THE ROYAL SOCIETY

Published online 18 March 2004

Biodiversity informatics: managing and applying primary biodiversity data

Jorge Soberón[1*] and A. Townsend Peterson[2]

[1]*Instituto de Ecología, National University of Mexico and Comisión Nacional de Biodiversidad, México, Periférico-Insurgentes 4903, Tlalpan, 14010 DF, Mexico*

[2]*Natural History Museum and Biodiversity Research Center, University of Kansas, Lawrence, KS 66045, USA, (town@ku.edu)*

Recently, advances in information technology and an increased willingness to share primary biodiversity data are enabling unprecedented access to it. By combining presences of species data with electronic cartography via a number of algorithms, estimating niches of species and their areas of distribution becomes feasible at resolutions one to three orders of magnitude higher than it was possible a few years ago. Some examples of the power of that technique are presented. For the method to work, limitations such as lack of high-quality taxonomic determination, precise georeferencing of the data and availability of high-quality and updated taxonomic treatments of the groups must be overcome. These are discussed, together with comments on the potential of these biodiversity informatics techniques not only for fundamental studies but also as a way for developing countries to apply state of the art bioinformatic methods and large quantities of data, in practical ways, to tackle issues of biodiversity management.

Keywords: biodiversity informatics; species ranges; species niches; taxonomic capacity

1. INTRODUCTION

As biologists began studying what are called 'biodiversity patterns', the primary data were observations of presences and absences of species across space and time, combined with geographical information regarding climate, soil, geology and other features of the regions in which they are found. This focus on primary occurrence information began with the earliest of the classic naturalists, and continued right up to the present (Krishtalka & Humphrey 2000). This basis, of course, requires the collaboration of the entire systematic enterprise—without sound taxonomic information, description and understanding of species diversity patterns and distributions would be impossible.

In the past 10 years, advances in information technology (e.g. large-capacity electronic storage media, the Internet, the World Wide Web, distributional database technology) and in the policies of owners of primary data sources (e.g. large-scale digitization of data, creation of public-access databases) are creating a revolution in the way that biodiversity information is created, maintained, distributed and used (Bisby 2000; Oliver *et al.* 2000; Edwards *et al.* 2000; Krishtalka & Humphrey 2000; Krishtalka *et al.* 2002), with the potential of much more to come (Godfray 2002). Moreover, the amount, variety and resolution of spatially explicit electronic data that can be used to describe environments (e.g. RS data available via the Internet) are similarly growing at a staggering pace. Very roughly (Faundeen 2003; figure 1), for the first Landsat family series (MSS 1), *ca.* 4 Terabytes of data were archived in the period from 1972 to 1982. In the next 20 years, Landsat (TM 4 and 5) accumulated *ca.* 140 Terabytes of data archived. For the past 3 years, S. Dech (personal communication) of the Deutsches Zentrum für Luft und Raumfahrt estimates a 60% (from 70 to 200 Terabytes) per year growth of RS data for the main European RS families. Growth rates for the RS data enterprise are probably exponential or more than exponential but are difficult to estimate owing to the overlap in availability of different sensors.

In any case, RS data are now essential for conservation science and other applications (Green *et al.* 1987; Stomes & Estes 1993; Veitch *et al.* 1995; Danks & Klein 2002; Turner *et al.* 2003; Kerr & Ostrovsky 2003) owing to their unique ability to characterize the Earth's surface from different perspectives, resolutions and spectral dimensions. This allows, among other things, the finding of correlates for inferences and classifications.

Primary biodiversity data—principally in the form of specimen information—is now also becoming accessible at an accelerated speed. Increasing numbers of museums and herbaria are computerizing data associated with natural history specimens (Krishtalka & Humphrey 2000). In addition, in many cases, high-resolution images keyed to tabular data and providing additional dimensions of access to specimens are also being created (Bisby 2000; Edwards *et al.* 2000; Oliver *et al.* 2000). In many cases, these datasets are being made available through the Internet. Excellent examples include: the New York Botanical Garden, the Museum of Vertebrate Zoology, the University of California at Berkeley, the Missouri Botanical Gardens and the Instituto Nacional de Biodiversidad, Costa Rica, but

* Author for correspondence (jsoberón@xolo.conabio.gob.mx).

One contribution of 19 to a Theme Issue 'Taxonomy for the twenty-first century'.

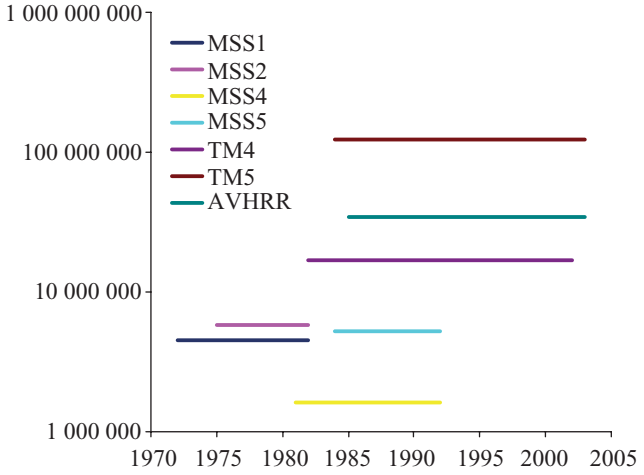

Figure 1. Growth of two families of environmentally related satellite data (AVHRR and Landsat). Horizontally, each line represents the span of activity of the different images. The vertical value gives the size (in Megabytes) of the accumulated (over the lifespan of the sensor) information radiated to receptors. The total value at the beginning of 2003 was 256 Terabytes. The MSS and TM images belong to the Landsat family of satellites, and the AVHRR to NOAA climatologic satellites.

the list is growing very fast. Also, several centralized databases provide access to information held in jointly created specimen databases. Fishbase, for example (http://www.fishbase.org/home.htm) offers data from hundreds of thousands of specimens. This initial commitment to sharing data and providing open access to data is an important step towards greater information access in the biodiversity world.

More profoundly still, since 1998, several distributed biodiversity information networks have provided a new class of access to biodiversity information. In particular, two specialized search engines, The Species Analyst (http://speciesanalyst.net) and REMIB (http://www.conabio.gob.mx/remib/remib.html) have solved key problems that plagued earlier, single-database implementations. These facilities provide access to distributed databases, which means that the data remain at the institutions where the voucher specimens are housed, thus maintaining the connection between primary documentation (specimens) and the information product (the database). Nevertheless, the contents of these dispersed databases are shared virtually via specialized Internet access engines. The Species Analyst and REMIB now connect databases of hundreds of collections, and serve data associated with millions of specimens (figure 2). Still better access is now permitted by a next-generation integrating technology (DiGIR; http://digir.sourceforge.net/), which has now been implemented fully for the first time in the MaNIS project (http://elib.cs.berkeley.edu/manis/) and will become the standard protocol of the collections associated to the GBIF.

These achievements represent not only the solution to challenging technical problems (like allowing simultaneous access to independent databases with different formats, database managers and operating systems) but mainly the willingness of institutions and data caretakers to allow free and open access to databases under their care (for some caveats see Graves (2000)).

Hence, overall, the world of information available for addressing questions related to biodiversity and ecological landscapes is changing dramatically. The possibilities for the study of spatial patterns of biological diversity, for both basic and applied purposes, are changing beyond recognition. In contrast to past decades, information access is less and less of a consideration, and analytical and computing capacities are becoming more of a concern.

2. BIODIVERSITY INFORMATICS

These new applications belong to the emerging field that we can term BI. As the word bioinformatics is now applied universally to genomics and proteomics applications, a new term may be needed to describe applications at the organismic level. Biodiversity Informatics then includes the application of information technologies to the management, algorithmic exploration, analysis and interpretation of primary data regarding life, particularly at the species level of organization. Biodiversity Informatics analyses and applications are characterized by a number of novel features.

Investigators are increasingly able to use large quantities of biodiversity data—that is, analyses are now frequently based on records numbering 10^4 or more (Colwell & Coddington 1994; Gioia & Pigott 2000; Rahbek & Graves 2000; Peterson *et al.* 2002*a*; Jetz & Rahbek 2002).

Analyses can now be performed across large areas (e.g. 10^5 or more km^2) at resolutions of 10 km^2 or smaller (Egbert *et al.* 2002; Peterson 2004; Soberón *et al.* 2004), rather than the customary 10^4 km^2 or more resolution that used to be normally the case (e.g. Cook (1969), Rapoport (1982) and Roberts *et al.* (2002), among many others).

The development and application of new methodologies designed to assess completeness of databases (Prendergast *et al.* 1993; Soberón & Llorente 1993; Colwell & Coddington 1994; Murguía & Villaseñor 2000; Petersen *et al.* 2003) now permits quantitative evaluation of adequacy and robustness of data used as inputs for biodiversity analyses.

Tools for inferring ecological niches and predicting distributional areas of species from occurrence information and electronic information characterizing ecological landscapes (Nix 1986; Austin *et al.* 1990; Walker & Cocks 1991; Carpenter *et al.* 1993; Mladenoff *et al.* 1995; Jones & Gladkov 1999; Manel *et al.* 1999*a*,*b*; Stockwell & Peters 1999; Gioia & Pigott 2000; Skov 2000; Guisan & Zimmermann 2000; Peterson *et al.* 2002*b*) now permit powerful and predictive inferences about the geographical dimensions of biodiversity. For a recent review of the closely related topic of geographical information systems technology to entomology, see Noonan (2003).

The combination of the elements listed above represents one emerging field within the larger area of BI (Bisby 2000). Basically, this combination permits estimation of fundamental ecological niches of species by means of detection of nonrandom associations between known occurrences and ecological landscapes (ICBP 1992; Miller 1994; Scott *et al.* 1996, 2002; Soberón *et al.* 1996; Umminger & Young 1997).

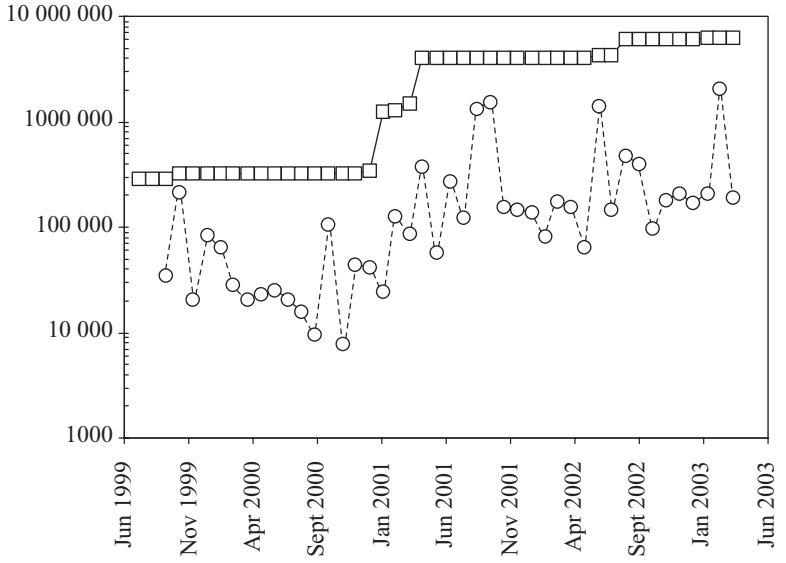

Figure 2. Growth of volume of specimen data made available (squares) and accessed (circles) via the Mexican network of specimen data (REMIB).

Under further assumptions related to the relative importance of biotic interactions and historical effects on species' distribution and dispersal, species' geographical distributions can be estimated. In this way, questions related to species' distributions, patterns of species richness (intersections of distributions), coexistence of taxa, locations of hot spots, complementarity of sites in terms of species representation, and so on, become amenable to formal, quantitative, data-intensive treatment. The use of formal, algorithmic exploration of large amounts of primary biodiversity data is what constitutes the field of biodiversity informatics, and the focus on calculation of niches and distributions may define a subfield within the larger subject.

3. APPLICATIONS OF BIODIVERSITY INFORMATICS TO BIOGEOGRAPHICAL QUESTIONS

Some applications of this new conjunction of data and methodologies focus on issues of basic science, like the study of evolutionary processes, causes of range limitation or species' reactions to changing environments (Peterson *et al*. 1999, 2002*a*; Anderson *et al.* 2002; Peterson 2003, 2004). In Peterson *et al.* (1999), niches were modelled for 37 pairs of species. This was done by using a genetic algorithm (GARP; Stockwell & Peters 1999) to search for regions in the map that are 'similar', in terms of annual precipitation, average temperature, elevation and potential vegetation to those regions where the species has been reported. Species presence data comes from extensive museum databases. The hypothesis was that related taxa should share niche features, thus confirming theoretical predictions about niche conservatism. Indeed, that was found, to a high degree of statistical significance, by reciprocal predictions among related and unrelated pairs of species. In a climate change-related application (Peterson *et al.* 2002*a*) the fundamental ecological niches of 1870 species of Mexican birds, mammals and butterflies were estimated using GARP again (Stockwell & Peters 1999), and the resulting niches were projected to future climates, obtained from general circulation models. Several analyses were then performed on likely changes of distribution areas under several scenarios of dispersal capabilities. The results highlight the relevance of mountain chains for conservation, as turnover of species is lower in mountainous areas than in the central plains of Mexico.

Still other applications focus on management issues, like biodiversity exploration (Lobo *et al.* 1997; Jones & Gladkov 1999; Soberón *et al.* 2004) or location of protected areas (Rebelo & Siegfried 1992; Csuti *et al.* 1997; Godown & Peterson 2000; Peterson *et al.* 2000; Kelley *et al.* 2002; Burgess *et al.* 2002; Chen & Peterson 2002), assessment of the potential for pest damage to crops (Sánchez-Cordero & Martínez Meyer 2000) or evaluation of possible routes for invasive species or diseases (Honig *et al.* 1992; Richardson & McMahon 1992; Scott & Panetta 1993; Higgins *et al.* 1999; Soberón *et al.* 2001; Peterson & Robins 2003; Peterson *et al.* 2003), to name just a few examples.

To provide an example of analysis of an invasive species, considerable recent concern has been caused by the cactus moth *Cactoblastis cactorum*, which as an invader could be catastrophic to certain cactus species, particularly the *Platyopuntia*. We drew location data for *C. cactorum* from scientific collections of the Smithsonian Institution, and used them to estimate hemispheric niche dimensions in terms of climatic variables (by application of the FLORA-MAP software; see Jones & Gladkov 1999). The geographical display of regions of high similarity (on the basis of the climatic variables chosen) to those where the species has been observed provides a prediction of the potential distribution for the species in North America. Then, geographical distributions of species of *Platyopuntia* cacti were obtained by first modelling their niches using the GARP algorithm via 5099 observational data provided by several herbaria (see Acknowledgements). The niches so obtained were then reduced by biogeographical reasoning supervised by experts on the group. Those two procedures yielded individual distributional ranges for 60 species of *Platyopuntia* on the North American continent. Overlay of

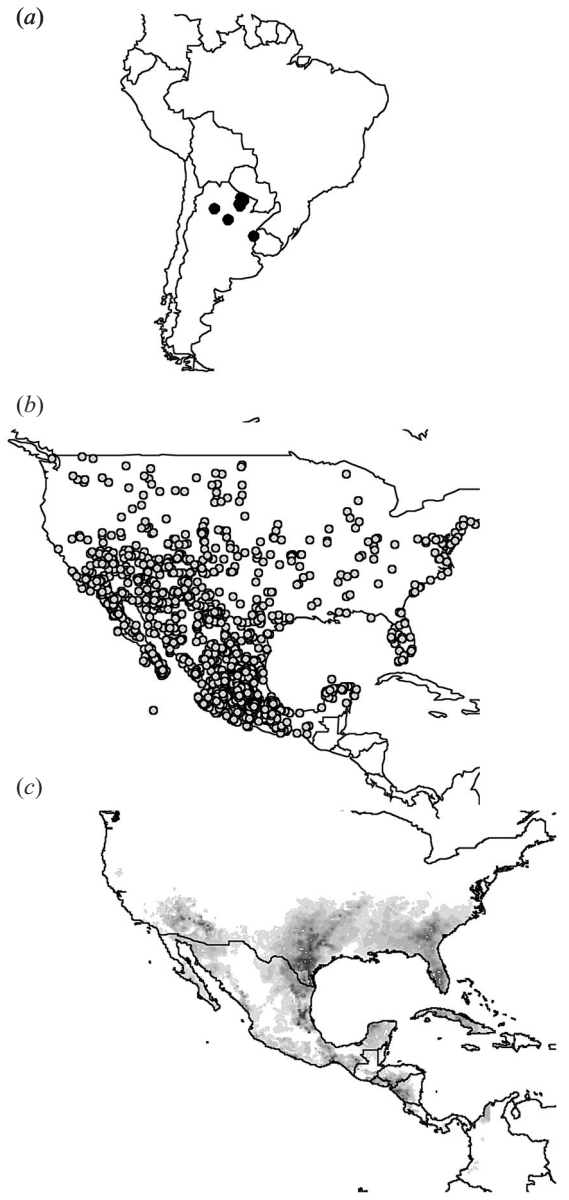

Figure 3. (*a*) Localities of *Cactoblastis cactorum* specimens amenable to precise georeference. Specimens in the Smithsonian Natural History Museum. (*b*) Localities for 5099 specimens in the subgenus *Platyopuntia* in Mexico and the USA. See Acknowledgments for provenance. (*c*) Bioclimatic surface, based on the *C. cactorum* specimens and calculated with the FloraMap software based on a principal component analysis involving three environmental variables distributed over a 12-month period.

Cactoblastis' niche with the distributions of its host plants thus provided a first approximation to understanding the potential route of invasion by *Cactoblastis* into the deserts of the southwestern USA and northern Mexico (figure 3).

4. CHALLENGES AND LIMITATIONS

Although of great potential, significant challenges do exist for this new world of BI. In the first place, presence data include significant biases in the spatial and temporal distribution of collecting efforts and in its overall quality (Soberón *et al.* 1996; Reynolds 1998; Peterson *et al.* 1998; Wilke *et al.* 1999). The dynamic nature of taxonomy means that databases that are not maintained actively may soon be outdated, with synonyms comprising 10–30% or more of names in many databases (Gaston & Mound 1993; Alroy 2002 and see figure 4). The ageing nature of many collections, owing to inattention or to lack of recent material, makes collections data challenging to interpret in light of ongoing land use changes (Remsen 1995; Winker 1996). Hence, in a time when gigabytes of primary biodiversity information are becoming available to all, issues related to quality control are more crucial than ever (Reynolds 1998; Soberón *et al.* 2002).

Although the problems mentioned above can be found in many biological databases, the heterogeneous origin of Web-assembled databases makes quality control even more important. As the origin of the data is heterogeneous, record quality may be uneven and numerous procedures must be used to detect and correct problems. Some of the more common problems include the following.

(i) Specimens may have wrong identifications. This error is quite frequent, and yet can be extremely difficult to detect and correct. Without expert participation in inspection and determination of the original specimens, only very obvious mistakes will be detected (see figure 5 for a clear example of a misidentified specimen that is also an obvious geographical outlier). Obviously, data from poorly determined collections should be used only with care when developing biodiversity analyses (see also Gotelli 2004). More generally, records from such collections should be flagged clearly or perhaps even not opened to search and query by nonprofessionals.

(ii) Outdated taxonomy. An additional suite of problems arises from the evolving nature of biological taxonomy: species identified correctly at one point in time as species X may later be assigned a different 'correct' name; splitting one species into several, changing generic affinities, etc., all may create such situations. Consultation of taxonomic authority files combined with geographical information about the geographical distributions of species to which those names refer, may permit identification of such names. For example, in preparation of the *C. cactorum* example (Soberón *et al.* 2001), we made use of a small database (5099 records) of *Opuntia* cacti drawn from several institutions in the USA and in Mexico (see Acknowledgements). A considerable number of specimens were listed under outdated names (e.g. *O. schotti*, which is now considered as *Grusonia schotii* (Engelm.)). A first check against an updated authority file for the Cactaceae of Mexico (Guzmán *et al.* 2003) detected many such inconsistencies; detailed geographical inspection was necessary to detect other problems, leaving the database relatively clean taxonomically. Unfortunately, in general, obtaining reliable, updated taxonomic authority files is a major problem for most taxa. The large taxonomic information services (e.g. Species 2000, The Integrated Taxonomic Information Service, Missouri Botanical Garden's Tropicos, the Index Kewensis and others) remain far from complete (Bisby 2000; Nic Lughadha 2004)—their

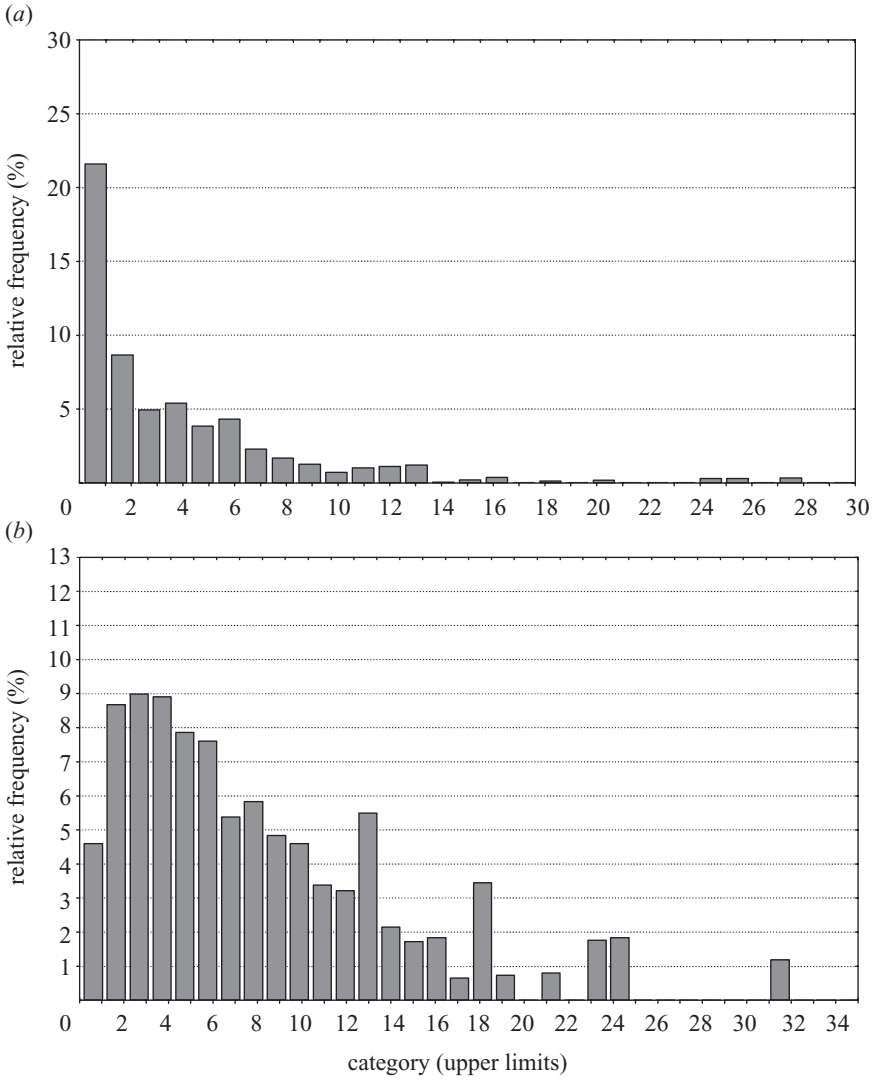

Figure 4. Distribution number of synonyms in two catalogues of Mexican species. (*a*) Poaceae (from Panero (2003)); (*b*) Cactaceae (from Guzmán *et al.* (2003)).

completion will remove one major obstacle for proper quality control of distributed specimen databases.

(iii) Faulty georeferencing (figure 6). Frequently, the identification and textual description of the collection locality may be correct but the geographical coordinates assigned to that site may be erroneous. Faulty georeferencing can be detected by means of consistency analyses, in which verbal descriptions of locality are checked against the geographical coordinates. At present, only a small minority of localities in museum databases are properly georeferenced, which, of course, raises the more basic question of how to add georeferences to specimen data quickly and efficiently. One important example is that of the MaNIS project, a community effort to integrate and georeference data from mammalian specimens in 17 museums: out of the 296 737 localities in the original pool of localities, only *ca*. 92 000 localities still remain to be georeferenced; the rate of advance is *ca*. 12 specimens per hour (J. Wieczorek, personal communication) The National Commission on Biodiversity of Mexico has obtained about two million georeferenced specimens, either georeferencing in-house or by cooperation with taxonomists and experts in museums and herbaria.

About 70–80% of specimen label data can be georeferenced by simple techniques and the use of gazetteers; the remaining localities may either prove impossible to reference or feasible only via the participation of experts familiar with the actual collectors. Recently, and most interestingly, much of this process is becoming automated in projects like BioGeoMancer (http://georef.nhm.ku.edu/)—recognition of locality strings has been made 'smarter', and interpreted locality descriptors are then compared with national or worldwide gazetteer databases; in this way, the bottleneck steps in the georeferencing process are automated, and human participation is focused more at the level of supervision and error checking, making the process considerably more efficient.

5. CONCLUSIONS

The massive storehouse of distributed, raw biodiversity data that the Internet is enabling will set the stage for how biodiversity patterns are analysed in the future. Abundant examples already demonstrate the rich potential of such

Figure 5. A map of registers for *Opuntia chlorotica* was presented to a specialist (Hector Hernandez, Instituto de Biología, National University of Mexico). The specimens for the obvious outliers in southeastern Mexico were checked and found to be misidentifications by the specialist.

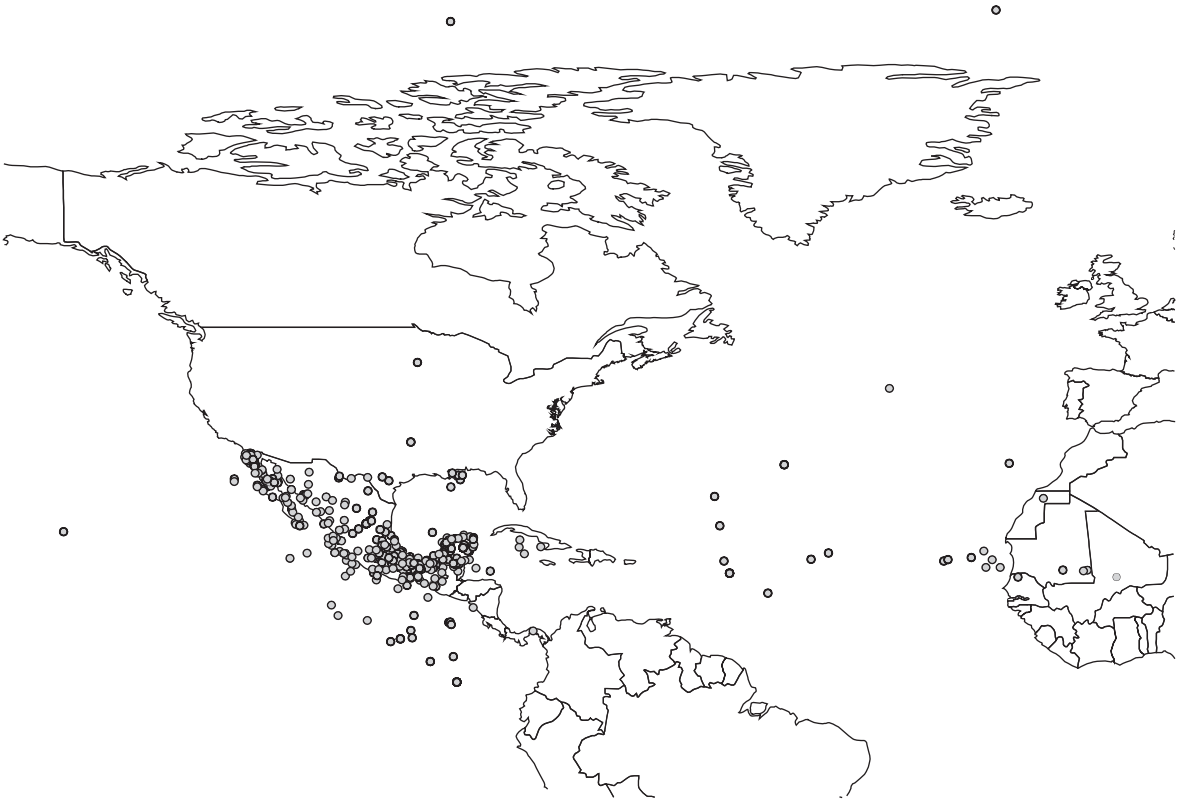

Figure 6. All the points depicted in this figure are specimens of terrestrial plants of Mexico with a faulty georeferencing. Some of them are obvious, like the Polar, mid-Atlantic or African ones. Others require careful comparison with standardized maps of Mexico, with accepted polygons for states and municipalities (often there are conflicts between states on the precise location of their borders), and a scale, projection and datum that allow consistent comparisons.

data when analysed and interpreted in the context of geospatial information as part of the nascent field of BI. Nevertheless, the demands that these technological advances will put on the shoulders of the taxonomic and systematics communities are significant—in fact, without a strong and active taxonomic community, BI will never be more than a clever set of software tools lacking a substantial factual basis.

Detection of problems associated with synonyms, misidentifications, georeferencing inconsistencies, outdated

taxonomy, and so on depend on the existence and enthusiastic participation of an active community of taxonomists. More importantly, these advances depend critically on adequate support to the fundamental infrastructure of the museums and herbaria of the world—these institutions provide the key infrastructure of the world of knowledge regarding biodiversity and every day are more endangered by cost-cutting bureaucrats.

Of course, we hope that the exploration of world biodiversity will continue and will gain new strength. Recent promising initiatives, unfortunately, have not achieved full success. For example, the most recent attempt to discover and describe the remaining species, the All Species initiative, was delayed owing to the change in world economy. The Global Taxonomy Initiative of the Convention on Biological Diversity was launched, but has not been funded (see Samper 2004). Failure to carry through with these initiatives is worrisome, as much remains to be explored, and many critical elements of biodiversity remain to be discovered.

It is clear that 'DNA Taxonomy' (Hebert *et al.* 2003; Tautz *et al.* 2003; Pennisi 2003) will add speed to the exploration of biodiversity, although there is still debate about how many of the claims of its proponents are real or feasible (Seberg *et al.* 2003). In any case, one major lesson learned from specimen-based BI analysis is that proper vouchering and georeferencing of specimens is the *sine qua non* of macroecological and biogeographical analysis. Simply 'DNA barcoding' specimens, without precise (as precise as possible, and with modern technology this is metres) reference to the locality may serve to tackle many problems in systematics, evolutionary biology and other fields, but will probably leave out whole categories of analysis.

Biodiversity Informatics is adding value to taxonomic activity in ways probably not foreseen even 10–20 years ago, and is becoming indispensable in developing countries striving to manage their biodiversity adequately. Indeed, in developing countries, national taxonomic institutions are often small and under-funded. Many large countries—and particularly those that qualify as 'megadiverse' (Mittermeier & Goetsch 1997)—are nevertheless essentially unexplored for many taxonomic groups (figure 7). For such countries, a practical answer to the lack of national taxonomic efforts or institutions is to refer to existing information and knowledge, using the array of techniques described above to improve insight. If or when the wealth of information that is held in world natural history museums and herbaria is available efficiently to those countries, the way biodiversity is managed will change radically.

To this end, enormous activity is required on the part of the museums and herbaria. Requisites include the following.

(i) Museums and herbaria must continue the enterprise of collecting new biodiversity material, which should be the richest in information content (e.g. precise geographical coordinates, detailed digital imagery or sound recordings, DNA profiling). In general, our biodiversity resources are ageing, and not only are new phenomena not represented, but a baseline of highest-quality information associated with recent specimens is in general lacking.

(ii) The museums must increase the pace for releasing good-quality raw data. Good quality means well determined (low misidentification rates), georeferenced and quality assessed and corrected (inconsistency checks performed). The cost of the above steps is significant, on the order of US $1–10 per specimen, without counting the base costs of building, curating and preserving the collections, as well as just running the institutions. The creation of the GBIF (http://www.gbif.org/), with the purpose of promoting that the world's biodiversity data will become freely and universally available, should take the preceding efforts a step forward by providing technical and financial support and by leveraging national resources and commitment to its objective.

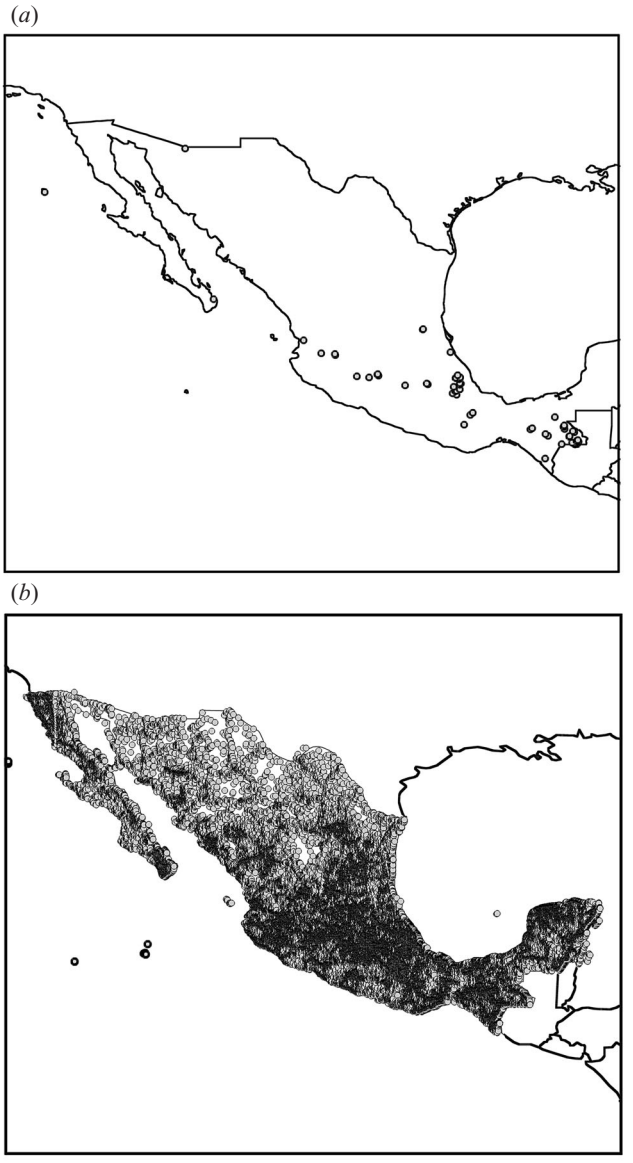

Figure 7. An example of two taxonomic groups with contrasting coverages in computerized databases. The Bryophites database (*a*) contains 1587 specimens from 18 museums. The Angiosperms of Mexico database (*b*) contains 721 175 specimens from 143 museums.

(iii) Investment in development of user-friendly, industrial-strength analytical tools should be increased. Current tools are in the 'artisan' stage, requiring suites of programs and file types to perform a single analysis, and often requiring hours and even days of computer time. Many existing software programs cannot handle the huge data matrices involved in the new world of BI, with its high spatial resolution and multispecies analyses. An effort of tool development similar to that applied to genomic and proteomic informatics will be required. Eventually, efforts will have to start to hasten the convergence of the two 'bioinformatics': the one oriented to suborganismic levels, and the one associated to species and ecosystems. Without doubt DNA barcoding will provide a powerful link between those two fields.

(iv) Training in BI will be a must. University curricula in this field are currently lacking—what is needed is a melding of aspects of biology, computer science and geography, a combination that is not often considered in university programmes.

In summary, BI consists of much more than the particular databases, tools and applications that we have mentioned in this review. Rather, BI is a sweeping, fundamental area of inquiry of which present analyses have touched only the smallest part. Many exciting and far-reaching innovations and steps forward remain to be developed, which will open new doors to funding, activity and further discovery.

Along with the novelty of the field are challenges as well. Assuring that appropriate credit is given where due (e.g. to collectors and curators), and protecting institutional ownership rights to data and their descendent products, may represent significant complications for the development of this field (Graves 2000). The BI community will have to develop and adopt the rules of behaviour that enhance the sharing of data, while preventing the proliferation of free-riders.

More fundamentally, these developments will involve the evolution of taxonomy and systematics beyond their traditional borders. Instead of just producing the traditional systematic revisions and phylogenetic treatments, BI activities will increasingly draw taxonomists and systematists into analyses and studies focused on issues of interest to agriculture, public health, invasive species and conservation. Although a distraction from the usual tasks of taxonomists and systematists, these issues are nevertheless key in developing BI into a field that will make the case for continuing and increasing support for the taxonomic and systematic enterprise.

The authors gratefully recognize the help of G. Jimenez, of the Instituto de Ecologia of Universidad Nacional Autonóma de México, Mexico, and P. Koleff and M. Schmidt of Conabio for their help in obtaining data or producing the figures. The Smithsonian Institution provided the label data from their specimens of *Cactoblastis*. The herbaria at the National University of Mexico, the National School of Biological Sciences of Mexico, the two herbaria (EB and XAL) of the Ecology Institute at Xalapa, Veracruz, the University of Sonora, the San Diego Natural History Museum, the University of Texas at Austin, the Missouri Botanical Garden, the California Academy of Sciences and the Smithsonian Institution provided *Platyopuntia* data. Cecilia Ayon helped with the revision of the last version.

REFERENCES

Alroy, J. 2002 How many named species are valid? *Proc. Natl Acad. Sci. USA* **99**, 3706–3711.

Anderson, R. P., Laverde, M. & Peterson, A. T. 2002 Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *Oikos* **93**, 3–16.

Austin, M. P., Nicholls, A. O. & Margules, C. R. 1990 Measurement of the realized qualitative niche: environmental niches of five *Eucalyptus* species. *Ecol. Monogr.* **60**, 161–177.

Bisby, F. A. 2000 The quiet revolution: biodiversity informatics and the Internet. *Science* **289**, 2309–2312.

Burgess, N. D., Rahbek, C., Williams, P., Larsen, F. W. & Balmford, A. 2002 How much of the vertebrate diversity of sub-Saharan Africa is represented by recent conservation proposals? *Biol. Conserv.* **107**, 327–339.

Carpenter, G., Gillison, A. N. & Winter, J. 1993 Domain: a flexible modeling procedure for mapping potential distributions of animals and plants. *Biodivers. Conserv.* **2**, 667–680.

Chen, G. & Peterson, A. T. 2002 Prioritization of areas in China for biodiversity conservation based on the distribution of endangered bird species. *Bird Conserv. Int.* **12**, 197–209.

Colwell, R. K. & Coddington, J. A. 1994 Estimating terrestrial biodiversity through extrapolation. *Phil. Trans. R. Soc. Lond.* B **335**, 101–118.

Cook, R. 1969 Species density of North American Birds. *Syst. Zool.* **18**, 63–84.

Csuti, B. (and 10 others) 1997 A comparison of reserve selection algorithms using data on terrestrial vertebrates in Oregon. *Biol. Conserv.* **80**, 83–97.

Danks, F. S. & Klein, D. R. 2002 Using GIS to predict potential wildlife habitat: a case study of muskoxen in northern Alaska. *Int. J. Remote Sensing* **23**, 4611–4632.

Edwards, J. L., Lane, M. A. & Nielsen, E. S. 2000 Interoperability of biodiversity databases: biodiversity information on every desktop. *Science* **289**, 2312–2314.

Egbert, S. L., Martinez-Meyer, E., Ortega-Herta, M. A. & Peterson, A. T. 2002 Use of datasets derived from time-series AVHRR imagery as surrogates for land cover maps in predicting species' distributions. *Proc. IEEE 2002 Int. Geosci. Remote Sensing Symp.* (*IGARSS*) **4**, 2337–2339.

Faundeen, J. 2003 National satellite land remote sensing data archive report. USGS/EROS Data Center, US Geological Survey, Sioux Falls.

Gaston, K. G. & Mound, L. A. 1993 Taxonomy, hypothesis and the biodiversity crisis. *Phil. Trans. R. Soc. Lond.* B **251**, 139–142.

Gioia, P. & Pigott, P. 2000 Biodiversity assessment: a case study in predicting richness from the potential distributions of plant species in the forests of southwestern Australia. *J. Biogeogr.* **27**, 1065–1078.

Godfray, C. 2002 Challenges for taxonomy. *Nature* **417**, 17–19.

Godown, M. E. & Peterson, A. T. 2000 Preliminary distributional analysis of U.S. endangered bird species. *Biodivers. Conserv.* **9**, 1313–1322.

Gotelli, N. J. 2004 A taxonomic wish-list for community ecology. *Phil. Trans. R. Soc. Lond.* B **359**, 585–597. (DOI 10.1098/rstb.2003.1443.)

Graves, G. R. 2000 Costs and benefits of Web access to museum data. *Trends Ecol. Evol.* **15**, 374.

Green, K. M., Lynch, J. F., Sircar, J. & Greenberg, L. S. Z. 1987 LANDSAT remote sensing to assess habitat for migratory birds in the Yucatan Peninsula, Mexico. *Vida Silvestre Neotrop.* **1**, 27–37.

Guisan, A. & Zimmermann, N. E. 2000 Predictive habitat destruction models in ecology. *Ecol. Model* **135**, 147–186.

Guzmán, U., Arias, S. & Davila, P. 2003 *Catálogo de las Cactáceas Mexicanas*. México D F: Universidad Nacional Autónoma de México y Comisión Nacional de Biodiversidad.

Hebert, P., Cywinska, A. S., Ball, L. & de Waard, J. R. 2003 Biological identifications through DNA barcodes. *Proc. R. Soc. Lond.* B **270**, 313–321. (DOI 10.1098/rspb.2002.2218.)

Higgins, S. I., Richardson, D. M., Cowling, R. M. & Trinder-Smith, T. H. 1999 Predicting the landscape-scale distribution of alien plants and their threat to plant diversity. *Conserv. Biol.* **13**, 303–313.

Honig, M. A., Cowling, R. M. & Richardson, D. M. 1992 The invasive potential of Australian banksias in South-African fynbos—a comparison of the reproductive potential of *Banksia ericifolia* and *Leucadendron laureolum*. *Aust. J. Ecol.* **17**, 305–314.

ICBP (International Council for Bird Preservation) 1992 *Putting biodiversity on the map: priority areas for global conservation*. Cambridge: International Council for Bird Preservation.

Jetz, W. & Rahbek, C. 2002 Geographic range size and determinants of avian species richness. *Science* **297**, 1548–1551.

Jones, P. G. & Gladkov, A. 1999 *FloraMap: a computer tool for predicting the distribution of plants and other organisms in the wild*. Cali, Colombia: Centro Internacional de Agricultura Tropical.

Kelley, C., Garson, J., Aggarwal, A. & Sarkar, S. 2002 Place prioritization for biodiversity reserve network design: a comparison of the Sites and ResNet software packages for coverage and efficiency. *Divers. Distrib.* **8**, 297–306.

Kerr, J. T. & Ostrovsky, M. 2003 From space to species: ecological applications for remote sensing. *Trends Ecol. Evol.* **18**, 299–305.

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

Krishtalka, L., Peterson, A. T., Vieglais, D. A., Beach, J. H. & Wiley, E. O. 2002 The Green Internet: a tool for conservation science. In *Conservation in the Internet age: strategic threats and opportunities* (ed. J. N. Levitt), pp. 143–164. Washington, DC: Island Press.

Lobo, J. M., Lumaret, J. P. & Jay-Robert, P. 1997 Taxonomic databases as tools in spatial biodiversity research. *Ann. Soc. Entomol. Fr.* **33**, 129–138.

Manel, S., Dias, J. M., Buckton, S. T. & Ormerod, S. J. 1999*a* Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *J. Appl. Ecol.* **36**, 734–747.

Manel, S., Dias, J. M. & Ormerod, S. J. 1999*b* 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.

Miller, R. I. 1994 *Mapping the diversity of nature*. London: Chapman & Hall.

Mittermeier , R. & Goetsch, C. 1997 *Megadiversidad: Los países biológicamente más ricos del mundo*. México D. F.: CEMEX.

Mladenoff, D. J., Sickley, T. A., Haight, R. G. & Wydeven, A. P. 1995 A regional landscape analysis and prediction of favorable gray wolf habitat in the northern Great Lakes region. *Conserv. Biol.* **9**, 279–294.

Murguía, M. & Villaseñor, J. 2000 Estimating the quality of the records used in quantitative biogeography with presence-absence matrices. *Ann. Bot. Fenn.* **37**, 289–296.

Nic Lughadha, E. 2004 Towards a working list of all known plant species. *Phil. Trans. R. Soc. Lond.* B **359**, 681–687. (DOI 10.1098/rstb.2003.1446.)

Nix, H. A. 1986 A biogeographic analysis of Australian elapid snakes. In *Atlas of elapid snakes of Australia* (ed. Australian Government Publishing Service), pp. 4–15. Canberra: R. Longmore.

Noonan, G. R. 2003. GIS Technology. A powerful tool for entomology. *Insight. A Milwaukee public museum series in natural history*, vol. 1, pp. 1–98. See http://www.mpm.edu/cr/insight/Papers/Noonan_GIS_review.pdf

Oliver, I., Pik, A., Britton, D., Dangerfield, J. M., Colwell, R. K. & Beattie, A. J. 2000 Virtual biodiversity assessment systems. *Bioscience* **50**, 441–449.

Panero, J. 2003 Catalogo de Autoridades de Asteraceae de México. University of Texas at Austin. Bases de datos SNIB-CONABIO. Proyecto AE 024. Mexico D.F.

Pennisi, E. 2003 Modernizing the tree of life. *Science* **300**, 1692–1697.

Petersen, F. T., Meier, R. & Nykjaer Larsen, M. 2003 Testing species richness estimation methods using museum label data on the Danish Asiilidae. *Biodivers. Conserv.* **12**, 687–701.

Peterson, A. T. 2003 Projected climate change effects on Rocky Mountain and Great Plains birds: generalities of biodiversity consequences. *Global Change Biol.* **9**, 647–655.

Peterson, A. T. 2004 Predictability of the geography of species' invasions via ecological niche modeling. *Q. Rev. Biol.* **78**, 419–433.

Peterson, A. T. & Robins, C. R. 2003 When endangered meets invasive: ecological niche modeling predicts double trouble for spotted owls, *Strix occidentalis*. *Conserv. Biol.* **17**, 1161–1165.

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

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

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

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

Peterson, A. T., Papes, M. & Kluza, D. A. 2003 Predicting the potential invasive distributions of four alien plant species in North America. *Weed Sci.* **51**, 863–868.

Prendergast, J. R., Wood, S. N., Lawton, J. H. & Eversham, B. C. 1993 Correcting for variation in recording effort in analysis of diversity hotspots. *Biodivers. Lett.* **1**, 39–53.

Rahbek, C. & Graves, G. R. 2000 Detection of how macroecological patterns in South American hummingbirds is affected by spatial scale. *Proc. R. Soc. Lond.* B **267**, 2259–2265. (DOI 10.1098/rspb.2000.1277.)

Rapoport, E. H. 1982 *Areography. Geographical strategies of species*, 1st edn. Oxford: Pergamon.

Rebelo, A. & Siegfried, W. R. 1992 Where should nature reserves be located in the Cape Floristic Region, South Africa? Models for the spatial configuration of a reserve network aimed at maximizing the protection of floral diversity. *Conserv. Biol.* **6**, 243–252.

Remsen, J. V. 1995 The importance of continued collecting of bird specimens to ornithology and bird conservation. *Bird Conserv. Int.* **5**, 145–180.

Reynolds, J. H. 1998 *WCMC Handbooks on biodiversity information management*, vol. 7. Data management fundamentals. Cambridge: The World Conservation Monitoring Centre.

Richardson, D. M. & McMahon, J. P. 1992 A bioclimatic analysis of *Eucalyptus nitens* to identify potential planting regions in Southern Africa. *S. Afr. J. Sci.* **86**, 380–387.

Roberts, C. (and 11 others) 2002 Marine biodiversity hotspots and conservation priorities for tropical reefs. *Science* **295**, 1280–1284.

Samper, C. 2004 Taxonomy and environmental policy. *Phil. Trans. R. Soc. Lond.* B **359**, 721–728. (DOI 10.1098/rstb.2004.1476.)

Sánchez-Cordero, V. & Martínez-Meyer, E. 2000 Museum specimen data predict crop damage by tropical rodents. *Proc. Natl Acad. Sci. USA* **97**, 7074–7077.

Scott, J. K. & Panetta, F. D. 1993 Predicting the Australian weed status of southern African plants. *J. Biogeogr.* **20**, 87–93.

Scott, J. M., Heglund, P. J. & Morrison, M. L. 2002 *Predicting species occurrences: issues of accuracy and scale*. Washington, DC: Island Press.

Scott, M., Tear, T. & Davies, F. 1996 *Gap analysis. A landscape approach to biodiversity planning.* Bethesda, MD: The American Society for Photogrammetry and Remote Sensing.

Seberg, O., Humphries, C. J., Knapp, S., Stevenson, S. W., Petersen, G., Scharff, N. & Andersen, N. M. 2003 Shortcuts in systematics? A commentary on DNA-based taxonomy. *Trends Ecol. Evol.* **18**, 63–65.

Skov, F. 2000 Potential plant distribution mapping based on climatic similarity. *Taxon* **49**, 503–515.

Soberón, J. & Llorente, J. 1993 The use of species accumulation functions for the prediction of species richness. *Conserv. Biol.* **7**, 480–488.

Soberón, J., Arriaga, L. & Lara, L. 2002 Issues of quality control in large, mixed-origin entomological databases. In *Towards a global biological information infrastructure,* Technical Report 70 (ed. H. Saarenmaa and E. S. Nielsen). Copenhagen: European Environment Agency.

Soberón, J., Llorente, J. & Benítez, H. 1996 An international view of national biological surveys. *Ann. Miss. Bot. Gard.* **83**, 562–573.

Soberón, J., Golubov, J. & Sarukhan, J. 2001 The importance of *Opuntia* in Mexico and routes of invasion and the impact of *Cactoblastis cactorum* (Lepidoptera: Pyralidae). *Florida Entomol.* **84**, 486–492.

Soberón, J., Dávila, P. & Golubov, J. 2004 Targeting sites for biological collections. In *Seed storage: turning science into practice* (ed. H. Pritchard & M. Way). Richmond: Kew Royal Botanic Gardens.

Stockwell, D. R. B. & Peters, D. P. 1999 The GARP modelling system: problems and solutions to automated spatial prediction. *Int. J. Geogr. Inform. Syst.* **13**, 43–158.

Stomes, D. M. & Estes, J. E. 1993 A remote sensing research agenda for mapping and monitoring biodiversity. *Int. J. Remote Sensing* **14**, 1839–1860.

Tautz, D., Arctander, P., Inelli, A., Thomas, R. H. & Vogler, A. P. 2003 A plea for DNA taxonomy. *Trends Ecol. Evol.* **18**, 70–74.

Turner, W., Spector, S., Gardiner, N., Fladeland, M., Sterling, E. & Steininger, M. 2003 Remote sensing for biodiversity science and conservation. *Trends Ecol. Evol.* **18**, 306–314.

Umminger, B. L. & Young, S. 1997 Information management for biodiversity: a proposed U.S. national biodiversity information center. In *Biodiversity II: understanding and protecting our biological resources* (ed. M. L. Reaka-Kudla, D. E. Wilson & E. O. Wilson), pp. 491–504. Washington, DC: Joseph Henry Press.

Veitch, N., Webb, N. R. & Wyatt, B. K. 1995 The application of geographic information systems and remotely sensed data to the conservation of heath land fragments. *Biol. Conserv.* **72**, 91–97.

Walker, P. A. & Cocks, K. D. 1991 HABITAT: a procedure for modeling a disjoint environmental envelope for a plant or animal species. *Global Ecol. Biogeogr. Lett.* **1**, 108–118.

Wilke, L., Cassis, G. & Gray, M. 1999 Quality control in invertebrate biodiversity data compilations. In *The other 99%: the conservation and biodiversity of invertebrates* (ed. W. Ponder and D. Lunney), pp. 147–153. Mosman Transactions of the Royal Zoological Society of New South Wales.

Winker, K. 1996 The crumbling infrastucture of biodiversity: the avian example. *Conserv. Biol.* **10**, 703–707.

GLOSSARY

AVHRR: advanced very high resolution radiometer
BI: biodiversity informatics
DiGIR: distributed generic information retrieval
GBIF: Global Biodiversity Information Facility
MaNIS: Mammal Networked Information System
MSS: multispectral scanner
RS: remotely sensed
TM: thematic mapper