Biodiversity Informatics, 1, 2004, pp. 1-13

GLOBAL BIODIVERSITY INFORMATICS: SETTING THE SCENE FOR A "NEW WORLD" OF ECOLOGICAL MODELING

V. P. CANHOS, S. SOUZA, R. GIOVANNI AND D. A. L. CANHOS
Centro de Referência em Informação Ambiental (CRIA), Av. Romeu Tortima 388, Barão Geraldo 13084-791, Campinas, S.P., Brasil

Abstract.--Recent developments in information and communication technology are allowing new experiences in the integration, analysis and visualization of biodiversity information, and are leading to development of a new field of research, *biodiversity informatics*. Although this field has great potential in diverse realms, including basic biology, human economics, and public health, much of this potential remains to be explored. The success of several concerted international efforts depends largely on broad deployment of biodiversity informatics information and products. Several global and regional efforts are organizing and providing data for conservation and sustainable development research, including the Global Biodiversity Information Facility, the European Biodiversity Information Network, and the Inter-American Biodiversity Information Network. Critical to development of this field is building a biodiversity information infrastructure, making primary biodiversity data freely and openly available over the Internet. In addition to specimen and taxonomic data, access to non-biological environmental data is critical to spatial analysis and modeling of biodiversity. Adoption of standards and protocols and development of tools for collection management, data-cleaning, georeferencing, and modeling tools, are allowing a quantum leap in the area. Open access to research data and open-source tools are leading to a new era of web services and computational frameworks for spatial biodiversity analysis, bringing new opportunities and dimensions to novel approaches in ecological analysis, predictive modeling, and synthesis and visualization of biodiversity information.

Key words.—biodiversity informatics, standards, protocols, primary data, specimens

Biodiversity information is critical to a wide range of scientific, educational, and governmental uses, and is essential to decision-making in many realms. Initiatives to integrate data into viable resources for innovation in science, technology, and decision-making are being developed as local, regional, and global initiatives. A formidable challenge lying ahead is integration of these initiatives into an organized, well-resourced, global approach to build and manage biodiversity information resources through collaborative efforts.

The existence of biodiversity data resources from different fields of knowledge, available to all interested, and the strong demand to integrate, synthesize, and visualize this information for different purposes and by different end users is leading to the development of a new field of research that can be termed *biodiversity informatics*. This emerging field represents the conjunction of efficient use and management of biodiversity information with new tools for its analysis and understanding.

This field has great potential in diverse realms, with applications ranging from prediction of distributions of known and unknown species (Raxworthy *et al.*, 2003), prediction of geographic and ecological distribution of infectious disease vectors (Beard *et al.*, 2002; Costa *et al.*, 2003; Peterson *et al.*, 2002c; Peterson and Shaw, 2003), prediction of species' invasions (Peterson and Vieglais, 2001; Peterson, 2003), and assessment of impacts of climate change on biodiversity (Peterson *et al.*, 2002b; Siqueira and Peterson, 2003; Thomas *et al.*, 2004). This potential nonetheless remains largely unexplored, as this field is only now becoming a vibrant area of inquiry and study.

HISTORICAL CONTEXT

Political Framework

The set of agreements signed at the 1992 Rio de Janeiro "Earth Summit" included two binding conventions, the United Nations Framework Convention on Climate Change (UNFCCC) and the Convention on Biological Diversity (CBD). While the CBD aims at conservation and sustainable use of biological diversity, ensuring benefit-sharing, UNFCCC targets stabilization of atmospheric concentrations of greenhouse gases, primarily through negotiation of global agreements such as the Kyoto Protocol. UNFCCC is also providing guidance to governments on practical ways to adapt to changing climates. Successful implementation of both conventions is highly dependent on combined efforts of countries and international organizations, integration of distributed information systems, and deployment of biodiversity informatics.

Exchange of information "from all publicly available sources, relevant to the conservation and sustainable use of biological diversity" including "results of technical, scientific and socio-economic

research" is addressed in CBD Article 17. "Biological diversity" is defined as "the variability among living organisms from all sources including, *inter alia*, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems." To support access to integrated data ranging from genes to ecosystem diversity and information ranging from scientific and technical to socioeconomic research, CBD is implementing the Clearing House Mechanism (CHM), an Internet-based network promoting technical and scientific cooperation and exchange of information.

The many ways in which climate and biodiversity interact are leading researchers and policy-makers to the conclusion that working on those issues together will be more effective than dealing with them separately. At a practical level, this understanding will mean that initiatives emanating from UNFCCC and CBD will need to use integrated approaches as much as possible[1].

The Roots of Biodiversity Informatics

Australia has been a leading country in biodiversity informatics. Since the mid-1970s, Australian herbaria have been digitizing their data cooperatively. The Environmental Resources Information Network (ERIN), was established in 1989 to provide geographically-related environmental information for planning and decision-making. Also in 1989, HISPID, Herbarium Information Standards and Protocols for Interchange of Data, a standard format for interchange of electronic herbarium specimen information, developed by a committee of representatives from all Australian herbaria, was first published. ERIN's experience set an example for several other initiatives, such as Mexico's Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO), Costa Rica's Instituto Nacional de Biodiversidad (INBio), and Brazil's Base de Dados Tropical (BDT). INBio and CONABIO both became fully engaged in biodiversity informatics after the exchange of experience with ERIN experts in 1993.

In the early 1990s, researchers from diverse fields of expertise held meetings, and began the Biodiversity Information Network – Agenda 21 (BIN21) initiative. This group established what was called a Special Interest Network with a Virtual Library. BIN21 was set up as an informal, collaborative, distributed network consisting of a series of participating "nodes" aiming at complementing existing or planned actions. BIN21 was actively involved in discussions of the CHM to the CBD, and produced a document proposing the structure being used today, composed of focal points and thematic networks[2]. At the time, owing to technological limitations, what was envisaged was creation of directories of people, institutions, and data sources.

In 1998, a research project was launched at the University of Kansas Natural History Museum and Biodiversity Research Center: the Species Analyst (TSA). TSA's main objective was to develop standards and software tools for access to world natural history collection and observation databases. This project was one of the first networks to draw on distributed data sources from biological collections worldwide, setting an example for other initiatives, which were attracted by examples based on an associated modeling tool, the Genetic Algorithm for Rule-set Prediction (GARP), originally developed by David Stockwell (Stockwell and Noble, 1991).

Global and Regional Efforts

Several global and regional efforts are aiming at organizing data stakeholders and making data available for conservation and sustainable development research.

Global Biodiversity Information Facility (GBIF)[3].—GBIF had its genesis in a recommendation from a working group of the Megascience Forum of the Organization for Economic Cooperation and Development (OECD), although GBIF currently has no direct ties to OECD. GBIF was founded in March 2001 and participation is open to any interested country, economy, or recognized international organization that agrees to make scientific biodiversity information available. Although GBIF intends eventually to incorporate all levels of biodiversity data (molecules to ecosystems), its initial phase is focusing on species- and specimen-level information. Its work program emphasizes four priority areas: (1) digitization of natural history collection data, (2) data access and database interoperability, (3) electronic catalogue of names of known organisms, and (4) outreach and capacity building. The GBIF data portal, which connects participant nodes and other data providers, came on-line in February 2004, and by October 2004 was serving over 40 million records.

Working with its partners, GBIF is promoting development and adoption of standards and protocols for documenting and exchanging biodiversity data. GBIF has partnered with the

[1] http://www.scidev.net/Dossiers/index.cfm.

[2] http://www.bdt.org.br/bin21/wks95/cbm_doc.html.
[3] http://www.gbif.org/.

Catalogue of Life initiative to speed up development of a global authority file for the approximately 1.75 million named species on Earth, including synonyms and vernacular names. In addition, GBIF is tackling several sociological and policy issues. GBIF has developed a Memorandum of Cooperation with the CBD, and is a key promoter of the database-related components of the CBD's Global Taxonomy Initiative and Global Strategy for Plant Conservation. GBIF provides seed-money grants to help digitize specimen and observational information, has partnered with UNESCO to start a network of academic chairs in biodiversity informatics, and gives a yearly prize to a scientist who creatively combines systematics and informatics research. GBIF is holding several consensus-building workshops and is contracting white papers on topics such as intellectual property rights (Muller, 2004) and data-sharing with countries of origin (Canhos *et al.*, 2004).

The European Effort.—Several programs aimed at development of a platform for European biodiversity data are being carried out with funds from the European Commission. The Consortium of European Taxonomic Facilities (CETAF)[4] is a network of Europe's largest taxonomic facilities including natural history museums, botanical gardens, and other biological collections. Together, CETAF institutions hold very substantial zoological, botanical, paleobiological, and geological collections, and provide resources for the work of thousands of researchers in a variety of scientific disciplines. CETAF was instrumental in establishment of the *Fauna Europaea* thematic network. With over 400 contributing specialists, *Fauna Europaea* is compiling lists of more than 115,000 non-marine animal species, as well as information on their distribution, and is a complement to the European Register of Marine Species (ERMS). In the botanical area, sister projects are the European Flora and the Euro-Mediterranean Plant Diversity (Euro+Med Plant Database). Several CETAF members have received support through the European Commission's Framework Programmes to enhance transnational access to their collections, equipment, and expertise.

The European Natural History Specimen Information Network's (ENHSIN) goal is to enable the development of a shared, interoperable infrastructure of natural history specimen databases in European institutions[5]. BioCASE[6] the Biological Collection Access Service for Europe, is a web-based information service providing researchers with unified access to biological collections from 35 institutions in 30 European countries plus Israel, leaving control of the information with the collection holders. The European Network for Biodiversity Information (ENBI)[7] began in January 2003, and is a three year thematic network project funded by the European Union. ENBI is the European contribution to the GBIF. Currently, the network has 66 members from 24 European countries, including the GBIF national nodes of member countries.

Inter-American Biodiversity Information Network (IABIN)[8].—IABIN is an Internet-based forum for technical and scientific cooperation that seeks to promote coordination among Western Hemisphere countries in collection, sharing, and use of biodiversity information relevant to decision-making and education. IABIN projects include development of a network for sharing invasive species information, a survey of New World specimen holdings in European collections, and creation of a gateway to the information resources of IABIN participants. Ongoing IABIN efforts include establishing cooperative linkages with national, regional, and global initiatives, and developing projects that build capacity in participating countries and institutions.

A Global Environment Facility (GEF) grant funded development of a plan for "Building the Inter-American Biodiversity Information Network." According to this plan, in the next five years, IABIN will work on interoperability and data access; data content creation; and information products for decision-making. Cross-cutting issues include standards and protocols to ensure compatibility of diverse data sources, and agreement as to which data models to use. An IABIN catalogue service will be developed by the United States' National Biodiversity Information Infrastructure (NBII), and the project implementation plan also recommends development of thematic networks on specimens, species, ecosystems, invasive species, pollinators, and protected areas. The IABIN Secretariat is being established at the City of Knowledge Foundation in Panama.

BUILDING THE BIODIVERSITY INFORMATION INFRASTRUCTURE

Most biodiversity informatics initiatives are focusing on species and specimen data as the first necessary information component of a global comprehensive data network on biodiversity. Nonbiotic environmental data and ecological data

[4] http://www.cetaf.org/.
[5] http://www.nhm.ac.uk/science/rco/enhsin/.
[6] http://www.biocase.org/.
[7] http://www.enbi.info/forums/enbi/index.php.
[8] http://www.iabin.net/.

are increasingly being used in biodiversity informatics for modeling distribution patterns of species and populations, and therefore also need to be addressed.

Primary Biodiversity Data

Specimen collections are the primary research archives documenting biological diversity on Earth. The 2.5-3.0 billion specimens available in biological collections worldwide document identities, habitats, histories, and spatial distributions of the roughly 1.75 million described species of life. The specimen vouchers and associated information provide a fundamental resource for biological systematics. Return on investments made during 250 years of global biological inventories can be realized dramatically through digitization and integration of information about species and specimens. Less than 10% of worldwide specimens are available in the electronic domain (Krishtalka and Humphrey, 2000).

Species Information

Access to consistent, scientifically credible taxonomic information is essential to many activities, including natural resource and waste management for sustainable use, environmental monitoring, regulation, and biotechnology development. Storage and retrieval of biological data requires high-quality, well-documented, and continuously updated sources of taxonomic information. A basic concept in databasing is the use of controlled vocabularies to assure that one can find, relate, and retrieve information that refers to a particular class of objects. This seemingly simple concept, of course, may become a nightmare when databasing specimen data. As such, prompt access to digital authoritative data on taxonomic and vernacular names becomes critical.

For broader applications, information systems integrating disparate biodiversity databases worldwide need a core dictionary of names for mediating queries across datasets—the species name may often be the only field in common between databases. The niche for nomenclatural data and taxonomic authority files has brought numerous taxonomic databases onto the Internet, including local fauna and flora checklists, and "taxon-by-taxon" databases such as the Integrated Taxonomic Information System (ITIS)[9] and the UNESCO-IOC Register for Marine Organisms (URMO)[10] (Bisby, 2000). To accelerate creation of a comprehensive index of all of life, in June 2001, two large initiatives, Species2000[11] and ITIS, joined forces and formed the Catalogue of Life (CoL) Program. The goal of this effort is gathering the ~1.75 million of species names thought to exist to serve the world's need for names information. CoL is composed of synonymic species lists prepared by taxonomic experts in various groups of organisms.

CoL is envisaged to become a common standard for referencing biodiversity data of all sorts, including the World Catalogue of Plants. CoL is being prepared in two forms—an annual checklist published as a new edition yearly, and a dynamic checklist provided as an online, distributed resource based on contributing databases in real time. The annual checklist has been available since 2001 on the Internet and on CD-ROM, and is now widely used in biodiversity data portals around the world. The fourth CD-ROM release of the annual checklist contains a searchable database of more than 800,000 scientific and common names (nearly 342,000 species and subspecies) of viruses, microorganisms, protists, fungi, animals, and plants (Bisby *et al.*, 2004). Beside the major task of accelerating integration of taxonomic-sector databases, major efforts are being invested in development of the Species2000 meta-database, peer review and quality control programs, and the software support program (e.g., new versions of SPICE[12] for Species2000, and experimental versions of LITCHI[13] software for monitoring taxonomic disparities).

Non-biological Environmental Data

Efforts to improve understanding of environmental patterns, their variability, their changes over time, and their implications for human welfare and decision-making, depend critically on the quality, accessibility, and usability of diverse environmental and related social science data. A key scientific and technical challenge, therefore, is improving access to existing and emerging sources of environmental, biological, and socioeconomic data, and to improve integration of these data in support of disciplinary and interdisciplinary research efforts and applications, and related policy-making initiatives (Canhos *et al.*, 2004). Terrestrial environmental data fall into three basic categories: terrain, climate, and substrate, and all are fundamental in ecological niche modeling. Terrain refers to surface morphology and includes parameters such as elevation, slope, and aspect. Climate data summarize patterns and variation in atmospheric characteristics, and substrate data

[9] http://www.itis.usda.gov/.
[10] http://www2.eti.uva.nl/database/urmo/default.html.
[11] http://www.species200.org/.
[12] http://spice.sp2000europa.org/SPICE/.
[13] http://litchi.biol.soton.ac.uk/.

include soils, lithology, surface geology, hydrology, and landforms. Climate change projections constitute another key suite of environmental datasets. For many regions, they are only available at global scales (Intergovernmental Panel on Climate Change, IPCC[14]), and hence may not have the precision required for modeling at local scales. Integrating regional climate models (RCMs) may represent a promising frontier to be explored.

Demand for environmental data by the biological community involved with ecological niche modeling is quite recent. It is important to enable open access to data by sharing data formats and by developing open-source tools for data conversion, visualization, and analysis. Challenges inherent are in managing data resources effectively for optimal access and use, and for developing rational rules and structures for that process. It is also important to address technical aspects concerning interoperability of environmental data across software and hardware systems, and, in the case of niche modeling, to develop tools for automated dataset preparation.

Preparation of environmental layers ranks among the most time-consuming and computer-intensive areas of the informatics enterprise. An interesting initiative focused on delivery of detailed, worldwide climate data is WORLDCLIM[15], and data are publicly and freely available. These layers are at scales ideal for species' modeling for both local area (30") and continental (2.5' and 5') analyses, and represent a major improvement over layers previously available.

Promotion of scientific cooperation and partnerships between researchers and institutions working with biological and non-biological environmental data is just beginning, and is expected to expand in the near future. This interaction is essential for better understanding of the environment to enable sustainable economic and social development.

Specimen and Strain Data Networks

Numerous specimen data networks are being developed to make this primary information freely and openly available. A particularly exciting development is the Distributed Generic Information Retrieval (DiGIR[16]) protocol that is being adopted by networks around the globe, including the former *Species Analyst* network, *species*Link in Brazil, and GBIF. Many regional and local initiatives are organizing specimen data and making it openly available on the Internet. In general, these systems are converging into a global data retrieval system, using compatible protocols.

The *Species Analyst* (TSA)[17], launched in 1998 as a prototype network initially based on Z39.50 information transfer protocol evolved into several taxon-based networks. These networks, in general funded by the National Science Foundation (NSF), include FishNet[18] (29 ichthyological collections), Mammal Networked Information System (MaNIS[19]; 17 North American mammal collections), and HerpNET[20] (37 North American herpetological collections). This network serves approximately 65 million records of specimens in 120 institutions; in the near future, the Ornithological Information System (ORNIS[21]; North American ornithological collections, including vocal recordings, egg and nest collections) will contribute nearly 4 million more specimen records. The entire TSA network is most recently shifting to a DiGIR platform for information transfer.

The *World Biodiversity Information Network* (REMIB)[22] was created in accordance with the Declaration of Oaxaca (November 1993) that established a Mexican Biological Information Network to interconnect biological data banks in the country. REMIB is known today as the World Biodiversity Information Network, serving specimen information not only from Mexico, but from 140+ countries. The network serves more then 6 million specimen records, and is based on a proprietary protocol.

The *European Natural History Specimen Information Network* (ENHSIN)[23], is a thematic network funded by the European Commission. Its aim is to enable development of a shared, interoperable infrastructure of natural history specimen databases in European institutions. Seven European institutions compose the initial phase of the project: Natural History Museum, London; Royal Botanic Gardens, Kew; Zoological Museum, University of Copenhagen; Museo Nacional de Ciencias Naturales, Madrid; Muséum National d'Histoire Naturelle, Paris; Botanischer Garten und Museum Berlin-Dahlem; and Universiteit van Amsterdam.

Common Access to Biotechnological Resources and Information (CABRI[24]) is an online service of

[14] http://www.ipcc.ch/.
[15] http://biogeo.berkeley.edu/worldclim/worldclim.htm.
[16] http://digir.sourceforge.net/.
[17] http://speciesanalyst.net/.
[18] http://habanero.nhm.ku.edu/fishnet/.
[19] http://elib.cs.berkeley.edu/manis/.
[20] http://herpnet.org/.
[21] http://www.specifysoftware.org/Informatics/informaticsornis/.
[22] http://www.conabio.gob.mx/remib_ingles/doctos/remib_ing.html.
[23] http://www.nhm.ac.uk/science/rec/enhsin/.
[24] http://www.cabri.org/.

European Biological Resource Centre (BRCs) catalogs. The 9 CABRI centers include 26 collections covering most of known diversity of human and animal cells, bacteria, fungi, yeasts, plasmids, animal and plant viruses, and DNA probes. Their holdings total nearly 140,000 deposits (100,000 of which are currently included in electronic catalogues), representing approximately half of deposits in Europe [European Culture Collection Organization (ECCO) membership, 1995]. An interesting feature was the definition of the minimum, recommended, and full data set for each type of organism. Brazil subsequently adopted a modification of this dataset definition for its information system of biotechnological collections, SICol[25].

The *Australian Virtual Herbarium* (AVH[26]), is a collaborative project of the Australian botanical community, providing integrated on-line access to herbarium datasets. In addition to making available taxonomic, distribution, and occurrence information associated with all 6.5 million herbarium specimens through a simple online GIS application, the ultimate aim is to provide a complete, integrated flora information system as a tool for scientific research, environmental decision-making, and public information. AVH is based on a distributed, heterogeneous system, so that the data reside with and are managed and controlled by the collections custodians. Each herbarium has a portal to receive requests and deliver data from its institutional database(s). A shared AVH query interface at each herbarium polls all participating herbaria and delivers to users a single integrated result set. The design philosophy of the AVH is based on information standards developed by and for the botanical community, open architecture, public domain software, and free availability of the application and information management structures for use in other projects.

The *Brazilian Biota/FAPESP Virtual Institute of Biodiversity Program*[27] was officially launched in 1999, and aims to collect, organize, and disseminate biodiversity information regarding São Paulo State, defining mechanisms for the conservation and sustainable use of biodiversity. The program funds nearly 50 projects, involving approximately 400 professional scientists and an equal number of students working at research institutions both locally and abroad. Major components of the integrated information system include *Sin*Biota[28], a system developed to support geo-spatial analysis and visualization of observational data, and *Species*Link[29], a distributed information system integrating primary specimen data from biological collections. The project is also developing tools to help visualize, clean, and use these data sets[30].

TRENDS AND DEVELOPMENTS IN BIODIVERSITY INFORMATICS

Biodiversity data are generated by thousands of researchers worldwide. The result is a global distributed array of heterogeneous biodiversity information resources. Integrating these data resources has been a goal of biodiversity informatics practitioners for more than 10 years. Following developments in information and communication technology, integration of distributed databases can now be virtual and dynamic, processed "on-the-fly," according to users' needs.

Standards and Protocols

To integrate data from distributed sources in real time, not only is technology necessary, but standards and protocols are essential. The International Union of Biological Sciences' Taxonomic Database Working Group (TDWG[31]) has been working on biodiversity data standards since 1994. One of TDWG's task groups, in a joint initiative with the Committee on Data for Science and Technology (CODATA[32]), is working on a standard called Access to Biological Collection Data (ABCD[33]). ABCD aims to define global formats for data exchange and retrieval from diverse biological collections. In practical terms, a highly structured XML schema is defined in agreement with representatives from different projects and communities. The current version contains reusable data types and specific elements related to each biological discipline, ranging from paleontological collections to collections of living organisms. Its impressive number of elements and the intricate structure provides a fairly complete representation of the biodiversity data universe.

More recently, distributed database networks of biological collections have taken a different approach. The Darwin Core[34] defines a simpler set of fields common to all taxonomic groups. The result is a simple XML schema with a small number of elements covering basic and essential information. Its simplicity allows straightforward

[25] http://sicol.cria.org.br.
[26] http://avh.calm.wa.gov.au/.
[27] http://www.biota.org.br/.
[28] http://sinbiota.cria.org.br/.
[29] http://splink.cria.org.br/.
[30] http://splink.cria.org.br/tools.
[31] http://www.tdwg.org/.
[32] http://www.codata.org/.
[33] http://www.bgbm.org/TDWG/CODATA/default.htm.
[34] http://speciesanalyst.net/docs/dwc/.

configuration and easier interfaces, although it does not satisfy the needs of *full*, detailed data exchange. Darwin Core was conceived as a foundation for functional data exchange in the DiGIR and Species Analyst projects. Other schema extensions were originally planned in a modularized strategy to accommodate broader data sets and full information transfer.

ABCD and Darwin Core are both federated schemas that need protocols to be used in networked systems. DiGIR is an XML-based protocol capable of working with configurable federated schemas. Use of generic query elements and customizable record structures enables not only decoupling from semantics, but also the possibility of wrapper software encapsulating heterogeneous data providers.

DiGIR was a project of the University of Kansas Natural History Museum and Biodiversity Research Center, California Academy of Sciences, and Museum of Vertebrate Zoology in Berkeley. The main motivation was to replace the Z39.50 protocol, and to unify diverse networks in a single technology. Part of its strategy was the use of open standards and protocols (e.g., HTTP, XML, and UDDI). A collaborative environment was set up following an open-source development model. DiGIR has been adopted by several distributed networks, including GBIF, MaNIS, OBIS, and *species*Link, but its original inability to work with a completely independent XML-federated schema (e.g., ABCD) has lead to a derivation of the protocol.

BioCASE is also implementing a networked biological collection information system for Europe, consisting of 31 national nodes providing metadata on collections and collection databases through daily XML-based updating mechanisms. These metadata are used to populate a database constituting a registry for clients accessing unit-level databases. BioCASE is using the ABCD XML schema as an exchange format for collection information. These two protocols (DiGIR and BioCASE) evolved separately, and at the moment have significant differences. A recent collaborative initiative involving GBIF and TDWG is working toward development of a single standard protocol.[35]

Collection Management Tools

During the last decade, several software packages have been developed to aid in digitization and data management for natural history collections. Many of these packages lack controlled vocabularies, and most fall short in collating region- or ecosystem-specific information (Chavan and Krishnan, 2003). To assess available software solutions for collection management and digitization for natural history collections institutions, GBIF recently contracted the Botanic Garden and Botanical Museum, Berlin-Dahlen, Germany, to carry out a survey. The report provides a comparative analysis of commercially available and free-of-charge software solutions to capture, organize, and manage specimen data. Elements of analysis included availability, cost, limitations to scalability, computing platform limitations, and data import/export capabilities (Berendsohn *et al.*, 2003[36]).

Georeferencing Tools

Georeferencing biodiversity data is absolutely necessary to biogeographic studies. The time and cost of geocoding large museum and herbarium datasets at first seems prohibitive to many institutions. However, according to Soberón and Peterson (2004), 70-80% of specimen label data can be georeferenced via simple techniques using convenient, automated online gazetteers. Remaining localities may either prove impossible to reference or may be feasible only via participation of experts familiar with the actual collectors.

BioGeoMancer[37] is a georeferencing tool specifically designed for text-to-coordinate conversion of locality data. It currently encompasses natural language processing (geo-parsing) to interpret descriptive localities, place-name lookup to register localities with known geographic coordinates, and ambiguity analysis to self-document uncertainties in resulting geographic descriptions. Results can be returned in HTML, XML, or as a graphic map. Its development is the result of a partnership between the University of Kansas Natural History Museum and Biodiversity Research Center, Peabody Museum of Natural History at Yale University, and Museum of Vertebrate Zoology in Berkeley. This initiative marks progress, not only in addressing the volume of georeferenced specimen data, but also in understanding the fundamental challenges of georeferencing. The spatial results obtained by this automated tool are comparable in quality to those done by trained specialists.

BioGeoMancer provides a significant step towards automated geo-referencing. Next versions will include a web services interface using the Simple Object Access Protocol (SOAP), which will

[35] http://www.cria.org.br/protocols/newprotocol.pdf.

[36] http://circa.gbif.net/irc/DownLoad/.

[37] http://www.biogeomancer.org/.

enable cross-platform interoperability between applications. Having such an interface will make it possible for existing applications to easily integrate automated georeferencing services.

Data-cleaning Tools

Use of biodiversity data in biogeographic studies has imposed an extra focus on issues of data quality. When digitizing specimen data, data quality considerations include errors in taxonomic identification, geocoding, and in the transcription process itself. Errors are common and are to be expected, but cannot be ignored. Good understanding of errors and error propagation can lead to active quality control and management improvement (Chapman, 2004).

The heterogeneous origin of the distributed biodiversity databases makes quality control even more important. Procedures must be used to detect and flag errors or potential errors. This problem becomes even more crucial when one considers all digitization processes being carried out by most museums and herbaria in the world, with legacy data covering the past 100-300 years. Historical data cannot be replaced by new surveys even if necessary funds were available, owing to loss of biodiversity and habitat changes and the unique nature of each organism that constitutes a specimen. Geocoding historical data can also be very complex, and is an additional potential source of errors (Soberón and Peterson, 2004).

Emerging web-based tools for validating georeferences, taxonomic identifications, and collection dates (or at least flagging records with high probabilities of error) are leading to development of complex automated data-validation tools. The need for tools capable of detecting geographic or ecological outliers, incorrectly georeferenced localities, and misidentified specimens, such that doubtful records are flagged for later checking by examination of specimens, is great. New tools will provide users with summary files that can easily be linked with master collections databases to update database records. The *species*Link and ORNIS projects are developing a number of data cleaning tools that are being tested and evaluated by scientific collections (Canhos *et al.*, 2004).

Modeling Tools

In general, existing biodiversity data do not provide sufficient coverage for direct, detailed environmental decisions. Modeling—or some inferential step—is thus needed for identifying and filling data gaps, planning future research, assessing conservation priorities, and providing information for environmental decisions. Modeling ecological niches for prediction of geographic distributions of species (Peterson *et al.*, 2002a; Soberón and Peterson, 2004) is a growing field in large-scale ecology and biodiversity informatics.

The general idea of ecological niche modeling (ENM) involves use of biodiversity information (species' occurrences) to characterize species' ecological requirements (the ecological niche). These models, which are in the middle stages of exploration and understanding, provide ability to integrate across diverse spatial and temporal scales, and can be applied to understanding diverse features of species' distributional ecology, including responses to environmental change, potential impacts on human economy, and public health. The field is clearly in its infancy, and yet already shows enormous potential to yield new classes of understanding.

Museum and herbarium data from specimens that traditionally have been used for taxonomic studies now present themselves as the basis for biogeographic studies. Integrating primary biodiversity data with environmental information via ENM, researchers can predict impacts of global climate change on terrestrial and marine biodiversity (Siqueira and Peterson, 2003; Thomas *et al.*, 2004); conservationists can find gaps in networks of conservation reserve systems (Rodrigues *et al.*, 2004); and agricultural researchers and health specialists can analyze insect collections to predict the timing and spread of pests and diseases (Peterson, 2003).

Many modeling tools and techniques can be used for ENM, such as BIOCLIM (Nix, 1986), generalized linear models (GLM; Austin *et al.* 1994), generalized additive models (GAM; Yee and Mitchell, 1991), regression and classification tree analyses (CART; Breiman *et al.* 1984), genetic algorithms (Stockwell and Peters, 1999), and artificial neural networks (ANN; Olden and Jackson 2002; Pearson *et al.* 2002), among others. Which method one uses may depend on the number of points available, type of environmental variables, availability of absence data, purpose to which the model is going to be put, and personal preferences and experience (Chapman, 2004).

Comparisons between techniques do exist (Thuiller, 2003; Manel *et al.*, 1999b), but are hard to accomplish. Apart from explicit differences among algorithms, each algorithm is usually implemented by a different tool which has its own restrictions regarding data input and output. When performing comparisons, one should preferably ensure that each algorithm runs under the same conditions and using the same input. However, recent efforts are

developing generic frameworks to support development and testing of modeling algorithms. BIOMOD (Thuiller, 2003) and openModeller[38] have followed this approach. BIOMOD provides four techniques to predict spatial distributions (GLM, GAM, CART and ANN), and includes accuracy testing for each result. OpenModeller is an open source library being developed as part of the *species*Link project. It is entirely based on open source software to accomplish tasks like reading different map file formats, converting between coordinate systems, and performing calculations. The current package includes several ENM algorithms (BIOCLIM, Climate Space Model, GARP, and Euclidean distance techniques), and includes a SOAP and a command line interface, and a desktop inteface is available as a plugin for the QuantumGIS project.

These initiatives are providing researchers with the required tools to compare modeling methodologies easily, and to spend more time analyzing and interpreting results. Algorithm developers can concentrate on algorithm logic when using frameworks that take care of handling input data and making projections. Moreover, in the near future, generic libraries like openModeller will be able to perform tasks in a distributed fashion, including running analyses separately in remote cluster processors via web services or Grid paradigms.

Web Services and Computational Frameworks

As the Internet grows in scope, technology, and speed, emerging technological concepts such as Web services and grid computing are changing the way that software development is done, as they allow sharing of computing power and software know-how across the global community. Web services enable application-to-application interaction, making functions written to deal with specific tasks available to programs running on any machine connected to the network. This technology can play an important role in the area of biodiversity software production in a cooperative way, as it makes existing know-how available to other groups.

If a group develops, for example, a tool to produce maps efficiently and attractively, it can make this result available on the Internet as a service for others needing to produce maps with their own data, but having no expertise in doing so. The service can be easily and transparently incorporated in the new tool. CRIA is presently developing a service for map production[39] based on the University of Minnesota's MapServer[40]. Another good example would be a Web service that answers questions about names of organisms—their correct spelling, validity, synonymy, etc. Incorporating such services in collections management software, for example, many collections' databases would benefit from standardized and clean nomenclature. The Web services approach makes such infrastructure scalable and available to both large and small institutions.

Integration of disparate data sets has led to qualitative advances in understanding geographic and ecological patterns in biodiversity. This initial integration has been achieved "by hand," but more efficient solutions are now becoming available. An important initiative based on the idea of an analytical framework is the Science Environment for Ecological Knowledge (SEEK[41]). Its goals are to make fundamental improvements in how researchers gain global access to ecological data and information; locate and utilize distributed computational services; and exercise powerful new methods for capturing, reproducing, and extending the analysis process itself. The project involves a multidisciplinary team of computer scientists, ecologists and technologists from the international, multi-institutional Partnership for Biodiversity Informatics (PBI). SEEK and related projects are working on a next generation of analytical tools, which will provide visual and automated environment in which researchers can build their own scientific workflows by selecting and connecting specific components. Semantic mediation will be key in determining which components can be used in each situation.

BiodiversityWorld[42] is another initiative based on emerging technologies. It is a three-year e-Science Pilot Project funded by the Biotechnology and Biological Sciences Research Council (BBSRC) to create a Grid-based problem-solving environment for studying biodiversity. Grid computing is a form of networking, but unlike conventional networks that focus on communication among devices, grid computing harnesses unused processing cycles of all computers in a network for solving problems too intensive for any stand-alone machine. BiodiversityWorld will also provide new analytical tools making use of resources connected to the Grid environment. Planned analyses include case studies for bioclimatic modeling with climate change scenarios, assessment of biodiversity richness, and phylogenetic and biogeography researches.

[38] http://openmodeller.sourceforge.net.
[39] http://www.cria.org.br/mapcria/.
[40] http://mapserver.gis.umn.edu/.
[41] http://seek.ecoinformatics.org/.
[42] http://www.bdworld.org/.

CHALLENGES AND OPPORTUNITIES

Substantial increases in computing capacity are enabling vast quantities of digital data to be put to use for multiple research purposes by many institutions. Technological developments are enabling exchange and integration of data and information systems, and are promoting a new framework for international collaboration and cooperation. Optimal international exchange of data, information, and knowledge will contribute decisively to advancement of scientific research and innovation in biodiversity informatics. Fostering broader access, open access, and wide use of biodiversity research data will enhance the quality and productivity of biological science systems worldwide. Open and unrestricted data access will promote scientific progress, facilitate training of researchers, and maximize value derived from public investments in data collection and archival efforts. Although legal, technical, and cultural restrictions exist and must be discussed and overcome, the challenge is enormous and opportunities are manifold.

Many challenging topics have already been addressed in this paper, such as adoption of common standards and protocols, and development of a global biological name service. Others are equally important, and are addressed below.

Open Access to Digital Data

Research advances depend on availability of diverse and rich databases from multiple public and private sources, and their openness to easy recombination, search, and processing. The overall principle is that full and open exchange of scientific data—the "bits of power" on which the health of the scientific enterprise depends—is vital for advancing progress and maximizing social benefits accruing from science worldwide (CODATA, 1997). Intellectual property laws in most countries have never allowed protection of data, and countries like the United States even have laws that specify that government data are in the public domain. Indeed, although a long sociological tradition exists among scientists to share and disseminate data, great pressures to protect data nonetheless exist (see, e.g., Directive 96/9/EC[43] of the European Parliament and of the Council of 11 March 1996 on legal protection of databases). Some in the field even among the scientists still consider data a source of potential revenue to be exploited, rather then a public good to be shared. Recently at a meeting entitled "Science, Technology and Innovation for the 21st Century"[44], OECD ministers recognized the value of sharing publicly-funded research data, and adopted a declaration entrusting the OECD to work towards commonly agreed principles and guidelines on access to research data from public funding.

GBIF has recently held a Meeting of Experts to discuss biodiversity data, databases, and intellectual property rights. A white paper was produced, with several recommendations, including that a policy of making data openly accessible to all, and in this way addresses the issue of data repatriation in the most positive manner (Muller, 2004). A GBIF study on primary biodiversity data-sharing with countries of origin (Canhos *et al.*, 2004) concluded that proper attribution, custodianship (i.e., each contributing museum retains ownership of its records), acknowledgement, and control of data delivery can be much more important to biological collections than considerations of intellectual property rights.

The major restrictions to open data-sharing are now coming mainly from developing countries, based on interpretations of CBD terms regarding access and benefit-sharing (Chavan and Krishnan, 2003). Nevertheless, moves to restrict access to primary biodiversity information will hurt developing countries more than others. For historical reasons (collecting expeditions, museum facilities, technological developments), primary datasets, both biological and environmental data are housed mainly in developed countries.

The existence of sensitive data cannot serve as an excuse for broad withholding of data. Vast amounts of biodiversity information are not sensitive, and can be shared to the benefit of all. It is important to document the benefits of sharing data to scientists and to institutional administrators and policy makers. Without access to primary biodiversity data, scientific studies carried out on regional or global scales like the extinction risk assessment carried out by Thomas *et al.* (2004) would not be possible.

Capacity Building and Outreach

Incorporation of recent advances in biodiversity informatics in research and maintenance activities is still restricted to a relatively few institutions around the globe. The global impact of deployment of the expanded data infrastructure and emerging tools is yet to be seen. Capacity building is not an easy issue, and yet is fundamental in consolidating this emerging field of knowledge, not only in developing countries and economies in transition, but also in industrialized countries. In addition to development of innovative mechanisms of training young scientists, such as the GBIF-UNESCO Biodiversity Informatics Chairs, special international programs

[43] http://www.legaleuropei.org/corgiueu/database.htm.
[44] http://www.oecd.org/document/.

are needed to address implementation of large-scale research projects involving development and consolidation of biodiversity informatics programs.

Long-term Archiving

Several important reasons exist for preserving and archiving scientific data. An important concept is that knowledge creation is a cumulative process. Science is based on hypotheses, and new hypotheses may change the relative importance of existing data. Therefore, availability of data for re-analysis and re-use is fundamental.

Data archiving includes the practices and procedures that support collection, long-term preservation, low-cost access to, and dissemination of, science and technology data (CODATA, 2002). Long-term preservation of digital data presents a variety of challenges. The more obvious technical challenges include the constant evolution of hardware and software, and the risk of systems becoming obsolete. On the other hand, positive developments include declining storage costs and developments in data management technology. Understanding that access to reliable scientific data includes both attention to acquisition of data and preservation and archiving of scientific data, CODATA established a task group on "preservation and archiving of scientific and technical data in developing countries." The main purpose of this group is to develop best-practices guidelines on preserving and archiving scientific data.

According to Hodge and Frangakis (2004), organizations are focused on capturing and acquiring digital information, rather than on preservation or permanent access. While many institutional repositories are committed to long-term preservation and access, the technical and metadata aspects required are not yet well incorporated into their systems. Open standards being developed for interoperability hold promise as a basis for preservation formats. Open formats are working toward hardware and software independence, and the potential for using these formats for preservation should be investigated further. Partnerships will be increasingly important as they have the benefit of providing some measure of redundancy, sustainability, and sharing of costs of preservation.

LOOKING AHEAD

As anticipated by Krishtalka and Humphrey (2000), substantial developments are being consolidated in the field of biodiversity informatics (Soberón and Peterson, 2004). Science is growing in size and complexity, and is becoming more cooperative and cumulative. More and more, the global science system is involving larger and more interdisciplinary teams of scientists, and is using larger instrumental facilities in all areas of knowledge. As information processed becomes more precise and systematically structured, new developments in information and communication technologies (ICT) are playing significant roles in science and innovation. Use of ICT is broadening the scope and scale of science, and is bringing new opportunities for international collaboration (Schroeder, 2003)

Similar to the impact of genomics and proteomics in the development of suborganismal bioinformatics, the huge amount of digital primary biodiversity data being released will have a tremendous impact in the development of biodiversity informatics. Biodiversity data access through new software tools, web services, and architectures will bring new opportunities and dimensions to novel approaches in ecological analysis, predictive modeling, and synthesis and visualization of biodiversity information.

Innovation and infrastructure developments will greatly reduce long-term data capture costs in the broader biodiversity community. Modular, configurable, open-source Web services will provide interoperability and scalability in distributed environments. By wrapping image processing, image-to-text conversion, and data markup capabilities into distributed, interoperable web services, greater efficiency, portability, and scalability will be achieved. It is expected that before the end of this decade, worldwide natural history collections will be contributing hundreds of millions specimen records into Internet-accessible data servers. Good scientific information is fundamental for sound environmental decision-making, and design of mechanisms to link scientific research to the decision-making process is no easy matter (Reid, 2004). Biodiversity informatics will directly benefit environmental education programs, resource management, conservation, and biomedical and agricultural research.

Development of interfaces with global environmental initiatives will be fundamental to promote coordination and avoid duplication of efforts. In July 2003, the Earth Observation Summit was held at Washington, D.C. (USA), with the goal of promoting development of a comprehensive, coordinated, and sustained Earth observation system among governments and the international community to understand and address global environmental and economic challenges. As an immediate result, an *ad hoc* Group on Earth Observations (GEO) was established to prepare a 10-year implementation plan for building such a system.

REFERENCES

Austin, M.P., A.O Nicholls, M.D., Doherty and J.A Meyers. 1994. Determining species response functions to an environmental gradicnt by means of a beta function. J. Veg. Sci. 5:215-228.

Beard, C.B., G. Pye, F.J. Steurer, Y. Salinas, R. Campman, A.T. Peterson, J.M. Ramsey, R.A. Wirtz and L.E. Robinson. 2002. Chagas disease in a domestic transmission cycle in southern Texas, USA. Emerging Inf. Dis. 9:103-105.

Berendsohn, W., A.Güntsch and D. Röpert. 2003. Survey of existing publicly distributed collection management and data capture software solutions used by the world's natural history collections. Global Biodiversity Information Facility, Copenhagen.

Bisby, F.A. 2000. The quiet revolution: Biodiversity informatics and the Internet. Science. 289:2309-2312.

Bisby, F.A., R. Froese, M.A. Ruggiero and K.L. Wilson. 2004. Species2000 & ITIS Catalogue of Life: Indexing the world's known species (CD-ROM). Species2000. Los Baños, Philippines.

Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone. 1984. Classification and regression trees. Chapman and Hall, New York.

Canhos, D.A.L., A.D. Chapman and V.P. Canhos. 2004. Study on data-sharing with countries of origin. Global Biodiversity Information Facility, Copenhagen.

Canhos, D.A.L, P. Uhlir and J.M. Esanu (editors). 2004. Access to Environmental Data: Summary of an Inter-American Corkshop. Committee on Data for Science and Technology, Paris.

Canhos, V.P., D.A.L. Canhos, S. Souza, M.F. Siqueira, M. Muñoz, R. Giovanni, A. Marino, I. Koch, R.L. Fonseca, C.Y. Umino, B. Cruz and A.P.S Albano. 2004. Sistema de informação distribuído para coleções biológicas: A integração do *Species Analyst* e *Sin*Biota. Relatório Técnico Anual. FAPESP, São Paulo, Brazil.

Chapman, A. D. 2004. Technical Report, March 2003-2004. Biota/FAPESP, Centro de Referência em Informação Ambiental, Campinas, Brazil.

Chavan, V. and S. Krishnan. 2003. Natural history collections: A call for national information infrastructure. Cur. Sci. 84:34-42.

CODATA. 1997. Bits of power: Issues in global access to scientific data. National Academy Press, Washington, D.C.

CODATA. 2002. Workshop on Archiving Scientific & Technical (S&T) Data. Pretoria, South Africa. Committee on Data for Science and Technology, Paris.

Costa, J., A.T. Peterson and C.B. Beard. 2002. Ecological niche modeling and differentiation of populations of *Triatoma brasiliensis* Neiva, 1911, the most important Chagas disease vector in northeastern Brazil (Hemiptera, Reduviidae, Triatominae). Am. J. Trop. Med. Hyg. 67:516-520.

Hodge, G. and E. Frangakis. 2004. Digital preservation and permanent access to scientific information: the state of the practice. A report sponsored by International Council for Scientific and Technical Information (ICSTI) and CENDI US Federal Information Managers Group.

Krishtalka, L. and P.S. Humphrey. 2000. Can natural history museums capture the future? BioSci. 50:611-617.

Manel, S., J.M. Dias and S.J. Ormerod. 1999. Comparing discriminant analysis, neural networks, and logistic regression for predicting species distributions: A case study with a Himalayan river bird. Ecol. Mod. 120:337-347.

Muller, M. R. 2004. An analysis of the implications of intellectual property rights (IPR) on the Global Biodiversity Information Facility (GBIF). Global Biodiversity Information Facility, Copenhagen, Denmark.

Nix, H.A. 1986. A biogeographic analysis of Australian elapid snakes. Pp. 4-15 *in* Atlas of Elapid Snakes of Australia (R. Longmore, ed.). Australian Government Publishing Service, Bureau of Flora and Fauna, Canberra.

Olden, J. D. and D. A. Jackson. 2002. Illuminating the 'black box': A randomization approach for understanding variable contributions in artificial neural networks. Ecol. Mod. 154:16.

Pearson, R.G., T.P. Dawson, P.M. Berry, P.A. Harrison. 2002. SPECIES: A spatial evaluation of climate impact on the envelope of species. Ecol. Mod. 154:289-300.

Peterson, A.T. 2003. Predicting the geography of species' invasions via ecological niche modeling. Q. Rev. Biol. 78:419-433.

Peterson, A.T. and K.P. Cohoon. 1999. Sensitivity of distributional prediction algorithms to geographic data completeness. Ecol. Mod. 117:159-164.

Peterson, A.T. and D.A. Vieglais. 2001. Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. BioSci. 51:363-371.

Peterson, A.T. and J.J. Shaw. 2003. *Lutzomyia* vectors for cutaneous leishmaniasis in southern Brazil: Ecological niche models, predicted geographic distributions, and climate change effects. Int. J. Parasitol. 33:919-931.

Peterson, A.T., D.R.B. Stockwell and D.A. Kluza. 2002. Distributional prediction based on ecological niche modeling of primary occurrence data. Pp. 617-623 *in* Predicting Species Occurrences: Issues of Scale and Accuracy (J.M. Scott et al., eds.). Island Press, Washington, D.C.

Peterson, A.T., J. Soberón, and V. Sánchez-Cordero. 1999. Conservatism of ecological niches in evolutionary time. Science 285:1265-1267.

Peterson, A.T., M.A. Ortega-Huerta, J. Bartley, V. Sánchez-Cordero, J. Soberón, R.H. Buddemeier and D.R.B. Stockwell. 2002. Future projections for Mexican faunas under global climate change scenarios. Nature 416:626-629.

Peterson, A.T., V. Sánchez-Cordero, C.B. Beard and J.M. Ramsey. 2002. Ecologic niche modeling and

potential reservoirs for Chagas disease, Mexico. Emerging Inf. Dis. 8:662-667.

Raxworthy, C.J., E. Martínez-Meyer, N. Horning, R.A. Nussbaum, G.E. Schneider, M.A. Ortega-Huerta and A.T. Peterson. 2003. Predicting distributions of known and unknown reptile species in Madagascar. Nature 426:837-841.

Reid, W. 2004. Bridging the science-policy divide. PLoS Biol. 2:169-171.

Rodrigues, A.S.L., S.J. Andelman, M.I. Bakarr, L. Boitani, T.M. Brooks, R.M. Cowling, L.D.C. Fishpool, G.A.B. de Fonseca, K.J. Gaston, M. Hoffmann, J.S. Long, P.A. Marquet, J.D. Pilgrim, R.L. Pressey, J. Schipper, W. Sechrest, S.N. Stuart, L.G. Underhill, R.W. Waller, M.E.J. Watts and X. Yan. 2004. Effectiveness of the global protected area network in representing species diversity. Nature 428:640-643.

Sánchez-Cordero, V. and E. Martínez-Meyer. 2000. Museum specimen data predict crop damage by tropical rodents. Proc. Nat. Ac. Sci. USA 97:7074-7077.

Schroeder, P. 2003. Digital research data as a floating capital of the global system. Pp. 7-11 *in* Promise and Practice in Data Sharing. OECD Report. NIWI-KNAW, Amsterdam.

Siqueira, M.F. de and A.T. Peterson. 2003. Consequences of global climate change for geographic distributions of cerrado tree species. Biota Neotropica 3(2). electronic journal, http://www.biotaneotropica.org.br/.

Soberón, J. and A.T. Peterson. 2004. Biodiversity informatics: managing and applying primary biodiversity data. Phil. Trans. R. Soc. Lond. 359:689-698.

Stockwell, D.R.B. and D.P. Peters. 1999. The GARP modelling system: Problems and solutions to automated spatial prediction. Int. J. Geogr. Inf. Sci. 13:143-158.

Stockwell, D.R.B. and I.R. Noble.1991. Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. Math. Comp. Simul. 32:249-254.

Thomas, C.D., A. Cameron, R.E. Green, M. Bakkenes, L.J. Beaumont, Y.C. Collingham, B.F.N. Erasmus, M.F. de Siqueira, A. Grainger, L. Hannah, L. Hughes, B. Huntley, A.S. van Jaarsveld, G.F. Midgley, L. Miles, M.A. Ortega-Huerta, A.T. Peterson, O.L. Phillips and S.E. Williams. 2004. Extinction risk from climate change. Nature 427:145-148.

Thuiller, W. 2003. BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. Glob. Change Biol. 9:1353-1362.

Yee, T.W. and N.D. Mitchell. 1991. Generalized additive models in plant ecology. J. Veg. Sci. 2:587-602.