Research Directions in Biodiversity Informatics

John L. Schnase

Earth and Space Data Computing Division
NASA Goddard Space Flight Center
Greenbelt, MD 20771
USA
schnase@gsfc.nasa.gov

Abstract

This paper provides an introduction to the major research directions in biodiversity informatics. The biodiversity enterprise is a vast and complex information domain. I describe the need to build infrastructure for this domain, major research thrusts needed to improve its work practices, and areas of research that could contribute to the advancement of the field. I emphasize that the science of biodiversity is fundamentally an information science, worthy of special attention from the computer and information science communities because of its distinctive attributes of scale and socio-technical complexity.

1. Biodiversity

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 species richness of habitats *per se* — is perhaps the single most important factor influencing the stability and health of ecosystems [2, 3].

It is not surprising, then, that information about biodiversity forms the basis of one of our most important knowledge domains, vital to a wide range of scientific, educational, commercial, and government uses. Unfortunately, most biodiversity information now exists in forms that are not easily accessed or used. From traditional paper-based libraries to scattered databases of varying size and physical specimens preserved in natural-history collections throughout the world, our record of biodiversity is uncoordinated and poorly integrated, and large parts of it are isolated from general use. We lack the technologies needed to effectively gather, analyze, and synthesize these data into new discoveries. As a result, this information is not being used as effectively as it could by scientists, resource managers, policy-makers, or other potential client communities. The good news is that research activities are being conducted around the world that could improve our ability to manage biodiversity information, and the emerging field of biodiversity informatics is attempting to meet the challenges posed by this domain.

2. Biodiversity Informatics

Until recently, little attention has been paid to computer and information science and technology research in the biodiversity domain. Those working in the field of biodiversity informatics — many cross-trained or cross-teamed in computer science and biology — are attempting to change that. If we are to keep pace with our need for

Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000

quality information about the living systems of our planet, we must produce mechanisms that can efficiently manage petabytes of a whole 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 and synthesize them, and present the results in an understandable and usable manner.

Despite encouraging advances in computation and communication performance in recent years, we are able to perform these activities only on a very small 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, we are clearly just embarking on a whole new world of computer-mediated exploration. We can, however, make rapid progress if the computer and information science and technology research community becomes focused on the challenges posed by the biodiversity research community.

3. Managing Complexity

The single most important factor influencing the nature of work in biodiversity informatics is the problem of complexity. Living systems are complex, and they are complex at many levels. The computational challenges associated with work on cellular processes — such as decoding the structure of DNA — are enormous. But DNA-level functions are only part of an elaborate web of biotic and abiotic interactions that span from molecules to cells, and from organisms to populations and entire ecosystems. Knowledge about biodiversity is a vast and complex information domain.

This 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. 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 information 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, 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, images, sound, and video. The range of other databases with which biodiversity datasets must interact is also broad, including geographical, meteorological, geological, chemical, physical, and genomic datasets. The mechanisms used to collect and store biodiversity data are almost as varied as the natural world that they document. In addition, biological data can be politically and commercially sensitive and can entail conflicts of interest. User's skill levels are highly variable, and training in this field is not well developed.

Because of these complexities, humans still play a crucial role in the processing of biological data. Biological information is not as amenable to automatic correlation, analysis, synthesis, and presentation as many other types of information, such as that in radioastronomy, where there is more coherent global organization and the problems being studied are often 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 creating higher-order understanding from dispersed datasets is a fundamental intellectual process in the biodiversity 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.

4. Research Directions

4.1 Information Infrastructure

The total volume of biodiversity and ecosystem information is almost impossible to measure. We do know that whatever the total, only a fraction has been captured in digital form. The natural history museums in the US alone, for example, contain at least 750 million specimens, the vast majority of which have not been recorded in databases. The same holds for the published record, where most biodiversity and ecosystem 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 enterprise into a digital world — to create the content for a global biodiversity digital library — by digitizing the existing corpus of scholarly work on a large scale.

Such an infrastructure would place challenging demands on network hardware services and on software services related to authentication, integrity, and security. Needed are both a fuller implementation of current technologies and consideration of tools and services in a broader context related to the use of online biological resources. Since biodiversity research is a global enterprise, this infrastructure must be designed to detect and adapt to various degrees of accessibility of resources connected to the Internet.

A fully digital, interactive biodiversity information service will require substantial computational and storage resources. Many information-retrieval and data mining 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 common to this field. As described earlier, large storage capacities are required for the myriad new datasets being populated by an expanding array of sensors.

In an accompanying paper in these proceedings, Cotter and Bauldock describe some of the capacity-building activities that are currently underway, the most prominent of which are work on the US National Biological Information Infrastructure (NBII) program and the international effort to create a Global Biodiversity Information Facility (GBIF). Importantly, both infrastructure efforts propose programs of basic and applied research in biodiversity informatics.

4.2 Process Improvement

New approaches to data management must be developed to handle biodiversity information. Massive datasets can lead to the collapse of traditional approaches in database management, statistics, pattern recognition, personal-information management, and visualization. Many of the interesting questions that users of biodiversity information would like to ask are "fuzzy," and the data needed to answer them must come from multiple sources that will be inherently different in structure and conceptually incompatible, and the answers might be approximate.

Major advances are needed in methods for knowledge representation and interchange, database management and federation, navigation, modelling, and data-driven simulation; in approaches to describing large, complex networked information resources; and in techniques to support networked information discovery and retrieval in large-scale distributed systems. In addition to near-term operational solutions, new approaches are 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.

Also much needed are software applications that provide more natural interfaces between humans and databases than are now available. We must refine and augment the interactions between people and machines, expand the role of agentry in information systems, and discover more powerful and more natural ways of navigating this complex scientific record.

4.3 Process Reinvention

Biologists have identified approximately 1.5 million living species of all kinds of organisms, but vast arrays of species remain to be discovered. The grand total for all life is currently estimated to fall somewhere between 10 and 100 millions species [3]. There is little doubt that the Earth's biodiversity is declining. By all estimates, we are in the midst of the sixth major extinction event of the planet's history, this one the primary result of human modifications to the environment. The Nature Conservancy has estimated that one-third of the plant and animal species in the US are now at risk of extinction. The problem is a monumental one, and forces us to consider how we should respond.

Given this context, it is disturbing to realize that the fundamental work practices of the biodiversity sciences — largely unchanged over the past two centuries — are utterly unable to keep pace with rate of habitat destruction and species loss. Species discovery is still largely a manual activity requiring field collection of specimens, months or years of laboratory analysis, and time-consuming publishing activities. Conservation practices alone cannot solve the problem. If we hope to ever fathom the Earth's biodiversity, the biodiversity enterprise must reinvent itself — develop wholly new approaches to dealing with global-scale problems in a rapidly-changing, information age. Herein lies some of the most interesting informatics research challenges. There are at least three broad categories of research of particular relevance to this reinvention effort.

Collaboration In-the-Large — Bowker, in an accompanying paper in these proceedings, has described how deeply interdisciplinary and collaborative work in the biodiversity sciences can be. As an example, the Flora of North America (FNA) project is attempting to produce a comprehensive study of all the naturally occurring species of plants in North America. Surprisingly, such a study of North American plants has never been done before. FNA is a long-term publishing project involving as many as 1000 botanists distributed across North America and Europe. The result will be 30-plus printed volumes produced by Oxford University Press and an online version of the flora. FNA is an example of one of the largest coordinated, scientific publishing activities ever funded by the National Science Foundation.

The point here is that collaboration — often among very different scientific communities — is an important feature of work in the biodiversity sciences and presents an opportunity to do fundamental research on collaborative systems that is deeply embedded in real-world activities. The major task of building robust databases in biodiversity is facilitating interdisciplinary communication — and this communication cannot just be a desired outcome, it must be designed into the data collection and representation work from the outset. Both the biodiversity and computer and information sciences domains have much to gain by paying attention to collaborative systems research in this area.

Instrumented Species Discovery — An important open question is the extent to which we might be able to instrument the discovery and monitoring of biodiversity. Think for a moment of how weather and climate prediction have been revolutionized by the capacity to detect, analyze, and inform scientists — as well as the general public — about salient attributes of the Earth's atmosphere on an ongoing, real-time basis. A similar capability for detecting and monitoring the status of biodiversity could likewise revolutionize this science.

I can only speculate about the candidate technologies that might make this possible. Certainly there is a role for mobile computing and the enhanced *in situ* collection of data about organisms in their habitats. MEMS sensors — microelectromechanical systems — offer the prospect of low-cost, high-resolution detection of a potentially unlimited range of environmental attributes, including temperatures, rainfall, chemicals, and sounds. The algorithms and software architectures required to manage dense MEMS arrays or mobile computers in exotic settings have yet to be invented, and this is quickly becoming the focus of much research.

Space-based remote sensing is entering a new era with the deployment of high-resolution optical instruments, hyperspectral sensors, and laser- and radar-based sensors. We really do not understand the full capacity of these instruments and how they might help the biodiversity community, but some interesting prospects are emerging. For example, Ritchie and Olff [1] have developed a synthetic theory of biodiversity that predicts relations between species richness and productivity more effectively than previous models. Their mathematical rules are based entirely on spatial scaling laws and notions about how organisms acquire resources in space. Is this a potential "hook" that would allow us to measure the carrying capacity — or perhaps even the species richness — of an ecosystem from space? After all, we can compute the fractal dimension of landscapes using satellite data. It is too early to tell, but clearly new areas of investigation are opening up that could have profound implications for the enterprise.

Computational Exploration — Biology is a science strongly influenced by historical events. Unlike much of physics or chemistry, one cannot predict what happens at time *t+1* by knowing only the conditions at time *t*. Evolution and environmental history impose "ecological memory" on living communities, introduce time lags in ecological processes, and constrain the trajectories of community composition in ways that are poorly understood. Ecological memory is encoded in the genetic structure of species and the current structure of biological communities. It affects how communities assemble, and it may affect the likelihood that they can be restored once dissembled. These attributes reveal a fundamental property of living systems: they are computational systems, encoding and storing data and programs in biopolymers (such as DNA) and executing the genetic algorithm against elements in a complex biotic and abiotic context.

The implication here, I believe, is that the dominate mode of discovery for the biodiversity enterprise must increasingly become model- and computation-driven. We cannot get a handle on these processes any other way, because they are simply too complex. We need a unified way of incorporating the spatial and temporal context of ecological interactions. This suggests an important role for research on artificial life systems, evolutionary computation, and adaptive and complex systems approaches to exploring biodiversity. The research opportunities here are unlimited.

5. Conclusion

In this paper, I have basically tried to make the point that the science of biodiversity is fundamentally an information science, and one worthy of special attention from the computer and information science communities because of its 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.

6. References

[1] Ritche, Mark E. and Han Olff. Spatial Scaling Laws yield a Synthetic Theory of Biodiversity. *Nature* 400: 557, 1999.

[2] Schnase, John L., Meredith A. Lane, Geoffrey C. Bowker, Susan Leigh Star, and Abraham Silberschatz. Building the Next-Generation Biological-Information Infrastructure. In *Nature and Human Society: The Quest for a Sustainable World*. Washington: National Academy Press, 2000.

[3] Tilden, David. Causes, Consequences, and Ethics of Biodiversity. *Nature* 405: 208, 2000.

[4] Wilson, Edward O. *The Diversity of Life.* Cambridge: Belknap Press, 1992.