

Biodiversity Informatics: The Challenge of Rapid Development, Large Databases, and Complex Data

Meredith A. Lane
Academy of Natural Sciences
1900 Benjamin Franklin Pkwy
Philadelphia, PA 19103
USA
lane@ansp.org

James L. Edwards
National Science Foundation
4201 Wilson Blvd
Arlington, VA 20560
USA
jledward@nsf.gov

Ebbe S. Nielsen
CSIRO Entomology
GPO Box 1700
Canberra ACT 2601
Australia
ebben@ento.csiro.au

Abstract

There are high expectations in all sectors of society for immediate access to biological knowledge of all kinds. To fully exploit and manage the value of biological resources, society must have the intellectual tools to store, retrieve, collate, analyze, and synthesize organism-level and ecological scale information. However, it currently is difficult to discover, access, and use biodiversity data because of the long history of “bottom-up” evolution of scientific biodiversity information, the mismatch between the distribution of biodiversity itself and the distribution of the data about it, and, most importantly, the inherent complexity of biodiversity and ecological data. This stems from, among many factors, numerous data types, the nonexistence of a common underlying (binary) language, and the multiple perceptions of different researchers/data recorders across spatial or temporal distance or both. The challenge presented to the computer science and information technology community by the biodiversity and ecological information domain is worthy of all the time and talent that can be

brought to bear, because the continued existence of the species *Homo sapiens* depends upon gaining an understanding of this spaceship Earth and our fellow passengers upon it.

1. Introduction

The economic prosperity and, indeed, the fate of human societies are inextricably linked to the natural world. Because of humanity’s dependence on natural systems, information about biodiversity and ecology is vital to a wide range of scientific, educational, commercial, and governmental uses. Biodiversity (in the sense of the totality of all species) and ecosystems are themselves interdependent. Ecosystems and the diversity of species they support underpin our lives and our economies in very real, though often underappreciated, ways. The living things with which we share the planet provide us with clean air, clean water, food, clothing, shelter, medicines, and aesthetic enjoyment. Yet, increasing human populations and their activities are disturbing species and their habitats, disrupting natural ecological processes, and even changing climate patterns on a global scale. These are greater stresses on the natural world than humanity has ever generated in the past. Since biodiversity is arguably the most precious resource on Earth, it is becoming more and more important that we actively conserve biodiversity and protect natural ecosystems in order to preserve the quality of human life. As human populations and their demands on the natural world grow, our accumulated knowledge about biodiversity and the environment will become ever more important in the effort to develop a sustainable world.

Recognition of this has led to the National Biological Information Infrastructure in the United States, to the Environmental Resources Information Network in

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Data Base Endowment. To copy otherwise, or to republish, requires a fee and/or special permission from the Endowment.

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

Australia, and to a number of regional biodiversity information networks (NABIN, IABIN, EIONet, and others). Indeed, the recommendation by an international working group established by the Global Science Forum (formerly Megascience Forum) of the Organization for Economic Cooperation and Development (OECD) that the nations of the world establish and maintain a Global Biodiversity Information Facility (GBIF), which is poised to become a reality in early 2001, is a direct outgrowth of both concern about the environment and the economy, and the acknowledgment that the complexity of biodiversity and ecological datasets reflects the complexity of natural systems. It has become apparent that practitioners in the computer science and information technology fields must become as invigorated by and invested in the biodiversity and ecological information domain as are the biologists, who collect, generate, query, and interpret the data.

2. Biodiversity and Ecology Information

Biodiversity itself is distributed all over the Earth, with concentrations primarily in developing countries. In contrast, scientific biodiversity knowledge is concentrated in major centers in developed countries. To be useful in the management and use of biodiversity, biodiversity information should be available when and where it is needed. At present, it is more likely that information on the plants of a particular part of Africa is stored in an herbarium in Europe, for example. Because it is not immediately at hand, biodiversity information is often not applied in policy or management decisions that affect the organisms involved. At present, scientists and others find it is difficult to discover, access, and use biodiversity data because of the long history of “bottom-up” evolution of scientific biodiversity information, the mismatch between the distribution of biodiversity itself and the distribution of the data about it, and, most importantly, the inherent complexity of biodiversity and ecological data. This complexity stems from inclusion of multiple data types, the nonexistence of a common underlying (binary) language, and the multiple perceptions of different researchers/data recorders across spatial or temporal distance or both. In contrast, in disciplines that have emerged very recently, such as genomics (which has a history measured in mere decades in comparison to the centuries of history of biodiversity science), researchers have been able to capitalize on modern information technology to capture the data in digital form and make the data more readily accessible from the very beginning of their science.

In order to comprehend and sustainably utilize the biodiversity resources of the world, humankind must learn how to exploit massive data sets, learn how to store and access them for analytic purposes, develop methods to cope with growth and change in data, and make it possible to “repurpose” previously existing data. Many

issues surround efforts to make biodiversity and ecological information electronically accessible and usable. We must unlock the knowledge and economic power lying dormant in the masses of biodiversity data that we have on hand that is stored in static media (for the most part in print, on paper). We need to bring the results of a great deal of earlier research that are now only found in static media into electronic format. We need information science research that is focused on biodiversity and ecosystems datasets and information—their structures, their complexity, their fuzziness, and their clarity.

3. Biological Informatics

There are high expectations in all sectors of society for immediate access to biological knowledge of all kinds. The desire to call upon the data in these databases, collate the information from various sources, analyze the combined data, and make predictions and correlations is growing exponentially. Therefore, one of the greatest needs in biology (and between biology and related sciences) lies in the area of information management and provision over networks; that is, biological informatics.

Biological informatics integrates biological, computational, networking, and organizational research, and is concerned with the development, sharing, and analysis of biological datasets, which are usually very large and complex. Examples of such datasets include results from the Human Genome Project, from research in biotechnology and medicine (including pharmaceuticals and neurobiology), and from environmental research such as biodiversity and ecosystem ecology. To date, “bioinformatics” as a term has been recognized as important mostly in the area of genome research. However, as genome data are accumulated, it will become more and more apparent that those data, without the context of all sorts of other information about the organism (physiological, ecological, etc.) from which the genome was sampled, will be limited in their usefulness. In the context of the full informational milieu, however, the usefulness of each of the several different kinds of data will be increased. To fully exploit and manage the value of biological resources, society must have the intellectual tools to store, retrieve, collate, analyze, and synthesize organism-level and ecological scale information. In short, biodiversity and ecological informatics inevitably will become increasingly important.

4. Biodiversity and Ecological Informatics

If all the biodiversity information that we currently have stored on static media were digitized and as freely available online as is that from the Hubble telescope or GenBank, biodiversity informatics would be more similar to astronomy and molecular biological informatics in its

present capabilities. This would lead to economic benefits from mining biodiversity itself for currently unknown and therefore untapped resources, as well as those benefits that can be derived from mining existing data.

A major challenge within ecological informatics is extrapolation from data sampled from an ecosystem to the totality of that ecosystem, about which queries may be posed. Generalizations are critical to ecological analyses, because thorough, complete-coverage data collection, given paltry funding levels compared to astronomy, medicine, or other fields, would be prohibitively expensive. Meeting this challenge has allowed ecological informatics to enable integration of traditional field ecology with modern technology. Thus, it is now possible consistently to link scientific research to ecological planning and management. Diverse sources of information, from field ecology to satellite images, are beginning to be brought to bear on a host of practical problems, from land use planning to global change.

Environmental informatics may be viewed as a merging of biodiversity and ecological informatics with geographic information systems (GIS) and other environmental data. Using information sources and software algorithms developed in Australia and elsewhere, it is possible to generate predictive models of environmental change. Increases in functionality, automation, and refinement of predictive capabilities are much needed.

One issue that is particularly critical to biological informatics is the establishment of data standards, and the development of metadata standards. This is true within all disciplines, but biodiversity and ecology are particularly far behind in this area. There are hardware technologies that are also needed, but the development of these is more nearly keeping pace with demand, with the exception of network bandwidth. Discipline-specific data exchange, interdisciplinary and requirements for data exchange, and the computer science of data exchange and integration, along with the definition of scientific and technical metadata issues (as well as the anticipation of future needs in these areas), are important challenges. Major advances are needed in methods for knowledge representation and interchange, database management and federation, navigation, modelling, and data-driven simulation; in effective approaches to describing large, complex networked information resources; and in techniques to support networked information discovery and retrieval in extremely large-scale distributed systems.

A major scientific consideration in biodiversity science is the need to bring 25 decades' worth of accumulated information into an electronically available format. The data labels on hundreds of millions of specimens (probably as many as three billion) in natural history collections around the world must be digitized, and hundreds of years' worth of scientific publication (unlike many fields of science, data and information—unless false or otherwise disproven—never

go out of date) must be prioritized and brought online. Rapid means of data entry must be designed and implemented to capture all of this from static media or to rescue it from the unscrupulous. It is important to reuse data derived from earlier studies. Not only is it impractical to recreate the research projects that resulted in data published in print, but many of those studies cannot be replicated because anthropogenic modifications have severely disturbed the habitat of the organisms involved, driven the study populations or species to extinction, or both.

In addition, new kinds of data are being generated by satellite imagery and other measures of nonbiological, global phenomena—phenomena that have significant influence upon biodiversity. Great forward strides could be made in the understanding of the biological world, for instance, if informatics techniques were developed to make it possible to correlate historical information with newly collected satellite data; if molecular genetic datasets could be linked to species-documentation datasets such as those held by natural history collections; and if neurobiological, physiological, chemical, and other sorts of datasets could be correlated with taxonomic and ecological ones.

To index it all, a catalogue of the names of known organisms is a critical element that must be developed. No complete listing of all the names for organisms exists. The information necessary to compile such a list does exist, but it is scattered throughout millions and millions of pages of printed journals and books. A database of this type would be useful in and of itself, but it is absolutely critical to the progress of biodiversity and ecological informatics. This is because the scientific name of an organism has been handled as the index marker for that organism for hundreds of years. Searching the literature for information on the breeding habits of the yellow-naped parrot will avail the researcher little unless s/he knows that the scientific name is *Amazona ochrocephala* ssp. *auropalliata*; in addition, to garner all the information that might be available, it is important to know that the same creature has also been called *Amazona auropalliata*, because some publications might have used that synonym. The electronic catalogue must make it possible for even the naïve user of a biodiversity and ecological information system to retrieve all information relevant to a query, even if that query is incompletely or poorly structured.

Critical thought must be given to mechanisms for across-scale computing. That is, we need to be able to pose queries that will require combinations and correlation of data that have been measured at scales that may differ by orders of magnitude. One such query might be: *What is the likelihood of survival of the red knot (a bird species that yearly migrates from the Southern tip of South America to the northern reaches of North America and back again) given the diminishment of the Delaware Bay horseshoe crab population?* The answer to this

question will require combination and correlation of (among other datasets) data on single specimens of red knots, the horseshoe crabs on whose eggs they feed mid-migration; NEXRAD data along the migration route; remote sensing data on the 30-meter scale; vegetational and faunal community sample data; trend data on accumulation of excess nitrogen and phosphorus, heavy metals, and pesticides over watershed and local scales over a multiyear period; and gene sequence data from population-size samples of birds, crabs, and other creatures that are part of the Delaware Bay ecosystem. This question might be followed by: *How great would be the positive effect (for the horseshoe crabs and therefore for the red knots) if the states of Pennsylvania, Maryland, and Delaware severely or moderately restricted the application of pesticides or fertilizers during the migration season? Banned it altogether? The year before?* The answers to these latter questions, of course, require additional input from datasets not initially included in the analysis. Users ultimately will require on-the-fly flexibility, so that they can reshape and redirect queries depending on initial results.

All of this informatics capability is needed because we are losing, at an ever-increasing rate, both species that we know (there are probably 1.8 million known, named species) and (approximately) ten times as many that we don't know. These species live and interact within communities; each interaction is multifactorial. Those interactions are the source of the emergent properties of ecological systems. Emergent properties then give rise to

further interactions; there are an unknown number of levels of interactions and emergent properties in ecosystems. Datasets that contain measurements of ecosystem elements probably don't (yet) reflect the interactions, even though the dataset size may be enormous. We need to find ways to incorporate the living nature of the natural system within our information systems. This challenge is worthy of the attention of all the computer science and information technology talent that can be brought to bear, because the continued existence of the species *Homo sapiens* depends upon gaining an understanding of this spaceship Earth and our fellow passengers upon it.

6. References

- [1] Lane, M. A. 1999. *Weaving a Web of Wealth: Biological Informatics for Industry, Science and Health*. Report from the Australian Academy of Science, Canberra, Australia. 40 pp.
- [2] President's Committee of Advisors on Science and Technology Panel on Biodiversity and Ecosystems. 1998. *Teaming With Life: Investing in Science to Understand and Use America's Living Capital*. 96 pp.
<http://www.whitehouse.gov/WH/EOP/OSTP/Environment/html/teamingcover.html>
- [3] Web site for the Global Biodiversity Information Facility (GBIF) <http://www.gbif.org/>