Biodiversity Informatics Infrastructure: An Information Commons for the Biodiversity Community

Gladys A. Cotter

U.S. Geological Survey
300 National Center
Reston, VA 20192
USA
gladys_cotter@usgs.gov

Barbara T. Bauldock

U.S. Geological Survey
300 National Center
Reston, VA 20192
USA
barbara_bauldock@usgs.gov

Abstract

This paper provides an overview of efforts to create an informatics infrastructure for the biodiversity community. A vast amount of biodiversity information exists, but no comprehensive infrastructure is in place to provide easy assess and effective use of this information. The advent of modern information technologies provides a foundation for a remedy. Biodiversity informatics infrastructures are being called for at national, regional, and global levels, and plans are in place to coordinate these efforts to ensure interoperability. The paper reviews some essential requirements and some challenges related to building this infrastructure.

1. Introduction

A vast amount of information on biological resources (plants, animals, and ecosystems) exists throughout the world today. It has been collected by government agencies, universities, museums, and private organizations. This information is diverse and includes biological specimens, journal articles, videos, numeric data, satellite images, and audio files. Some of the information such as biological specimens were collected by early explorers and scientists and are maintained in natural history museums. Journal articles exist in libraries throughout the world, often in paper format. Other information such as numeric and visual representations are maintained at high performance computer facilities and are fully digitized. A significant portion of the information remains in the hands of the scientists who originated it. The fact that biodiversity information is collected and stored in diverse forms, formats and locations has proved a serious obstacle to our ability to correlate and synthesize the information to create new knowledge.

The biodiversity information which exists today has economic value and represents an investment of billions of dollars worldwide. Unfortunately, a comprehensive infrastructure that would allow this information to be easily accessed and effectively used so that society can reap a return on its investment does not yet exist. Often, people who need help to answer a question or solve a problem are unable to ascertain if the information required even exists. Therefore questions go unanswered, problems go unsolved, or money is wasted re-collecting information that already exists but is not accessible.

The advent of the Internet and the World Wide Web has created a technological environment in which it is possible to link people and information in unprecedented ways. We have the opportunity to apply this technology to develop an information infrastructure that will enable us to unlock the wealth of biodiversity information that exists around the world. We can in effect choose to create an interoperable global biodiversity information "Commons" that will bring together people, information, and analytical capabilities that can accelerate the process of knowledge discovery, deliver answers to natural resource management and research questions, and affect the quality of life on earth for the good.

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

2. Community Calls to Action - National, Regional, Global

The opportunities that technological advancements have laid before us to create a biodiversity information infrastructure complement the calls to action from different sectors of the community to create this infrastructure. In 1996, the Megascience Forum of the Organization for Economic Cooperation and Development (OECD) established a Working Group on Biological Informatics to further the vision of interconnected, interoperable, global biological informatics. Biodiversity informatics was a prime focus of this Group which recommended the establishment of a Global Biodiversity Information Facility (GBIF). GBIF will comprise the expertise and products of efforts going on in all participating countries and will facilitate the development of standards including an electronic catalog of the names of known organisms.

Also globally, the United Nations' Convention on Biological Diversity has established a Clearing-House Mechanism (CHM) on the World Wide Web, which provides an opportunity for nations to share biodiversity information and technology.

We hear calls to action at hemispheric levels as well. For example, development of an Inter-American Biodiversity Information Network (IABIN) is an initiative resulting from the 1996 Summit of the Americas on Sustainable Development. IABIN's goal is to promote greater coordination among Western Hemisphere countries in the collection, communication, and exchange of biodiversity information to support decision-making and education.

At a sub-hemispheric level, groups of countries are getting together as in the case of the North American Biodiversity Information Network (NABIN), an initiative of the North American Commission for Environmental Cooperation which seeks to promote open access to biodiversity data and collaborations among scientists.

At the national level, national academies of science, presidential committees, and natural resource groups are calling for a commitment to build a biodiversity information infrastructure. In its report, "Teaming with Life: Investing in Science to Understand and Use America's Living Capital," the President's Committee of Advisors on Science and Technology (PCAST) stated:

> The economic prosperity and, indeed, the fate of human societies are inextricably linked to the natural world. Because of this, information about biodiversity and ecosystems is vital to a wide range of scientific, educational, commercial, and governmental uses. Unfortunately, most of this information exists in forms that are not easily used. ... There exists no comprehensive technological or organizational framework that allows this information to be readily accessed or used effectively by scientists, resource managers, policy makers, or other potential client communities [1].

The Committee called for the development of a next-generation National Biological Information Infrastructure (NBII), the goal of which would be to promote the use of biodiversity and ecosystems information in management decisions, in education and research, and by the public. Similar activities are occurring in countries around the globe.

Perhaps most encouraging are the community-building activities that are coming from the biodiversity community itself- the universities and non-governmental organizations, the museums, the government agencies, the private sector, and the public citizens who have started working together and who share a vision to make the biodiversity informatics Commons a reality.

3. Essential Elements for the Global Commons

In order to implement this global "Commons" for the biodiversity information community, we need to work together on several critical, interrelated pieces. Content- the biodiversity information itself- is the most important piece. Schemas for organizing available information, technologies to enable its use, and rules of conduct to protect its value are the other components which must be addressed collaboratively to realize the benefits of our global association.

3.1 Content

Biodiversity information exits in both digital and non-digital forms and is held in a variety of institutions and agencies worldwide. The first step in building content for our commons, then, is to discover what data and information are already available in the community and to work toward incorporating them into the knowledge base. As the integrated knowledge base grows, gaps in the information required to support biodiversity activities will become apparent. These gaps then become challenges to the research community and may help individuals and institutions direct their research initiatives to areas of potential high return on research dollar investment. Knowing where our knowledge or understanding falls short may also suggest areas where multi-institutional collaborations on a particular biodiversity issue might be most appropriate.

3.2 Knowledge Organization

Knowledge flowing into our biodiversity commons must be organized in useful ways to facilitate the discovery and retrieval of a comprehensive selection of information pertinent to the question at hand. Library science has a long history of the development and maintenance of schemas for the organization of information, and the lessons of that discipline should guide us as we sort through, discuss, and choose for adoption the various ways of organizing electronically accessible information.

Biodiversity information can be organized thematically, such as by taxon, issue, etc., or geospatially, such as by region or ecosystem. It can be organized chronologically, to show trends over time or provide historical snapshots. Fortunately, digital technologies allow us to store information in ways such that all of these organizational approaches can be applied to our knowledge space concurrently, allowing users to select the schema most appropriate for their requirements.

Knowledge organization requires tools to facilitate information discovery and retrieval. Controlled vocabularies in general, and standard taxonomies in particular, allow information seekers to "speak the same language"- literally - as information providers, which increases retrieval precision. Much work remains to be done, however, in the area of thesauri development to agree on a set of terms to describe the various parameters central to discussions of biodiversity and to provide multi-lingual access to those terms.

Standard taxonomies, including scientific names, synonyms, common names in various languages, and information about the authorities on which the standard taxonomies are based, are central to the identification and retrieval of biodiversity information by species. IT IS, the Integrated Taxonomic Information System, is an example of such a standard system. ITIS was first developed by a group of U.S. Federal government agencies for internal use. Now ITIS has been expanded to include other institutions and agencies as partners. The system has been incorporated into biodiversity networks in Canada and Mexico and represents a substantial contribution to the Species 2000 global taxonomic information system initiative.

Metadata provide standardized descriptions of the biodiversity databases, datasets, and information products in our knowledge base. These descriptions convey concisely such things as subject matter; how, when, where, and by whom the data were collected; how to access the database or information product; and person(s) or institution(s) to contact for more information. Use of a consistent metadata format allows users to compare and contrast different distributed data and information sources quickly and easily and to locate and choose those which best meet their needs.

We must examine metadata standards currently in use and determine or develop a recommended standard for biodiversity information. Because much of the information of interest to the biodiversity community is georeferenced, our community should build on the standards work of the spatial data community. In the United States, for example, we have developed a biological extension to the Geospatial Metadata Content Standard developed by the Federal Geographic Data Committee. This standard, including the biological extension, is currently being considered for adoption by the International Standards Organization.

3.3 Information Technology Tools

Information technologies are advancing rapidly and provide us with many of the capabilities required to implement our vision. Information technology provides us with tools to digitize information and store it in accessible systems; discover and retrieve data pertinent to the issue at hand; analyze data from diverse, distributed databases; input these data to decision-support, modelling or other management systems; and promote interaction among colleagues through collaboratoria, Internet-based communications facilities which enable discussion, document development and revision, and decision-making in real time. The biodiversity community must assess the IT tools available and customize them as necessary for our particular requirements. We must also develop an IT research agenda for the biodiversity infrastructure. Both of these efforts would be most effectively addressed through multi-institutional and multi-national collaborations.

3.4 Rules of Conduct

A final aspect which we must address collectively might be termed "Rules of Conduct" for the community. Digital technologies allow data and information to come together in ways not previously possible. Internet-based technologies enable not only the scientific community but the global public at large, sometimes a mixed blessing. Open access is a fundamental principle in most biodiversity networking initiatives; however, in many areas, some agreement must be made concerning the level of detail of information which will be generally available through our networks. Locations of endangered species is an obvious example.

We must also seek concurrence on rules for the protection of privacy, intellectual property rights, and the value of information, and we must consider the impact of existing and proposed national and international law touching on these areas. In many instances, these legal and institutional issues may be greater barriers to information-sharing than the technical issues.

4. Complementary Networking Initiatives

Implementation of a global biodiversity information Commons as envisioned here is already underway through a series of nested networking initiatives. The "nest" of initiatives in which an institution participates varies from country to country or region to region. In the United States., biodiversity informatics is included in our NBII effort. NBII has been developed through collaboration among Federal, state and local governments, academic institutions, non-governmental organizations, inter-agency groups, and commercial enterprises to provide increased access to the nation's biological resources.

4.1 NBII's Role in Other Networking Initiatives

The NB1I then is the U.S. component in regional and global information networks including NABIN, IABIN, CHM, and the proposed GBIF. Each of these multi-national network initiatives had a different genesis, but all share common principles: open access to scientifically credible biodiversity information; interoperable data systems linking geographically dispersed resources; data ownership remaining with the data providers; and respect for intellectual property rights in the data.

By recognizing the shared goals and by designating the NBII as a focal point for these various initiatives, the United States seeks to ensure complementary development and elimination of duplicate effort in the creation of information, tools and policies relevant to all of these initiatives. Furthermore, through collaboration across networks, scarce network implementation resources can be leveraged to meet the objectives of more than a single initiative.

4.2 The "Species Analyst" Collaboration

An excellent example of such a multi-network initiative is the development and promulgation of Species Analyst, an Internet-based query system (search engine) which accesses dozens of databases compiled by universities, natural history museums, conservation organizations, and other groups and agencies. Species Analyst, developed by the University of Kansas, allows concurrent searching of databases of specimen information from collections located throughout the world and subsequent analysis of these data using computer applications such as Microsoft Excel and ESRI's ArcView GIS. The system will be linked to the San Diego Supercomputer Center, which will be able to perform predictive and geospatial analysis of the data. The resulting distributional maps are used to make educated guesses about the whereabouts of rare or poorly known species, or where an invasive species might gain a foothold in a new area.

The Species Analyst began as a project of NABIN, sponsored by the Commission on Environmental Cooperation. The U.S. National Science Foundation provided funding for the development of a prototype and additional enhancements. The University of Kansas and the World Bank subsequently provided funding to train IABIN participants on the tool, thereby expanding its use from North America to the Western Hemisphere. An effort is now underway to apply Species Analyst to a multi-national invasive species project. Efforts continue to recruit additional institutional participants worldwide. The Species Analyst success story shows the value of coordination and collaboration within the community.

5. The Way Ahead - Two Success Factors

New information technologies will enable us to accomplish extraordinary things by working in a distributed but coordinated information paradigm. But realization of the vision for a global biodiversity Commons is dependent upon the continuing desire of the biodiversity community to bring it into being. It is critical that the community stay together, keep growing, express its requirements with a unified voice, and develop innovative partnerships among all sectors of society to work toward this common goal.

Funding for this venture is also a critical need and must be identified soon. Although some requirements can be met through partnerships among existing efforts, there is a need for new funding targeted at priority gaps in content, technology, and infrastructure identified by the community. The PCAST report referenced earlier points out that the investment needed to create a biodiversity information infrastructure is small compared to expenditures on data gathering, and that failure to create the infrastructure will result in missed opportunities to generate new knowledge from existing data. The recent conference summary, "Weaving a Web of Wealth, Biological Informatics for Industry, Science, and Health," asserts that biological informatics can result in the generation of wealth [2]. But the investment must precede the harvest.

Although there is a growing acknowledgement in many sectors of society that a biodiversity information infrastructure needs to be funded, the debate continues on how, when, and from where this funding will appear. We must work together to articulate and demonstrate through our fledgling efforts toward a biodiversity informatics Commons the value that such a capability will provide not only to scientists and researchers but to decision-makers, educators, and the general public, those constituencies to which our potential funding sources are accountable. By supporting the efforts of the biodiversity informatics community, those most directly responsible for stewardship of the earth's natural resources support their own efforts to preserve and protect the living capital of the planet.

6. References

[1] President's Committee of Advisors on Science and Technology, PCAST Panel on Biodiversity and Ecosystems. *Teaming with Life: Investing in Science to Understand and Use America's Living Capital.* Washington, DC: Executive Office of the President of the United States, 1988.

[2] Lane, Meredith. *Biological Informatics: Weaving a Web of Wealth for Industry Science and Health.* Canberra: Australian Academy of Science, 1998.