A global distributed biodiversity information network: building the world museum

by A. Townsend Peterson, David A. Vieglais, Adolfo G. Navarro Sigüenza & Marcos Silva

SUMMARY

Biodiversity information is not presently managed in a manner that enables or encourages broad, efficient, or novel uses. New technologies that permit integration of biodiversity information stored at institutions worldwide into a single biodiversity information facility, however, have the potential to change this situation. Information integrated from diverse institutions and available in quantity greatly empowers a diversity of novel products that amply demonstrate the importance of the information resource. We discuss the promise and the opportunity, as well as the impediments to assembling a global distributed biodiversity information network—effectively a 'world museum', built from the world's biodiversity and available to all freely and openly.

Introduction

This paper serves to present a major initiative to share biodiversity information on a global scale. This effort takes the form of a distributed data network implemented over the Internet, permitting full access to biodiversity information to all. The network can be seen as a move towards open collaboration by biodiversity scientists, and simultaneously as repatriation of information about biodiversity from the countries holding that information to the countries where that biodiversity exists in nature. The first element of this network was made openly available for worldwide access and use in April 1999—called *Species Analyst*—and is accessible at http://speciesanalyst.net/. It is a project that has grown out of the perception of the need for much-improved information services in the biodiversity world, and has been adopted and driven forward by a number of institutions around the world. The purpose of this paper is to outline the dimensions of the data network, stimulate discussion about its implementation, and promote participation by institutions across Europe.

History

Biodiversity exploration has seen many biases and imbalances geographically. The vast storehouse of biodiversity—species diversity and unique taxa—is in the tropics. The storehouses of biodiversity information and collections, however, are in the Northern Hemisphere, principally in the United States and Europe.

The scientific exploration and documentation of biodiversity began in earnest in the 1700s. The early explorations were initiated by Europeans, later joined by investigators from the United States. These explorations were often carried out without participation of scientists from the host countries, and information was often not returned to or shared with scientists in those countries. Lack of collaboration with, or training of, scientists in biodiversity-rich countries further widened the gap

in information and expertise. Hence, specimens and scientific literature important to understanding world biodiversity became concentrated in countries where relatively little diversity was actually present.

The present

Biodiversity data are currently available in a dispersed system based on institutional and national boundaries. Most members of the community of data caretakers (curators in natural history museums and herbaria) are more than willing to provide information; nevertheless, the system is simply inefficient and difficult to access. For this reason, biodiversity studies have not taken full advantage of the information in existence.

Most biodiversity information is stored in the form of scientific collections in museums and universities across North America and Europe. This information is often not computerised, and is considered the property of individual institutions. Hence, access to each collection must be handled on a one-by-one basis, making access to the totality of information in existence a laborious task.

When assistance or access to data is requested, most biodiversity information caretakers make information and specimens readily available (Navarro-Sigüenza *et al.* 2002). Requests are accepted by mail or by electronic mail, or visitors are received in most collections; in some cases, data are made available freely over the Internet (Soberón 1999), but some voices still speak strongly against this idea (Graves 2000). The distributed nature of the collections, with important specimen holdings in 10–20 countries for most regions of interest, makes such communications or travel difficult or even prohibitive for most biodiversity information users, particularly those from where the information originated.

To illustrate the importance of complete biodiversity information, reference can be made to the avian dataset for Mexico under preparation as the *Atlas of the distribution of Mexican birds* (Peterson *et al.* 1998). Here, the contents of 43 scientific collections were assembled in a centralised database, totalling more than 300,000 specimen records for the country. The largest single collection held only 16% of the total, so the emergent properties of the large dataset were not realised until the contents of numerous collections were aggregated. Since its assembly, however, this dataset has been instrumental in advances both in conservation and in basic biology regarding Mexican birds (Peterson *et al.* 1999, Navarro-Sigüenza & Peterson 2000, Navarro-Sigüenza *et al.* 2002, Peterson *et al.* 2002a).

In recent years, several programmes aimed at improving the state of biodiversity informatics have emerged. *Species2000* (http://www.sp2000.org.) and the *Integrated Taxonomic Information System* (ITIS) (http://www.itis.usda.gov/plantproj/itis) aim to produce a catalogue of all named *species* in the world, but manage only information related to the *names* of the species, not to their distributions, occurrences or ecology. At present, four Internet-based facilities provide a novel form of 'distributed' access to biodiversity data: databases located at the institutions that 'own' the data (and the specimens most often associated with them) are probed by Internet-based search

algorithms. These facilities include *The Species Analyst*, as well as Red Mundial para la Información de la Biodiversidad (REMIB) (http://www.conabio.gob.mx), the Virtual Australian Herbarium (http://www.rbgsyd.gov.au/HISCOM/Virtualherb/virtualherbarium.html) and the European Natural History Specimen Information Network (ENHSIN) (http://www.nhm.ac.uk/science/rco/enhsin) all provide broad access to such primary data. Most recently, the Global Biodiversity Information Facility (GBIF) was formed to seek means of integrating biodiversity information on a global scale, and has adopted much of the distributed model as the basis for its work.

At present, the information services in the world of biodiversity information is woefully inefficient. Data exist in impressive quantities for many taxonomic groups, yet this information is too often difficult to access. Because of the numerous impediments to access, biodiversity data are too frequently not included in studies and reports that would benefit immensely from their inclusion. This information gap leads both to an information undernourishment in many policy decisions, and a lack of appreciation of the immense value of biodiversity information held in, for example, natural history museums. Large-scale biodiversity conservation studies, although focusing on exactly the information in question, are often based only minimally, or secondarily, on biodiversity data. For this reason, such studies lack analytical power and information completeness, and the results often reflect this failing. For these reasons, we propose the expansion of the worldwide distributed biodiversity information network to include a broad sampling of European institutions, thus uniting data sources in North America and Europe.

The Species Analyst

Technology

The most complete representation of global biodiversity can be found in the world's natural history museums. Although records for museum collections are increasingly maintained in electronic format, their use has been hindered by the lack of standard methods for search and retrieval from disparate databases. Using ANSI/NISO Z39.50, a standard for information retrieval that has proved successful in the bibliographic and geospatial domains, and a newer information transfer protocol called XML, it is now possible to search and retrieve information from biological collections connected by the Internet.

The Species Analyst is a set of software extensions that enable Z39.50 searches from a web interface, as well as from applications such as Microsoft Excel and ESRI's ArcView GIS. Users may query multiple collection databases simultaneously and, in a matter of seconds, obtain information directly into a client application in a form suitable for further analysis. This suite of capabilities provides an infrastructure that allows seamless search, retrieval and analysis of a wealth of biodiversity data that has hitherto been impossible. Although at present data are served 'as is'—that is, names are provided as the owner institution has them, and are not at present

integrated, translated or otherwise standardised; eventually, connections with efforts designed to assemble up-to-the-minute names catalogues will provide this next generation of data access.

The Species Analyst currently provides seamless access to more than 50 million specimens in 84 datasets housed at 65 institutions. Committed to participation are an additional 69 institutions with datasets principally focused on mammals, reptiles and amphibians. Hence, the total of biodiversity data now committed to participation is well over 130 institutions and approximately 55 million species-occurrence records. Additional millions of species-occurrence records are served via REMIB, Virtual Australian Herbarium, and ENHSIN, making for a substantial quantity of biodiversity information available worldwide to all users, and most vouched by specimens in scientific collections. The customary figure cited for total specimens in world natural history museums is about three billion, and best calculations are that about 5% of those specimens are now computerised (Krishtalka & Humphrey 2000); by this reasoning, about 150 million specimen records exist in electronic form, about 30% of which are served or slated to be served by *The Species Analyst*.

Applications of the distributed biodiversity network

The present situation of compartmentalised data and inefficient access constitutes a critical bottleneck in biodiversity applications. Once data are united in large, inclusive sets, many new possibilities open up for analysis, leading to new dimensions of understanding. These new possibilities can be referred to as 'emergent properties' of large biodiversity datasets, and further underline the need to enable the information currently present in natural history museums, as well as to continue building natural history museum collections. Three examples focused in the area of biodiversity conservation are presented below.

Endangered and poorly known species

In contrast to endangered species in many developed countries (Godown & Peterson 2000), many species of conservation concern are so poorly known or so rare that basic distributional information is unavailable. This problem is even more frequent in taxonomic groups less well known than birds. These species clearly present a most difficult challenge for biodiversity conservation, given that even the most basic information is often lacking.

An excellent example of a poorly known bird species is the Slaty Finch *Haplospiza rustica* of tropical America. Although more frequently encountered in the South American portion of its distribution, its known occurrences in Mexico and northern Central America are vanishingly few (Howell & Webb 1995). Populations are so poorly known that study and documentation of their taxonomic status, ecological characteristics and conservation status are essentially impossible.

Among scientific specimens of Slaty Finches, two Mexican point localities are available (Jalapa, Veracruz, in the 1860s; Volcán Tacaná, Chiapas, in the 1950s)

from Mexico. Modelling the ecological requirements of the species using GARP (Stockwell & Noble 1992, Stockwell 1999, Stockwell & Peters 1999, Peterson *et al.* 2002b), a map of potential distributional areas for the species can be developed (Fig. 1). Although this map is very general, and probably overly inclusive given a possible connection with stands of bamboo, it provides a useful first step in outlining areas for search and inventory for the species in field efforts. In this particular case, a recent record (G. Escalona-Segura unpubl. data.) coincided exactly with one predicted potential distributional area (see arrow, Fig. 1). Using better-known species, these predictive methodologies have been put to more rigorous, statistical tests (Peterson *et al.* 2002b), providing convincing evidence that distributional models can be developed for many rare and poorly-known species.

Conservation prioritisation

Once the possibility exists of predicting geographic distributions of species with precision, a clear next step is that of predicting distributions of suites of species of

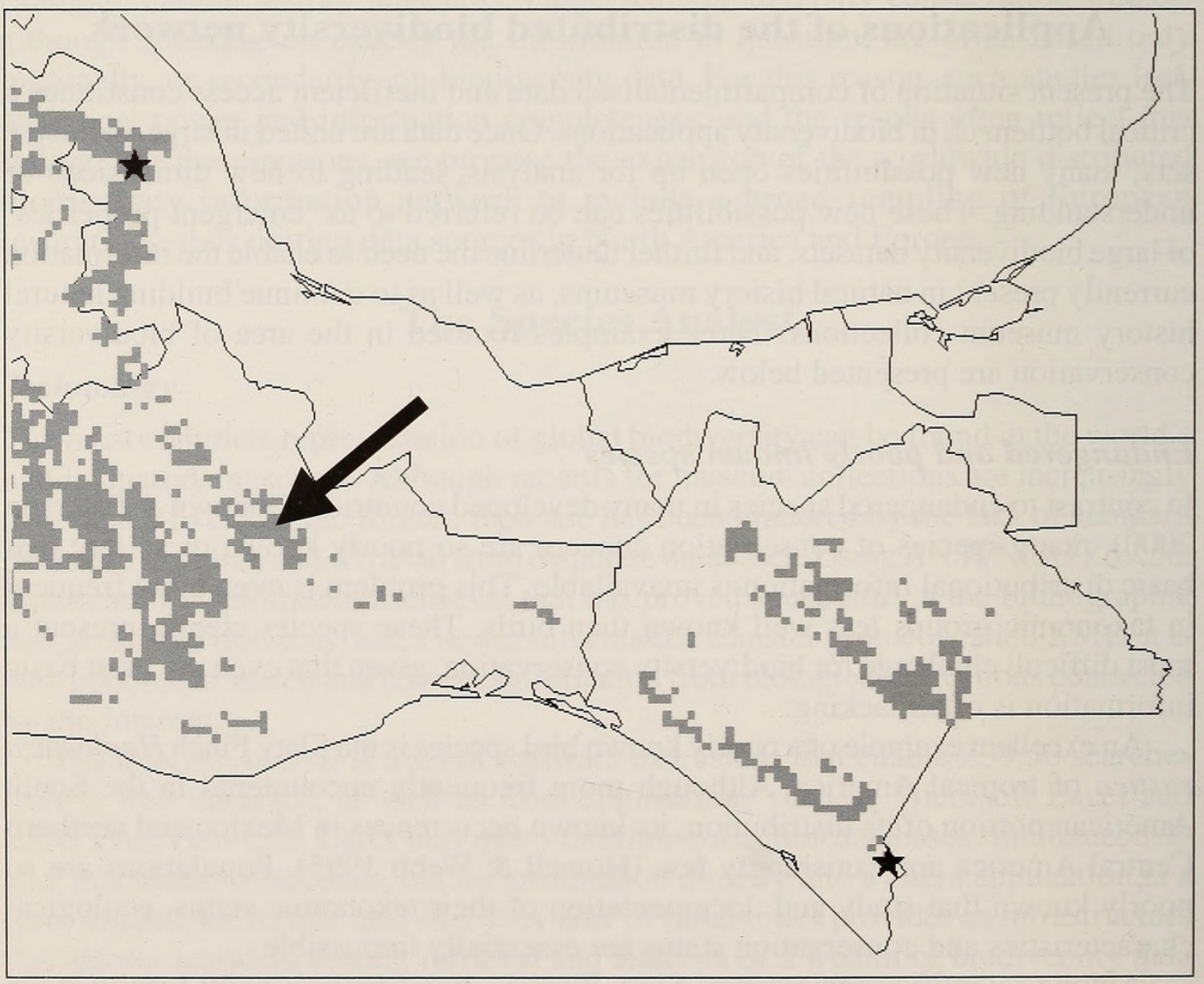

Figure 1. Distributional prediction (grey shading) for Slaty Finches *Haplospiza rustica* in Mexico based on two known specimen localities (black stars), showing third site discovered in 1995 (arrow).

Figure 2. Richness of 16 species endemic to western Mexican tropical dry forests, showing increasing species richness as darker shades of grey. Locations of two biosphere reserves are shown with stars, coinciding with primary concentration of 12 species; the arrow indicates a secondary concentration, in which all four remaining species are represented.

special interest, and taking their joint distribution as conservation priorities. For instance, one might model the distributions of all endangered or endemic species in a region, and then identify resulting foci of species co-occurrence as priority areas. This approach has important advantages over past approaches (e.g. Bojórquez-Tapia *et al.* 1995), in that individual species' distributions are the input into the decision-making process, allowing the design of truly inclusive reserve systems, perhaps using algorithms aimed at maximising complementarity among areas (Peterson *et al.* 2000).

As an example of these possibilities, Kluza & Peterson (unpublished) modelled the geographic distributions of all 16 bird species endemic to the dry tropical forests of south-western Mexico. Individual species' distributions were overlaid, and a map of endemic species richness was created (Fig. 2). One area on this map in northern Guerrero (labelled 'A') represented a peak of species richness, including 12 of the 16 species; this area coincided well with two existing biosphere reserves. The remaining four species, however, were not protected by any biosphere reserves, and co-occurred in north-western Oaxaca. This area, in particular if avian patterns are coincident with those in other taxonomic groups, could be taken as a clear priority area for conservation action (Kluza & Peterson unpublished).

Global climate change and conservation planning

As a further demonstration of the flexibility and promise of the data and analytical approaches described herein, a final level of complexity can be added, taking into account the influence of global climate change on species' geographic distributions. Global climate change, rather than being simply 'global warming', rather represents a series of reorganisations of climatic processes, and therefore requires a series of complex modelling efforts. This work provides a first step towards a robust methodology.

In the GARP modelling process employed above, distributional predictions are derived from models of distribution in ecological dimensions, effectively a model of the ecological niche of the species (Peterson *et al.* 2002b). To the extent that ecological niches present stable sets of constraints on species' distributions over moderate periods of time (e.g. Peterson *et al.* 1999), distributions of species in a changing environment may be taken as the distributions of their ecological niches through those changes. Projections of these niche models to modeled future climates provide estimates of future potential distributions of species (Peterson *et al.* 2001, Peterson *et al.* 2002a).

As an illustration, we have modelled the potential future distribution of West Mexican Chachalaca *Ortalis poliocephala* under two scenarios of change in the Hadley simulation (Pope *et al.* 2002) of global climate change (50 yr of 0.5–1.0%/yr CO_2 increase, with and without sulphate aerosol forcing). The present geographic distribution (Fig. 3) is more or less continuous along the western coast of Mexico. Through the simulated scenario of climate change, however, although coastal portions of the species's modelled distribution remain intact, the interior portions of Mexico become largely uninhabitable for the species. Of species that we have modelled similarly (Peterson *et al.* 2001, 2002a), we see a diversity of responses, including extinction, fragmentation and expansion. The conservation prioritisations presented above, then, would have to be adjusted to take into account these modified expectations of species' distributions over very short horizons of time.

The proposal

The goal

The principal objective of the *Species Analyst* and related efforts to build a global distributed biodiversity information system is to spark collaboration and cooperation among biodiversity scientists via open access to biodiversity data. The project will effectively end the present compartmentalised system, in which access to information is on an institution-by-institution basis, and move the field towards worldwide integration—a virtual 'world museum', in which barriers to information access disappear. Information taken from countries over several centuries will become openly accessible to all, effectively constituting repatriation of biodiversity data. Matching improved access to information with open sharing of expertise and software

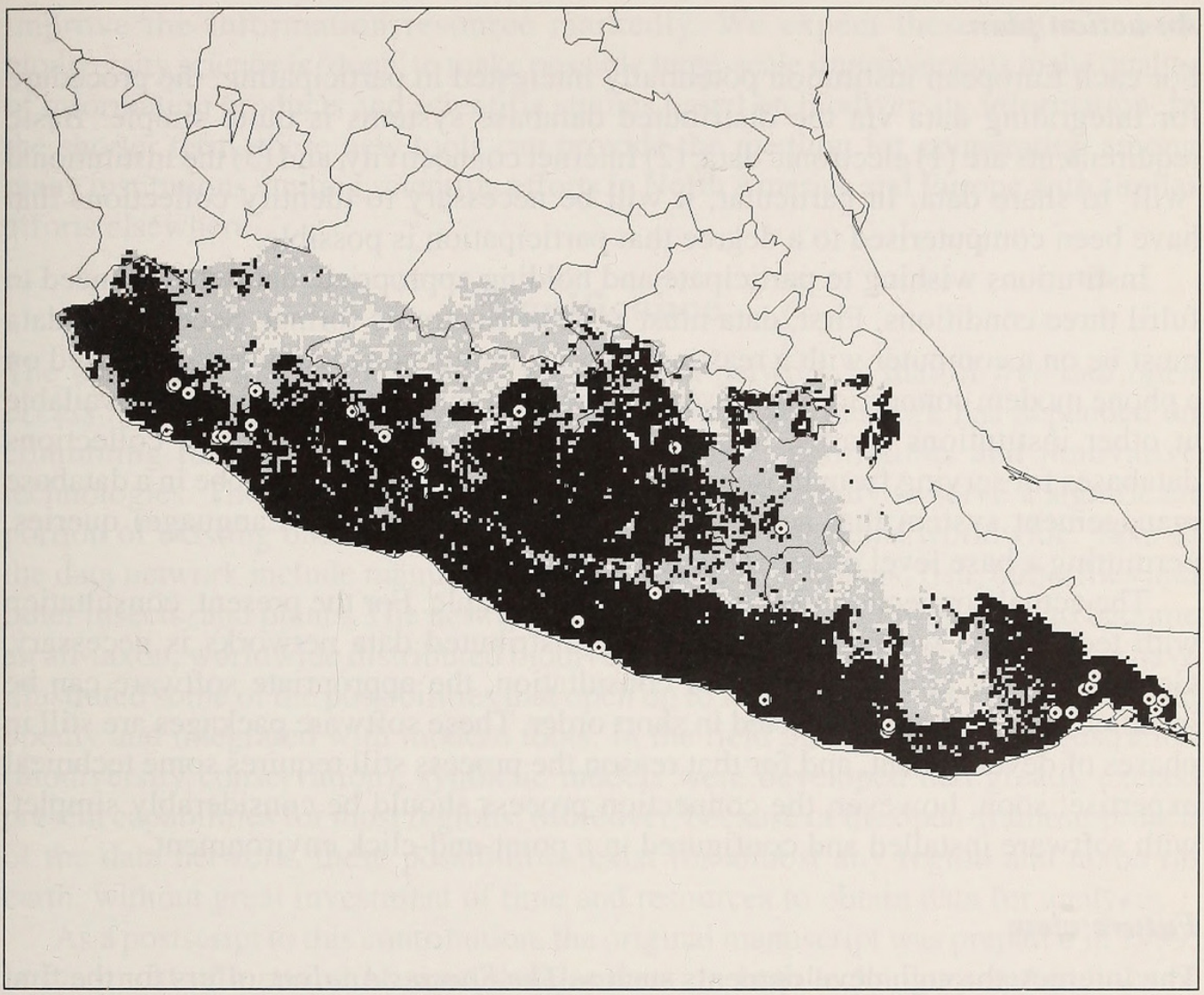

Figure 3. Projected effects of global climate change on the geographic distribution of the West Mexican Chachalaca *Ortalis poliocephala*: present distribution is shown in grey and black, and projected post-change distribution is in black; specimen-based distributional points are shown as dotted circles.

will lead to a qualitative leap forward in ability to use biodiversity information effectively.

Whereas most North American institutions are participating, at least in part, in *The Species Analyst*, few European institutions are currently linked to the network. Although European distributed data initiatives are beginning (e.g. ENHSIN), they lag behind the American efforts in serving a major portion of the biodiversity information stored in European institutions. Clearly, this difference derives from a variety of reasons. However, prominent among them is the fact that very few European natural history museum collections are at present computerised, obviously making serving data impossible. (Of course, it should be pointed out that several important American bird collections are also not computerised—e.g. American Museum of Natural History—or are only partially computerised—e.g. U.S. National Museum of Natural History—so the contrast is not entirely black *versus* white!)

An action plan

For each European institution potentially interested in participating, the procedure for integrating data via the distributed database systems is quite simple. Basic requirements are (1) electronic data, (2) Internet connectivity, and (3) the institutional 'will' to share data. In particular, it will be necessary to identify collections that have been computerised to a degree that participation is possible.

Institutions wishing to participate and holding appropriate datasets will need to fulfil three conditions. First, data must exist in electronic format. Second, the data must be on a computer with a reasonably fast Internet connection (i.e. not based on a phone modem connection) (in cases in which fast Internet connections are available at other institutions nearby, it may be possible to deposit copies of collections databases for serving from those institutions). Finally, the data must be in a database management system that accepts SQL (Standardised Query Language) queries, permitting a base level of access to the information.

The actual connection process is relatively simple. For the present, consultation with technicians working for one of the distributed data networks is necessary. Generally in an afternoon or so of consultation, the appropriate software can be installed, and data are connected in short order. These software packages are still in phases of development, and for that reason the process still requires some technical expertise; soon, however, the connection process should be considerably simpler, with software installed and configured in a point-and-click environment.

Future view

The Internet, through developments such as *The Species Analyst*, offers for the first time the possibility of widespread integration of information in existence worldwide about biological diversity. These advances make possible a rapid shift from the previous situation (closed institutional resources) to an exciting new one—information can be integrated across regional, national, taxonomic and institutional boundaries to provide a resource never before possible. Already, this improved access to information has been instrumental in stimulating novel approaches to predicting species invasions (Peterson & Vieglais 2001), design of endangered species conservation efforts (Godown & Peterson 2000), design of efforts to combat agricultural pests (Sánchez-Cordero & Martínez-Meyer 2000), understanding ecological niche evolution (Peterson *et al.* 1999), and predicting the effects of global climate change on species distributions (Peterson *et al.* 2001, 2002a), etc. In this way, the totality of biodiversity information can be applied to critical questions, such as biodiversity conservation, natural resource management, land use planning, and others.

These steps remove a critical impediment—that of access to information—and allows scientists and decision-makers worldwide to proceed to new levels of understanding. Data served over the distributed network can be improved via iterative sweeps of standardisation, adding value, and error detection, which can serve to

improve the information resource markedly. We expect these shifts in how biodiversity science is 'done' to make possible large-scale improvements in the quality of information products and scientific studies based on biodiversity information. In the shorter term, these new tools can provide the medium for cooperation among many institutions, uniting scientific efforts in North America and Europe with similar efforts elsewhere.

Significance

The basic principle of the *Species Analyst* data network is that of *free and open access to data and technology*. The development of the network has depended on combining this philosophy with careful political negotiations and innovative technologies. The result is a network that has grown rapidly to serve a significant portion of existing data in natural history museum collections worldwide. Taxa in the data network include mammals, birds, reptiles, amphibians, fish, butterflies and other insects, and plants. The network is growing rapidly, and is on track to become an all-taxon, worldwide distributed biodiversity data network. In this paper, we have illustrated some of the possibilities that open up to investigators once data are shared openly and integrated with modern tools. In the field that we chose for illustration (biodiversity conservation), synthetic models were developed that greatly exceed present capabilities for most regions. Moreover, because of the open-frame approach of the data network, these possibilities exist for almost any region and taxon on earth, without great investment of time and resources to obtain data for analysis.

As a postscript to this contribution, the original manuscript was prepared in 1999, in the midst of a period of rapid development, in both technology and politics of biodiversity information. On the technological side, many important advances have been made in making the software solutions broadly applicable, stable, fast and reliable. Several distributed biodiversity information networks now exist, and many are collaborating on a next-generation technology that should allow integration of all of the networks into a single global distributed biodiversity information network. On the political side, some aspects have changed dramatically, and others have *not* changed dramatically. On the side of change, the initiation of the Global Biodiversity Information Facility has emphasised the need for participation in data-sharing efforts within each member country, and has sparked many exciting steps forward in the assembly of efficient biodiversity information services around the world. On the side of *no* change, some workers in the field—including curators at important and large natural history museums—continue to resist the idea of bringing natural history museums into the information age. One can only hope that these persons and their ideas can evolve as the idea of global sharing of such important information becomes more and more the rule, and no longer the exception.

References:

Bojórquez-Tapia, L. A., Azuara, I., Ezcurra, E. and Flores V., O. A. 1995. Identifying conservation priorities in Mexico through geographic information systems and modeling. *Ecological Applications* 5: 215-231.

Godown, M. E. & Peterson, A. T. 2000. Preliminary distributional analysis of U.S. endangered bird species. *Biodiversity and Conservation* 9: 1313-1322.

Graves, G. R. 2000. Costs and benefits of Web access to museum data. *Trends in Ecology and Evolution* 15: 374.

Howell, S. N. G. & Webb, S. 1995. *A guide to the birds of Mexico and northern Central America.* Oxford Univ. Press.

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

Navarro-Sigüenza, A. G. & Peterson, A. T. 2000. Western Mexico: a significant center of avian endemism and challenge for conservation action. *Cotinga* 14: 42-46.

Navarro-Sigüenza, A. G., Peterson, A. T. & Gordillo-Martínez, A. 2002. A Mexican case study on a centralised database from world natural history museums. *CODATA Journal* 1: 45-53.

Peterson, A. T., Egbert, S. L., Sánchez-Cordero, V. & Price, K. P. 2000. Geographic analysis of conservation priorities using distributional modelling and complementarity: endemic birds and mammals in Veracruz, Mexico. *Biol. Conserv.* 93: 85-94.

Peterson, A. T., Navarro-Sigüenza, A. G. & Benítez-Díaz, H. 1998. The need for continued scientific collecting: a geographic analysis of Mexican bird specimens. *Ibis* 140: 288-294.

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

Peterson, A. T., Sánchez-Cordero, V., Soberón, J., Bartley, J., Buddemeier, R. H. & Navarro-Sigüenza, A. G. 2001. Effects of global climate change on geographic distributions of Mexican Cracidae. *Ecological Modelling* 144: 21-30.

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

Peterson, A. T., Stockwell, D. R. B. & Kluza, D. A. 2002b. Distributional prediction based on ecological niche modeling of primary occurrence data. Pp.617-623 in. Scott, J. M., Heglund, P. J. & Morrison, M. L. (eds.) *Predicting species occurrences: issues of scale and accuracy.* Island Press, Washington, D.C.

Peterson, A. T. and Vieglais, D. A. 2001. Predicting species invasions using ecological niche modeling. *BioScience* 51: 363-371.

Pope, V. D., Gallani, M. L., Rowntree, V. J. & Stratton, R. A. 2002. *The impact of new physical parametrizations in the Hadley Centre climate model - HadAM3.* Hadley Centre for Climate Prediction and Research, Bracknell, Berks, U.K.

Sánchez-Cordero, V. and Martínez-Meyer, E. 2000. Museum specimen data predict crop damage by tropical rodents. *Proceedings of the National Academy of Sciences USA* 97: 7074-7077.

Soberón, J. 1999. Linking biodiversity information sources. *Trends in Ecology and Evolution* 14: 291.

Stockwell, D. R. B. 1999. Genetic algorithms II. Pp.123-144 in Fielding, A. H. (ed.) *Machine learning methods for ecological applications.* Kluwer Academic Publishers, Boston.

Stockwell, D. R. B. & Noble, I. R. 1992. Induction of sets of rules from animal distribution data: a robust and informative method of analysis. *Mathematics and Computers in Simulation* 33: 385-390.

Stockwell, D. R. B. & Peters, D. P. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *Internatn. J. Geographic Information Systems* 13: 143-158.

Addresses: A. Townsend Peterson, Natural History Museum, The University of Kansas, Lawrence, Kansas 66045 U.S.A.; David A. Vieglais, Natural History Museum, The University of Kansas, Lawrence, Kansas 66045 U.S.A.; Adolfo G. Navarro Sigüenza, Museo de Zoología, Facultad de Ciencias, Universidad Nacional Autónoma de México, Apartado Postal 70-399, México, D.F. 04510 Mexico; Marcos Silva, North American Biodiversity Information Network, Commission for Environmental Cooperation, 393 Saint-Jacques Street, Suite 200, Montreal, Quebec, Canada, H2Y 1N9