

Summary of GBIF Accomplishments 2001 - 2005

The Global Biodiversity Information Facility (GBIF) came into existence as of 1 March 2001, with the US having been the first signatory to the international Memorandum of Understanding that established it as a free-standing, not-for-profit organization. The GBIF mission is “making the world’s biodiversity data freely and universally available via the Internet for the benefit of science, society and a sustainable future.” In the context of the GBIF program of work to date, “biodiversity data” means data about the occurrences of species in place and time. These come predominantly from labels on specimens in natural history collections and programs of observation of organisms (e.g. bird counts), although they are highly and complexly interconnected with data from all levels of biological organization and areas of investigation.

In its Strategic Plan for 2001 – 2006, several overarching goals were laid out for GBIF’s beginning phase. One of these was that all nations have access to key data necessary to understand their own national biodiversity. As of this writing, GBIF mediates at least some data for every country on Earth except San Marino and the Vatican. Among the 100+ million records currently accessible, there are about 20 million that are “repatriated” from the countries of the institutions that hold biodiversity specimens to the countries of origin of those specimens by virtue of their accessibility via GBIF.

It was also promised that scientific correlations and insights, and biodiversity discovery and research, would be facilitated by GBIF, which indeed is turning out to be the case (Shao et al. 2002, Krupnick & Kress 2003, May & Ragia 2003, Stewart et al. 2003, Astley 2004, Brown et al. 2004, Graham et al. 2004, Hortal et al. 2004, Lebeda et al. 2004, Meier & Dikow 2004, Parr et al. 2004, Somervuo & Harma 2004, Becker et al. 2005, Hortal & Lobo 2005, Malchus & Warén 2005, Perlmutter 2005, Perlmutter & Greene 2005, Rowe 2005, Halpin et al. 2006, Lange & Andersen 2006, Rissler et al. 2006, Sonké et al. 2006, Stevens et al. 2006, Veron et al. 2006, Raymond et al. *in ms*). Interestingly, GBIF and the information science challenges that it faces in working toward its goal of a fully interoperable, globally distributed biodiversity data-provision system have themselves provided subjects for researchers in the computer, information and social sciences (e.g. Aubert & Bayar 2000, Vignes-Lebbe 2000, Lane & Shattuck 2002, Morris et al. 2002, Scoble 2002, Saarenmaa & Nielsen 2002, Stevenson et al. 2003, Aisbett et al. 2004, Beaman et al. 2004, Breckling & Reuter 2004, Caplan & Haas 2004, Dalcin 2004, Rennolls et al. 2004, Wiczorek et al. 2004, Castro et al. 2005, Dawyndt et al. 2005, Delgado et al. 2005, Fazey et al. 2005, Graham & Kennedy 2005, Huettmann 2005, Kennedy et al. 2005, Leong et al. 2005, Villa 2005, Hussey et al. 2006, Rajbhandari et al. 2006, Baker & Bowker *in press*). In the proposal for NSF 0301149, it was predicted that GBIF-mediated digitized biodiversity data could be used to solve real-world problems, which again is proving to be true (Simmonds et al. 2001; James 2002; Pohja 2002; Crossin et al. 2003; Laihonon et al. 2003; Cotter et al. 2004; Helimo 2004; Samper 2004; Silva 2004; Arzberger et al. 2004a, 2004b; Anderson 2005; Arizona Game & Fish Department 2005; Hockland 2005; Iwata & Chen. 2005; Moret 2005; Shivas et al. 2006; Anonymous 2006).

As indicated by the mission statement, one of the founding principles of GBIF is that data should be openly shared. The Strategic Plan 2001 - 2006 stressed that GBIF should work toward having its concepts and principles of data-sharing be accepted and adopted globally. To that end, the Governing Board adopted a *Recommendation on Open Access to Biodiversity Data* (16 Jan 2006)¹. This document is one of a large number of similar statements issued worldwide (a boolean search on Google for “(Open Access) AND (Statement OR Declaration)” on 8 Aug 2006 resulted in 303 M hits), but is specifically directed at science agencies within GBIF member countries, asking them to make open access to the resulting data a condition for funding biodiversity studies. With respect to Intellectual Property Rights (IPR) issues associated with biodiversity data, GBIF has developed Data Use and Data Sharing

• ¹ It is not necessary to refer to this or any other GBIF document mentioned herein to review this proposal. However, should the reviewer wish to do so, this one and all others subsequently cited (such as the *Third Year Review* and the *Strategic Plan*) can be accessed at http://www.gbif.org/GBIF_org/GBIF_Documents/.

Agreements that are emulated by several of its Nodes. These Agreements were developed based on a white paper commissioned by GBIF, and using guidance from a GBIF-sponsored workshop on IPR issues held in 2004. In addition, a group of *pro bono* legal experts on IPR has been established; it met for the first time in September 2006 to consider the ramifications of various types of licensing for GBIF data-sharing, among other issues.

Outcomes projected by GBIF's Strategic Plan 2001 - 2006 (organized according to the Themes that are used in the *GBIF Plans for 2007 – 2011: From Prototype towards Full Operation*), and their status as of August 2006, are as follows:

THEME 1: INFORMATICS

Facilitate searching of several biodiversity databases simultaneously, including reporting of the combined, interoperable results.—Instituted as of the Launch of the Prototype Data Portal in February, 2003. Data are currently emanating from ~180 interoperable data providers and ~800 databases (which are being shared using several different versions of Darwin Core and ABCD, as well as Catalogue of Life partnership [CoLP] data) in 35 countries, integrated through the global index of the prototype data portal.



Enable digital access to collections data, library information, other databases and the like.—The careful thought that is going into the engineering of the information architecture represented by the next-generation GBIF data portal will facilitate transparent linkages among collections, library, image and other resources, as well as across information domains.

Year 1: Hold workshops to develop standards for interoperability.—GBIF works closely with a subcommittee of CODATA, the Taxonomic Databases Working Group (TDWG) to develop cross-community standards, and has held several workshops that have encouraged broad community participation in TDWG processes. Since 2005, GBIF has been administering a grant of 1.5M from the Moore Foundation to TDWG/GBIF to ensure that these standards are developed in a consistent, flexible and reusable way, and exploit developments in XML, RDF and ontologies.

Year 2: Further develop standards, and implement them among database providers.—The GBIF UDDI registry opened in 2003, and installation packages for data providers were crafted and distributed. GBIF has been active in the development and distribution through its portal of several key TDWG standards, including DiGIR and TAPIR, ABCD and Darwin Core, the Taxon Concept Schema and Structured Descriptive Data.

Year 2: Hammer out initial developments for handling name services.—Still hammering, but initial steps have been taken. The Taxon Concept Schema will facilitate the handling of names. Significant progress is anticipated early in Phase 2.

Year 3: Plan for inter-communication of species- and specimen-level databases with molecular and ecological databases.—A link exchange with GenBank has been put in place; a workshop was held with IPGRI and BioMoby (IPGRI's SINGER database being integrated with GBIF network as this proposal is written, and GBIF services will soon be accessible via BioMoby), and GBIF has been included as a resource in [BioBar](#). GBIF is participating in the planning of the GISIN architecture and is cooperating with the SEEK project.

Year 4: Implement cross-domain communication capabilities.—Data from museum communities, citizen observation networks, and environmental agencies is integrated via the data portal. Linkages to agricultural genetic data are accomplished. Data standards (via TDWG) have been re-engineered to make them easier to use in the context of the semantic web. GUIDs (specifically LSIDs) will simplify

cross-referencing between data objects; these are to be implemented according to plans made during two workshops held in conjunction with NASCent during 2006.

Year 5: Achieve interoperability and robust search capabilities across multiple information domains.—The prototype data portal integrates primary data and includes datasets that have some image data. The new data portal will integrate nomenclatural and taxonomic data with specimen/observation data sets and wider link-outs to GenBank, IdentifyLife, etc. Two GBIF mirror sites (in Germany and Korea) were established in 2005, and another planned for the US in the near future.

In 2004, Dr. John McCarthy of Lawrence Berkeley Laboratory reviewed GBIF's information architecture and concluded, "**based on detailed review of GBIF's excellent technical documentation and first-hand use of their current on-line facilities... GBIF is well on its way toward becoming one of the premier examples of a successful federated database network.** Moreover, they have done so ... on a remarkably modest budget by using widely used modern software, protocols and standards."

THEME 2: CONTENT

Encourage collections community to come together to seek funding for distributed digitisation projects, and thus increase the pace of growth of digital data content.—Many seed money awards have encouraged the development of consortia of institutions, and the 2005 - 2006 awards were made *only* to consortia. Networked cooperation is becoming more and more common (e.g. MaNIS, HerpNET, ORNIS, antbase, GloBIS, etc.). Sharing specimen data through the GBIF network is identified as one of the major incentives for digitization by many collections and regional networks (Neill, 2006) and scientists state in print that they want to share their data via GBIF (e.g. Pausas et al. 2003, Dettai et al. 2004, Ryan & Smith 2004, Crous 2005, Savolainen et al. 2005, Faith & Baker 2006). Unfortunately, funding for digitization efforts continues to be a problem for many institutions.

Establish links among natural history museums to provide more thorough geographic and taxonomic coverage.—Thematic (e.g. HerpNET, MaNIS, OBIS, ORNIS et al.) and regional networks (e.g. IABIN, ENBI et al.) of institutions share data through GBIF, and as a result of growing community awareness of the benefits of doing so, more will be coming online (e.g. ALTERnet, MarBEF) as their internal information architectures are developed and as GBIF's web services accommodate their needs.

Develop best practices for digitisation and georeferencing within the collections community.—Through broad consultation and partnerships with other efforts (e.g. Biogeomancer, an NSF-supported development), recommendations as to best practices are being distributed throughout the GBIF community. In addition, GBIF has commissioned and makes available white papers on use of biodiversity data, data cleansing, and on quality assurance. Also, a software package to facilitate data cleansing has been developed (in part through a GBIF contract), and is freely available via GBIF.

Data providers are using the ECAT as the "authority file" for scientific names in their databases.—Since this goal was written, the capabilities of the central information architecture have made it possible to use multiple taxonomies, so that data providers are not forced to adopt any particular "authority file" in the sense meant here.

Biodiversity scientists both use and contribute to the ECAT during their regular work.—As with the previous goal, the information architecture works in such a way as to "harvest" names from specimen and observation databases, without specific effort on the part of data providers. Future developments of ECAT-related software will extend the harvesting to digital literature resources, and will also facilitate the work of taxonomists who can contribute to resolving among these multiple taxonomies.

ECAT facilitates queries by automatically providing synonyms of the name entered by the user.—For names of species that are included in the databases served through the CoLp (representing about half of the ~1.75 million accepted species) this has been accomplished. Efforts will intensify early in Phase 2.

Synonyms of a name entered by the user are automatically incorporated into search queries presented to GBIF.—This awaits completion of the new portal architecture and web services, as well as the tools that will allow user choice among taxonomies and the cascade of shifts in synonym relationships that will result from such choices. These activities are focal points of Phase 2.

Year 2: Document extent of already-digitised collections.—A more thorough survey is still being developed through the GBIF national Nodes. However, preliminary estimates based on reports from 36 of GBIF's member countries (not including the US) include:

- Specimen holdings already digitized: 105.2 M
- Non-digitized specimen holdings: 631.5 M
- Observational dataset records already digitized: 189.6 M
- Non-digitized observational data records: 376.9 M

Year 2 (et seq.): Make seed money awards to stimulate digitisation projects.—To date, there have been three competitions (2003, 2004 and 2005-2006). Total awards made:

- DIGIT: \$ 1,911,805 to 34 projects involving researchers in 26 countries
- ECAT: \$ 1,268,110 to 29 projects involving researchers in 25 countries

Year 2: Sign Memorandum of Cooperation with the Catalogue of Life partnership (CoLp).—Signed in December, 2003. In addition, Memoranda of Cooperation have been concluded with uBio, the Secretariat of the Convention on Biological Diversity (CBD) and the UN Food and Agriculture Organization (FAO).

Year 3: 40% of all species (names and synonyms) included in ECAT.—As of Year 5, 50% of the ~1.75 M species on Earth have been scrutinized by taxonomists for synonymies and other nomenclatural anomalies. This resource is available via the GBIF data portal.

Year 4 et seq.: Steadily increase name availability; add vernacular names and bibliographic information.—The increase to date has actually been steep, and it is possible to get to bibliographic information for some names via global species databases (GSDs). Some vernacular names are available, especially via the connection among GBIF, GenBank and the Consortium for the Barcode of Life (CBOL). More emphasis will be put on vernacular names during the latter part of GBIF Phase 2.

Year 4 or 5 (approximately): Begin to provide access to "SpeciesBank", a digital online information collation facility.—In 2005 (year 4), GBIF held a 50-person workshop entitled "SpeciesBanks: How Shall We Shape the Future?" There is currently lots of activity in this area by multiple initiatives (McArthur Foundation's MEL project, et al.). Based on guidance from the workshop, financial realities, and the multiplicity of players, GBIF will most likely best serve the community by playing a coordinating role over the next 5 years. The modularity of GBIF's information architecture will accommodate the additions that will be required, and GBIF will continue to work closely with TDWG, which has working groups that are developing standards for Structured Descriptive Data, Images, Natural Collections Descriptions and Taxonomic Names, all of which will be critical to a digital online information collation facility.

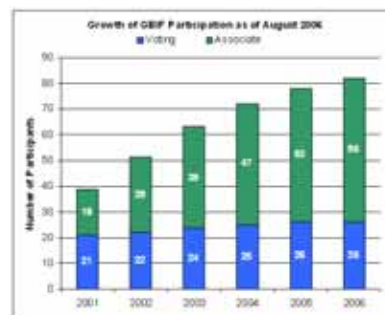
Year 5: Most GBIF Participant Nodes have worked with institutions in their countries to fund digitisation.—Funding for within-country digitization has been provided by: Argentina, Austria, Belgium, Japan, the Netherlands, Switzerland and the USA. In addition, workshops to encourage digitization have been held by Australia, Belgium, Colombia, Costa Rica, Denmark, France, India, Japan, the Netherlands, New Zealand, Nicaragua, Norway, Peru, Portugal, South Africa, Spain, Sweden and the USA.

The NSF Advisory Committee for GPRA Performance Assessment 2005 stated (NSF 05-210, p. 40) that **GBIF's "revolutionary capability for sharing a treasure of unique data collected from across the entire planet will promote scientific collaboration and dramatically improve fundamental understanding of the state of the world's biodiversity."**

THEME 3: PARTICIPATION

Increase GBIF Participation by 10% or more per year.—At the end of 2001, there were 39 Participants; as of August 2006 there are 82.

Bring representatives from developing countries to Governing Board meetings.—Over the five years of GBIF's existence, representatives from 37 countries² have been invited to attend (at GBIF expense) one or more of the 12 Governing Board meetings. Of these, five have become Participants in GBIF, and an four to six are currently considering becoming Participants and/or moving from Associate to Voting status.



Transmit research techniques and capacities ... to the developed and developing world.—Two Ecological Niche Modelling workshops have trained 44 people representing 38 GBIF Participants, including representatives of developing countries. The seed money awards and Nodes mentoring programs have helped to bridge the digital divide and establish collaborations across boundaries.

Enable links among taxonomic and ecological databases, particularly in the developing world.—GBIF is working (with funding from the Moore Foundation) to enable the Amazon Basin Biological Information Facility (ABBIF). It also works with the Inter-American Biodiversity Information Network (IABIN), the EC-funded Networks of Excellence (ENBI, EDIT, ALTERnet, MarBEF), and OBIS, a major GBIF partner for marine biodiversity. All include institutions in countries in the developing world or with emerging economies.

Provide training courses that involve the biodiversity science and information technology communities, establish national and regional training programs and develop a School of Biodiversity Informatics.—More than 12 training workshops have been conducted worldwide in English, French and Spanish, focusing on DiGIR technologies or on putting biodiversity data to use for decision-making (these have served 161 persons, who represented over 40 countries and 15 organizations). Some Participants (Argentina, Costa Rica, Spain, USA) have conducted within-country training programs based on GBIF workshop content. GBIF is working to increase emphasis on “training the trainers.” A GBIF partner, the International School of Biodiversity Studies (ISOBIS), has instituted a Summer School in Biodiversity Informatics, in which several GBIF staff serve as faculty.

Organize a pilot programme of focused support for developing countries.—The Capacity Enhancement Program for Developing Countries (CEPDEC) is in development; negotiations are underway with the foreign aid agencies of various countries for CEPDEC funding. In addition, a Nodes Liaison Officer position has been added to the Secretariat, as well as a program of Node-to-Node mentoring.

Promote the spread and development of the science of biodiversity informatics.—GBIF yearly awards the Ebbe Nielsen Prize, the only one in the world in the area of biodiversity informatics, and has sponsored four one- or two-day Annual Science Symposia that were focused on the topic, as well as special symposia at other international scientific meetings.

Confirmation that GBIF is a vital scientific and societal initiative that is being conducted in a financially and organizationally sound manner was provided by an independent, external review of GBIF's organizational framework and its contributions to science and society. This *Third Year Review* was conducted between April 2004 and February 2005 by a committee of six eminent international scientists, with representatives from CODATA and the international accounting firm KPMG. The Review concluded that **GBIF is “the right initiative with the right goals at the right time”** and that **“if it did not exist, it would need to be created.”**

² Bolivia, Brazil, Burkina Faso, China, Colombia, Congo, Cook Islands, Cuba, Ecuador, El Salvador, Estonia, Ethiopia, Hungary, Indonesia, Jamaica, Kenya, Latvia, Lithuania, Malaysia, Mongolia, Myanmar, Namibia, Nigeria, Palau, Philippines, Russia, Samoa, South Africa, Sri Lanka, Swaziland, Thailand, Uganda, Ukraine, Uruguay, Vanuatu, Venezuela, and Zambia.