URL: http://www.iupac.org/symposia/proceedings/phuket97/duncan.html

Information Systems and Biodiversity Databases: Trends and Challenges

Thomas Duncan

Museum Informatics Project, Information Systems and Technology, University of California, Berkeley, CA 94720

Abstract: During the last five years there has been an explosive growth of Internet information resources on biodiversity. However, biodiversity information systems have only begun to utilize the wide array of networked technologies currently available. Trends in World Wide Web (WWW) tools and services, multi-tier client server architectures, component software development environments, repository management and maintenance and image presentation will significantly change the way in which information systems are designed and implemented. These emerging technologies offer new possibilities for storing, presenting, and sharing nomenclatural, distributional, and phylogenetic data that are fundamental to understanding global biodiversity and tracking changes in it. Biodiversity information providers and systems developers must address the technical and institutional challenges such technologies offer and how these technologies will change the way in which taxonomic, biotic, and phylogenetic studies are conducted.

INTRODUCTION

The growth of the Internet and network-based information resources continues to significantly influence the way in which biodiversity information systems are developed and deployed. However, the rapid pace of technical change presents challenges to institutions and individuals that wish to develop and share biodiversity information. In this paper, I focus on how new and emerging technologies impact institutions and individuals within the museum and taxonomic communities. As the creators of and stewards for primary data on the classification and distribution of the world's biota, taxonomists and museum curators must devise more effective electronic methods to access these primary data in order to facilitate a broader understanding of biodiversity and its changes. In this paper, a brief review is presented on technology trends that will likely have a significant influence on the information technology infrastructure for biodiversity information systems. In addition to the information technology infrastructure, biodiversity information systems require a data infrastructure. Museum, taxonomic, and evolutionary

*Invited lecture presented at the International Conference on Biodiversity and Bioresources: Conservation and Utilization, 23–27 November 1997, Phuket, Thailand. Other presentations are published in *Pure Appl. Chem.,* Vol. 70, No. 11, 1998.

biologists and their institutions are beginning to develop projects to provide the data infrastructure that will facilitate biological inventories, monographic studies, phylogenetic analyses, and environmental planning and management. A number of technical and political challenges must be addressed to develop the technology and data infrastructure to support biodiversity information systems.

INFORMATION TECHNOLOGY TRENDS

The rapid growth in Internet-based technologies over the last five years has markedly changed the way in which biodiversity information resources are being developed and deployed. In this section, trends in five technology areas are outlined and implications for biodiversity information system development and deployment. Biodiversity information systems are only at the very early stages of utilizing these new technologies. It is expected that over the next few years, marked changes in the methods for presenting, accessing, and integrating biodiversity data will occur using these technologies as important components of an information technology infrastructure to support biodiversity information systems.

World Wide Web

Since the release of Mosaic in 1993, the number of web servers has grown at an unbelievable rate[1]. From approximately 200 http servers in 1993, the number of web servers now number in the millions and continuing to grow rapidly. The ease with which an individual or organization is able to place materials on the World Wide Web has enabled a wide range of biodiversity information providers to present their data and projects to a world-wide audience and for information users to more easily access current research about biodiversity[2].

The WWW however, has had limitations for database access and for the presentation of complex documents. With the development of the Extensible Markup Language (XML)[3] authoring and presenting complex documents will be possible. XML defines a class of data objects called XML documents and partially describes the behavior of computer programs that process them. XML is an application profile or restricted form of SGML, the Standard Generalized Markup Language. HTML, the Hypertext Markup Language in use to design WWW pages is also a subset of SGML but is limited in the complexity permitted in document authoring and presentation.

Although SGML has been in existence for over 20 years, its use has largely been limited to book publishers. The lack of network-based authoring and querying tools has limited its use on the WWW. Some notable examples of SGML use on the WWW are illustrated by the Encoded Archival Description[4] developed within the archival community to describe finding aids. Finding aids are documents that describe the content of an archival collection.

For database access through the World Wide Web, the introduction of the Java Database Connectivity (JDBC) standard as an extension of the Open Database Connectivity (ODBC) standard has been a significant development[5]. The use of Java and JBDC moves WWW access to databases from a stateless connection where each database request results in the creation of a new HTML page of results to a stated connection that allows greater interactivity with databases. Most taxonomic and museum database interfaces on the WWW provide basic querying capabilities but are still limited in the degree to which users can interact with these databases. Tools for Java development are just beginning to reach a mature stage.

n-Tier Client Server Architectures

Client server applications such as the WWW as well as other networked biodiversity information systems have been developed primarily as two-tier client server systems. In a two-tier client server application, both business services and the application logic can reside on either the client or the server. Most frequently in biodiversity information systems the server has only provided data access services. Due to this lack of specificity, two-tier client server applications do not foster reuse of applications, do not readily scale either in terms of numbers of users or increasing complexity, and thus are difficult to deploy and maintain. In an n-tier client server system[6], the system is divided into logical component levels called tiers. Ideally, the client handles only the presentation logic. One or more tiers are devoted to handling the business services and application logic. A separate tier handles data access services. Use of a modular multi-tier design allows applications and business rules to be managed at the server level. Modularity of design allows changes in application logic or modification of business rules without redeployment of these changes to each client system that uses the applications or business rules.

Component Software Development Environments

With the trend toward development of n-tier client server architectures, the growth of the WWW, the increasing use of object-oriented programming languages such as Java, component software models have become increasingly popular for application development[7]. A component model is a programming model that contains declarative language and application programming interfaces for the creation, execution and management of software components. A run-time software component is a dynamically bindable package of one or more programs managed as a unit and accessed through documented interfaces that can be discovered at run-time. Applications in this sense change from static pieces of compiled code to dynamic constructions composed of needed components. Changes in components do not necessarily require changes in the applications that use them. Components can be shared among applications and reused as needed. To manage components and applications that use them is provided by middleware tools such as object request brokers.

Image Compression

Presentation of high-resolution images on the WWW has been difficult due to the size of these images and the bandwidth needed to access an image. Most images are presented as compressed files where much of the detail contained in the original high-resolution images is lost in the compression process. The development of the Still Picture Interchange File Format (SPIFF) through the International Standardization Organization (ISO) provides a mechanism by which high resolution images on the WWW. Kodak has implemented this standard as the FlashPix format and Netimage as the Jpeg Tiled Image Pyramid (JTIP) format. The Internet Imaging Protocol (IIP), developed by Hewlett Packard[8] in collaboration with Kodak, Microsoft, and Live Pictures, Inc., provides access to files that conform to the SPIFF format. Pyramidal image compression one of many compression techniques being developed to address the problem of presentation of high-resolution images in a bandwidth efficient manner. ISO working groups are considering extensions of the JPEG standard to provide mechanisms for the evaluation of the content of images and the choice of an appropriate method for compression based on content.

Repository Management and Maintenance

During the last five years, digital library research has focused on the development of architectures for repositories of digital objects[9]. In this context, digital libraries are composed of digital objects (e.g.., images, databases, SGML documents, etc.) stored in repositories and retrieved from repositories through use of repository access protocols (e.g.., http, SQL, Z39.50). Digital objects contain digital data, associated metadata, and a unique identifier. Repositories have both archiving functions and access functions. Repositories require institutional management and serve projects that require long-term access to collections of digital objects

DATA INFRASTRUCTURE

In this section I outline some examples of information resources that constitute the initial attempts to present primary biodiversity data on the taxonomy and distribution of various groups of organisms. These primary data resources constitute the data infrastructure that will support the development of projects

Nomenclatural Data

Names, places of publication, and nomenclatural type information are currently available electronically for some groups of organisms. Examples include mammal species of the world[10] and Index Nominum Genericorum[11] through the Smithsonian Institution, the Gray Herbarium Index[12] of Western Hemisphere plants through Harvard University, Index Kewensis[13] through the Royal Botanical Garden, Kew. There are many other such projects underway for insects, birds, fish, diatoms, marine algae, and spiders, to name a few. Much of these data are not yet accessible electronically. Thus, there are many efforts underway to organize our existing nomenclatural knowledge. In addition, there has as yet been little effort to provide way to search and integrate these data. The facilities currently on the WWW provide for search by taxonomic name. No facilities are yet available for dynamically referencing names in a database to these standard nomenclatural references.

Museum Collections

During the last five years, museums throughout the world have begun to utilize the WWW for institutional advertisement, descriptions of programs and research activities, and for access to databases at these institutions[14]. Most museums are only beginning to present collection catalogs on the WWW. Retrospective conversion of catalog or specimen information is a time-consuming and expensive task. Maintaining these databases constitute a significant challenge of institutions that also have significant financial and human resource limitations in providing the basic curation of and space for the collections that are their responsibility. Without major new sources of funds, it is unlikely that a significant fraction of the world's holdings of the historical record of biodiversity as represented in museums will be available electronically.

Phylogenetic Data

Data and tree structures that describe the phylogenetic relationships among organisms have now been generated for a wide variety of organisms[15]. Two major projects are underway to organize these data on the WWW. The TreeBase project[16] at Harvard University is a prototype relational database being developed to manage and explore information on phylogenetic relationships. Its main function is to store published phylogenetic trees and data matrices. It also includes bibliographic information on phylogenetic studies, and some details on taxa, characters, algorithms used, and analyses performed. The database is designed to allow retrieval and recombination of trees and data from different studies, and it can be explored interactively using trees included in the database. TreeBase therefore provides a means of assessing and synthesizing phylogenetic knowledge.

The Tree of Life[17] is a project designed to contain information about the phylogenetic relationships and characteristics of organisms, to illustrate the diversity and unity of living organisms, and to link biological information available on the Internet in the form of a phylogenetic navigator. This project will provide a map to biological information that can be used by researchers, teachers, and students, and that, as it grows, it will provide a forum for those who wish to share knowledge about the diversity of earth's organisms. More than 1290 pages currently in the Tree of Life are housed in four countries. The Tree of Life is intended for individuals interested in locating information about a particular group of organisms, for biologists seeking identification keys, figures, phylogenetic trees, and other systematic information for a group of organisms, and for educators teaching about organismal diversity.

A GLIMPSE OF THE FUTURE

If the technologies discussed above are applied to the development of an information technology and data infrastructure in support of biodiversity projects it is possible to outline a glimpse of what the future would be under these conditions. One could image institutionally managed repositories of nomenclatural, specimen, and phylogenetic data, support for the electronic publishing of new taxa, monographs, and phylogenies, dynamic support for interoperation with conservation and molecular databases, an easy to use public interface, and decision support software for environmental analysis and management. This is a daunting challenge and will require much work, cooperation, funding, and trust within the taxonomic and museum community to accomplish. Some of the immediate challenges to be faced are outlined below

INSTITUTIONAL CHALLENGES

With rapidly changing technology and still limited access to primary biodiversity data, there are a number of challenges to biodiversity providers and users.

Network Access

A major challenge for globally accessible network-based biodiversity information systems is for those who wish to access these information resources to have connections to national and international networks. Depending on the part of the world in question, significant limitations remain in providing network connectivity to those who wish access to the Internet. Although most of North America and Europe, significant parts of Asia, and selected parts of Central and South America are connected, many parts of China, South America, Africa, and the countries of the former Soviet Union are still largely unconnected. One major challenge of the coming decade will be to extend Internet connectivity to reach a more significant portion of the world's population. Without access major portions of the world's population will be unable to participate effectively in the use of networked information resources such as biodiversity resources discussed here.

Prospective and Retrospective Information Systems

The data housed in museums and in the literature about the taxonomy and distribution of the world's biota are still largely inaccessible electronically. At the same time, there exists an increasing urgency to document the biodiversity in numerous poorly understood groups of organisms and areas. For the museum and taxonomic communities, this combination of factors presents a complex challenge. The design of information systems for taxonomic and distributional information to be captured from museum specimens or the literature is different from such systems to be used in a prospective sense.

Federated Databases and Database Interoperability

Although the WWW is powerful and has been employed to good use in the distribution of structured data by several major databases and projects, data integration and database interoperability is still a major problem. A major challenge in this regard is the development of federated biodiversity databases. In federated systems, participating databases maintain their databases in whatever manner they choose, but to participate in federation they make their data available over networks.,

Without a federated approach, biological databases face an increasingly complex problem of data integration. Meeting this challenge will require the development of technical and sociological processes that will allow multiple databases to interoperate functionally, while still maintaining much of their individual managerial autonomy. Horizontal partitioning of data across institutions makes the challenge of interoperability especially acute because achieving interoperability under these circumstances requires the development of semantic consistency among participating sites.

To be useful to the widest range of potential users, biodiversity information systems should be capable of functionally interoperating, at some minimum basic level, with many different information systems (such as nucleotide sequence databases, geographical place-name servers, etc.). Successful interoperation among a large and autonomous set of independent data sites can only occur if all sites use equivalent, generic tools to publish their holdings according to common protocols and syntaxes.

NOTES

1. For a history of the early development of the World Wide Web, see http://www.w3.org/History.html.
2. A general reference for biodiversity information on a global basis is found at the International Biological Information on the Internet web site at http://www.nbii.gov/iao/ibii.html

3. The specification for version 1.0 of XML is at www.w3.org/TR/WD-xml-090907.html/.
4. http://www.loc.gov/ead/ead.html
5. One of many dbms vendors developing JDBC tools for database access is Sybase at http://www.sybase.com/products/entcon/.
6. The Gartner Group provides online access to reports and forecasting on client server systems, in particular, and networked information systems, in general http://gartner6.gartnerweb.com/public/static/home/home.html.
7. Information on component software architectures, Java, object request brokers and WWW tools can be found at http://www.omg.org/ or http://www.javasoft.com/ or http://search.netscape.com/newsref/pr/newsrelease440.html.
8. For a discussion of IIP, pyramidal image formats, and Hewlett Packard imaging activities, see http://www.image.hp.com/index.html.
9. An architecture for digital library services is described by Kahn and Wilensky (http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html). This architecture has been implemented by the National Digital Library Program at the Library of Congress (http://www.dlib.org/dlib/february97/cnri/02arms1 .html) and is the basis for the development of digital library architectures for the National Digital Library Federation (http://lcweb.loc.gov/loc/ndlf/ndlfhome. html).
10. http://nmnhwww.si.edu/msw/
11. http://nmnhwww.si.edu/ing/
12. http://herbaria.harvard.edu/Data/data.html
13. http://www.rbgkew.org.uk/pitcom/pitcom-ik.html
14. A general reference on natural history museum activities on the WWW is located at the Biodiversity and Biological Collections Web Server at http://biodiversity.uno.edu/
15. A general reference for phylogenetic data on the WWW is http://www.ucmp.berkeley.edu/subway/phylo/phylodat.html#molecular
16. The TreeBase project is located at http://herbaria.harvard.edu/treebase.
17. The Tree of Life project is located at http://phylogeny.arizona.edu/tree/phylogeny.html