinois at Urbana-Champaign

Abraham Silberschatz
Information Sciences Research Center
Bell Laboratories, Lucent Technology

To appear in the proceedings of the National Academy of Sciences
Forum on Nature and Human Society: The Quest for a Sustainable World,
October 27 - 30, 1997, Washington, DC.

The grand challenge for the 21st century is to harness the accumulating knowledge of Earth's biodiversity and the ecosystems that support it. To accomplish this, we must mobilize biological information — assemble it, organize it, and deliver it with dramatically increased capacity. We must elevate the global biological information infrastructure to a new level of capability — a "next generation" — that will allow people to share on a world-wide basis the knowledge created by biodiversity and ecosystems research.

Realizing the urgency of this task, the President's Committee of Advisors on Science and Technology, through its Panel on Biodiversity and Ecosystems, recently coordinated a review of the United State's National Biological Information Infrastructure.[1] Over a six-month period in 1997, people from a broad cross section of the public and private sectors contributed their insights, experiences, concerns, and hopes. What emerged was a renewed understand of the importance of biological information to all aspects of human society. It also became clear that much remains to be done to assure that this information is complete and usable. While the purpose of the review was to develop recommendations to build capacity in the United States, many of the Panel's findings address global concerns of relevance to biodiversity research wherever it occurs. In this paper, we provide a summary of the Panel's report, a view of what a "next generation" biological information infrastructure might encompass, and suggestions about how it might be achieved.

Background

In the United States, the National Biological Information Infrastructure (NBII) is the primary mechanism whereby biodiversity and ecosystems information is made available to all sectors of society. It is the biological component of the National Information Infrastructure and, as such, is the framework that connects US activities to the global biodiversity and ecosystems research enterprise. Its meaning is expansive and intended to convey the idea that an information infrastructure is comprised of more than just computers, networks, and the like, but also the information, policies, standards, and people who use it. Initiation of the NBII was one of the primary recommendations made by the National Academy of Sciences National Research Council in their 1993 report, "A Biological Survey for the Nation."[2]

Since our fate and economic prosperity are so completely linked to the natural world, information about biodiversity and ecosystems — as well as the infrastructure that surrounds it — is vital to a wide range of scientific, educational, commercial, and governmental uses. Unfortunately, most of this information now exists in forms that are not easily accessed or used. From traditional, paper-based libraries to scattered databases and physical specimens preserved in natural history collections throughout the world, our record of biodiversity and ecosystem resources is uncoordinated, and large parts of it are isolated from general usage. It is not being used effectively by scientists, resource mangers, policy makers, or other potential client communities.[3,4]

Fortunately, research activities are being conducted around the world that, if leveraged, could improve our ability to manage biological information. In the United States, the Human Genome Project is producing new medical therapies as well as developments in computer and information science. Geographic Information Systems (GIS) are expanding the ability of federal agencies to conduct data-gathering and synthesis activities more responsibly while creating opportunities for commercial partnerships that can lead to new software tools. The National Spatial Data Infrastructure is improving the management of geographic, geological, and satellite datasets; the Digital Libraries projects are beginning to produce useful results for some information domains; and the High-Performance Computing and Communications initiative has enhanced certain computation-intensive engineering and science areas.

Unfortunately, little attention has been paid to computer and information science and technology research in the biodiversity and ecosystems domain. We must produce mechanisms that can efficiently search through terabytes of Mission to Planet Earth satellite data and other biodiversity and ecosystems datasets, make correlations among data from disparate sources, compile those data in new ways, analyze and synthesize them, and present the results in an understandable and usable manner. Despite encouraging advances in computation and communications performance in recent years, we are able to perform these activities on only a very small scale. We can, however, make rapid progress in these areas if the computer and information science and technology research community becomes focused on the needs of the biodiversity and ecosystems research community.[5]

MANAGING COMPLEXITY

Knowledge about biodiversity and ecosystems is a vast and complex information domain. The complexity arises from two sources. The first of these 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. These 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 in biodiversity and ecosystems information is sociologically generated. The sociological complexity includes problems of communication and coordination — among agencies, among divergent interests, and across groups of people from different regions, from different backgrounds (academia, industry, government), and with different views and requirements. The kinds of data humans have collected about organisms and their relationships vary in precision, accuracy, and in numerous other ways. Biodiversity data types include text and numerical measurements as well as images, sound, and video. The range of other databases with which biodiversity datasets must interact is also broad, including geographical, meteorological, geological, chemical, and physical databases. The mechanisms used to collect and store biological data are almost as varied as the natural world they document. Additionally, biological data may be politically and commercially sensitive, and entail conflicts of interest. User skill levels are highly variable, and training in this area is not yet well developed.

Because of these complexities, humans still play a crucial role in the processing of biological data. It is not as amenable to automatic correlation, analysis, synthesis, and presentation as many other types of information, such as in the field of radioastronomy where there is more coherent global organization, and the problems being studied are frequently conducive to automatic analysis. In biodiversity research, 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 extracting higher-order understanding from dispersed datasets is a fundamental intellectual process, yet 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 of this scale are commonplace in biodiversity and ecosystems research.[6]

In order for a biological information infrastructure to be effective, it must provide the means to manage complexity. It must allow scientists to extract new knowledge from the aggregate mass of information generated by the data gathering and synthesis activities of other scientists. It must use the power of computers to facilitate the queries, correlations, and processing activities that are impossible for humans to perform alone. And it must deliver this functionality within a physically and intellectually accessible framework. This means developing ways of delivering the information to a wide range of users, with differing skills, ages, and investment in the material.

We are only beginning to develop a vocabulary to describe these large-scale, synthetic, information-processing activities. Some sociologists use the term "distributed cognitive system" to emphasize the role of humans within a synergistic, information-processing network.[7] "Data mining" is a term that is often used by the database community. Whatever the name, these activities form only a part of a larger process of knowledge discovery that includes the large-scale, interactive storage of information (known by the unintentionally uninspiring term "data warehousing"), cataloging, cleaning, preprocessing, transformation, verification, and reduction of data, as well as the generation and use of models, evaluation and interpretation, interpersonal communications, the evolution of sophisticated user interfaces, and finally consolidation and use of the newly extracted knowledge. These processes will become increasingly important if we are to use what we know and expand our knowledge in useful directions.

At present, the NBII provides little support for these activities. At best, it can be used to access information in databases held by federal agencies and other institutions around the country. Once accessed, however, the task of organizing, integrating, and interpreting the information remains, for the most part, a laborious, manual process. The development of computational tools for the biodiversity and ecosystems enterprise lags behind other sciences. Important classes of information are missing (fewer than 1% of the specimens in our natural history collections have been databased!), and existing databases are uneven in the types of information that they hold. It is difficult for individual scientists to publish their data electronically in meaningful ways. Standards for information exchange have not been widely adopted. We have no mechanism for archiving data over generations of use and generations of technologies. And the power of communication networks to build communities remains largely untapped. In summary, the NBII is currently neither a system nor an infrastructure: it is a cumbersome and brittle patchwork — presenting as many obstacles to scientific work as it does opportunities. It is clearly time to transform it into a coherent and empowering capability.

THE NEXT GENERATION

We envision a "next generation" National Biological Information Infrastructure, NBII-2, that would address many of the concerns described above. The overarching goal of NBII-2 would be to become a fully accessible, distributed, interactive digital library. NBII-2 would provide an organizing framework from which scientists could extract useful information — new knowledge — from the aggregate mass of information generated by various data gathering activities. This would be accomplished by using the power of computers and communications networks to augment the processing activities that now require a human mind. It would make analysis and synthesis of vast amounts of data from multiple datasets easier and more accessible to a variety of users. It would also serve management and policy decision-making, education, recreation, and the needs of industry by presenting data to each user in a manner tailored to that user's needs and skill level.

We envision NBII-2 as a distributed facility that would be considerably different than a "data center," considerably more functional than a traditional library, considerably more encompassing than a typical research institute. Unlike a data center, NBII-2's objective would be the automatic discovery, indexing, and linking of datasets rather than the collection of all datasets on a given topic into one facility. Following the best practice of traditional libraries, this special library would update the form of storage and upgrade information content as technologies evolve. Unlike a typical research institute, this facility would provide services to research going on elsewhere, while its own staff would conduct biodiversity and ecosystems research and research in biological informatics. The facility would offer "library" storage and access to diverse constituencies.

The core of NBII-2 would be a “research library system” that would comprise at least five regional nodes, sited at appropriate institutions (national laboratories, universities, museums, etc.) and connected to each other and to the nearest telecommunications providers by the highest bandwidth network available. In addition, NBII-2 would seamlessly integrate all computers — laptops, workstations, fileservers, and supercomputers — capable of storing and serving biodiversity and ecosystems data via the Internet. The providers of information would have complete control over their own data, but at the same time have the opportunity to benefit from (and the right to refuse) the data indexing, cleansing, and long-term storage services of the system as a whole.

NBII-2 would be:

- The framework to support knowledge discovery for the nation’s biodiversity and ecosystems enterprise and would involve many client and potential-client groups;

- A common focus for independent research efforts, and a global, context for sharing information among those efforts;

- An accrete-only, no-delete facility from which all information would be available online — twenty-four hours a day, seven days a week — in a variety of formats;

- A facility that would serve the needs of (and eventually be supported by partnership among) government, the private sector, education, and individuals;

- An organized framework for collaboration among federal, regional, state, and local organizations in the public and private sectors that would provide improved programmatic efficiencies and economies of scale through better coordination of efforts;

- A commodity-based infrastructure that utilizes readily available, off-the-shelf hardware and software and the products of digital libraries research wherever possible;

- An electronic facility where scientists and others could “publish” biodiversity and ecosystem information for cataloging, automatic indexing, access, analysis, and dissemination;

- A place where intensive work on how people use large information systems would be conducted, including studies of human-computer interaction, the sociology of scientific practice, computer-supported cooperative work, and user interface design;

- A place for developing the organizational and educational infrastructure that will support sharing, use, and coordination of massive data sets;

- A facility that would provide content storage resources, registration of datasets, and “curation” of datasets (including migration, cleansing, indexing, etc.);

- An applied biodiversity and ecosystems informatics research facility that would develop new technologies and offer training in informatics;

- A facility that would provide high-end computation and communications to researchers and institutions throughout the country.

This facility would not be a purely technical and technological construct, but rather would also encompass sociological, legal, and economic issues within its research purview. These would include intellectual property rights management, public access to the scholarly record, and the characteristics of evolving systems in the networked information environment. The human dimensions of the interaction with computers, networks, and information will be particularly important areas of research as systems are designed for the greatest flexibility and usefulness to people.

The needs that the research nodes of NBII-2 must address are many. A small subset of those needs includes:

- New statistical pattern recognition and modeling techniques that can work with high dimensional, large-volume data;

- Workable data cleaning methods that automatically correct input and other types of errors in databases;

- Strategies for sampling and selecting data;

- Algorithms for classification, clustering, dependency analysis, and change and deviation detection that scale to large databases;

- Visualization techniques that scale to large and multiple databases;

- Metadata encoding routines that will make data mining meaningful when multiple, distributed sources are searched;

- Methods for improving connectivity of databases, integrating data mining tools, and developing better synthetic technologies;

- Methods for improving large-scale project coordination and scientific collaborations;

- Ongoing, formative evaluation, detailed user studies, and quick feedback between domain experts, users, developers and researchers;

- Methods for facilitating data entry and the digitization of large amounts of irregularly structured information;

- Ways of engaging society in the pursuit of global information sharing.

None of these problems is unique to biodiversity research. However, there is an urgent need to address these questions within the biodiversity domain, since research has demonstrated that there can be no domain-independent solutions. We cannot "borrow" discoveries wholesale from other disciplines; we must work through these problems ourselves.[8] In order to comprehend and utilize our biodiversity and ecosystem resources, we must learn how to exploit massive data sets, learn how to store and access them for analytic purposes, and develop methods to cope with growth and change in data. The NBII-2 envisioned here can be the enabling framework that unlocks the knowledge and economic power lying dormant in the masses of biodiversity and ecosystems data that we have on hand now and will accumulate in the future.

INFRASTRUCTURE REQUIREMENTS

The total volume of biodiversity and ecosystems information is almost impossible to measure. We do know that whatever the total, only a fraction has been captured in digital form. Our natural history museums, for example, contain at least 750 million specimens, the vast majority of which have not been databased. The same holds for the published record, where most biodiversity and ecosystems information still resides in paper-based journals, books, field notes, and the like. Clearly, one of the most important infrastructure issues is to move the biodiversity and ecosystems enterprise into a digital world — create the content for the NBII-2 digital library — by digitizing on a large-scale the existing corpus of scholarly work.

The NBII-2 digital library will place challenging demands on network hardware services, as well as software services relating to authentication, integrity, and security. Needed are both a fuller implementation of current technologies, such as digital signatures and a public-key infrastructure for managing cryptographic key distribution, and a consideration of tools and services in a broader context related to library use. For example, the library system may have to identify whether a user is a member of an organization that has some set of access rights to an information resource. As a national and international enterprise that serves a very large range of users, the library must be designed to detect and adapt to varying degrees of accessibility of resources connected to the Internet.

A fully digital, interactive library system such as NBII-2 will require substantial computational resources, although little is known now about the precise scope of the necessary resources. In many areas that are critical to digital libraries, such as knowledge representation and resource description, or summarization and navigation, even the basic algorithms and approaches are not yet well defined, making it difficult to project computational requirements. We do know that many existing information retrieval techniques are intensive in their computational and input-output demands as they evaluate, structure, and compare large databases in a distributed environment. Distributed database searching, resource discovery, automatic classification and summarization, visualization, and presentation are also computationally intensive activities that are likely to be commonplace in the NBII-2 digital library.

Finally, NBII-2 will need massive storage. Even though the library system we are proposing would not set out to accrue datasets in order to become the repository for all biodiversity data — after all, many other federal agencies have their own storage facilities, and various data providers will want to retain control over their own data — large amounts of storage on disc, tape, optical, and an array of other future storage technologies will still be required. As research is conducted to produce new ways to manipulate large datasets, these will have to be sought out, copied from their original source, and stored for use in the research. And, in serving its long-term curation function, NBII-2 will accumulate substantial amounts of data for which it will be responsible, including redundant datasets that will have to be maintained in order to insure against loss.

RESEARCH AGENDA

New approaches to managing information must be developed in the context of NBII-2. Faced with massive datasets, traditional approaches in database management, statistics, pattern recognition, personal information management, and visualization collapse. For example, a statistical analysis package assumes that all the data to be analyzed can be loaded into memory and then manipulated. What happens when the dataset does not fit into main memory? What happens if the database is on a remote server and will never permit a naive scan of the data? What happens if queries for stratified samples are impossible because data fields in the database being accessed are not indexed so the appropriate data can be located? What if the database is structured with only sparse relations among tables, or if the dataset can only be accessed through a hierarchical set of fields?

Furthermore, challenges often are not restricted to issues of scalability of storage or access. For example, what if a user of a large data repository does not know how to specify the desired query? It is not clear that a Structured Query Language (SQL) statement — or even a program — can be written to retrieve the information needed to answer a query such as "show me the list of gene sequences for which voucher specimens exist in natural history collections and for which we also know the physiology and ecological associates of those species." Many of the interesting questions that users of biodiversity and ecosystems information would like to ask are of this type: they are "fuzzy", the data needed to answer them must come from multiple sources that will be inherently different in structure and conceptually incompatible, and the answers may be approximate.

Major advances are needed in methods for knowledge representation and interchange, database management and federation, navigation, modeling, and data-driven simulation; in approaches to describing large complex networked information resources; and in techniques to support networked information discovery and retrieval in extremely large scale distributed systems. In addition to near-term operational solutions, new approaches are also needed to longer-term issues such as the preservation of digital information across generations of storage, processing, and representation technology. Traditional information science skills such as thesaurus construction and indexing must be elaborated upon and scaled to accommodate large information sources. We need to preserve and support the knowledge of library and information science researchers, and help scale up the skills of knowledge organization and information retrieval.

Also much needed are software applications that provide more natural interfaces between humans and databases than are now available. For example, a valuable data cleansing activity might be to "show the data relating to all specimens in our natural history collections whose likelihood of being mislabeled exceeds 0.75." Assuming that certain cases in the database can be identified as "labeled correctly" and others "known to be mislabeled," then a training sample for a data mining algorithm could be constructed. The algorithm would build a predictive model and retrieve records matching that model rather than a structured query that a person might write. This is an example of a much needed and much more natural interface between humans and databases than is currently available. In this case, it eliminates the requirement that the user adapt to the machine's needs rather than the other way around. We must refine and augment the interactions between people and machines, expand the role of agentry in information systems, and discover more powerful and natural ways of navigating the scientific record.

In return, computer and information science and technology research in the biodiversity and ecosystems domain is likely to yield discoveries of value to other areas.[9] Certainly, nowhere do we find the problems of heterogeneous database federation more challenging than in the life sciences. A fully implemented digital library for biology would include everything from ideas to physical objects, and enormous amounts of information in every media type imaginable. Research on global climate change, habitat destruction, and the discovery of species are among the most distributed of our scientific activities, creating extraordinary opportunities to learn about computer-mediated project coordination and communication. 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, perhaps more than in any other circumstance.

ACTION PLAN

We have laid out the case for building a fully digital, interactive, research library system for biodiversity and ecosystems information, and the basic requirements of and goals for the library and its research and service. But how much will it cost, and how long will it take to build?

We estimate that each of the regional nodes that will form the core of NBII-2 will require an annual operating budget of at least $8 million and probably more. Minimally, supporting five such nodes would require at least $40 million per year, an amount that is a fraction of the funds spent nation-wide each year to collect data (conservatively estimated at $500 million for federal government projects alone). As with the Internet itself, the federal government should provide the "jumpstart" for this new infrastructure by investing heavily in its formative stages. Part of the investment should be devoted to developing incentives for the participation of private-sector partners. Gradually, support and operation of the infrastructure should be shared by non-governmental participants, just as has happened with the Internet.

The planning and request-for-proposals process should be conducted within one year. Merit review and selection of sites should be complete within the following six months. The staffing of the sites and initial coordination of research and outreach activities should take no more than a year after initial funding is provided. The "lifetime" of each facility should not be guaranteed for more than five years, but the system must be considered a long-term activity, so that data access is guaranteed in perpetuity. Evaluation of the sites and of the system should be regular and rigorous, although the milestones whereby success can be measured will be the incremental improvements in ease of use of the system by students, policy-makers, scientists, and others. In addition, an increasing number of public-private partnerships that fund the research and other operations will indicate the usefulness of accessible, integrated information to commercial and governmental interests.

Conclusion

In the 21st century, work will be increasingly dependent on rapid, coordinated access to shared information. Through the shared digital library of NBII-2, scientists and policy makers will be able to collaborate with colleagues across geographic and temporal distances. They will use the library to catalog and organize information, perform analyses, test hypotheses, make decisions, and discover new ideas. Educators will use its systems to read, write, teach, and learn. In traditional fashion, intellectual work will be shared with others through the medium of the library — but these contributions and interactions will be elements of a global and universally accessible library that can be used by many different people and many different communities. By increasing the effectiveness of information, NBII-2 is likely to lead to scientific discoveries, advance existing areas of study, promote disciplinary fusions, and enable new research traditions. And most important, it could help us protect and manage our natural capital so as to provide a stable and prosperous future.

References

1. President's Committee of Advisors on Science and Technology (PCAST). 1998. Teaming with Life: Investing in Science to Understand and Use America's Living Capital. Report to the President of the United States from the PCAST Panel on Biodiversity and Ecosystems.

2. National Research Council. 1993. A Biological Survey for the Nation. Washington: National Academy Press.

3. National Performance Review. 1997. Access America: Reengineering Through Information Technology. Report of the National Performance Review and the Government Information Technology Services Board.

4. National Research Council. 1997. Bits of Power: Issues in Global Access to Scientific Data. Washington: National Academy Press.

5. Robbins, R. J. 1996. Bioinformatics: Essential infrastructure for global biology. Journal of Computational Biology 3(4):465-478.

6. Schnase, J. L., D. L. Kama, K. L. Tomlinson, J. A. Sánchez, E. L. Cunnius, and N. R. Morin. 1997. The Flora of North America digital library: A case study in biodiversity database publishing. Journal of Network and Computer Applications 20:87-103.

7. Hutchins, E. 1995. Cognition in the Wild. Cambridge: MIT Press.

8. Star, S. L., and K. Ruhleder. 1996. Steps toward an ecology of infrastructure: Design and access for large information spaces. Information Systems Research 7(1):111-134.

9. Spasser, M. A. (forthcoming). Computational Workspace Coordination: Design-in-Use of Co-operative Publishing Services for Computer-Mediated Collaborative Publishing. Unpublished Ph.D. dissertation. University of Illinois, Urbana-Champaign.