VIEWPOINT

The Quiet Revolution: Biodiversity Informatics and the Internet

Frank A. Bisby

The massive development of biodiversity-related information systems on the Internet has created much that appears exciting but chaotic, a diversity to match biodiversity itself. This richness and the arrays of new sources are counterbalanced by the maddening difficulty in knowing what is where, or of comparing like with like. But quietly, behind the first waves of exuberance, biologists and computer scientists have started to pull together in a rising tide of coherence and organization. The fledgling field of biodiversity informatics looks set to deliver major advances that could turn the Internet into a giant global biodiversity information system.

There is a resonance between the needs of biodiversity science and the opportunities for globalization and interoperability provided by the Internet. One is that biodiversity workers are distributed all over the globe, literally dotted about in every country and on every island. A second arises from our interdependence. Global events and global syntheses in biodiversity have an impact on all of us. People who set conservation priorities do not just access local information, they need to understand the whole; they need information, for instance, from neighboring regions and from climatically similar lands in distant continents. But third, and most important, the science of global biodiversity studies depends critically on high-level concepts—biomes, ecosystems, phyla, floras and faunas, hotspots, genetic erosion, the impact of aliens—abstractions put together by synthesizing the myriad observations and studies by local observers, local teams, and local institutions. Hence, a central goal in biodiversity informatics is to develop systems that permit interoperability and knowledge synthesis across wide arrays of local systems, and to embed them in global knowledge architectures such as Species 2000 (*1*) and the Global Biodiversity Information Facility (GBIF) (*2*). Here, I give a brief picture of the research, techniques, and developments that are bringing these goals within reach.

Interoperability. One priority is to draw together basic biodiversity accession records from dispersed sites. How could we access the vast number of plant and animal records dispersed in the museums and herbaria of the world? The utility of doing this was first demonstrated by the Australian government's original Environmental Resources Information Network (ERIN) system, now part of Environment Australia Online (*3*), albeit by centralizing plant and animal distribution records. ERIN led the way by making the combined data available for Australia-wide Geographic Information System (GIS) analysis and modeling.

A number of interoperative systems are approaching the tasks originally offered by ERIN for its centralized data, but with the powerful possibility of extending to data from a vast range of autonomous institutions around the world. The Biological Collection Information Service for Europe (BioCISE) program (*4*) has established an extensive metadata system holding information centrally on the contents and locations of various collections. The idea is that intelligent software will lead users to this information, which will be retrieved using common interfaces. The University of Kansas team is developing its Species Analyst system (*5*), which interacts directly with an array of herbarium and museum accession databases. The Z39.50 protocol is used to locate and return records, and these are transformed into Extensible Mark-up Language (XML) for use by World Wide Web browsers and analytical software. The Z39.50 search profile used corresponds to the Darwin Core metadata standard (*6*) being developed informally among U.S. institutions. A request (such as for specimen records for a particular species) goes to all museums and herbaria selected by the user, and the dispersed databases return data, for instance, giving latitude, longitude, and date for every matching specimen. The assembled data set from mixed sources is then available for analysis using GIS mapping and modeling routines at the San Diego Supercomputer Center. Similar goals are being pursued by the TaxaServer group in Australia (*7*) and by the European Natural History Specimen Information Network (ENHSIN) team in Europe (*8*).

A second area for networking and interoperability is the taxonomic framework itself. Again, there are centralized models from the 1990s where organizations bring together taxonomic treatments from authors and institutions to provide a centrally collated system. It now seems agreed that these taxonomic frameworks should be constructed "taxon-by-taxon" as in Species 2000 (*1*), the Integrated Taxonomic Information System (ITIS) (*9*), and the UNESCO-IOC Register of Marine Organisms (URMO) (*10*), thus avoiding the "flora-by-flora" work of integrating systems in which the taxonomies overlap, a contrast illustrated in Fig. 1. Only the International Organization for Plant Information (IOPI) Global Plant Checklist (*11*), perhaps because of well-developed flora databases, is attempting the flora-by-flora route (*12*).

Species 2000. Species 2000 (*1*) is a global program to compile a "catalog of life" using distributed networking on the Internet. It has the ambitious aim of creating a uniform and validated index to the world's known species for use as a practical tool in inventorying and monitoring biodiversity worldwide. The index will be used to provide (i) electronic baseline species lists for use in inventorying projects worldwide, (ii) the index for an Internet digital library of species databases worldwide, (iii) a reference system for comparison between inventories, and (iv) a comprehensive worldwide catalog for checking the status, classification, and naming of species.

The comprehensive index of all known plants, animals, fungi, and microorganisms is being constructed by accessing a distributed array of taxonomic indexes, one for each group of organisms. These are known as global species databases (GSDs), represented by boxes in the primary array of the Species 2000 architecture (Fig. 2). The taxonomic database organizations starting the program already provide such indexes for viruses, bacteria, archaea, corals, algae (red, green, and brown), cephalopods, crustaceans, scarabaeid

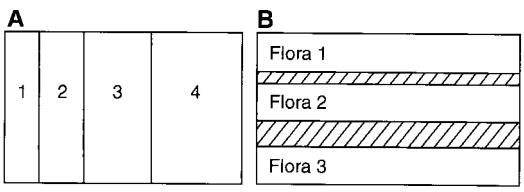

Fig. 1. Comparison between global taxonomies assembled from taxonomic treatments for complete taxa (taxon-by-taxon, no overlap) (**A**) and from floras or faunas (in this example, flora-by-flora, with overlaps) (**B**). [Adapted from (*12*)]

Centre for Plant Diversity & Systematics, School of Plant Sciences, University of Reading, Reading RG6 6AS, UK. E-mail: f.a.bisby@reading.ac.uk

beetles, tineid and geometrid moths, weevils, fishes, birds, mammals, some groups of fungi, mosses, and angiosperms (including fagales, legumes, and umbellifers). Organizations with databases covering a further 85 major groups are joining the program, and it is projected that existing database projects may provide for about 55% of known species. Partner programs participating in Species 2000 are URMO (*10*), ITIS (*9*), and the IOPI Global Plant Checklist (*11*).

Current estimates are that about 1.75 million species are "known" in the sense that they have been described and named by taxonomists. At least 150 global species databases, each initially covering 10,000 to 25,000 species, will be needed for all to be included. Species 2000 proposes to stimulate completion of the array of taxonomic databases. It will seek resources both to complete the existing databases and to help establish new databases to cover the gaps, thought to account in total for about 45% of species. The present prototypes on the Web site (Fig. 3) are to be replaced with enlarged systems for both the Dynamic Checklist and the Annual Checklist during late 2000. In early 2001 they should reach the "critical mass" of about 300,000 species covered.

An important development is the provision of onward species links as part of the digital library, depicted as lines to the secondary array of databases in Fig. 2. Once a species has been located, onward links will be provided to rich data sources for that species in a variety of conservation, molecular, germplasm, or ecological databases in different countries. Prototype onward links are already available from some of the start-up components of Species 2000, such as the International Legume Database & Information Service (ILDIS) LegumeWeb (*13*), FishBase (*14*), and the Bacteriology Insight Orienting System (BIOS) (*15*).

The SPICE for Species 2000 project (*16*) is carrying out the enabling research for the interoperability system behind the Species 2000 Dynamic Checklist. The purpose is to poll an array of up to 200 global species databases on the Internet, one for each group of organisms, to provide a functional virtual catalog of all known species. The scalability of the system, heterogeneity of the databases, stability of dispersed academic sites, and autonomy in their management all contribute to making this a challenging assignment. SPICE 1 is under trial and uses Common Object Request Broker Architecture (CORBA) object brokering to link to the array of GSDs. They are linked to SPICE 1 either within CORBA (an "undivided" wrapper) or via a two-part or "divided" wrapper using Common Gateway Interface (CGI)/XML. Species 2000 is also testing an Annual Checklist at the International Center for Living Aquatic Resources Management (ICLARM), Philippines, as a stable reference (updated once a year) to be made available as a CD-ROM "catalog of life" as well as on the Internet.

Species 2000 was established by the International Union of Biological Sciences (IUBS), the ICSU (International Council for Science) Committee on Data for Science & Technology (CODATA), and the International Union of Microbiological Societies (IUMS). It is endorsed by the United Nations Environment Programme (UNEP) and is associated with the Clearing House Mechanism of the UN Convention on Biological Diversity. It is planning to work closely with GBIF (*2, 17*).

During its current development phase, the Species 2000 project team is eager to contact the custodians of global species databases covering any group of organisms worldwide with a validated taxonomic component, as well as regional biodiversity systems wishing to connect to and from the digital library.

Taxonomic backbones. A remarkable element of the new biodiversity systems is the central role being played by taxonomy. Like it or not, Latin names and the skilled handling of synonymy provide the indexing key to much of the organism data and the links to data provided by associated disciplines such as genomics, ethnobiology, and natural products. Of course, most users of biodiversity systems are not primarily interested in the taxonomy per se. They want to find the right data for the right organism, preferring the Latin names and taxonomic complications to remain out of sight, and to locate exactly what they want, as if by magic.

The niche for "taxonomic backbones" has brought a variety of taxonomic databases onto the Web. The early ad hoc systems were driven initially by need and only later by the efforts of taxonomists, such as the catalogs of the World Conservation Monitoring Centre (WCMC), NAPRALERT (Natural Products Alert, Chicago), the U.S. National Center for Biotechnology Information (NCBI), the USDA Genetic Resources Information Network (GRIN), and the System-wide Information Network for Genetic Resources (SINGER) (*18*). More recently the taxonomists have generated regional works equivalent to faunas and floras, such as FloraBase (plants of Western Australia) (*19*) and the European Register of Marine Species (ERMS) (*20*), and to monographs or global catalogs such as CephBase (world cephalopods) (*21*) and ILDIS LegumeWeb (world legumes) (*13*). Indeed, the recent completion of one of these [FishBase (*14*) with 25,000 species, 19 August 2000] is a cause for celebration, the completion of a major sector of the "catalog of life."

Making these taxonomic systems suitable for their new Internet role has proved nontrivial and has generated some fascinating bioinformatics research into knowledge models and functionality. Central here are the structural relation between names and taxa (one species may have had more than one name) and the need to traverse between alternative taxonomies in use for the same organisms. The fluid nature of nomenclature and classification has continued to infuriate system administrators, and doomed attempts to freeze taxonomy continue to this day.

At one level the names-and-taxa problem is solved by synonymic indexing, but the preoccupation of taxonomists with names has meant that this has been slow to be intro-

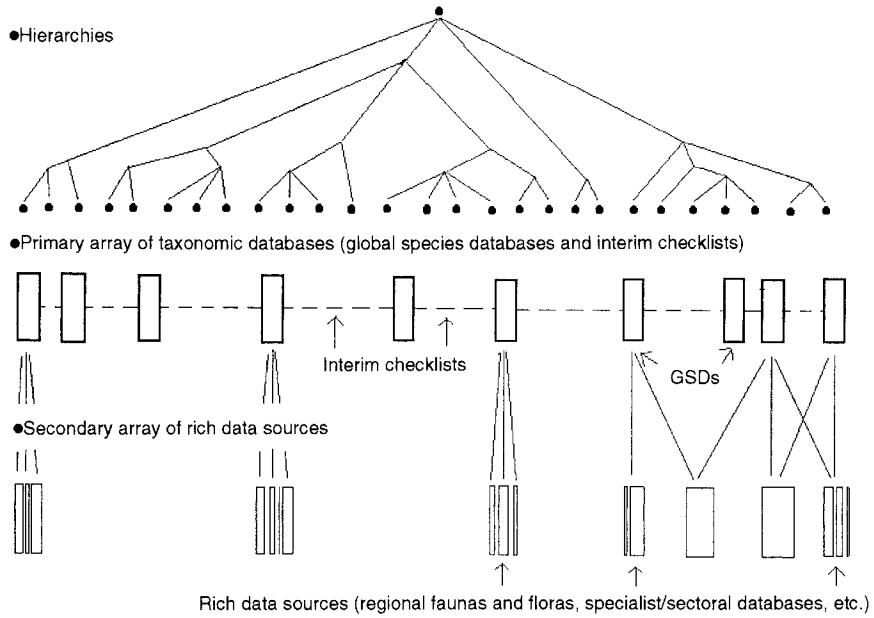

Fig. 2. The Species 2000 architecture for building a "catalog of life" from an array of global species databases and for providing onward links to other data sources. [From (*32*)]

duced. The ALICE software (*22*) and ILDIS LegumeWeb (*13*) were among the first to present a taxon interface, for instance for a species, in which a single accepted name and all relevant synonyms are shown together for one species, as has always been the case in printed floras, faunas, and monographs. Automated synonymic indexing means that a user who makes an inquiry under any name treated as a synonym for a species is taken directly to the taxon page for that species.

Synonymic indexing suffers one major drawback: The codes of nomenclature permit the same name to be used for differing "breadths" or circumscriptions of species, provided they all overlap in containing the type to which the name is attached. This can lead to data for differing definitions of a species being badly confused, in particular being confused with data for segregate taxa. In theory this problem has been resolved by Beach *et al*. (*23*), who point out that it is not the accepted name "*Genista sylvestris* L." but rather the accepted treatment "*Genista sylvestris* L. as used by Flora Europaea (Tutin *et al*. 1964 *et seq*.)" that should be the unit of indexing. Zhong *et al*. (*24*) and Berendsohn (*25*) have developed operative models based on this idea, but these have yet to reach generic software packages. A further problem arises when inquiries are made using that small fraction of names that are pro parte synonyms, misapplied names, or homonyms given without author; in all three cases, the name is ambiguous and neither the system nor the user may know which of two species is referred to. Such ambiguities become frequent when common names are used.

The UK Bioinformatics Initiative has funded research on two very different approaches. In one approach, Pullan *et al*. (*26*) have modeled taxonomy treating each taxon as defined by the set of specimens examined and included, somewhat in the style of specimen citations in some floras. This enables them to connect different sets of specimens to alternative taxonomies, and thus to build a precise "cross-map" between them. The approach has an attractive practicality at the herbarium level, but it also illustrates the gargantuan, perhaps impossible, task of attempting to document the cross-maps between all conflicting taxonomies created by different taxonomists at different sites. Previously, ILDIS had created a cross-map based on circumscriptions between 2254 legume species known in the former Soviet Union and the 1200 species to which they were thought to be equivalent in ILDIS LegumeWeb. The published CD-ROM *Legumes of Northern Eurasia* (*27*) may be the first published taxonomic work with functional alternative taxonomies on the screen, but the bad news was that it took several taxonomists 2 months to create the cross-map.

A different approach has been taken by the Logic-based Integration of Taxonomic Conflicts in Heterogeneous Information Systems (LITCHI) consortium (*28*), which is applying rule-based systems to glean as much taxonomic intelligence as possible from automated examination of the taxa, synonymy, and annotations in one synonymic checklist, and comparison with a second treatment. This approach is based on the idea that disturbances in the association of accepted names, synonyms, and pro parte synonyms can demonstrate which species may possibly be identical in two treatments, which species are definitely different, and which species may be identical in circumscription but different in name (such as species moved to another genus) (*29*). The existing LITCHI software uses its rule-based conflict detection engine to detect disparities between two taxonomic treatments. Further developments involve a number of "taxonomically intelligent" processes. Potentially the most valuable of these would enable taxonomically intelligent species links to be made on the Web, so that onward links from one system to another may locate the matching species even though it is listed under a quite different name.

Together, these developments add up to a considerable advance: We are well on the

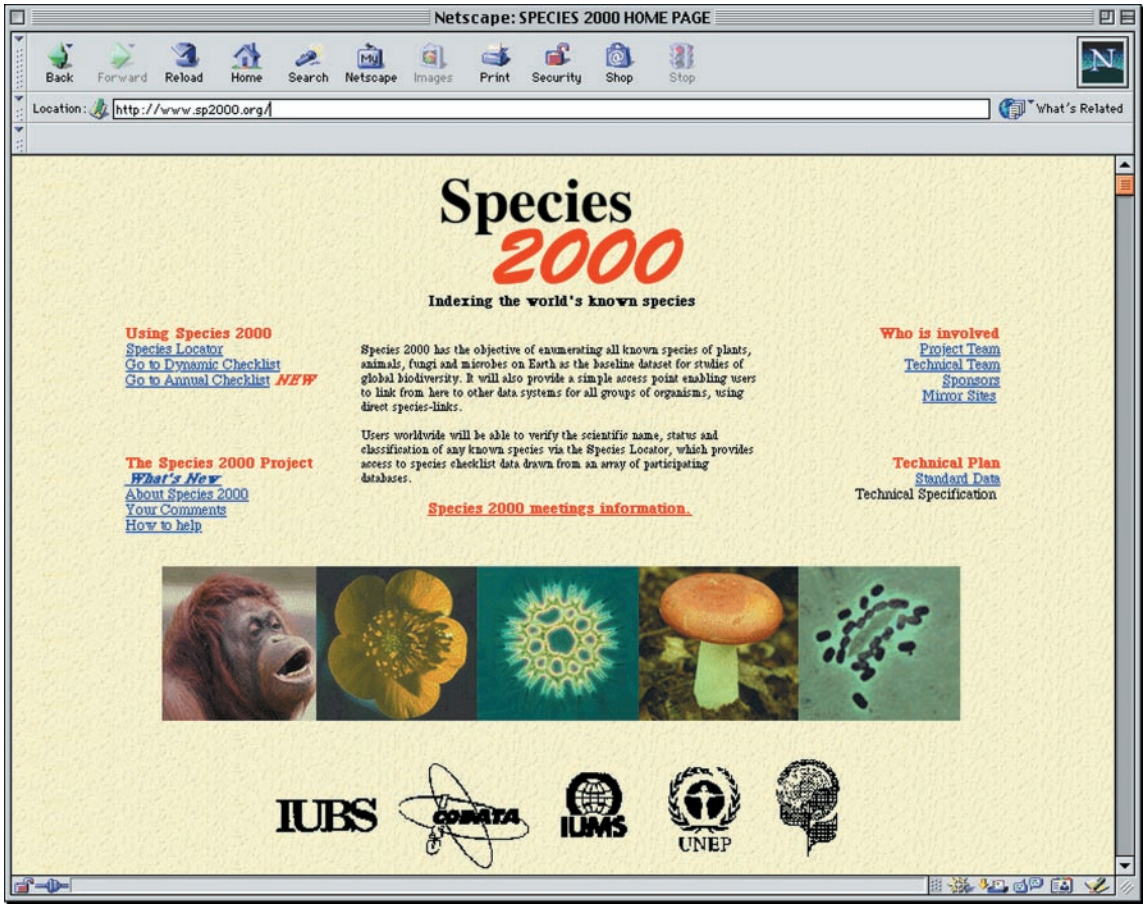

Fig. 3. Species 2000 home page, with access to the Dynamic Checklist and the Annual Checklist.

way to creating the "invisible magic" needed to locate and track data for precisely the right organism in desktop biodiversity systems on the Internet.

Robots and knowledge integration. The MultiFlora system (*30*) being developed at the University of Manchester and the Natural History Museum, London, uses Information Extraction (IE) to seek information on the same species from parallel unstructured text resources. Information from the various sources discovered is then returned, using XML, and assembled into a single database, with some interesting features for handling variable and conflicting data. The idea is that redundancies between sources may allow the system to create accurate databases despite some of the shortcomings of IE techniques.

A further biodiversity analytical system is in development at the Natural History Museum, London. The WORLDMAP system (*31*) can be used with distribution data sets to plot measured species biodiversity distribution patterns and to highlight hotspots and areas of endemicity. Of interest is the array of biodiversity measures provided, including the much debated taxic measures that incorporate distances over the phylogeny.

Startling as all these developments may be, they might just be the tip of an iceberg, preceding undreamt-of models in the coming century. Certainly there are those who expect the Internet, as seen by biologists, to become one giant global biodiversity information system. Even biologists who spend a lifetime of travel and fieldwork cannot observe the whole. But as an abstraction, could global biodiversity come to exist, modeled and visualized, on the Internet as nowhere else?

References and Notes

1. Species 2000 (www.sp2000.org).
2. J. L. Edwards, M. A. Lane, E. S. Nielsen, *Science* **289**, 2312 (2000).
3. Environment Australia Online (www.environment.gov.au/search/search.html).
4. BioCISE project (www.bgbm.fu-berlin.de/biocise/default.htm#).
5. The Species Analyst (http://habanero.nhm.ukans.edu).
6. Darwin Core metadata standard (http://habanero.nhm.ukans.edu/Z.X).
7. R. Leow and K. Taylor, in *Proceedings of the 12th International Conference on Scientific and Statistical Database Management,* O. Gunther, Ed. (IEEE Computer Society Press, Los Alamitos, CA, 2000), pp. 25–38.
8. ENHSIN (www.nhm.ac.uk/science/rco/enhsin).
9. ITIS (www.itis.usda.gov).
10. URMO (www2.eti.uva.nl/database/urmo/default.html).
11. IOPI Global Plant Checklist (http://bgbm3.bgbm.fu-berlin.de/iopi/gpc).
12. F. A. Bisby, in *Designs for a Global Plant Species Information System,* F. A. Bisby, G. F. Russell, R. J. Pankhurst, Eds. (Oxford Univ. Press, Oxford, 1993), pp. 145–157.
13. ILDIS LegumeWeb (www.ildis.org).
14. FishBase (www.fishbase.org).
15. BIOS (www-sp2000ao.nies.go.jp/bios/index.html).
16. Species 2000 Interoperability Coordination Environment (SPICE for Species 2000) project (www.systematics.reading.ac.uk/spice).
17. GBIF (www.gbif.org).
18. SINGER (www.singer.cgiar.org).
19. FloraBase (http://florabase.calm.wa.gov.au).
20. ERMS (http://erms.biol.soton.ac.uk).
21. CephBase (www.cephbase.dal.ca).
22. ALICE Software (http://dialspace.dial.pipex.com/town/square/fd95).
23. J. H. Beach, S. Pramanik, J. H. Beaman, in *Advances in Computer Methods for Systematic Biology: Artificial Intelligence, Databases, Computer Vision,* R. Fortuner, Ed. (Johns Hopkins Univ. Press, Baltimore, 1993), pp. 241–252.
24. Y. Zhong, S. Jung, S. Pramanik, J. H. Beaman, *Taxon* **45**, 223 (1996).
25. W. G. Berendsohn, *Taxon* **44**, 207 (1995).
26. M. Pullan, M. Watson, J. Kennedy, C. Reguenaud, R. Hyam, *Taxon* **49**, 55 (2000).
27. G. P. Yakovlev, Y. R. Roskov, A. K. Sytin, S. A. Jezniakowsky, *Legumes of Northern Eurasia* [CD-ROM] (ILDIS Northern Eurasia, St. Petersburg, 1998).
28. LITCHI Project (http://litchi.biol.soton.ac.uk).
29. A. C. Jones *et al*., in *Proceedings of the 12th International Conference on Scientific and Statistical Database Management,* O. Gunther, Ed. (IEEE Computer Society Press, Los Alamitos, CA, 2000), pp. 3–13.
30. MultiFlora (www.cs.man.ac.uk/ai/MultiFlora).
31. WORLDMAP (www.nhm.ac.uk/science/projects/worldmap).
32. F. A. Bisby and P. M. Smith, Eds., *Species 2000 Project Plan* (Species 2000 and UNEP, Southampton, UK, 1996).

33. I thank J. Heald and S. Brandt for comments on the text and R. White, J. Beach, M. Scoble, and J. Croft for assistance with particular items.

VIEWPOINT

Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop

James L. Edwards,[1]* Meredith A. Lane,[2] Ebbe S. Nielsen[3]

Data about biodiversity are either scattered in many databases or reside on paper or other media not amenable to interactive searching. The Global Biodiversity Information Facility (GBIF) is a framework for facilitating the digitization of biodiversity data and for making interoperable an as-yet-unknown number of biodiversity databases that are distributed around the globe. In concert with other existing efforts, GBIF will catalyze the completion of a Catalog of the Names of Known Organisms and will develop search engines to mine the vast quantities of biodiversity data. It will be an outstanding tool for scientists, natural resource managers, and policy-makers.

Biodiversity is distributed all over the Earth, with the highest concentration in tropical regions, especially in developing countries, and in the oceans. In contrast, scientific information about biodiversity is largely concentrated in major centers in developed countries, especially in the scientific collections of the world's natural history museums, herbaria, and microorganismal repositories. At present, it is more likely that information on the plants of a particular part of Africa is stored in an herbarium in Europe, for example, than in its source country. Approximately 3 billion specimens of organisms of all types are held in the natural history collections of the world (*1*). Each of these specimens has associated data, including, at the minimum, the scientific name of the specimen, when and where it was collected, and by whom. Many specimens also have other kinds of associated information, including pointers to other physical samples derived from the specimen (e.g., frozen tissues, DNA extracts, hosts, parasites), photographs, recordings of mating calls or other behavior, the field notes of the collector(s), and a wide range of other data.

In toto, then, there is an enormous amount of information already collected about the world's biodiversity. However, to date most of this information has not been digitized. Thus, in most cases, the only way a potential user can find out about the data is to travel physically to the place where the specimen is housed or to contact the repository where a relevant specimen may be housed and ask to borrow it (and its associated data).

The sustainable use and management of biodiversity will require that information about it be available when and where that information is needed by decision-makers and scientists alike. Because biodiversity information is not immediately at hand, it is often not applied in policy or management decisions that affect the organisms involved, nor is that information readily accessible by

[1]Directorate for Biological Sciences, National Science Foundation, Arlington, VA 22230, USA. [2]Academy of Natural Sciences, Philadelphia, PA 19103, USA. [3]Australian National Insect Collection, CSIRO Entomology, Canberra ACT 2601, Australia.

*To whom correspondence should be addressed. E-mail: jledward@nsf.gov