Gilman, E., King, N., Peterson, T., Chavan, V., Hahn, A. In Press 2009. Building the Biodiversity Data Commons - The Global Biodiversity Information Facility. IN Maurer, L. (Ed.) *ICT for Biodiversity*.

BUILDING THE BIODIVERSITY DATA COMMONS THE GLOBAL BIODIVERSITY INFORMATION FACILITY

Eric Gilman[1], Nicholas King[2], A. Townsend Peterson[3], Vishwas Chavan[2], Andrea Hahn[2]

[1] Global Biodiversity Information Facility & University of Tasmania, EricLGilman@gmail.com
[2] Global Biodiversity Information Facility
[3] Biodiversity Institute, University of Kansas

Abstract: The Global Biodiversity Information Facility (GBIF) is an inter-governmental organization mandated to construct a biodiversity informatics research infrastructure to enable free and open access to biodiversity data worldwide - the 'biodiversity data commons.' Since its inception in 2001, GBIF has developed the most comprehensive portal to primary biodiversity data in the world, currently enabling access to >177M biodiversity records and >1M species names, served by 291 publishers in 35 countries. The GBIF infrastructure provides tools for data owners to publish Internet-accessible copies of their data in an internationally-agreed, standardized format to ensure interoperability among datasets. Through training, access to international experts, and mentoring programmes, GBIF builds the capacity of national and regional institutions to become active, fully functional 'biodiversity information facilities,' as part of the coordinated, globally distributed GBIF network. Numerous studies demonstrate the utility of primary species-level point occurrence data in improving our understanding of the main drivers of changes and losses of global biodiversity. The example of invasive alien species is described to demonstrate how GBIF-enabled data are contributing to this research domain. However, progress to date in mobilizing data and integrating online biodiversity datasets remains inadequate for many priority applications, owing to thematic gaps in content, such as coverage of some ecosystem types, taxonomic groups, regions, and time periods. GBIF-enabled data are biased taxonomically towards better-studied groups, especially birds, and over two-thirds of GBIF-enabled, georeferenced records come from just three countries (USA, Sweden, UK); dataset publication in Africa, Asia, and Oceania is particularly underrepresented. The overriding bottleneck hampering progress in filling gaps is not informatics tools, which are reasonably well-developed. Instead, the principal barriers are capture of biodiversity data currently in non-digital formats, insufficient institutionalization of incentives for publishing data, and inadequate enforcement of existing relevant policies. Operationalizing a proposed GBIF-enabled 'Data Publishing Framework,' in combination with needed changes in policy and legal frameworks, will address many of the obstacles to biodiversity dataset publication, discovery, and use.

Keywords: Biodiversity informatics; data publishing; Global Biodiversity Information Facility

1 Introduction

'Biodiversity informatics,' a term first used only in the early 1990s, is an emerging field that includes efforts to make global biodiversity information resources available in convenient digital formats, and to develop effective tools for their analysis and understanding [Edwards et al., 2000; Guralnick and Hill, 2009]. The global community has recognized the importance of free and open access to biodiversity data for many years. In 1992, recognising wide disparities in access to such data, the Rio Earth Summit's Agenda 21 noted that [UNGA, 1992]:

> *… the gap in the availability, quality, coherence, standardization and accessibility of data between the developed world and the developing world has been increasing, seriously impairing the capacities of countries to make informed decisions concerning environment and development.*

and that

> *… there is a lack of capacity, particularly in developing countries, and in many areas at the international level, for the collection and assessment of data, for their transformation into useful information and for their dissemination.*

In 1999, a meeting of the Organization for Economic Co-operation and Development Committee for Scientific and Technological Policy at Ministerial Level called for the formation of an intergovernmental organization to coordinate the standardization, digitization, and global dissemination of world biodiversity data, leading to the formation of a new intergovernmental body, the Global Biodiversity Information Facility (GBIF), in 2001 [OECD, 1999].

GBIF is the sole global biodiversity informatics organization established via an intergovernmental agreement. Based on a non-binding Memorandum of Understanding (MoU) between countries, GBIF's mission is to "... make the world's biodiversity data freely and openly available via the internet." GBIF membership currently includes 50 countries and 40 international organizations, where, under the MoU, Voting Participants are countries that have signed the MoU and make a financial contribution, and as a result have voting rights on the GBIF Governing Board, as well as other advantages such as access to capacity-building programmes; Associate Participants are either countries that are not yet making financial contributions to GBIF or international or intergovernmental organizations or economies that are willing to observe the MoU provisions, i.e., contribute to the GBIF mission. Fig. 1 shows the current distribution of GBIF country members (but does not show member organizations). The majority of international organizations and initiatives working in the field of biodiversity informatics are already Associate Participants, including the Encyclopedia of Life (Chapter 2), an international project compiling a webpage for every species, the Long Term Ecological Research Network (Chapter 1), a broad network of sites under intense ecological study, and the Taxonomic Data Working Group, an international professional association of individuals that develops standards and protocols for sharing biodiversity data, many of which are used by GBIF to achieve its goals (e.g., Darwin Core and Access to Biological Collections Data standards, Access Protocol for Information Retrieval, and Life Science Identifiers protocols).

GBIF promotes the free dissemination of biodiversity data, and does not assert any proprietary rights to the data made accessible via the GBIF network. GBIF is thus neither a data repository nor data aggregator; instead, data publishers retain all rights and responsibilities associated with the data that they choose to make freely accessible via the GBIF network, and determine what access and use restrictions, if any, they choose to apply.

The GBIF infrastructure provides an Internet-based, globally-distributed network of interoperable databases containing primary biodiversity data – data records placing a particular taxon at a particular place and point in time. These primary data are drawn from information associated with natural history collections and field observations of organisms. GBIF provides tools for data owners to publish Internet-accessible copies of their datasets in an internationally-agreed, standardized format to ensure interoperability among resources. In addition, a system is under development by which users will be able to better discover diverse biodiversity data resources, including a metadata cataloguing system, the Global Biodiversity Resources Discovery System (GBRDS), which will enable discovery of data-holders and datasets globally. Finally, GBIF is developing a global architecture of taxonomic nomenclature by which to resolve differences in scientific and common names between integrated datasets.

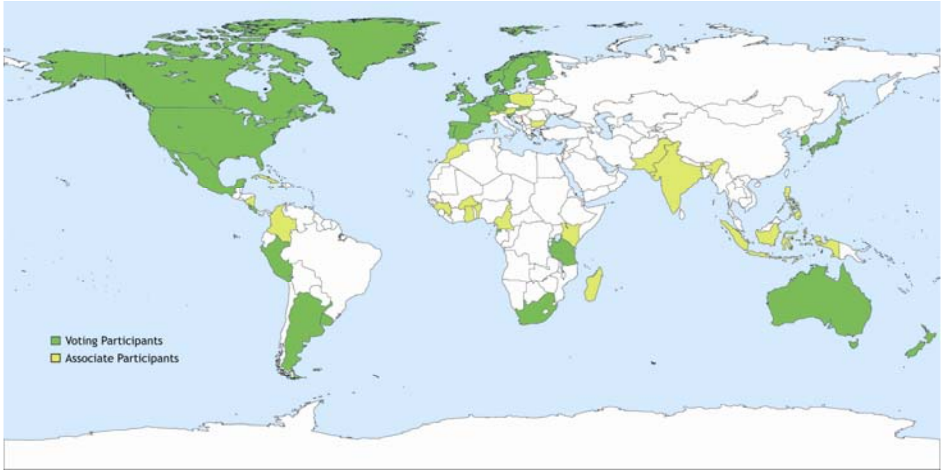

Fig. 1 GBIF 30 Voting and 20 Associate Participant countries (July, 2009)

2 BIODIVERSITY INFORMATICS: A CRITICAL TOOL TO UNDERSTAND AND CONSERVE BIODIVERSITY

Rationale for Global Integration of Biodiversity Datasets

Recognition of a growing biodiversity crisis in the late 1980s resulted in wide acknowledgment of the need for broad-scale understanding and analysis of the dimensions, distribution, and processes comprising global biodiversity [Wilson, 1988]. Biodiversity data form the foundation for understanding the status and trends in biodiversity. Because human threats to biodiversity extend across broad spatial and temporal scales, biodiversity and ecosystem monitoring, forecasting, and risk assessments require data to be organized in a globally-accessible, integrated infrastructure. Furthermore, by pooling research-grade datasets, this allows repeated use in perpetuity, multiplying returns on investment expended in collecting data.

Integrating datasets establishes more equitable access to research-grade information: researchers in developed but biodiversity-poor countries have historically had relatively easy access to biodiversity information, while researchers in biodiversity-rich, yet less-developed countries have had limited access [Canhos et al. 2004; Gaikwad and Chavan, 2006; Collen et al., 2008]. Hence, an additional rationale for making research-quality data freely available is that these data are a public good produced in the public interest, and often with public funding, i.e., information on biodiversity is effectively the intellectual property of the home country, such that barring access to this information violates the public trust [Arzberger et al., 2004]. Technology permitting free and open access to primary biodiversity data, and for the discovery of biodiversity data resources via rich metadata, can help resolve this digital divide by building the biodiversity data commons.

Applications of GBIF-enabled Datasets

Primary biodiversity data are a recognized important source of information for biological research [e.g., Yesson et al., 2007], for example, for modeling species' current and forecasted distributions [Sánchez-Cordero and Martinez-Meyer, 2000], assessing changes (e.g., changes in population sizes, distribution, species composition/richness) and losses (e.g., species extirpations and extinctions) in biodiversity [Ponder et al., 2001; Jarvis et al., 2008; van Zonnevald et al., 2009], taxonomic revisions [Pennisi, 2000], and compiling lists of threatened species [Shaffer et al., 1998].

Primary occurrence data are used to document patterns of species' occurrence across landscapes, which can be used in ecological niche modeling (ENM). Here, primary occurrence data are analyzed to obtain estimates of species' ecological requirements, generally without data documenting 'absence' [Soberón and Peterson, 2005]. ENM depends on two main data inputs: (i) occurrences of the species across a particular landscape, and (ii) digital raster datasets describing relevant environmental parameters. Numerous algorithms exist for estimating niches from these data inputs [e.g., Elith et al., 2006; Ward et al., 2009; Thuiller et al., 2009], although techniques for their comparison are not well developed [Peterson et al., 2008]. Niche models, under certain sets of assumptions, can be used to estimate species' current distributions, or to anticipate changes in distributions in response to environmental change [Sánchez-Cordero and Martínez-Meyer, 2000; Pearce and Boyce, 2006]. ENM approaches have been used to estimate distributions of poorly-known bird species [Peterson et al., 2002], estimate the distributional potential of invasive species [Nyári et al., 2006], and predict distributional responses to climate change [Pearson and Dawson, 2003; Jarvis et al., 2008; van Zonnevald et al., 2009].

ENM approaches have particular potential for the study of invasive alien species [Peterson, 2003], given that species' ecological characteristics tend to be conserved over moderate periods of evolutionary time [Peterson et al., 1999]. However, the most significant bottleneck in such applications is access to occurrence data: the access to increasing data volumes enabled by GBIF provides a solution. Indeed, GBIF-enabled data resources for eukaryote species on the global "100 Worst Invaders" list [Lowe et al. 2000] range from three for the termite *Coptotermes formosanus*, to 1,039,459 for the Eurasian Starling (*Sturnus vulgaris*), with a mean data density of 14,809 records per species (Fig. 2). As niche modeling algorithms typically require >~20 unique occurrence points for robust model development [Stockwell and Peterson, 2002; Wisz et al., 2008], GBIF-enabled data make possible analysis across a broad swath of potential invasive species, including 83 of those included on the "100 Worst Invaders" list. GBIF-enabled data have already been used for numerous analyses of species' invasions [e.g., Christenhusz and Toivonen, 2008; Ebeling et al., 2008; Giovanelli et al., 2008; Laufer et al., 2008; Rödder et al., 2008; Second and Rouhan, 2008; Beaumont et al., 2009]. GBIF nomenclatural resources also permitted authentication and understanding of synonymy for all 100 names via linkages with Species 2000, TROPICOS, and others, and identification of 5,310 recent literature publications relevant to the question within seconds of making the query.

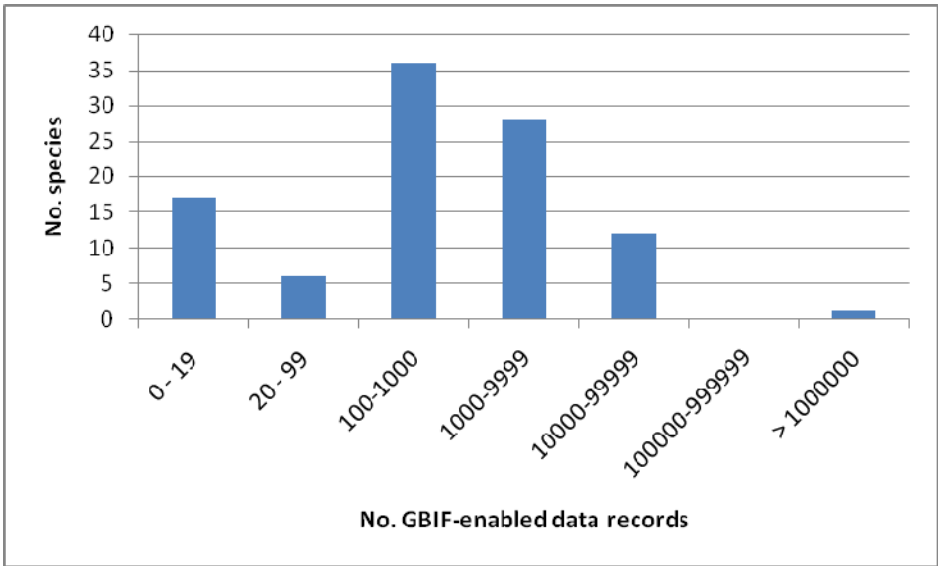

Fig. 2 Numbers of GBIF-enabled data records (April 2009) for the "100 Worst Invaders" list.

3 INTEGRATION OF GBIF SERVICES

Combined, GBIF's data portal, the system enabling discovery of biodiversity data resources, and the global names architecture (when completed and integrated), will contribute greatly to the informatics infrastructure for understanding and conserving global biodiversity. Lists of names of species that address changes in taxonomy, synonyms (different names applied to the same species), homonyms (one name referring to different taxa), misspellings of names, and other problems, are a prerequisite to querying large integrated data portals effectively. Simply put, a researcher seeking to obtain all data records from a portal of integrated datasets corresponding to a species that underwent a change in scientific name would either require this knowledge so as to query for both names, or the portal infrastructure would need to account for it automatically. To address this need, GBIF is developing a Global Names Architecture, with the aim of organizing taxonomic information in a dynamic system [Remsen, 2008]. The resulting names resolution service will improve discovery of and access to data, and improve existing and establish new global, regional, and thematic checklists. This architecture will:

- Create an authority list of nomenclature, providing a common framework for publishing, discovering, and accessing taxonomic data;
- Use this framework to inventory and index authoritative global, regional, and thematically-defined taxonomic and nomenclatural checklists into a dynamic checklist of names; and
- Develop a complete inventory of all scientific and common names used to refer to taxa by creating a global index of names tied to primary occurrence data, scientific publications, and other information about taxa.

The discovery of relevant datasets through metadata, this in addition to the functionality of the names architecture, and free and open access to primary data, can be critical to support biodiversity research. Metadata are critical to (i) enabling initial data discovery, (ii) determining whether pooling individual datasets is merited, and (iii) determining how individual datasets can best be integrated. Dataset-level metadata typically include information on the dataset's basic characteristics, ownership, and how to obtain further information. More detailed metadata will include information on how the data were derived, details of data quality, and technical details for access and use. A given study may require pooling of only datasets collected via consistent and standardized methods, information that is obtainable through metadata. Lacking metadata, indexed datasets may be pooled and analyzed in inappropriate manners, resulting in unsubstantiated and potentially misleading findings [e.g., Gilman et al., 2005]. In response to this need, the GBRDS will offer a single entry point for discovery of many forms of information about biodiversity resources, including biodiversity data, standards, and services, and allow the integration of other systems with the GBIF network [O'Tuama, 2009].

4 GBIF DATA PORTAL STATUS AND TRENDS

CONTENT AND VOLUME

Numbers of data records and data publishers accessible via the GBIF portal have been accumulating linearly since inception (Fig. 3). As of July 2009, >177M primary biodiversity records from 7,517 datasets have been integrated via the GBIF portal (8 July 2009; GBIF data portal query). Of GBIF-enabled data records, >64% (114M) are observational and 24% (42M) are ‘specimen-based’ (digitized records from natural history collections, excluding fossils). The remainder is a combination of records for germplasm, fossils, and captive and cultivated organisms. The earliest records published through the GBIF portal date from the early 1800s, so more than 200 years of species-occurrence data are integrated via the GBIF portal.

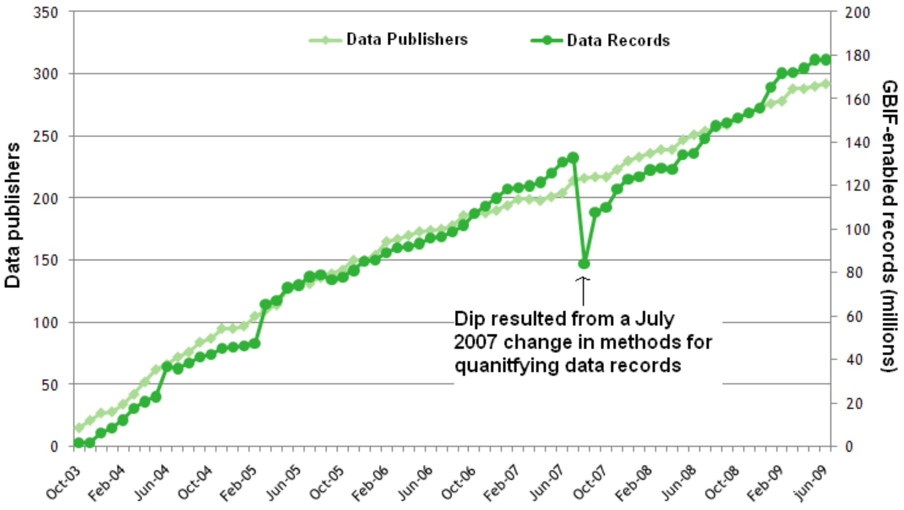

Fig. 3 Number of data publishers and individual species-level data records available through the GBIF data portal

In all, 291 data publishers are searchable in the GBIF portal (8 July 2009). Of GBIF’s 50 member countries and 40 member international organizations, 52 (35 countries and 17 organisations) have published datasets via GBIF. Of the >177M records included in these 7,517 datasets, however, 70% were published by only five members (USA 66M, Sweden 19M, UK 17M, Ocean Biogeographic Information System 12M, and France 11M). Similarly, GBIF-enabled records fall predominantly within North America and Europe, with over two-thirds coming from only three countries: 42, 15, and 13% in the USA, Sweden, and UK, respectively (Fig. 4). While GBIF dataset publishers are predominantly from North America and Europe, global species biodiversity is highest in the Tropics; hence, Africa, Asia, and Oceania are severely under-represented, contributing only 1.2%, 1.6% and 2.8% of total GBIF-enabled records, respectively, with only eight countries publishing datasets from these three regions combined.

While it is important to have all countries contribute to the global biodiversity data commons, because developed countries generally currently possess the majority of biodiversity data originating from developing countries, improving publication by developing countries would be insufficient to resolve the unevenness in distribution of published data records. Far greater publication of data held in developed countries is therefore required.

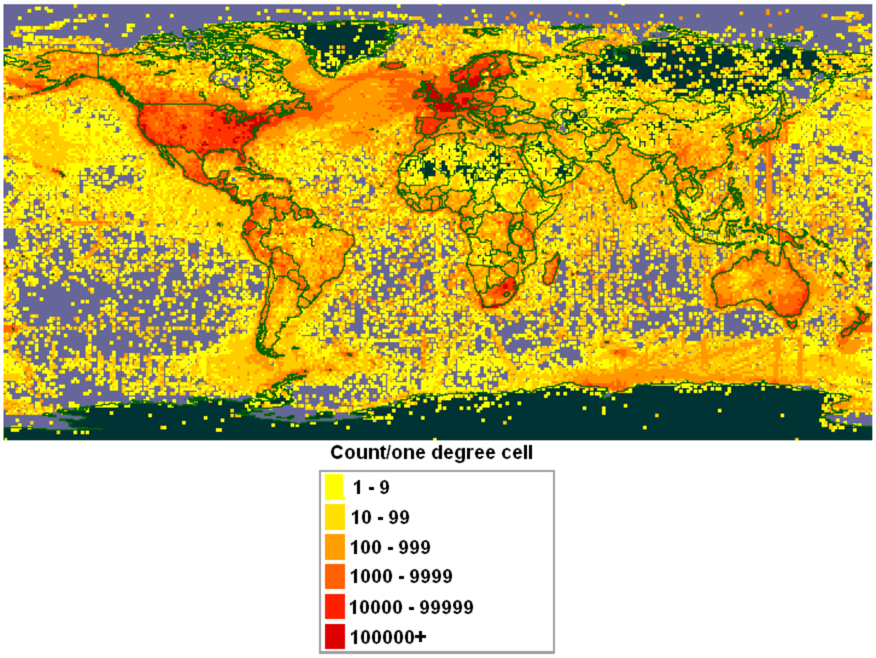

Fig. 4 Global distribution of >145M georeferenced, species-level, point occurrence data accessible through the GBIF data portal (July, 2009)

Fig. 5 summarizes the volume of GBIF-enabled occurrence data by selected major taxonomic groups. Lacking information on species richness of GBIF-enabled data records for selected taxonomic groups, we resort to comparing the number of records published through GBIF in these selected taxonomic groups to the number of species in each group as a first order indicator of taxonomic coverage, updating a similar analysis conducted by Collen and Rist [2008]. Taxonomic coverage of GBIF-enabled data is biased towards well-studied groups, especially birds (0.6% of total described species, but comprising 38% of GBIF-enabled records), and to a lesser degree, mammals and fishes. Invertebrates are, in general, substantially under-represented (insects comprising >60% of total described species but only 9.5% of GBIF-enabled records). Alternatively, comparing the volume of GBIF-enabled records to the number of records likely in existence reveals that a tremendous volume of plant specimens in herbaria remain to be digitized and integrated.

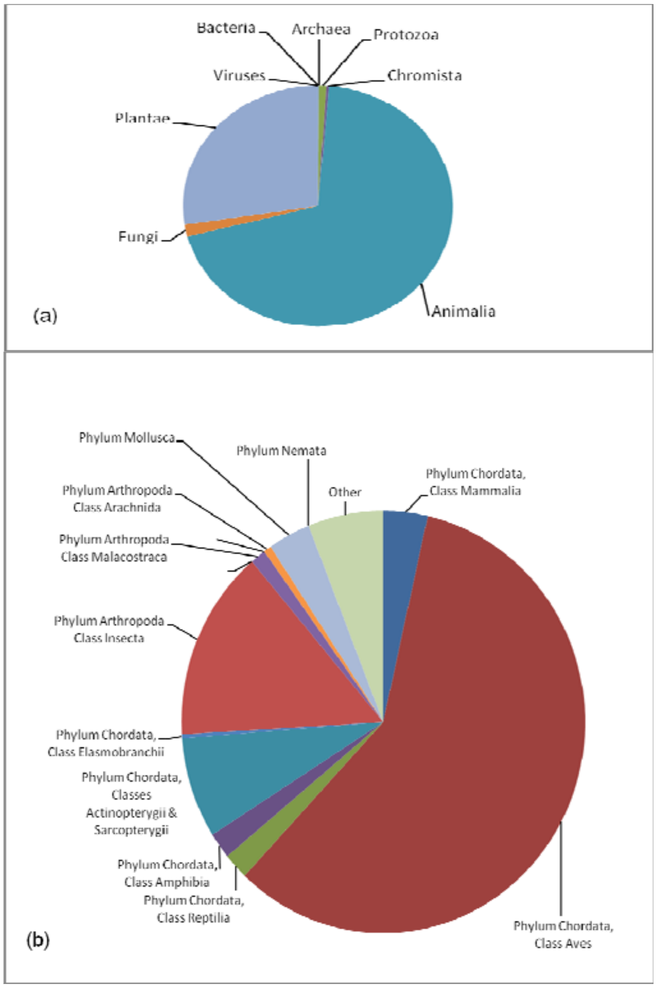

Fig. 4 Inventory of GBIF-enabled data records by (a) kingdom; (b) breakdown of the animal kingdom by selected phyla and classes (GBIF data portal, 8 July 2009).

WEB-BASED TOOLS

GBIF has produced a large range of web-based filters to allow searches, classification, aggregation, and disaggregation of the 100GB of GBIF-enabled occurrence data records. The currently available search filters are listed in Table 1.

Table 1 Search filters available via the GBIF data portal.

Taxonomy	**Geospatial**	**Datasets**	**Other details**
Scientific name	Country	Data provider	Occurrence date
Common name	Region (of 23)	Dataset name	Year range
Classification	Bounding box	Host country	Year
Type status	Latitude		Month
	Longitude		Institution code
	Elevation		Collection code
	Depth		Catalogue number
	Coordinate status (with/without coordinates)		Basis of record
	Coordinate issues (none/issues detected)		Image URL (present/absent)
	Protected area		

In addition, GBIF has supported the production of key informatics tools to meet the increasing need for interoperability between online biodiversity-related databases, to enhance capabilities for large-scale spatial analyses for a range of research applications [GBIF, 2009a]. GBIF has, for example, developed informatics tools to overlay GBIF-enabled point data on the Convention on Migratory Species Global Register of Migratory Species (GROMS) GIS shape file polygon data of species distributions (http://www.groms.de/), and overlay these on Google Map layers (Fig. 6) [GBIF, 2009b], as well as to integrate the database on protected area boundaries of the International Union for the Conservation of Nature’s World Database on Protected Areas (WDPA) with GBIF-enabled point occurrence data (Fig. 6; www.wdpa.org). These two integrated visualisation tools provide users with immediate graphical summaries of GBIF-enabled species-occurrence data to validate and quality control migratory species’ distributions, determine how much biodiversity data is available for an area of interest (in this case, within a selected protected area), determine if sufficient sample sizes exist to determine accurate measures of species richness, and, more generally, determine whether data can be used for specific research applications [Guralnick et al., 2007; GBIF, 2009a].

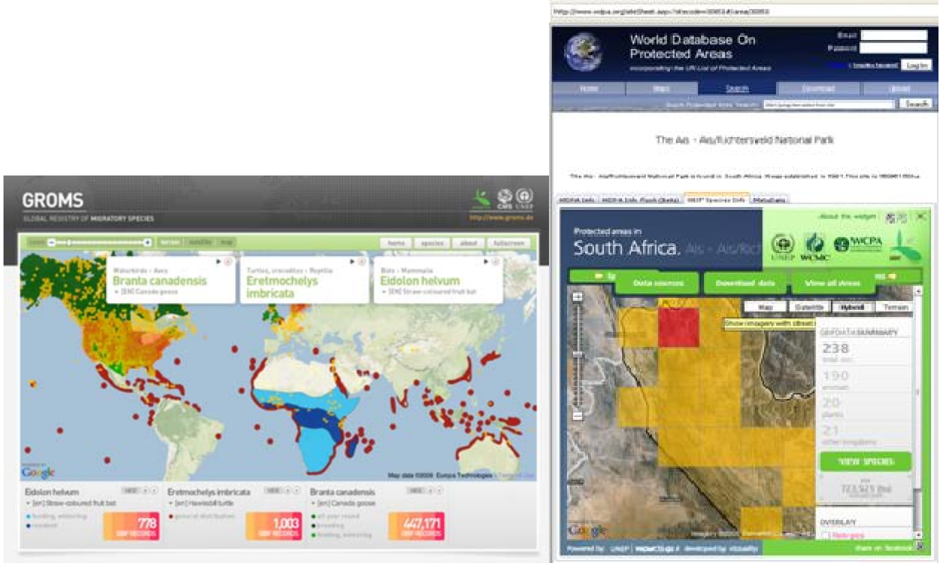

Fig. 5 Web screenshot examples of GBIF informatics tools enabling overlaying GBIF-enabled point data with species range polygons from the Global Register of Migratory Species (left) and with protected area boundaries from the World Database on Protected Areas (right)

DATA QUALITY

Data-quality concerns related to biodiversity databases include issues related to elimination of errors, incorrect taxonomic identifications, and errors in georeferencing, among others. GBIF has made considerable investment in development of tools for effective management, treatment, and reduction of these errors, particularly via support for development of error detection and data cleaning tools. A growing suite of tools is available to detect inconsistencies in geographic references, ecological outliers, nonstandard taxonomic names, and other data quality problems [Chapman 2005a, b]. GBIF has also invested in training courses and best-practices manuals designed specifically to eliminate problems in use of the integrated datasets for biodiversity analyses [Chapman 2005a, b].

Precise georeferences are essential for many applications of species-level occurrence data [Guralnick et al., 2006; Collen and Rist, 2008]. Georeferenced coordinates (e.g., latitude, longitude, elevation) are not always available, and in fact only a fraction of available data (particularly as regards data associated with natural history collection specimens) is georeferenced. While of the total >177M records enabled by GBIF, 82% (145.6M) have coordinates (GBIF data portal, 8 July 2009), this large proportion may be deceptive: georeferenced data show even more unevenness taxonomically and spatially than the broader pool. Retrospective georeferencing has seen impressive recent development, in which 'smart' software tools now facilitate the task greatly [Guralnick et al., 2006]. Modern retrospective georeferencing via the point-radius method includes careful documentation of estimated error [Wieczorek et al., 2004], with automated incorporation into record-level metadata.

Unfortunately, although collecting and reporting geographic coordinates and an estimate of error in positional accuracy is a standard method in modern biological monitoring surveys, this information has not been routinely captured in metadata of datasets published via GBIF, fundamental information for some research applications, where positions with high uncertainty are unsuitable for use in research employing fine spatial scales. Information on resolution is also needed to determine appropriateness for pooling datasets. Information on sampling effort has also not been routinely captured, information that is necessary to determine species abundance.

5 FUTURE WORK

Content Needs and Gaps

We highlight the following gaps and/or priority datasets for future integration:

- Datasets for underrepresented taxonomic groups, especially invertebrates; and
- Datasets from underrepresented regions, especially Africa, Asia, and Oceania.

A more quantitative, comprehensive inventory and gap analysis of GBIF-enabled datasets (e.g., to identify species richness by taxonomic group and region) will enable a stronger basis for the identification of priority gaps. Categories of gaps needing exploration include coverage by ecosystem types, such as mangroves [Gilman et al., 2008a]; taxonomic groups, such as certain plant taxa [e.g., legumes, Yesson et al., 2007] and marine apex predators subject to fishing mortality [Myers et al., 2007; Gilman and Lundin, 2009]; protected areas [Bertzky and Stoll-Kleemann, 2009]; and gaps in spatial and temporal coverage [Collen and Rist, 2008; Poloczanska et al., 2008].

Constraints and Solutions to Dataset Publication and Access

Primary biodiversity data are the information foundation for biodiversity science. Soberón and Peterson [2009] examined numbers of biodiversity records as a function of their year of origin, finding an exponential accumulation of biodiversity information, where, over the past two centuries, the number of records available per year has doubled approximately every 42 years. Based on these results, and given continued advances in information technology, real potential exists for substantial additional biodiversity data to be published in coming years. However, this potential will only be realized if a suite of substantial obstacles is addressed adequately.

Digital data capture is a major constraint, where for example, progress in digitizing natural history collections of museums, herbaria, and universities is limited by funding and human resources [Butler et. al., 1998; Pennisi, 2000; Chavan and Krishnan, 2003]. The ca. three billion specimens held in the world's estimated 6,500 natural history collections (excluding collections of microorganisms) represent the work of thousands of individuals carried out over centuries [Butler et.al., 1998; Chavan and Krishnan, 2003]. A substantial proportion of specimen datasets include long time series, spanning 200 years or longer, dating back to when ecosystems were relatively pristine, and potentially exceeding the temporal variability of factors that may significantly affect change and loss in biodiversity (e.g., life-spans of species, return period of important environmental disturbances, cyclical processes). Thus, these datasets enable the construction of baselines to measure anthropogenic-caused changes and losses in biodiversity [Jackson et al., 2001; Suarez, 2004; Gilman et al., 2008a]. However, as a consequence of their date of collection, most specimen data are not initially available in digital form (only ca. 5-10% of these records are digitized [Krishtalka and Humphrey, 2000]), and these non-digital data dominate information in existence for key groups such as insects and plants. Main obstacles to increasing the rate of digitisation of specimen data include:

(i) Digitisation is costly and labour-intensive;
(ii) Lack of funding for such tasks, perhaps owing to insufficient awareness of the research applications and demand for such data [GBIF, 2008]; and
(iii) Insufficient technologies to enable rapid, efficient, and cost-effective data capture.

A further major constraint on the efficacy of biodiversity informatics information resources are limitations on publishing datasets [Andelman et al., 2004; Roberts and Chavan, 2008; Chavan and Ingwersen, In Press]. The technological infrastructure is in place to publish and integrate datasets, but a need exists for policies by relevant bodies, including governments and funding agencies, to require making data freely available, and to enforce relevant existing policies sufficiently [Costello, 2009]. Need also exists for development of mechanisms for data citation and impact factors for data usage to provide public and professional recognition for data-sharing, providing data holders with additional incentives to make data available [Andelman et al., 2004; Roberts and Chavan, 2008; Chavan and Ingwersen, In Press]. Some impediments to open access to research data include [Costello, 2009; Gaikwad and Chavan, 2006; Chavan and Ingwersen, In Press]:

(i) Concern that other researchers may use published data to “scoop” planned research by the data owner or custodian;
(ii) Lack of awareness of where and how to publish data;
(iii) The perception that dataset publishing is too difficult or time consuming;
(iv) Concern that the data will be used improperly;
(v) Concern that the data will be used for commercial gain. Some governments have cited concern over the risk of ‘biopiracy,’ monopolization of genetic resources and traditional knowledge [Greene, 2004], as a reason for refraining from making biodiversity data freely available. Information on medicinal plants or commercial fishery data may be confidential, and in some cases are required to be amalgamated or to reduce spatial resolution of geographic references prior to public disclosure [Bosselman, 1995; Dutfield, 2000; WCPFC, 2007; Magnuson Stevens Fishery Conservation and Management Act, 16 U.S.C. 1801 *et seq*. Section 402(b)], which precludes some research applications;
(vi) Concern over the risk of harming sensitive species through revealing locations of occurrence;
(vii) Worry that intellectual property rights, including ownership, authorship, or control of the data, will be lost; and
(viii) Lack of informed consent and confidentiality.

Operationalizing the proposed GBIF 'Data Publishing Framework,' in combination with needed changes in policy and legal frameworks, and securing long-term sustainable financing, could address many of the obstacles to biodiversity dataset discovery and publication [Penev et al., 2009; Chavan and Ingwersen, In Press]. Furthermore, GBIF is in the process of developing the infrastructure to enable publishing polygon and raster-based biodiversity data and developing informatics tools to enable integration of biodiversity data with name-based areas (e.g., protected areas and other fine scale areas). For instance, enabling publication of secondary, taxonomic-based polygon data for species' distributions enables inventory and gap analysis of available GBIF-enabled primary data and validation and quality control for the polygon range data. Raster-based records can be primary data, where sampling designs were grid-based. Or raster-based records can be secondary data, where primary data are aggregated, such as with data from commercial fisheries, or for range maps based on interpretations of primary data, which again support gap analyses and validation. Enabling the publication of these new data types promises to augment biodiversity data publication, fill prioritized gaps, and augment spatial analytical capabilities.

Amending standards for metadata to capture information on sampling effort, data collection methods, and estimate of positional error would provide the requisite information to determine if pooling of various databases is suitable, observe species composition/richness, and observe species abundance/population sizes.

More generally, data mobilization often requires data owners, custodians, funding entities, and management authorities to deviate from longstanding practices. A change in culture and behavior of scientists, managers, and funders is needed to facilitate multi-scale data integration [Andelman et al., 2004; Chavan and Ingwersen, In Press]. Finally, sustainable levels of financing for data discovery and mobilization are needed to meet priorities for content and sample size.

REFERENCES

Andelman, S.J., Bowles, C., Willig, M., Waide, R. 2004. Understanding environmental complexity through a distributed knowledge network. *BioScience* 54(3): 240-246.

Arzberger, P., P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhlir and P. Wouters. 2004. Promoting access to public research data for scientific, economic and social development. *Data Science Journal* 3: 135 – 152.

[Beaumont et al., 2009] Beaumont, L. J., L. Hughes, and A. J. Pitman. 2008. Why is the choice of future climate scenarios for species distribution modelling important? *Ecology Letters* 11: 1135-1146.

Bertzky M., Stoll-Kleemann S. 2009. Multi-level discrepancies with sharing data on protected areas: What we have and what we need for the global village. *Journal of Environmental Management* 90: 8-24.

Bosselman, K. 1995. Plants and politics: the international legal regime concerning biotechnology and biodiversity. *Colo J Int Environ Law and Polic*y 7.

Butler, D., Gee, H. and Macilwain, C. 1998. Museum research comes off list of endangered species. *Nature* 394: 115-117.

Canhos, V. P., S. de Souza, R. de Giovanni, and D. A. L. Canhos. 2004. Global biodiversity informatics: Setting the scene for a "new world" of ecological forecasting. *Biodiversity Informatics* 1: 1-13.

Chapman, A. D. 2005a. *Principles and Methods of Data Cleaning, Version 1.0.* Global Biodiversity Information Facility, Copenhagen.

Chapman, A. D. 2005b. *Principles of Data Quality, Version 1.* Global Biodiversity Information Facility, Copenhagen.

Chavan, V., Krishnan, S. 2003. Natural history collections: A call for national information infrastructure. Current Science 84(1): 34-42.

Chavan, V., Ingwersen, P. In Press. Towards a data publishing framework for primary biodiversity data: Challenges and potentials for the biodiversity informatics community. *BMC Bioinformatics*.

Christenhusz, M. J. M., and T. K. Toivonen. 2008. Giants invading the tropics: the oriental vessel fern, *Angiopteris evecta* (Marattiaceae). Biological Invasions 10:1215-1228.

Collen, B., M. Ram, T. Zamin, and L. McRae. 2008. The tropical biodiversity data gap: Addressing disparity in global monitoring. *Tropical Conservation Science* 1: 75-88.

Collen, B., Rist, J. 2008. *Testing the Usability of GBIF Data for use in 2010 Biodiversity Indicators. Streamlining European 2010 Biodiversity Indicators (SEBI 2010): Developing a methodology for using bats as indicator species; and testing the usability of GBIF data for use in 2010 biodiversity indicators*. Prepared under European Environment Agency Tender EEA/BSS/07/008. European Environment Agency, Copenhagen.

Costello, M. 2009. Motivating online publication of data. *BioScience* 59: 418-427.

Dutfield, G. 2000. *Developing and Implementing National Systems for Protecting Traditional Knowledge: A Review of Experiences in Selected Developing Countries*. UNCTAD Expert Meeting on Systems and National Experiences for Protecting Traditional Knowledge, Innovations and Practices, 30 October – 1 November 2000, Geneva. 28 pp.

Ebeling, S. K., E. Welk, H. Auge, and H. Bruelheide. 2008. Predicting the spread of an invasive plant: combining experiments and ecological niche model. *Ecography* 31: 709-719.

Edwards, J., Lane, M., Nielsen, Ebbe. 2000. Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science* 289: 2312-2314.

Elith J, Graham C., Anderson R, Dudík M, Ferrier S, Guisan A, Hijmans R,Huettmann F, Leathwick J, Lehmann A, Li J, Lohmann L, Loiselle B, Manion G,Moritz C, Nakamura M, Nakazawa Y, Overton J, Peterson A, Phillips S, Richardson K, Scachetti-Pereira R, Schapire R, Soberon J, Williams S, Wisz M, Zimmermann N. 2006. Novel methods improve prediction of species distributions from occurrence data. *Ecography* 29: 129-151.

Gaikwad, J., V. Chavan. 2006. Open access and biodiversity conservation: Challenges and potentials for the developing world. *Data Science Journal* 5: 1-17.

GBIF. 2009a. *Global Biodiversity Information Facility Development of Informatics Tools to Support Large-Scale Spatial Analyses of Species-level Biodiversity Occurrence Data*. By E.L. Gilman, T. Robertson, M. Chaloupka. Global Biodiversity Information Facility, Copenhagen.

GBIF 2009b. *Global Registry of Migratory Species Project Technical Report*. Prepared by De La Torre, J., under GBIF Grant 2008-IT-RIA-3. Global Biodiversity Information Facility, Copenhagen.

GBIF. 2008. *Global Strategy and Action Plan for Mobilisation of Natural History Collections*. Global Biodiversity Information Facility, Copenhagen.

Gilman, E., J. Ellison, N. Duke, C. Field. 2008. Review: Threats to mangroves from climate change and adaptation options. *Aquatic Botany* 89: 237-250.

Gilman, E., Lundin, C. 2009. Minimizing Bycatch of Sensitive Species Groups in Marine Capture Fisheries: Lessons from Commercial Tuna Fisheries. IN: Grafton, Q., Hillborn, R., Squires, D., Tait, M., Williams, M. (Eds.). *Handbook of Marine Fisheries Conservation and Management*. Oxford University Press.

Giovanelli, J., C. Haddad, and J. Alexandrino. 2008. Predicting the potential distribution of the alien invasive American bullfrog (*Lithobates catesbeianus*) in Brazil. *Biological Invasions* 10: 585-590.

Greene, S. 2004. Indigenous people incorporated? Culture as politics, culture as property in pharmaceutical bioprospecting. *Current Anthropology* 45: 211-237.

Guralnick, R.P., Hill, A., Lane, M. 2007. Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters* 10: 663-672.

Guralnick, R.P., Wieczorek, J., Hijmans, R.J., Beaman, R., and the Biogeomancer Working Group, 2006. Biogeomancer: automated georeferencing to map the world's biodiversity data. *PLoS Biol*. 4: 1908-1909.

Guralnick, R., and A. Hill. 2009. Biodiversity informatics: Automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics* 25: 421-428.

Jackson, J., Kirby, M., Berger, W., Bjorndal, K., Botsford, L., Bourque, B., Bradbury, R., Cooke, R., Erlandson, J., Estes, J., Hughes, T., Kidwell, S., Lange, C., Lenihan, H., Pandolfi, J., Peterson, C., Steneck, R., Tegner, M., Warner, R. 2001. Historical overfishing and the recent collapse of coastal ecosystems. *Science* 293: 629-638.

Jarvis, A., Lane, A., Hijmans, R. 2008. The effect of climate change on crop wild relatives. *Agriculture, Ecosystems and Environment* 126: 13-23.

Krishtalka, P., Humphrey, G. 2000. Can natural history museums capture the future. *Bioscience* 50: 611-617.

Laufer, G., A. Canavero, D. Núñez, and R. Maneyro. 2008. Bullfrog (*Lithobates catesbeianus*) invasion in Uruguay*. Biological Invasions* 10: 1183-1189.

Lowe, S., M. Browne, S. Boudjelas, and M. de Poorter. 2000. 100 of the World's Worst Invasive Alien Species: A Selection from the Global Invasive Species Database. Invasive Species Specialist Group. Global Invasive Species Programme, International Union for the Conservation of Nature, Invasive Species Specialist Group, Gland, Switzerland.

Myers, R.A., Baum, J.K., Shepherd, T.D., Powers, S.P., Peterson, C.H., 2007. "Cascading effects of the loss of apex predatory sharks from a coastal ocean." Science 315(Mar.): 1846-1850.

Nyári, A., C. Ryall, and A. T. Peterson. 2006. Global invasive potential of the House Crow (*Corvus splendens*) based on ecological niche modelling. *Journal of Avian Biology* 37: 306-311.

OECD. 1999. *Final Report of the OECD Megascience Forum Working Group on Biological Informatics*. Organisation for Economic Co-operation and Development, Paris.

O'Tuama, E. (Compiler). 2008. *Metadata Requirements for Datasets Delivered via the Global Biodiversity Information Facility Network*. Global Biodiversity Information Facility, Copenhagen.

Pearce J, Boyce M (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology* 43: 405-412

Pearson, R. G., and T. P. Dawson. 2003. Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? *Global Ecology and Biogeography* 12: 361-371.

Penev, L., Erwin, T., Miller, J., Chavan, V., Moritz, T., Griswold, C. 2009. Publication and dissemination of datasets in taxonomy: ZooKeys working example. *ZooKeys* 11: 1-8.

Pennisi, E. 2000. Taxonomic revival. *Science* 289: 2306-8.

Peterson, A. T. 2003. Predicting the geography of species' invasions via ecological niche modeling. *Quarterly Review of Biology* 78: 419-433.

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

Peterson, A. T., M. Papeş, and J. Soberón. 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modelling. *Ecological Modelling* 213: 63-72.

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

Poloczanska, E., A.J. Hobday & A.J. Richardson. 2008. Global database is needed to support adaptation science. *Nature* 453: 720.

Ponder WF. 2001. Evaluation of museum collection data for use in biodiversity assessment. *Conserv. Biol*. 15: 648–657.

Remsen, D. 2008. *The Global Names Architecture (GNA): An Integrated and Federated Approach to Enabling Discovery and Access to Biodiversity Information*. Global Biodiversity Information Facility, Copenhagen.

Roberts, D., and V. Chavan. 2008. Standard identifier could mobilize data and free time. *Nature* 453: 449-450.

Rödder, D., M. Solé, and W. Böhme. 2008. Predicting the potential distributions of two alien invasive housegeckos (Gekkonidae: *Hemidactylus frenatus*, *Hemidactylus mabouia*). *North-western Journal of Zoology* 4: 236-246.

Sanchez-Cordero, V., Martinez-Meyer, E., 2000. Museum specimen data predict crop damage by tropical rodents. Proceedings of the National Academy of Sciences of the United States of America 97, 7074-7077.

Second, G., and G. Rouhan. 2008. Human-mediated emergence as a weed and invasive radiation in the wild of the CD genome allotetraploid rice species (*Oryza*, Poaceae) in the Neotropics. *PLoS ONE* 3: e2613.

Shaffer, H.B., Fisher, R.N., Davidson, C., 1998. The role of natural history collections in documenting species declines. *Trends in Ecology & Evolution* 13: 27-30.

Soberón, J., and A. T. Peterson. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiversity *Informatics* 2: 1-10.

Soberón, J., and A. T. Peterson. 2009. Monitoring biodiversity loss with primary species-occurrence data: Toward national-level indicators for the 2010 Target of the Convention on Biological Diversity. *AMBIO* 38: 29-34.

Stockwell, D. R. B., and A. T. Peterson. 2002. Effects of sample size on accuracy of species distribution models. *Ecological Modelling* 148: 1-13.

, A. 2004. The value of museum collections for research and society. *BioScience* 54: 66-74.

Thuiller W, Lafourcade B, Engler R, Araujo M (2009) BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography* 32: 369-373.

UNGA. 1992. Report of the United Nations Conference on Environment and Development, Rio de Janeiro, 3-14 June, 1992. A/CONF.151/26 (Vol.1). United Nations General Assembly, New York.

van Zonnevald, M., Jarvis, A., Dvorak, W., Lema, G., Leibing, C. 2009. Climate change impact predictions on *Pinus patula* and *Pinus tecunumanii* populations in Mexico and Central America. *Forest Ecology and Management* 257: 1566-1576.

Ward G, Hastie T, Barry S, Elith J, Leathwick J (2009) Presence-only data and the EM algorithm. *Biometrics* 65: 554–563.

WCPFC. 2007. *Conservation and Management Measure for the Regional Observer Programme*. Conservation and Management Measure 2007-01. Western and Central Pacific Fisheries Commission: Palikir, Federated State of Micronesia.

Wieczorek, J., Guo, Q., Hijmans, R. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *Int. J. Geographical Information Science* 18: 745-767.

Wilson, E. O., Ed. 1988. Biodiversity. National Academy Press, Washington, D.C.

Wisz, M. S., R. Hijmans, J. Li, A. T. Peterson, C. H. Graham, and A. Guisan. 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions* 14: 763-773.

Yesson, C., P. W. Brewer, T. Sutton, N. Caithness, J. S. Pahwa, M. Burgess, W. A. Gray, R. J. White, A. C. Jones, F. A. Bisby, and A. Culham. 2007. How global Is the Global Biodiversity Information Facility? *PLoS ONE* 2: e1124.