Research and Societal Benefits of the Global Biodiversity Information Facility

JAMES L. EDWARDS

In recent months, articles have been published both questioning and extolling the value of digitizing museum specimen data. Making a point for the doubters, Wheeler and colleagues (2004) stated that "some naively see the information technology challenge as liberating data from cabinets. The reality is that for all but a few taxa, much data is outdated or unreliable. Many specimens represent undescribed or misidentified species. Rapid access to bad data is unacceptable; the challenge is not merely to speed data access but to expedite taxonomic research" (p. 285).

On the other side, Suarez and Tsutsui (2004) described many uses for specimen data, for taxonomy as well as for a wide range of other areas, and noted that "the benefit of these collections to society must be maximized by stepping up the rate at which this information is entered in databases and made accessible" (p. 73). Likewise, Raxworthy and colleagues (2003) showed how both "old" museum data (specimens collected before 1978) and newer data were accurate predictors of reptile diversity in Madagascar. These researchers even used such data to make successful predictions about where new species of chameleons would be found.

Who has the upper hand in this debate? I believe some of the most persuasive current evidence for the value of museum specimen data comes from CONABIO, Mexico's Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (*www.conabio.gob.mx/*). Over many years, CONABIO has systematically gathered data about specimens of plants and animals collected in Mexico from natural history collections and herbaria all around the world. This as yet unparalleled database allows CONABIO scientists to undertake analyses ranging from systematic studies to optimal placement of biosphere reserves, and from analyses of the current and probable future effects of invasive species to predictions of the effects of climate change.

Globally, natural history collections and herbaria contain a far vaster amount of information, but because it is not dynamically accessible, even many taxonomists do not know it exists. The good news is that CONABIO's concept is now, in effect, being replicated worldwide through the Global Biodiversity Information Facility (GBIF), an Internet-accessible, interoperable network of biodiversity databases and information technology tools. In February 2004, GBIF went online with a prototype data portal (*www.gbif.net*) for simultaneously accessing data from the world's natural history collections, herbaria, culture collections, and observational databases. The portal is currently serving data from more than 50 providers of specimen and observational data, and from more than 20 names databases. It allows users to search for georeferenced specimen and observational data, check for common and scientific names for organisms (including synonyms), plot maps showing the known localities of specimens in the system, and retrieve lists of taxa by country. Although GBIF currently makes accessible only a small proportion of the world's digitized biodiversity data, by the end of 2004 it should be serving more than 100 million specimen and observational records.

One of GBIF's major goals is to help digitize biodiversity data from specimens in developed countries that were originally collected in other parts of the world, so that the data can be easily shared with the countries of origin. This impetus promotes scientific collaboration and helps to overcome the digital divide.

Participation in the GBIF consortium is open to any country or relevant international organization. The consortium currently consists of more than 65 participants (see the full membership list at *www.gbif.org/GBIF_org/participation*). As a freestanding entity, GBIF is not part of the Convention on Biological Diversity (CBD) or any United Nations body, although it cooperates closely with the CBD Secretariat, with UNESCO's Man and the Biosphere Program, and with the World Conservation Monitoring Centre of the United Nations Environment Programme. Among the nongovernmental organizations participating in GBIF are the Ocean Biogeographic Information System (O'Dor 2004), the World Conservation Union, and the World Federation of Culture Collections.

Each participant agrees to develop a computerized node (or nodes) for sharing its data. All control over the data available through the GBIF portal remains with the data providers. They are responsible for deciding what data to serve and what restrictions they wish to put on the use of those data. The system employs open-source software, including the DiGIR protocol for sharing data (*http://digir.sourceforge.net/*), and requires rigorous data standards. The backbone of the portal is the Electronic Catalogue of Names of Known Organisms (ECAT), which is being built in collaboration with several providers of name data and is intended to be an authoritative file detailing all of the world's species names. GBIF's plans call for catalyzing the completion of ECAT by the year 2011, 10 years after GBIF came into being. The GBIF node in the United States is the National Biological Information Infrastructure, but institutions

can share data with GBIF without going through the node; 15 US institutions currently serve more than 4.5 million specimen records directly to the GBIF portal.

Visitors to the portal, for their part, are required to accept a data use agreement, which stipulates that they consent to abide by the data restrictions of the providers and to publicly acknowledge the source of any data they mine from the portal and subsequently use.

GBIF is helping to develop a cataloguing system that will allow researchers to more easily explore global museum holdings. For poorly known groups of taxa, the proper identification of any one specimen may be less important than knowing that a particular museum has significant holdings for a particular clade that were collected from a geographic region over a particular period. I judge that this information will often expedite taxonomic research just as much as new collecting efforts do, because the combined resources of the world's natural history collections form an unsurpassed tool. For now, GBIF serves up only species- and specimen-level biodiversity data, but the consortium intends eventually to serve data interoperably at all levels, from molecules to ecosystems.

The challenges facing this ambitious scheme should not be underestimated. Wheeler and colleagues (2004) are probably right in their observation that many specimens are misidentified and that some georeferences for specimens are incorrect. I believe, however, that one of the best ways to expose those errors is to make the data visible, so that all qualified researchers can compare and correct them. This is why every record served through the GBIF portal contains a feedback button, which enables users to comment on the data and propose corrections. Eventually, GBIF may make these comments public so that users can participate in rating records, much as users of Amazon.com can evaluate books. This service should help taxonomists to identify outliers and correct mistakes.

Wheeler and colleagues, despite their cautious stance, acknowledge that the World Wide Web presents a perfect medium for exploring how to undertake collaborative taxonomic studies. In fact, this is to be one of GBIF's future areas of emphasis, along with promoting digital libraries of biodiversity literature. By working in these areas, GBIF will be helping to achieve Wheeler and colleagues' vision of "virtual monographs, revisions, floras and faunas that are living dynamic works rather than static documents."

Another motivating force for GBIF is the need to train personnel around the world in biodiversity informatics. An academic curriculum is needed to train people to mine the vast wealth of biodiversity information and put it to use in society. To this end, GBIF has joined forces with UNESCO to develop a network of GBIF–UNESCO Chairs in Biodiversity Informatics. The network will start with six chairs, four in developing countries and two in developed countries, to be named in 2005.

In an analysis of the lessons to be learned from the Human Genome Project, Collins and colleagues (2003) noted that much of the project's success was due to rapid and full release of genomic data. They observed that discovery-based science, built upon analysis of the genomic data, has enabled a wide range of new directions in biomedical science, and they stated that "other candidates for the 'big science' approach to biology include interagency and international approaches to biological database creation and maintenance" (p. 290).

I could not agree more. Science and society stand to gain much when species-level data are accessible online and can be easily and quickly combined with molecular and ecological data. Data mining will turn up gems of insight and understanding that cannot be predicted but are likely to lead to fruitful new directions for research. Such insights are vital if humanity is to learn to use its natural resources sustainably.

James L. Edwards (e-mail: jedwards@gbif.org) is executive secretary of the Global Biodiversity Information Facility, Universitetsparken 15, DK-2100 Copenhagen, Denmark.

References cited

Collins FS, Morgan M, Patrinos A. 2003. The Human Genome Project: Lessons from large-scale biology. Science 300: 286–290.

O'Dor R. 2004. A census of marine life. BioScience 54: 92–93.

Raxworthy CJ, Martinez-Meyer E, Horning N, Nussbaum RA, Schneider GE, Ortega-Huerta MA, Peterson AT. 2003. Predicting distributions of known and unknown reptile species in Madagascar. Nature 426: 837–841.

Suarez AV, Tsutsui ND. 2004. The value of museum collections for research and society. BioScience 54: 66–74.

Wheeler QD, Raven PH, Wilson EO. 2004. Taxonomy: Impediment or expedient? Science 303: 285.