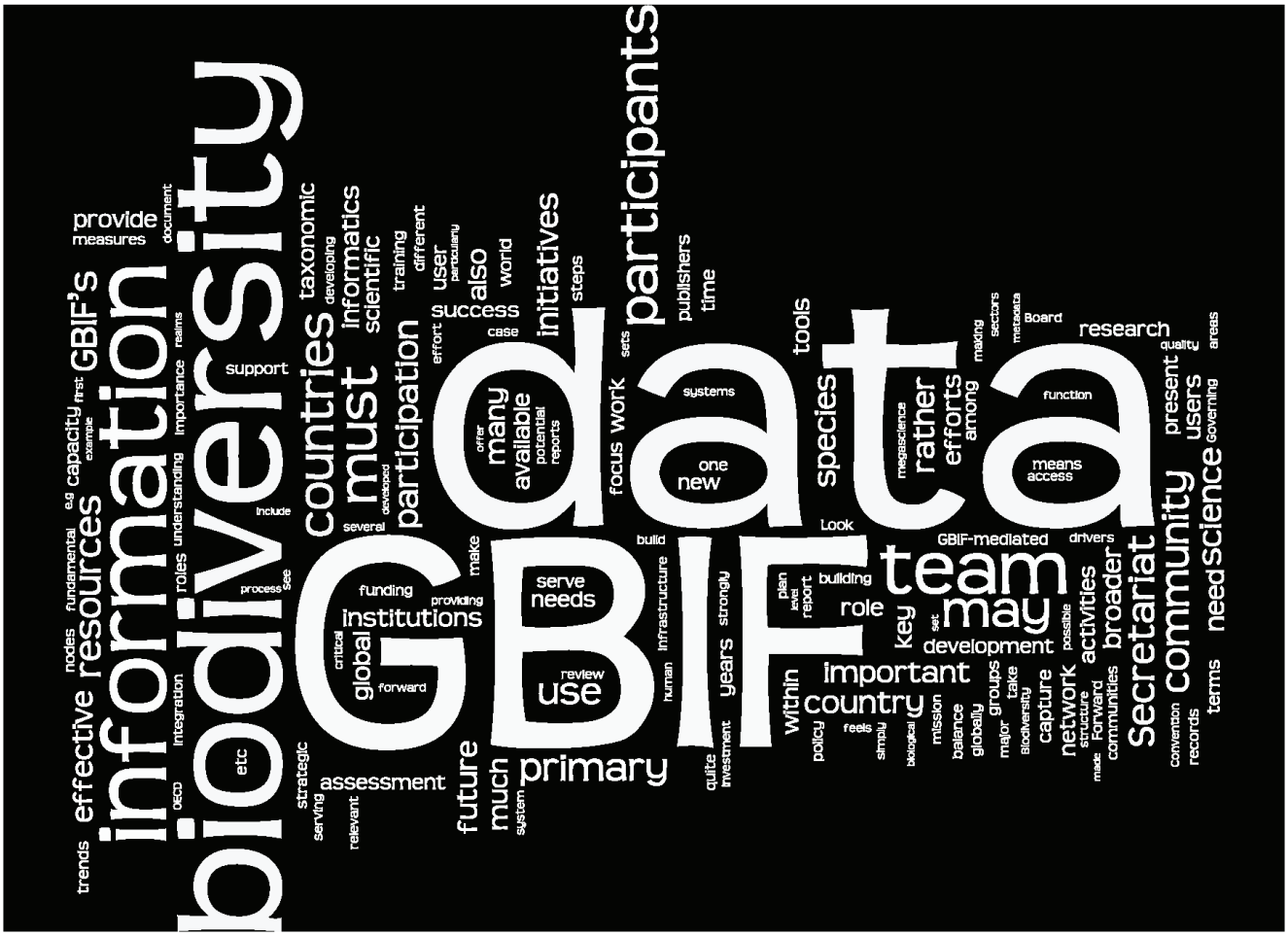
# GLOBAL BIODIVERSITY INFORMATION FACILITY

# 2010 FORWARD LOOK REPORT




*("word cloud" generated by [www.wordle.net](http://www.wordle.net) based on word composition of Forward Look Committee Report, showing emphasis on "data," "information," and "biodiversity")*

*A. Townsend Peterson (Chair)*
*Dora Canhos, Ulf Gärdenfors, Robert J. Scholes, and Yoshihisa Shirayama (Members)*
*Mark S. Graham and Francisco Pando (ex officio)*

17 August 2010

## Contents

# I. EXECUTIVE SUMMARY

The Forward Look Team (FLT) has reviewed extensive information, considered key documents, and engaged in discussions with many players within and outside of the GBIF[1] network. The FLT has contemplated major drivers in the biodiversity informatics world and the major challenges and opportunities for GBIF in the coming years. This report makes suggestions both at the horizon level (i.e., 5-20 years into the future) and more immediately, with the latter considerations aimed at making the horizon-level vision possible.

In the considered opinion of the FLT, the core focus of GBIF in its next phase should continue to be on mobilizing large quantities of known-quality primary biodiversity data, principally at the organism (species) level, that is as broadly representative, both geographically and taxonomically, as possible.

***This focus has several direct corollaries:***

- GBIF must prioritize the data *per se* over information that simply indicates the existence of data sets that are not otherwise accessible. Metadata-oriented initiatives within GBIF must lead directly to discovery of and access to high-quality primary biodiversity data, or must enrich primary data to improve their fitness for use;
- GBIF's focus on being the single most important infrastructure for primary biodiversity data at the organism level must be accompanied by strong, effective, and targeted links to data at the genetic, genomic, and ecosystem levels;
- GBIF should directly and primarily serve the science community, but with a view of how that community in turn uses biodiversity data to address societal needs; as such, GBIF must pay close attention to the needs of this primary user group in terms of their data requirements, data fitness for use, and the training and tools necessary to permit those uses;
- GBIF must take concrete steps toward effective monitoring of current trends in science and policy, such that it is maximally responsive and effective as a megascience data infrastructure; this role should reside within a "science committee" that does more science than at present.

[1] Throughout this document, when we use "GBIF," we refer to the entire network, including the Secretariat, Governing Board, full and associate participants, etc. When we refer to particular components of the network, we specify to whom we are referring.

***The top-level priorities and objectives for GBIF in the next decade should thus include:***

- Maintain and further develop a well-functioning, stable, fit-for-purpose and well-documented primary biodiversity information facility.
- Develop and champion a Global Names Architecture that allows for unambiguous and persistent identification of biodiversity entities at the organism level, that is robust and responsive to taxonomic changes, and that does not necessarily depend on prior scientific description of the species in the traditional sense.
- Evolve a governance system that is efficient, effective, and widely supported, and that makes appropriate provision for the diverse but interacting roles of data publishers, data users, and country representatives.
- Interact broadly with, and influence in positive ways, the process of developing and maintaining interoperability standards that affect the direct core functionality of GBIF and the community which it represents.
- Focus on expanding and improving the fitness-for-use of the data served, by documenting and improving its quality, adding geographic referencing, and expanding the geographic and taxonomic representation.

***GBIF should initiate or expand its activities in the following fields:***

1. A better and broader understanding of user needs, both in terms of data requirements and capacity needs in order to use the data effectively;
2. Metrics of GBIF functionality, such as use reporting and user feedback to data publishers, analyses to evaluate taxonomic and geographic gaps, indicators of data quality, and increased data interoperability. GBIF must invest in diverse and effective impact-oriented measures of success, shifting emphasis from how many data records are available to measures of use, incorporation in research, and incorporation in policy-making. Similarly, success of participation (e.g., training and capacity building, infrastructure development) should be measured in terms of impact, rather than of basic provision;
3. Engagement with major gene-sequence data initiatives to ensure interoperability, for instance by agreeing on standards for linkage between gene-sequence data and primary biodiversity data, as well as for taxonomic identifiers, georeferences, and other information to accompany the sequence;
4. Linking GBIF-enabled primary biodiversity data to the information systems and needs of research at the ecosystem level, where many societal needs and concerns regarding biodiversity are expressed;
5. Linking GBIF-enabled primary biodiversity data to the information systems and needs of research at the emerging taxonomic technologies, in particular the field of "environmental genomics," which will produce large amounts of biodiversity data in coming years that are not directly tied to traditional taxonomic identifiers.

***GBIF should consider reducing, refocusing, or halting activities in the following arenas:***

1. Technology development should not generally be conducted in-house (i.e., at the Secretariat), but rather should be fostered through community efforts, partnerships, contracts, or outsourcing;

2. Metadata (see detailed discussion and definitions below) should enrich primary data records, and promote full enabling of key resources; however, metadata that facilitate data discovery but that do not provide research-level access should not be a major focus of GBIF activities;

3. Secondary biodiversity data (i.e., biodiversity information products such as range maps or diversity maps that are derived from primary biodiversity data) should not be a major focus of GBIF activities.

***GBIF must establish a somewhat different set of strategic partnerships with related initiatives than it currently has, such that GBIF's role is respected, credited fully, and enabled profoundly. The key partnerships for GBIF to cultivate are:***

1. The broader science community that GBIF as a megascience initiative would ideally serve, and within which GBIF is not widely appreciated;

2. GEOSS, and in particular GEO BON, which can facilitate interoperability with other global environmental datasets, servicing societal benefit areas, and resolving, at a political level, issues of free and open data exchange;

3. DIVERSITAS and similar biodiversity scientific research networks, particularly those that go beyond the traditional GBIF support-base of the taxonomic community, to keep abreast with the rapidly evolving and broadening need for biodiversity data;

4. Science-policy interface organizations in the biodiversity sphere, such as the proposed International Platform for Biodiversity and Ecosystem Services (IPBES), which establish pathways by which GBIF may demonstrate the social utility of its services;

5. Organizations that provide secondary, derived, or value-added biodiversity information to the science community, the public, educational institutions, or policymakers, such as Encyclopedia of Life.

***To meet these objectives, GBIF will need to***

- Develop and implement tools, processes, capacity-building, mechanisms, reward systems, standards, partnerships, and infrastructure needed to make the above objectives a reality;

- Clarify the roles, rights, and responsibilities of country participants, data publishers, and data users within GBIF and its governance mechanisms, while attending to the complexities that arise where specific participants play one or more roles;

- Deal with the distinct and diverse needs of non-OECD and OECD participants (with special attention to the opportunities in mega-diverse, emerging mega-economies), as well as the benefits of bringing these different communities together in collective action.

***Specific GBIF targets and goals over the coming 5 years (2011-2015) would thus include***

- Significant improvements in fitness-for-use of GBIF-mediated data, manifested as

    - Closure of the gap in retrospective georeferencing GBIF-mediated species occurrence data, at a rate of at least 30% of ungeoreferenced records per year, such that over the coming 5-year period the gap is reduced by approximately 85%;

    - Implementation of data quality flagging, error detection, and error correction (when possible) modules available for all GBIF-mediated data resources;

    - Implementation of gap analysis modules for coverage in terms of taxonomy, geography, and time for GBIF-mediated data resources; closure of gaps in coverage manifested as significant year-to-year improvement in the gap measures;

- Successful enabling by means of digitization and publication via GBIF of global biodiversity information resources residing in major natural history museum and herbarium collections across North America and Europe that are not presently accessible;

- Broad uptake of GBIF data resources, infrastructure, and tools, manifested as

    - 50% yearly increase in data download rates across the entire GBIF network;

    - 50% yearly increase in citation rates of GBIF data usage in the scientific literature;

    - Broad uptake by members of the GBIF community, and particularly among those participating in GBIF training and capacity building initiatives;

- Development of an effective and responsive user needs assessment process that touches not only the participant focal points, but also the science and decision-making communities within each of the GBIF participant countries and organizations;

- Effective new partnerships across the biodiversity, conservation, genetic, genomic, and ecosystem realms, such that GBIF's unique role in providing primary biodiversity data, infrastructure, and tools is recognized fully, and enabled and enhanced (i.e., improved interoperability) by the new relationships;

- Successful incorporation of new country participants, particularly large and emerging economies and megadiverse countries (e.g., Russia, China, Brazil, India, Ecuador, Indonesia, Philippines).

# II. PREAMBLE

Any vision for the future will benefit enormously from a clear understanding of the past. This section thus provides a brief overview of the origin and first 10 years of the Global Biodiversity Information Facility (GBIF).

GBIF was born out of the OECD Megascience Forum (1999). Among the statements motivating the creation of an entity such as GBIF was the following:

> *It is the need to link these informatics resources (and the people who use them) into a synergistic, interoperable whole that makes biological informatics a megascience endeavor.*

GBIF was envisioned as a network of country and organization participants that develops a distributed system of interlinked and interoperable modules (database, software and networking tools, search engines, analytical algorithms, etc.), with a secretariat that would build coalitions among ongoing efforts, encourage new developments, provide mechanisms for coordinating separate national investments, and forge international agreements. The plan was for GBIF to have an administrative structure consisting of a secretariat, a governing board consisting of delegates from countries, *ad hoc* scientific and technical advisory groups, and close links with established organizations, such as DIVERSITAS, Species2000, and the Clearinghouse Mechanism of the Convention on Biological Diversity. The vision was that the benefits deriving from the existence of GBIF and its provision of abundant information on living organisms and ecosystems would accrue "not only to biological research, but to decision and policy makers, education, and society at large."

GBIF began its work in March 2001, with a mission to make the world's biodiversity data freely and universally available via the Internet, for the benefit of science, society, and a sustainable future. "Biodiversity," in the context of GBIF's work program, meant data regarding occurrences of species in space and time, as well as data regarding taxonomy. Hence, GBIF's purpose more specifically was to promote and implement the compilation, standardization, digitization, and global dissemination of the world's primary biodiversity data, in close collaboration with established programs and organizations that compile, maintain, and use other biodiversity information resources.

The following is not a comprehensive list of achievements between 2001 and 2006, but some highlights which struck the FLT as particularly notable:

- Establishment of a biodiversity information portal populated with rich biodiversity information;
- Increasing participation in biodiversity information sharing at global scales;
- Significant advocacy for free and open biodiversity data sharing;
- Collaboration with several initiatives to establish data models and communication protocols, contributing significantly towards biodiversity data interoperability;
- Making seed money awards to stimulate digitization and integration of biodiversity data, which was particularly successful in establishing new partnerships and participants;
- Signing a memorandum of cooperation with the Catalogue of Life partnership, and with several other key initiatives.

A review was conducted by independent scientific experts in 2004-2005, which concluded that "GBIF is the right initiative at the right time with the right goals and that if it did not exist it would need to be created." That review concluded that, as a public megascience infrastructure, GBIF had made important progress in promoting public access to biodiversity data, with appropriate links to the Convention of Biological Diversity, and that (considering its length of operation and funding) GBIF had made important progress. On the other hand, it acknowledged that, while GBIF had not reached a level of scientific importance and relevance as a megascience undertaking, the level of funding for both the GBIF Secretariat and nodes was low for it to become a genuine megascience endeavor.

Key recommendations from this earlier report and review included the following[2], that GBIF should:

- Be fully supported and continue as a megascience undertaking
- Encourage greater participation by the broader GBIF community in development of the yearly work program
- Develop a comprehensive benchmarking process to have a more thorough understanding of progress
- Divide the organization of outreach and capacity building into separate areas, having nodes help each other, instead of having the Secretariat as the focal point
- Develop a strategy for long-term support of nodes
- Identify new data publishers continuously, building an ever-larger data inventory
- Provide analytical tools through the portal
- Build a distributed user-support infrastructure
- Build an outreach strategy focused on specific user groups
- Consider the different relationships that it may have with its broader community: voting participants, observer participants, associate participants, affiliate participants, data publishers, donors, partners, and friends of GBIF
- Simplify both the MOU and Governance Structure
- Increase its level of funding to US$10M, 70% from basic contributions and 30% from voluntary, supplementary sources.

Based on the work of the review team, a strategic and operational plan for 2007-2011[3] was developed, led by the Secretariat with the contribution of several scientific advisory committees

[2] Full report available at http://www.gbif.org/fileadmin/Temp_for_New_Web_Site/3YR_full.pdf.

and participants. This plan has as its subtitle the phrase: *From Prototype towards Full Operation*. Its main purpose, therefore, was to provide a roadmap to enable GBIF to move toward full operational status by 2011. Fundamental directions laid out in the plan for 2001-2006 were maintained:

- Enable scientific research that has never before been possible
- Facilitate use of scientific data in biodiversity policy and decision making
- Make a whole world of biodiversity data that are currently exceedingly difficult to access freely and universally available via the Internet

The operational plan grouped tasks within major themes, with clearly stated goals and milestones:

- ***Content***
  - Complete >95% of the Electronic Catalogue of Names of Known Organisms
  - Promote digitization and sharing of primary data records about species, reaching between 500M and 1B records by 2011
  - Incorporate new data types and new tools that will facilitate web-enabled taxonomy by 2011
- ***Informatics***
  - Improve user friendliness and capabilities of the data portal system to both humans and machines
  - Enable Internet searching for biodiversity data across all levels of biological organization, from molecules to ecosystems
  - Provide tools for improving data quality and assessing its fitness for use
- ***Participation***
  - Support the capabilities of participant nodes (100% online presence by 2009)
  - Enhance the capacity of people to use and provide data
  - Increase participation in GBIF and have all mega-diverse countries involved by 2011
- ***Funding***
  - Increase the Secretariat's Budget from €3 million in 2006 to €10 million in 2011

A work program was approved in 2008 for 2009-2010[4], in which substantial changes from previous work plans were presented. The plan aimed to reduce dependency on the Secretariat, moving to a more distributed network of fully active participants. Emphasis on the relative roles of the Secretariat and Participants changed, whereby "participants take greater responsibility for implementation and promotion of the agreed work program with less reliance on the limited capacity and budget of the Secretariat." The GBIF Secretariat is described as a "service provider to the client," where the client is the GBIF network of participants.

The strategic plan for 2007-2011 had stated three major themes within which tasks were grouped: *Content*, *Informatics*, and *Participation*, themes believed to be central to its success. The

---

[3] Available at http://www2.gbif.org/strategic_plans.pdf.

[4] Available at http://www2.gbif.org/WP2009-10.pdf.

work plan for 2009-2010 reduced the activities to two thematic areas (*Informatics* and *Participation*), downgrading *Content* to a sub-item of *Informatics*, even though if GBIF does not have content, it has nothing. Such a step has the potential to reduce the significance of content as a major topic to be fully addressed by all.

More generally, GBIF presently serves just below 200M primary biodiversity records, and its taxonomic data resources exceed 1M scientific names; GBIF also has developed a substantial and significant role in development of globally accepted standards and protocols for biodiversity informatics. GBIF is in an important second phase, in which the institution must 'grow out' to become a permanent, useful, and efficient data backbone for biodiversity science and policy, focusing on providing an information infrastructure; community-developed tools, standards and protocols; and training and capacity-building, such that all stakeholders may take advantage of the information.

# III. PURPOSE OF THIS REPORT

GBIF had its first programmatic review in 2005 (GBIF Review Committee, 2005). A second review, 5 years after the first, accompanies this "forward look" report. GBIF is thus at an opportune moment to look forward, assessing its challenges and its opportunities over the coming decade. Many lessons can be learned from experience to date that can improve GBIF's probability of success in coming years. A forward look, such as this report, must learn from those lessons from the past and provide a vision of what GBIF could and should be doing in the future: who should be its partners, what should be its structure, and what data and technologies are crucial to its success. Such an integrative vision is the purpose of this report.

A team of five persons (the "FLT," see Appendix 1) was given a mandate by the GBIF Governing Board to undertake this assessment in October 2009, with the task to be completed by May 2010, so that it may be duly considered in advance of the next Governing Board meeting. Members of the FLT are not affiliated with GBIF, but are knowledgeable in broad sectors of the field of biodiversity informatics and the community of stakeholders and users that it must serve.

In accordance to the terms of reference (TOR; see Appendix 2) of the FLT, it should, based on performance assessment of GBIF's developed independently and on a horizon-level view of the broader field, build its assessment of future options into a series of concrete recommendations. Of necessity, given time constraints, the Review Team and the FLT had to work in parallel, but advance copies of the review have been provided to the FLT, along with detailed direct discussion. Additional reference documents used in creation of this report included (1) Summary of GBIF Accomplishments 2001-2005, (2) GBIF 3rd Year Review, GBIF Plans for 2007-2011: From Prototype Towards Full Operation, and the (3) GBIF Work Programme 2009-2010.

In accordance with the TOR, *the primary output from the team will be a high-level strategic document outlining the unique niche which GBIF should fill post-2011, with solid rationale for endorsement and further support from current and future Participants. If the GB endorses a further phase of GBIF, the Forward Look Report will form the core of a further Strategic Plan for GBIF post-2011*. As a consequence, we have focused this document not on the details of how to implement each suggestion that we offer, but rather on a view of priorities in the mid-term future and paths by which to achieve those objectives.

This document is thus the FLT's report to the GBIF Governing Board. Given the TORs, the FLT sees its mission as being to make and justify recommendations on broad and high-level components of GBIF's future, providing the foundation for a possible Strategic Plan for 2012-2017 or beyond. Expected outputs included in the TOR are:

- Key areas that GBIF should build upon and consolidate, as well as new areas to be addressed
- These key areas should be synthesized into a set of organizational recommendations
- Recommendations should address scientific, technical, technological, financial, human capacity, and political issues within a future strategic context
- The document must clearly define the separate roles of the Secretariat and the members of GBIF in any future activities

The document will be provided to the Governing Board in electronic format for commentary and review in May 2010, and the FLT Chair will present the findings in person to the Governing Board in October 2010.

# IV. MAJOR TRENDS AND DRIVERS

A critical component of any 'forward look' is an in-depth understanding of global trends and driving factors that are, in effect, the context for GBIF's activities and existence. Perhaps the most significant of these trends are external to GBIF—that is, they are the trends that will be motivating the broader biodiversity science and policy community to use, contribute to, and support GBIF over coming years. Other drivers are more intrinsic, relating to the data available and the community that GBIF serves. Clearly, then, GBIF must be cognizant of and responsive to these forces, if it is to serve the purpose for which it was created. Extrinsic trends and drivers—in the sense of what the broader community is "doing" and the evolving needs for biodiversity information, as fodder for reflection by GBIF and a context for the remainder of this report—include the following:

- ***Cybertaxonomy***

  According to the Convention on Biological Diversity, one of the greatest challenges to overcome is the so-called "taxonomic impediment." After >250 years of biotic inventory, no global checklist is available, and very few countries even have national lists (most are inaccurate and/or require updating). About 1.75M species have been described, out of an estimated 10M or more. Biological collections worldwide likely hold 2-3B specimens, but only a small proportion of this number is in digital format, much less integrated and served online. Online working checklists integrated with new information and communication technologies are now very possible. Besides making existing data available and interoperable, information systems must be transformed into powerful instruments to accelerate the process of expanding knowledge about species and on how ecosystems function. To be able to understand, predict, and perhaps influence the responses and capabilities of complex biological systems, interoperable information infrastructures must be in place to receive, store, manage, retrieve, and serve digital data to all interested users. Such data systems must also be used as an instrument to assist further field work in identifying what is recorded and/or collected, in promoting use of accepted standards and protocols for data collection, in establishing automatic data management procedures, and in leveraging significant steps forward in addressing the large-scale challenges faced by taxonomy.

- ***Pervasive, accelerating, and unprecedented biodiversity loss***

  The rich global heritage of millions of years of evolutionary diversification is currently being lost at a rate far in excess of its replacement. This loss is accelerating owing to human activities, principally habitat loss and pressure for human resource use, but increasingly exacerbated by climate change, environmental pollution, and invasion of alien species beyond their historical ranges. This "Biodiversity Crisis" places a premium on documenting fully, and understanding in detail, biodiversity in all of its dimensions in the shortest spans of time possible.

- ***Need for comprehensive and reliable data on the environment***

  Detailed and reliable data on biodiversity and the environment are becoming central to many international agreements and initiatives. For example, the Convention on Biodiversity set the 2010 Target for reduction of biodiversity loss; the World Summit on Sustainable Development in 2002 identified rich data as a key element in effective solutions; and the Group on Earth Observations (GEO) prioritizes development of such information resources. New initiatives are re-proposing and amplifying several of the same objectives. The importance of such data resources is recognized by

many international treaties, including the RAMSAR Convention on Wetlands, Convention on Migratory Species, Convention on International Trade in Endangered Species, and others. Finally, even the private sector is seeing the importance of such resources, given the need to evaluate development activities in terms of their environmental impacts. This panoply of efforts on the part of so many actors both opens many opportunities for GBIF, and places a premium on careful, strategic, and intelligent positioning of GBIF.

- ***The role of biodiversity in ecosystem function***

  This driver centers on the increasing appreciation of the fundamental biodiversity underpinnings of ecosystem function (Loreau *et al.*, 2001). From an ecosystem perspective, it is not just a question of which type of ecosystem is present, but rather what is the actual species composition. From the biodiversity perspective, it is not just a question of what species are present, but also what role they play in the functioning of the system. This factor moves GBIF-enabled biodiversity information from the 'nice to have' to the 'have to have,' emphasizing that biodiversity considerations are not simply aesthetic, but rather quite utilitarian in issues of human well-being.

- ***Human, animal, and plant health***

  Disease systems represent sets of interacting species, one or more of which cause damage to the health or well-being of humans, their domesticated livestock and crops, or other elements of biodiversity. Emerging diseases and novel pathogens are appearing and being detected with increasing frequency worldwide (Jones *et al.*, 2008), representing threats to human well-being. Disease transmission is a biodiversity phenomenon (pathogens, vectors, and hosts are all elements of biodiversity). The apparent increase in such outbreaks in recent years may be related to the pervasive disturbance of ecosystems: biodiversity may regulate disease emergence, if current theory on "dilution effects" proves robust (Ostfeld & Keesing, 2000). In a world of asymmetrical warfare, deliberate introductions of pathogens (i.e., bioterrorism) are an additional consideration, demanding capacity for swift and effective identification of origins and provenance of elements of biodiversity. The increasing recognition of the biodiversity aspects of disease represents a challenge for biodiversity science and opportunity for GBIF to take on considerable additional relevance.

- ***Open source initiatives and data democracy***

  Information is increasingly seen as a global commons, with attendant rights of open access. Under this view, information-rich countries and institutions can no longer consider biodiversity information as their exclusive property, but rather must recognize some degree of rights of countries of origin of the information and/or countries of potential impact. This trend is not a one-way street—i.e., it is not simply 'repatriation' of information from information-rich entities to information-poor entities—rather, considerable empirical evidence indicates that opening access to information increases its utility and value to the holders as well. Put more simply, since biodiversity information represents a 'sunk cost,' the cost-to-benefit equation can only be balanced by maximizing the use of the information. Finally, biodiversity information, in the form of the taxonomic literature, specimens, and other materials has an impressive half-life, in contrast to many other sectors of science.

- ***Data-intensive science***

  An emerging approach in science has been called "data-intensive science" (Hey et al., 2009). In more traditional approaches, patterns, processes, and regularities are subsumed by theoretical constructs that may take the form of simple equations with a few parameters (like Kepler's or Maxwell's equations) that encapsulate what it is known about important phenomena. In the data-

intensive approach, in contrast, thousands of digital objects comprising terabytes (perhaps petabytes) of data represent the relevant information for the problem, which is too complex or too noisy to encapsulate in a few simple parameters.

- ***Citizen science and participatory science***

  The global community is increasingly 'enabled' to participate in science. Considerable public interest, combined with the existence of field guides, regional summaries, Internet resources, and efficient communications technologies, makes biodiversity science particularly well-suited to incorporate citizen scientists powerfully and integrally, including in developing countries and for taxa that have not participated strongly in these dimensions to date. This step would require more profound appreciation of the citizen scientist as full partner in the science process, rather than a lesser element in the enterprise. A unique feature of biodiversity sciences is that lay experts have always been part of the enterprise, albeit often poorly acknowledged. The opportunity exists to take their contribution to a new level, and in the process build a whole new support base.

- ***Continued acceleration of technology for information management***

  Moore's Law indicates that computational power, memory, and bandwidth will continue to become cheaper in real terms, and indeed connectivity grows at a factorial rate. As a consequence, many considerations that might appear impractical or overly ambitious are actually quite reasonable within the scope of future planning. A corollary of this continual and massive improvement in computing capacity is the emergence of large-scale and well-funded global IT corporations, which frequently have interest in and need for engagement with public good activities such as biodiversity information. This interest may offer considerable opportunities to GBIF.

- ***Genomics technology developments***

  In tandem with the IT improvements have gone massive improvements in technologies for characterizing genomes—full-genome sequencing is now fairly quick and is becoming almost routine, robotic high-throughput sequencers are broadly available, and genomic information is growing immensely. Similarly, certain media (e.g., marine environments, soils) are now being surveyed for genetic elements (e.g., segments of DNA) and their derivatives (e.g., proteins), producing large amounts of novel data. These improvements are now running ahead of more traditional taxonomy, producing large amounts of "biodiversity data" that are tied not so much to traditional scientific names but to genomic characterizations—this trend has important implications (explored below) for coming years of GBIF as a megascience research platform in the biodiversity world.

In addition to the extrinsic factors, several intrinsic factors—that is, trends and patterns within the area of biodiversity informatics—must also be considered. These trends are important, as they have the potential to influence the future path of GBIF in terms of how certain steps will result (success or failure). Important intrinsic drivers include the following:

- ***Geographic balance in biodiversity information***

  Well-known in the biodiversity informatics world is a global mismatch between distribution of global biodiversity among countries and the distribution of biodiversity information among countries. That is, only in very rare cases can the so-called "megadiverse" countries count on mega-scale biodiversity information resources (Soberón et al., 1996). The distinction reflects, to a large degree, the historical 'north-south' divide in economic development and a parallel historical divide in development of systematics and scientific collections. This driver has manifold implications for GBIF activities: (1) the need for "data repatriation" to provide biodiversity-rich countries with information regarding their own natural heritage; (2) the need to incentivize retrospective digital capture and exchange of data held in large-scale institutions that are typically in biodiversity-poor countries, but that hold massive amounts of information on world biodiversity; (3) the need for prioritization of biodiversity information resources for digital capture and integration; and (4) the need to promote and facilitate *de novo* construction of useful biodiversity information resource capture in developing countries.

- ***Taxonomic balance in biodiversity information***

  Also well-known is that biodiversity information is strongly biased towards certain groups, particularly vertebrates, and that most major lineages are poorly known, poorly documented, and little studied, to the point that nine of every ten species on Earth may not even be described scientifically, let alone characterized in any detail (see classic analyses of Erwin, 1991). Not only is the information highly biased, but also the expertise and human resources by which more information can be generated—in short, the human infrastructure of taxonomy—is similarly biased towards certain groups, leaving many others without much expertise. The gap between information-rich and information-poor lineages is widening. All of these lineages merit study: many poorly-characterized lineages are nonetheless of economic, ecological, public health, or other importance to human well-being. Patterns of richness and endemism frequently do not coincide among lineages, so well-documented groups are poor proxies for the rest. The early success of GBIF in assembling huge amounts of biodiversity data was thanks in largest part to harvesting the low-hanging fruit of the information-rich lineages. The next phase will be more challenging, requiring effective strategies for prioritization of digital capture of existing data and effective development of new data. In addition, given the improved capacity for producing genomic information, which we suspect may come to dominate the future of biodiversity information, GBIF must prepare itself to deal with the emergence of abundant "biodiversity information" that does not rely on traditional taxonomic systems.

In sum, GBIF does not exist in a vacuum, nor does it have a unique and isolated niche such that it can exist without careful attention to important trends in the broader community. The above lists and descriptions are among the more relevant of those trends, and must be pondered carefully in outlining any future steps for GBIF.

# V. THEMATIC FOCUS AND TARGET COMMUNITIES

Perhaps the most important single suggestion that this "forward look" can offer to GBIF is that of the dire need to understand its own thematic focus and target communities much better. That is, an initiative such as GBIF cannot afford to imagine that it simply needs to "create the data resource and they will come." This approach, we believe strongly, will not serve GBIF well in coming years—rather, data, thematic emphases, and target communities (of people, that is) must be identified carefully and strategically, such that every expenditure is maximally effective, and such that the greatest possible benefits accrue to the broader biodiversity informatics community from GBIF's activities.

The OECD working group report on biological informatics (1999) that led to GBIF's establishment noted that 75% of the world's holdings in biological collections were located in OECD countries, so the initial emphasis was focused on OECD countries to guarantee basic data and funding. It envisioned that OECD countries would play a leading role, not only in digitizing and making their data openly available through a common framework, but in helping and even financing developing countries to do the same. These countries would pay for maintenance of a secretariat that would be responsible for the articulation of the network and for interlinking and making existing systems interoperable through common protocols and standards. Participation of non-OECD countries and other associate participants was also sought. After 10 years, of the 30 OECD member countries, 21 are voting members, 4 are associate members, and 5 have not joined GBIF. Of the 21 OECD voting members, only 3 are considered "megadiverse" countries. Of the 14 non-OECD megadiverse countries, on the other hand, 5 are voting members, 5 are associate members and 7 are not part of GBIF (Brazil, China, Democratic Republic of the Congo, Ecuador, Malaysia, Papua New Guinea, and Venezuela).

Country-based representation and participation must be sought and built continuously, with a focus on scientific development to eliminate barriers that prevent cooperation and providing a framework for data, information, and knowledge sharing. Other partnerships are fundamental to achieving this overall mission, which makes the overall structure much more complex perhaps than was first envisaged. In sum, then, GBIF's original design and many of its present activities assume that *countries* would all have a leading role in serving and using data as a megascience initiative, while international networks and initiatives would form a second group with which a close partnership would be established.

The reality of the situation, however, is more complex, as illustrated in Figure 1. Different individuals, institutions, or countries may play one, two, or three of the roles illustrated in the figure, to varying degrees and in different contexts. For instance, a large natural history museum in Europe or North America may in some cases only serve data (case A); however, a university down the street may serve no data, but may make extensive use of GBIF-mediated data in its research endeavors (case C); and both institutions may be located in a country that is

a GBIF member, but the GBIF focal point for the country neither serves nor uses GBIF-mediated data (case G). Some GBIF target groups combine two or all three of these roles: another museum or university may both serve and use data (case B), and some country participants serve data directly (case E), or use data from GBIF-mediated resources (case F). Too often, data-rich institutions in developed countries have made little or no effort to publish their data to the broader community whatsoever.

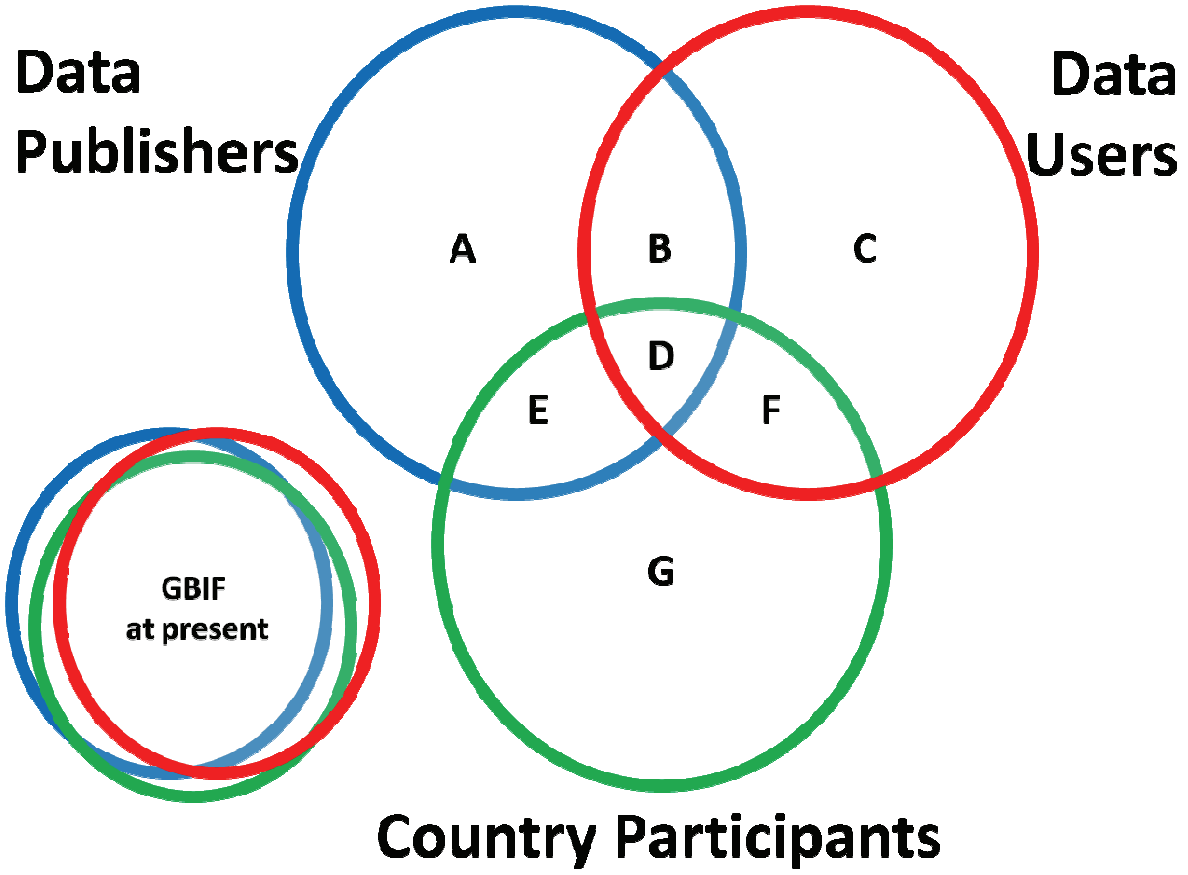



**FIGURE 1.** SCHEMATIC OF THREE TYPES OF PARTICIPANTS IN THE DATA-RELATED REALMS OF THE GBIF WORLD, INCLUDING COUNTRY PARTICIPANTS, DATA PUBLISHERS, AND DATA USERS. THE LETTERS ARE DISCUSSED IN THE TEXT AS ILLUSTRATIONS OF DIFFERENT EXAMPLE PARTICIPANTS. THE INSET PRESENTS A CARICATURE OF THE PRESENT CONFIGURATION OF THESE CIRCLES IN THE GBIF INITIATIVE, IN WHICH INSUFFICIENT DISTINCTION IS MAINTAINED AMONG THE THREE SUITES OF PARTICIPANTS.

The present-day GBIF treats these three roles as if they overlapped broadly (Figure 1, inset). A critical step for GBIF is to parse the roles of different target groups or stakeholders into these different dimensions. For example, some country participants may have few or no data to serve, but may be much more interested in training in data use; and some institutions may wish to become active data publishers and users, in spite of not being official country participants. These points are only illustrations, but serve to reveal the complexity of the GBIF community, and the degree to which GBIF has oversimplified its treatment of its target groups.

GBIF target communities are quite diverse. They may be governmental; quasi-governmental (based at state-supported universities, research institutes, and museums); or fully non-governmental organizations, which in turn are mostly not-for-profit operations, but may include commercial corporations. Eventually, GBIF may even extend its reach to a more direct interface with the general public. Thematically, the GBIF community includes taxonomy, systematics, evolutionary biology, ecology, biodiversity conservation, agriculture, transportation, forestry, landscape planning, forensics, food security, and public health. In environmental dimensions, GBIF's areas of relevance include simple focus on preservation of natural heritage, but also on maintenance of ecosystem services and natural resource-based economic sectors, and climate change remediation.

In general, though, the interaction between GBIF data and its use in socially-relevant ways is (appropriately) mediated via the scientific community, rather than being directly used, for instance, by policymakers. This point has important consequences for the way the data are accessed, presented, and documented. The FLT believes that this relationship between GBIF, its 'proximal users' (scientists), and 'ultimate users' (society) will (and should) persist within the next phase. Thus GBIF needs to have a clear understanding of the uses to which the data it serves are put, in order to anticipate their fitness-for-use, but the best source of such understanding are the scientific users who are themselves serving ultimate users. As such, GBIF would benefit enormously from representation at the annual meetings of the scientific societies that would most use and benefit from the GBIF megascience infrastructure.

The team feels strongly that an important dimension of GBIF's activities in coming years must speak to each of these communities *on its own terms*, perceiving and responding to where communities do and do not overlap, and what roles each community can and cannot play. The current arrangements implicitly are based on a tit-for-tat model: I will show you my data if you show me yours. The future may need to be more radically redistributive – from each according to its capacity, to each according to its needs. This point will be touched on in a later section that discusses how to measure success of GBIF efforts … when attempting to understand the degree to which some activity is meeting user needs, are simple numbers of data records made available or numbers of hits adequate measures of success? We strongly suspect that they are not, and suggest below a suite of alternatives.

Country participants must also realize the fundamental role that they must play as a component of GBIF, not only as data publishers and/or users, but as drivers of a megascience initiative. As such, country participants should appreciate the need to advocate for free and open access to biodiversity data, information and knowledge, enabling a collaborative scientific endeavor that never before was possible. Participation takes the form of investment in the global endeavor (i.e., GBIF membership dues), but also, and probably to a much larger extent, in supporting biodiversity informatics-related efforts within the country. The team suspects that improving avenues for country participants and the scientists resident in their countries to see real and tangible benefits in terms of information products (rather than just feeling good about having supplied data to the broader effort) may constitute an important step forward in this regard.

# VI. USER NEEDS AND DESIRES

GBIF deals—appropriately and correctly so—with primary biodiversity observations that are clearly traceable to their sources. GBIF itself does not *make* primary observations; rather, GBIF participating organizations and biodiversity science institutions around the world do so. It is thus essential for GBIF to maintain vital links to the original institutions that generate the data. This core mandate does not exclude generation and management of downstream information products ('secondary' or 'derived' data), provided that they link *integrally* back to the original primary data. This requires scrupulous metadata generation and transparent methodological reporting. The FLT believes that the "bread-and-butter" of GBIF and its source of comparative advantage should be primary biodiversity data[5], and not secondary biodiversity data (i.e., inferred information such as range maps, atlas data) or pointer metadata[6]; the FLT believes firmly that primary biodiversity data, and specifically at the organism level of biodiversity (i.e. specimen and quality-controlled observation records) will serve GBIF's mission much more firmly and uniquely than other types of data.

In the broadest and most general sense, GBIF users are all pretty much alike with respect to their needs: they want primary biodiversity data of dependably-known quality in terms of taxonomic identification and spatial and temporal accuracy of the observation. Thereafter, of course, the needs diverge—different user communities (see above) have radically different interests and desires. As a result of this diversity within the GBIF community, a user needs assessment process that takes into account all of the sectors identified in Figure 1 becomes an immediate priority for the organization—the results of this assessment should translate directly and tangibly into tools and training for different sectors of the GBIF community. Instead of only commissioning once-off, IT systems-oriented, and rather arid 'use-case analyses', the GBIF user-needs assessment should be continuous, integral, and conducted using all available tools: from establishing effective feedback mechanisms at all points in the data chain, allowing users to say what worked and what did not; to structured use-case analyses; to creation of opportunities to bring users onto the GBIF team for 'sabbatical' periods to develop capabilities on both sides. In

---

[5] Primary biodiversity data are here defined as data records placing a taxon in a place at a particular point in time, with a high degree of certainty, based on physical collection or direct observation. These data records are carefully distinguished from secondary biodiversity data records, which—though based on primary biodiversity data—are subject to inference, generalization, interpretation, or other human-mediated processing that distances them from the original, unitary records

[6] Here, definitions become crucial. Metadata are data that describe other data records. We distinguish "pointer metadata," which use simple and very general data fields (e.g., country, order or class, decade) to describe data set derivation and provenance and ascertain general data correspondence, but do not contain the full or even sufficient detail to be considered primary biodiversity data. "Record-level metadata," on the other hand, are a very different thing: data that document the provenance, characteristics, derivation, and quality of primary records—these metadata are crucial and critical parts of the data mediated by GBIF.

general, such greater interaction with GBIF country participants and associate participants has great potential to awaken greater interest and investment on the part of new users and partners, as well as to provide rich case studies of which the Secretariat can take advantage to "sell" the GBIF concept to additional new partners.

Some types of usage may be fomented by GBIF through partnerships with institutions that require species' occurrence data to produce analytical reports. Products such as reports based on GBIF portal data for IUCN's redlist assessment process, list of species occurring in protected areas, endemic or rare species of the world, lists of names per taxonomic group, per country etc. could be produced for further specialists' scrutiny. Other such secondary products could include range summaries of insect species that are vectors of disease, status reports of spread by invasive species, etc. However, the FLT emphasizes that these product "needs" should generally be identified *by GBIF participants and partners*, rather than the GBIF Secretariat.

GBIF should not second-guess the data quality needs of users by offering only one quality of data ('research grade'), or by applying unnecessarily restrictive data quality standards determined independently of the use. The key issue here is 'fitness-for-use', which is best decided on a case-by-case basis by the user. The role of GBIF is to flag the data quality in a reliable and easily-accessible way, so that the user is in a position to make an informed choice.

Another important priority is to add value to existing data. A first level is that of adding geographic references to data, following well-developed standard methodologies, which allows rich cross-linking of GBIF-enabled data to other data realms, such as remotely sensed data. In addition, data can be enriched by flagging species that are endangered, rare, endemic, invasive, etc., through data integration systems, and adding image-based documentation of important specimens. Quantitative information, such as abundance, biomass, habitat, water depth, and other value-added information that may accompany basic data records should be enabled by means of add-ons and annexes to the basic Darwin Core structure.

However, GBIF should not be tempted to be responsible for these other types of data, so as not to lose focus on primary biodiversity data and species-level taxonomic data. The overall aim in this case is to increase the usability and scope of species and occurrence data. Known quality is fundamental, so GBIF should aim at developing tools and mechanisms to make quality assessment reports available and to flag possible data inconsistencies that users can filter if necessary.

# VII. DATA PUBLISHERS

The institutions or individuals who publish data to the GBIF network are nearly as diverse as the potential users. They range from long-standing natural history museums, to groups organizing electronic data collection of observational data, to custodians of complex biodiversity information networks. The data that they publish to the GBIF network take two basic forms: primary biodiversity records (which include both biological collections data and observational data), and taxonomic data documenting the species-level constitution of biotas worldwide.

The team notes that other highly relevant data realms (e.g., species status data, phenotypic data, genomic data, geospatial data, remotely sensed data, ecosystems data, ecological data, climate data) also exist. However, the team emphasizes that those data realms are the focus of their own initiatives that are effectively GBIF's sister organizations. Rather than fall into the temptation to replicate those efforts, GBIF should focus on building effective partnerships with related initiatives, and work to assure effortless interoperability among the different data sets. This is the purpose of GEOSS, for example, and GBIF should make full and active use of that mechanism.

For data that *are* within GBIF's core scope of activity, tools and applications should be sought, enabled, and offered to help data publishers with detecting and correcting errors and thus improving and documenting data quality (nomenclature and georeferencing). Overall data assessment reports and record-level flagging for suspect data records or fields within data records should be made available online to help analyze the geographic and taxonomic balance of the data being served. These indicators can be used to measure GBIF's overall progress as to data content, and should be available to all interested, and data usage reports must be fed back to data publishers to show their relevance and importance. The experience with the vertebrate networks in the US is that data usage reports constitute invaluable indicators to data publishers of the value of their information resources to the broader community.

# VIII. DATA SERVED AND TOOLS DEVELOPED

The considerations presented above regarding user needs and data publishers lead the team to a series of guidelines for future steps in the realm of data resource development for both primary biodiversity data and species-level taxonomic data. The FLT emphasizes that all of these points are considered to be individually and collectively important and fundamental to the success of the GBIF mission. We envision that the future success of GBIF is proportional to the degree to which these elements are addressed.

- ***Pointer metadata versus primary biodiversity data*** – Pointer metadata are the focus of GBIF's Global Biodiversity Resources Discovery System (GBRDS), and are envisioned as information useful in detecting and prioritizing resources for capture and integration. While the FLT appreciates and supports the need for a GBRDS for these purposes, we emphasize that *pointer metadata themselves will seldom provide fertile soil for research*, and that serving pointer metadata will not by itself address issues of data access and data repatriation. Metadata are a means to an end, not an end in themselves. GBIF should keep a strict focus on serving primary biodiversity data, and use GBRDS only as a step towards full capture and integration of primary data, as described below.

- ***Fitness-for-use of data as an important focus*** – GBIF has, to date, taken a fairly hands-off approach to data quality and data enrichment. In contrast, some related activities (for instance, of the US-based vertebrate networks funded by NSF) have spent 70-80% of their effort on accurate georeferencing and development of techniques for detecting and flagging likely errors. GBIF has focused on serving large *amounts* of data. It should focus future efforts on understanding, documenting and improving the quality of those data (e.g., information regarding precision of georeferences and confidence of taxonomic identifications) and practices for incorporating such information back into the 'original' data sets efficiently and using it in diverse analyses. GBIF should also analyze and weigh the disadvantages of some sorts of data in terms of desirability for inclusion—for example, the FLT is not convinced about the general utility of data originating from environmental impact assessments. The utility of data always needs to be weighed against the costs of acquiring and maintaining it.

- ***Need for persistent identifiers***[7] – A handicap in the present state of biodiversity information is the lack of globally unique, persistent identifiers for primary biodiversity data records. Once incorporated into biodiversity data realms, such identifiers will enable many, exciting steps forward, but their absence is becoming nothing short of crippling, particularly as regards integrating different data realms. GBIF has fostered key discussions in this area. This feature should be accorded high priority for implementation, as it will be an important investment in the shortest time frames.

- ***Need for reduced delay in incorporating data improvements into original data sets*** – GBIF has discussed and explored a number of means for tracking information "back" to the original data set, and providing the possibility of flagging likely errors as dubious (with markup as to why), and

[7] A persistent identifier uniquely identifies a particular data record, and survives changes in the taxonomic arrangement or physical location of the record.

providing comments back to data curators. However, still, as yet, no working version has been implemented, so the priority of these steps remains high.

- ***Emphasis on interoperability*** – Biodiversity data do not exist in a vacuum, but rather co-occupy a thematic space with many other domains of information (e.g., genomic data, geospatial data, etc.). The most exciting research results from biodiversity data and the GBIF-mediated data archive are likely to come from making links among these information realms (Peterson *et al.*, 2010). As a consequence, it is critical to reexamine GBIF data architectures and formats with an eye to making data as readily interchangeable and interoperable as is feasible among domains of information, as is apparently occurring as part of GBIF participation in GEOSS efforts.

- ***Gap assessment and data adequacy analyses*** – One of the most serious challenges that GBIF is facing is that of assembling data resources to form a globally relevant information system, not just one that is relevant to data-rich regions of Europe and North America. The team sees assessing gaps (spatial, taxonomic, and temporal) and developing effective strategies for filling them as a critical priority, and recommends that GBIF survey the scientific literature for effective gap assessment tools and metrics, and apply them to GBIF-mediated data globally, by taxonomic group, to explore data density. Strategies for filling gaps identified can then be developed, including capture and integration of key data sets, even if those data sets are housed chiefly in Europe or North America[8]. Key data sets may be identified for prioritization for capture and integration via GBRDS, if it is able to 'grow out' fast enough to be useful, in the process.

- ***Balance one or a few taxonomic groups globally*** – Given the pervasive imbalance in the spatial distribution of biodiversity data, the team suggests that GBIF invest significant effort and resources in balancing data density for one or a few groups (e.g., in particular birds and some other vertebrates, in terrestrial realms; perhaps certain groups of plants and insects; in marine realms, perhaps certain groups of fishes) globally by means of prioritized retrospective data capture in tandem with citizen-science reporting facilities. As demonstrated by the research of the 2009 Ebbe Nielsen Prize winner, balance across space permits researchers to take much better advantage of data resources. Indeed, such an effort would lead to many "mega-science" analyses that would provide effective demonstrations of the quality science that could be developed were such a data resource to be created.

- ***Secondary data products must refer integrally back to primary data*** – Building primary data resources can be a thankless task, as the "fun" (and recognition, in terms of publication and citation) lies in the analyses and interpretation of those data. As a result, GBIF will always face the temptation to produce secondary data products, such as range maps, range loss estimates, species status assessments, regional biotic summaries, etc., which (as discussed above) should optimally spring from GBIF participant needs and demands. These secondary products are needed, and certainly attractive, but should always refer back in a repeatable and traceable manner to the primary data. Metadata attached to these secondary products will play a very important role in this aspect.

---

[8] The team notes a comment from one GBIF participant from a data-rich European country, to the effect that his institution needs more clear guidance regarding digitization and mobilization of resources—they *have* resources for data capture, but no guidance in how to prioritize. They are very willing to adjust their activities so as to maximize their contributions to the overall broader good, but are getting no guidance.

- ***Data deposit*** – GBIF has explored a very interesting potential, which is to mirror the journal-driven data deposit system of GenBank, but for primary biodiversity data instead of molecular sequence data. In the successful prototype, a Memorandum of Cooperation between GBIF and the journal *ZooKeys* demonstrated the potential of linking scholarly publication of biodiversity science contributions to data deposit in the GBIF distributed network. The team sees this step as highly interesting and creative, and suggests that GBIF explore means of extending this concept more broadly among key biodiversity science journals, particularly as persistent identifiers become pervasive among GBIF-enabled data.

- ***Tool development*** – GBIF should take steps to encourage new technologies (e.g., automated data capture, cloud computing solutions) that change the landscape of biodiversity informatics, permitting novel steps forward or alleviating impediments to data capture and integration. Although the FLT does not see a GBIF role in *developing* such technologies (which fall into the realm of biodiversity informatics *research*), GBIF should support such developments, and promote them once their utility in advancing the GBIF mission becomes clear.

- ***Portal development*** – The GBIF portal has certain basic capabilities, and has excellent advantages over other such portals in terms of query and reporting speed. In light of the research-support mission of GBIF, however, careful consideration of functionalities that research users desire regarding queries (e.g., Boolean searches), dataset filtering and exploration, and research-grade export will be quite important. Such attention to user needs will surely reap benefits in terms of greater usage and user satisfaction.

- ***Usage reporting*** – From the perspective of data publishers, a very important currency is that of how the data that are published are used by the community—that is, many data publishers are strongly interested in seeing extensive use and appreciation of their data resources. GBIF should implement (or complete the implementation of) such usage reporting such that country participants and individual participants alike can access this information flexibly and efficiently.

- ***Technology management*** – A final consideration in this realm is in management of GBIF personnel resources for technology development. The FLT would not wish to see GBIF invest heavily in technology development, as that falls much more in the category of biodiversity informatics research, and would detract from the mission of serving data to researchers and policy-makers. A particularly relevant example at the moment is that of shifting many GBIF provider services to a cloud-computing environment—the VertNet research team has recently obtained funding for such research, and GBIF has the luxury of observing their success, and learning from relevant lessons. In short, the team feels that GBIF should invest time and effort chiefly in technology that has been vetted thoroughly.

GBIF has made an important shift in recent years from the idea of subcommittees to the more flexible and responsive idea of task groups. Several highly useful reports have resulted from these groups, treating several points mentioned above. As a result, clear paths forward are now in sight for those issues (e.g., persistent identifiers). The FLT encourages GBIF to (1) publish these reports in the scientific literature (to avoid growing a gray literature, and to provide an independent peer-review), and (2) take aggressive steps towards transforming these compilations into definitive advances.

# IX. PARTICIPATION

The team sees several fundamental problems in the structure and function of GBIF participation. At present, nodes are fundamental within GBIF, but the implementation has been a political, and largely country-based, concept[9], rather than a functional one. We refer to Figure 1 in these comments: while country participants are a key element in the GBIF network, excessive focus on the 'country,' subsuming other forms of participation may devalue other players. The result is the present system, in which key data-holding institutions and potential data users are neglected, in favor of country participants that, while politically well placed, may neither provide nor use data from the GBIF network. Much broader use of and reliance on GBIF-mediated data in a global community of users will be a critical element in building broader bases of support for GBIF and improving its funding base. A revised GBIF vision of participation must therefore consider a fuller diversity and independence of participation.

At the same time, GBIF country participants need to make several shifts in their own role in this initiative. GBIF participants must work to develop vibrant "nodes," including capture of important biodiversity data sets, participation in and rich response to communiqués regarding user needs and accomplishments, and promoting use of and participation in the GBIF network within their countries or organizations. Finally, and critical to the survival of GBIF, is assuring prompt signing of Memoranda of Understanding and payment of GBIF membership fees, as the GBIF Secretariat at times has had to make difficult decisions regarding staffing and services for lack of membership fees already promised.

Although country participants are clearly important drivers of GBIF, given that they are also its funders, gaps exist between country participants and other sorts of participants, who have much to offer to the network in terms of crucial data resources and expertise. These participants provide resources of considerable value to the GBIF mission. As a consequence, it is important that GBIF reconsider its present structure, with an eye to promoting separately and collectively all of the sorts of participation shown in Figure 1.

The FLT believes that a key step will be the user needs assessment mentioned above. For instance, in recent GBIF meetings, besides Governing Board meetings and Science Symposium, activities offered to GBIF participants have centered on training in serving data. We received comments from several GBIF participants that indicated that these activities did not consider what *the participants desire* in terms of training and capacitation. An effective user needs assessment extending beyond the country focal points to the broader community would provide a much more detailed and informative picture of what particular participants need

[9] We note that, although GBIF admits associate participants that are not countries, these entities have no voting rights within GBIF, and as such are somewhat disenfranchised within the organization. While the team does not necessarily advocate a dramatic reworking of the structure of the organization, providing non-country participants with more of a voice in the organization is nonetheless important.

from GBIF, and what they can offer in return. In this example, many country participants may be much more interested in training in the *use* of GBIF-mediated information for purposes such as conservation planning, inventory planning, gap analysis, invasive species modeling, niche modeling, etc.

Among the most important challenges for GBIF in terms of participation is building investment and activity among the large, biodiversity-information-rich institutions located largely in Europe and North America, as these institutions hold crucial portions of the global knowledge bank regarding biodiversity not otherwise available. In a non-trivial proportion of cases, these institutions have not been as active as they could in capturing and mobilizing the data associated with their specimens, yet those data are critical to achieving a globally balanced biodiversity information resource within GBIF. As a consequence, building more effective relationships with such institutions will be crucial, potentially involving mobilizing resources to support such efforts, and possibly exploring creative solutions, such as mobilizing people-power from the biodiversity-rich countries to assist in data capture and improvement (see the example of Mexican bird data resources, Navarro-Sigüenza et al., 2003). Country participants have an important role here too, in providing the incentives for such institutions to participate in GBIF. The result will be a much-richer data resource that provides a much-clearer view of biodiversity pattern worldwide, of particular benefit to developing countries, where such data are often minimal or lacking.

A further challenge as regards building GBIF participation is that of promoting fuller investment and involvement by the scientific community, in particular that based in developing countries. The team urges GBIF to build such participation by means of creative initiatives, such as informational presentations at scientific meetings, and funding opportunities for visiting scientists to spend research-productive periods of time at the GBIF Secretariat or at GBIF-partner institutions. GBIF should also ensure that the expense of attending meetings does not come to constitute a permanent and pervasive barrier to fuller participation by developing countries in leadership roles within GBIF.

Finally, GBIF participation should be "built out," in terms of participation by countries and organizations, particularly the former at the level of full voting members. The team sees clear priority for megadiverse countries and emerging mega-economies (especially Russia, China, Brazil, India, Ecuador, Indonesia, Philippines). These steps, although clearly very challenging and time-consuming, are little short of critical for GBIF's success in meeting its goals.

# X. RELATION TO OTHER BI INITIATIVES

The field in which GBIF is playing is a very complex array of data streams, institutions, initiatives, and opportunities (see Figure 2). As a result, a critical element is a detailed and continuously-adapting understanding of "who is doing what," and which elements GBIF must be principally responsible for, is crucial. More specifically, GBIF must have a clear picture of how to relate to each other effort: certainly, GBIF's strength is in the primary biodiversity data and taxonomic data to which it provides access, and this strength should remain the central element in its existence. In many or most cases, cooperation and partnership with other efforts will be the optimal strategy. In a very few cases, careful analysis may indicate that competition to establish a dominant position is a better course, but such cases should be exceptions rather than the rule. The general ethos should be collaborative rather than competitive.

Of particular relevance as regards this arena is GBIF's relationship to large-scale, politically-mandated efforts, including GEO and its Biodiversity Observation Network (GEO BON), the Biodiversity Convention, and other biodiversity-related international agreements and initiatives. A careful and agile strategy for GBIF can use them to establish legitimacy and importance for biodiversity information, foster a global environment of data sharing, improve integration of biodiversity information with other data streams, and to resolve higher-level challenges and impediments.

In this line of thinking, the team recommends that GBIF assemble a detailed and carefully considered assessment of its "niche" in the broader picture of data relevant to biodiversity informatics. GBIF is *uniquely* a source of primary biodiversity data, but need not only serve occurrence and taxonomy data at the species level. However, it should expand from this foundation of comparative advantage only very carefully and strategically, and not extend into areas that other initiatives manage better or can dominate easily. Clearly, the GBIF leadership must understand the scope, intent and capacity of potentially competing initiatives and interact strategically with them. Examples include VertNet, SpeciesLink, VectorBase, LifeWatch, and GEO BON, at least to some extent. Other initiatives may *appear to* compete, especially to those not intimately familiar with each program, but carefully negotiated partnerships could clearly be of mutual benefit (e.g., Encyclopedia of Life, Barcode of Life, LifeWatch).

Two arenas were identified by the team for particularly careful thinking and strategizing.

- ***Genomics*** – First, as mentioned above among major external drivers, recent advances in genomics are likely to change the biodiversity world dramatically and fundamentally. This shift is already happening, and represents a significant opportunity for GBIF to lead in an important paradigm shift in biodiversity science; into the future, GBIF must play an integral, proactive, and positive role in managing and promoting such changes in biodiversity science—that is, managing the primary biodiversity data associated with the genomic information. In addition to potential competition for 'big initiative' funding, the genomics world is now producing large amounts of data related to

biodiversity, but not tied to particular scientific names. The team strongly recommends that GBIF initiate high-level conversations to position itself to take on a leadership role for the primary biodiversity data associated with this new world of genome-level data[10]. This suggestion is not the same as being the repository or server of genomics data: it is the function of building links between an established body of knowledge and an emerging one.

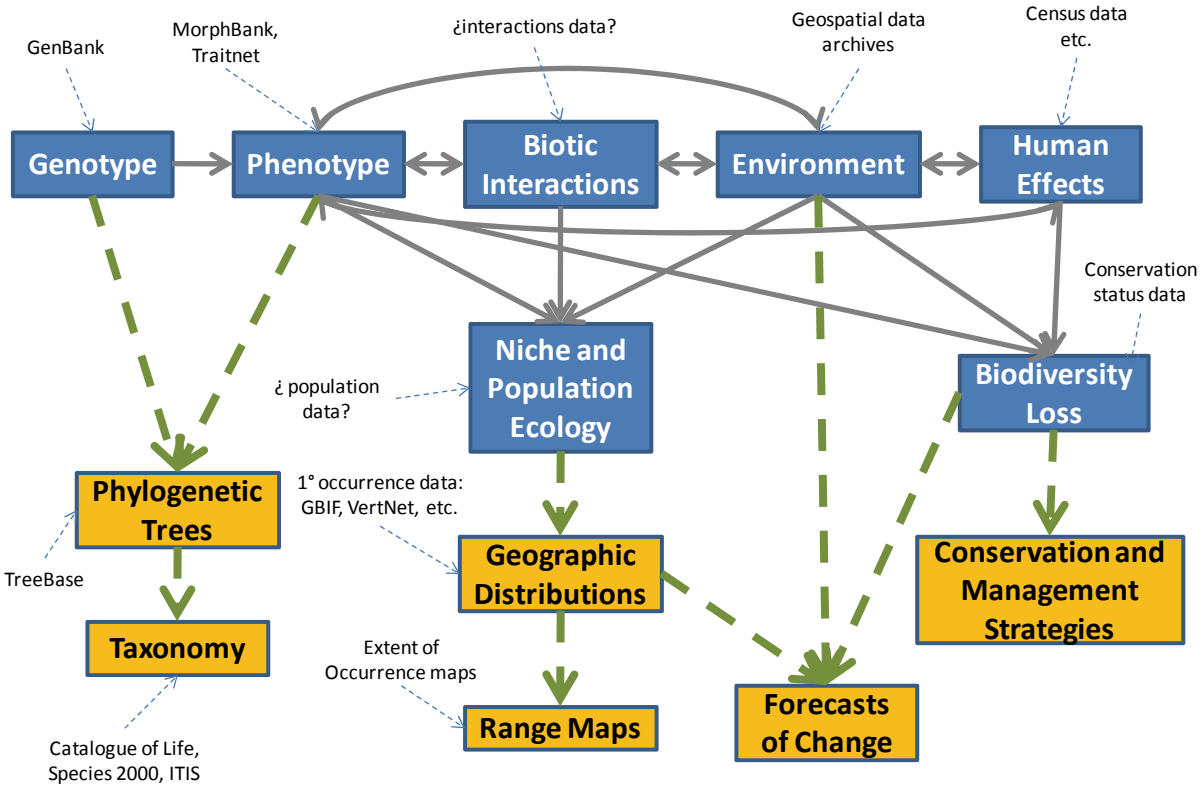



FIGURE 1. PARTIAL SUMMARY OF THE INFORMATION WORLD THAT SURROUNDS GBIF AND ITS EFFORTS TO SERVE BIODIVERSITY DATA. THE BLUE BOXES INDICATE ASPECTS OF BIOLOGY AND BIOLOGICAL PROCESSES. THE GOLD-COLORED BOXES REPRESENT INFORMATION PRODUCTS THAT GBIF AND OTHERS MANAGE. FINALLY, THE BLACK-LETTERED LABELS PROVIDE PARTIAL SUMMARIES OF INITIATIVES ASSOCIATED WITH EACH OF THE ISSUES, FROM PETERSON ET AL. (2010).

- ***Ecosystem function*** – A similar challenge for GBIF is to establish effective linkages with the community that focuses on ecosystems and their function. The present connections are via geographic coordinates, and are at best tenuous. The occurrence of an organism within the mapped extent of an ecosystem says little about its participation or role in that ecosystem. A linkage that is much more vibrant than the present one would require information on the interactions between species, and their ecological and physiological attributes. An "interactions database" and associated ontology permitting richer vocabularies is necessary to go beyond simple spatial association. At the same time, GBIF does need to provide, or at least link to, credible sources of ecosystem-level biodiversity information: the spatial location of ecosystems, and key properties such as dominance, productivity and protection status.

[10] Note that genomic data are not synonymous with gene-sequence data, as managed by GenBank.

Well-negotiated cooperative agreements such as those outlined above with other initiatives could establish GBIF's central role in biodiversity informatics. The team envisions GBIF working in tandem with the other initiatives to pipeline basic reports and analyses (species lists, distributional summaries, etc.) to the other group, of course with proper credit and attribution to GBIF and to the original data providers. Finally, GBIF should explore the potential for partnering with one or more private sector information technology enterprises, perhaps finding dimensions of mutual interest and benefit.

# XI. IMPLICATIONS FOR GBIF AS AN INSTITUTION

Adjusting the roles and types of GBIF participation to reflect the multidimensional framework shown in Figure 1 is a critical step. The present situation is quite monolithic: only a single way to participate exists (with minor variants), and little variation other than voting or non-voting exists. As discussed above, the FLT recommends parsing roles of publishers, users, and participants in view of their quite-different roles, needs, and resources.[11]

GBIF should strive to provide a data infrastructure that is alert and responsive to major trends and drivers in the biodiversity world and the broader conservation, environment, agriculture, public health, and natural resources communities. Nonetheless, the Secretariat should focus on providing a data infrastructure to support responses to those trends and drivers, rather than establishing its own directions intrinsically. These points suggest the need to build more vibrant relationships with user communities, in essence keeping a "finger on the pulse" of the broader biodiversity informatics arena. The FLT explored a variety of suggestions for GBIF on how to achieve this connection: a stronger "science input" role for the current Science Committee, or other mechanism to permit thinking and exploration along the frontier of the field of biodiversity informatics, such as "frontier think tanks" developed each year alongside the Science Seminar.

The FLT feels strongly that GBIF should revisit the idea of campaigns and/or seed money initiatives as means of strategic investment in intriguing ideas from the broader community—these efforts appear to the team to have been successful in stimulating new activity and new partners, yet they have been all but abandoned over the past year or two[12]. More generally, the user needs and gap assessments discussed above should guide these (and any and all) strategic investments, rather than have the choice of campaigns reflect the interests of the Secretariat or strong individual voices on the Governing Board. Any such strategic initiatives should be user- and participant-driven, scope-limited, time-bound, and with have clear goal orientation, and as such should be driven by some combination of the Science Committee and Nodes Committee. The team is not referring in particular to campaigns, but rather to the broader idea of providing a modicum of support for exciting opportunities from within the biodiversity informatics world.

---

[11] A minor—but potentially quite important—issue is that of time limits on Associate Participant status, which the team feels may function to move countries out of the picture entirely, rather than to encourage ascension to Full Participant status.

[12] Note that the Secretariat responded to this point: "The Campaigns are being given time to deliver on their 3-year goals before being fully evaluated." The team feels strongly that it is completely feasible to evaluate the Campaigns based on existing information, and that the concept of seed money or campaigns as means of supporting and responding to GBIF community interests and ideas is crucial to keeping GBIF responsive, active, and creative.

A final consideration from the team, again in line with Review Team thinking, is that the GBIF funding model may need some revisiting—quite frankly, because country contributions are relatively small, GBIF is quite vulnerable to one or a few countries 'dropping out,' and bringing down the entire network. Several country participants indicated to the team that country contributions may be so small as to make GBIF irrelevant to national policy… in effect, if you ask for too little, you become unimportant. GBIF must operate in a user-needs-responsive manner to ensure that a cost-benefit analysis by the countries will make such a commitment an easy decision.

The resources currently available to GBIF are a factor constraining the timely delivery of its mandate. That mandate is challenging, necessary and urgent. GBIF (including support to its participating institutions) is the most obvious and efficient way to achieve the core objective of mobilizing primary biodiversity data at the organism level. GBIF needs to be resourced at a level that will allow it to meet these objectives effectively, and sooner rather than later.

# XII. MEASURES OF SUCCESS

All of the discussions in the preceding pages become near-moot if GBIF has no effective and relevant means of tracking which efforts and expenditures are successful, and which are not. The first challenge in this regard is that of understanding the meaning of "success." The FLT feels strongly that measures of how much data GBIF offers to the community, while not irrelevant, are not sufficient to gauge success. Much more important is understanding what is *used* by the community and *incorporated* in fundamental ways into research, policy-making, decision-taking, and education. That is, which GBIF-enabled data, infrastructure, and tools change the way science is done and the way policy is set? The FLT recommends that GBIF incorporate measures of impact in any evaluation of success. A second general point is that measures that reflect the shrinkage of important gaps would be more useful than raw information on the number of records on offer. Some potential measures are listed below:

- ***Repeated gap assessments*** – Gap analyses are designed for the prioritization of resources for capture and integration; it is feasible to repeat such analyses at regular intervals. Such information will allow measurement of the rate at which gaps are closing, both in geographic and taxonomic dimensions.
- ***Cost-benefit analysis of GBIF's role*** -- What value is added to data as a consequence of GBIF participation? These measures would emphasize ease of access, value added via georeferencing and error detection, etc.
- ***Use rates, and their balance globally*** – To what degree are different sectors of the relevant communities accessing GBIF data, and how is this use distributed across sectors of the community, geographic regions, etc.?
- ***Feedback rates, and their balance globally*** – Once feedback mechanisms are well established in the GBIF data portals, to what degree are users participating in providing feedback to improve GBIF-mediated data, and how is this use distributed across sectors of the community, geographic regions, etc.?
- ***Citation rates, and their balance globally*** – Citation in published reports of scientific research provide one key measure of the importance of a data infrastructure. Similarly, citation as a key source in planning and prioritization initiatives (which may not be actually published) reflects importantly on the importance of the infrastructure. The balance of this use across geography and thematic communities is also of interest.

These above measures apply not just to data resources, but also to other GBIF activities such as informatics infrastructure developments and training and capacity building. That is, infrastructural success should not be measured by how many standards or software packets were published in a given year, but rather by the degree to which the community adopts those tools. Similarly, training success should not be measured in terms of how many courses were given and how many students trained, but real incorporation of the class material in daily activities of the trainees. Did the training really make a difference to how the trainee or his/her

institution functions? This sort of results-based evaluation, rather than simply numbers processed, has much more to offer in guiding future efforts.

Similar points can be made regarding nodes and their participation in GBIF. A more mature evaluation would ask how many nodes have offices and budgets, how many have information systems in place, how many are serving data, how are they interacting with in-country scientific communities regarding GBIF, and how much in-country training is going on? On a more fundamental level, GBIF can evaluate its progress in achieving global coverage of national participation, and particularly coverage of biodiversity information-rich nations, emerging major economies, and biodiversity-rich nations as participants.

A final comment is that ultimate impacts of infrastructural "backbones" such as GBIF typically happen a long time into the future, and frequently involves many players. As such, because impacts do not always lend themselves easily to quantitative evaluation, "telling the story" can be a powerful way of recording impact. Each annual report can document one or two of these case examples, to illustrate who is using GBIF-mediated data, protocols, and standards, and how: what changed as a result of GBIF existing and working and making information and tools available?

# XIII. LITERATURE CITED

Erwin, T.L. (1991) How many species are there? Revisited. *Conservation Biology*, **5**, 330-333

Gbif Review Committee (2005) *Gbif 3rd-year review report*. Global Biodiversity Information Facility, Copenhagen.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L. & Daszak, P. (2008) Global trends in emerging infectious diseases. *Nature*, **451**, 990-993

Loreau, M., Naeem, S., Inchausti, P., Bengtsson, J., Grime, J.P., Hector, A., Hooper, D.U., Huston, M.A., Raffaelli, D., Schmid, B., Tilman, D. & Wardle, D.A. (2001) Biodiversity and ecosystem functioning: Current knowledge and future challenges. *Science*, **294**, 804-808

Navarro-Sigüenza, A.G., Peterson, A.T. & Gordillo-Martínez, A. (2003) Museums working together: The atlas of the birds of mexico. *Bulletin of the British Ornithologists' Club*, **123A**, 207-225

OECD Megascience Forum (1999) *Final report of the working group on biological informatics*. Organisation for Economic Co-operation and Development, Paris.

Ostfeld, R.S. & Keesing, F. (2000) Biodiversity and disease risk: The case of lyme disease. *Conservation Biology*, **14**, 722-728

Peterson, A.T., Knapp, S., Guralnick, R., Soberón, J. & Holder, M.T. (2010) The big questions for biodiversity informatics. *Systematics and Biodiversity*

Soberón, J., Llorente B., J.E. & Benítez-Díaz, H. (1996) An international view of national biological surveys. *Annals of the Missouri Botanical Garden*, **83**, 562-573

# XIV. APPENDIX 1 FORWARD LOOK TEAM BIOGRAPHIES

The composition of the team is as follows:

- **A. Townsend Peterson** (Chair) – University Distinguished Professor and Curator in the Biodiversity Institute and the Department of Ecology and Evolutionary Biology of the University of Kansas. Ph.D. from the University of Chicago in 1990. He is an active researcher in biogeography, disease ecology and geography, ornithology, and biodiversity informatics.
- **Dora Canhos** (Member) –Associate Director of the Centro de Referência em Informação Ambiental (Reference Center on Environmental Information - www.cria.org.br), Brazil. She has been working with databases and online information systems since 1985, and has been involved with biodiversity information networks since 1992.
- **Ulf Gärdenfors** (Member) – Professor of Conservation Biology and Deputy Director at the Swedish Species Information Centre (ArtDatabanken), University of Agricultural Sciences, Sweden. Ph.D. in Systematic Zoology from Lund University. Has initiated and/or taken part in several bioinformatics endeavors at SSIC, such as the Swedish Taxonomy Initiative, national Red List assessments, databases for observations, taxonomy and ecology, and a Swedish LifeWatch network.
- **Robert J. Scholes** (Member) – Fellow, Council for Scientific and Industrial Research, South Africa; Chair of the Group on Earth Observation Biodiversity Observation Network; Vice-Chair DIVERSITAS. Ecologist with a PhD from University of the Witwatersrand (1987), works on issues of global change and ecosystem services.
- **Yoshihisa Shirayama** (Member) – Director and Professor of Field Science Education and Research Center, Kyoto University, and Director of Seto Marine Biological Laboratory, Kyoto University. D. Sc. from the University of Tokyo in 1982. Active researcher of marine ecology and conservation biology especially regarding deep-sea meiobenthos, and taxonomy of marine nematodes, kinorhynchs, and Loriciferans.
- **Mark S. Graham** (*ex officio*) – Director, Research Services, Canadian Museum of Nature. Ph.D. from Memorial University of Newfoundland and Labrador, 1985. Active museum executive, inactive researcher in environmental physiology and aquatic toxicology.
- **Francisco Pando** (*ex officio*) – Chair of the GBIF Participant Node Managers Committee of node managers. Head of the GBIF Spain node. Ph.D. from Universidad Complutense (Madrid) in 1994. Researcher in the Spanish Research Council, working on the taxonomy of Myxomycetes, global checklists of plants and mosses, information systems for collections management and floristic studies, and computer-aided identification systems.

# XV. APPENDIX 2 – FORWARD LOOK TEAM TERMS OF REFERENCE

## BACKGROUND

The Global Biodiversity Information Facility (GBIF) is now mid-way through its second 5 year term, for which the strategic goal 'from prototype towards full operation' was adopted. The MOU that is the foundation for this period was re-negotiated from the MOU of the first 5-year term following an external review of GBIF's progress in developing the prototype. The strong endorsement of GBIF from the 2005-6 Review, in summary, that GBIF should not only be continued, but should be grown to expand its influence and impact led to the development of a 2007-2011 Strategic Plan. The mandate provided therein is reflected in the growing membership, delivery of new products and services and increasing recognition of GBIF as a key global science research infrastructure for the mobilisation of, and access to primary biodiversity data for improved decision-making.

As GBIF has matured, so inevitably, have both GBIF and its external operating environment changed. This second phase of GBIF has seen an increase in Participation, linear growth in data mobilization through Participants and significant changes in personnel at the Secretariat. As GBIF prepares to move beyond 2011, the Governing Board (GB) during GB15 in Arusha, Tanzania in November 2008 endorsed the need for an updated review of performance to date against the objectives of the 5-year Strategic Plan, which would help shape development of a preview or 'Forward Look' for a possible next phase beyond 2011. Upon GB approval and endorsement for a next phase, the Preview would be expanded into a new Strategic Plan bound within a revised MOU and associated financial arrangements.

## FORWARD LOOK TEAM REQUIREMENTS

The Governing Board agreed that, given GBIF's maturation, a more limited Review to that initiated in 2005 would be satisfactory as a basis on which to build future plans and a new MOU, and that a greater focus should go into the 'Forward Look'. This two-stage process will see a separate Review Team provide a performance-assessment document with a set of key recommendations regarding GBIF's future focus, upon which the Forward Look Team can build their assessment of future options. Accordingly we seek to contract suitably qualified individuals/institutions to undertake this 'Forward Look'. The GBIF Secretariat will formally contract those selected.

The Forward Look Team (4+ persons) is required to have a combination of:

- a full understanding of the challenges and opportunities (scientific, governance, political, cultural, financial, technical) faced by international science and policy research infrastructure initiatives/organizations;
- a significant understanding of the realm of IT, and in particular, future technology trends;
- extensive career involvement in the field of biodiversity informatics and its potential contribution to solving relevant key issues facing humankind;
- strategic oversight and/or project management experience in delivering international/multilateral operational programmes aligned to strategic objectives.

To ensure relevance and focus, members of the Forward Look Team should have a good current involvement with and/or understanding of GBIF, its history, mandate, challenges and achievements. One member of the team will be designated as Chair.

The Executive Committee considers that the Forward Look will be largely a desk and brain-storming exercise based primarily on both the Review Report and recommendations and the sourcing of strategic insight, knowledge and inputs as required to complement the Team's own expertise and insights. The Primary output from the team will be a high-level strategic document outlining the unique niche which GBIF should fill post-2011, with solid rationale for endorsement and further support from current and future Participants. If the GB endorses a further phase of GBIF, the Forward Look Report will form the core of a further Strategic Plan for GBIF post-2011. Identification of a small number of external 'peer reviewers' by the Team to review a draft/s is highly recommended.

## FORWARD LOOK TEAM OUTPUTS

The Forward Look Team will provide a substantive written report giving a quantitative and qualitative assessment of the future potential role of GBIF, clearly articulating the key areas that the initiative should both build upon and consolidate, as well as what new areas should be addressed. These must be synthesized into a strategically coherent set of core organizational recommendations for consideration by the individual GB members in their own assessments of a next phase of GBIF. The recommendations should address scientific, technical, technological, financial, human capacity and political issues within a future strategic context; they must be presented in such a way that national-level decision-makers can easily understand and utilize the information to make an informed decision on their future involvement in, and support for, GBIF. It must thus clearly define the separate roles of the Secretariat and the members of GBIF in any future activities. The Team Chair will be required to present the findings in person to the Governing Board in October 2010.

## PROPOSED TIMELINE AND DELIVERABLES

- *Establish Forward Look Team:* Team members agreed after consultation with the Executive Committee and contracts signed *(mid-August)*
- *Kickoff Meeting and initial information gathering:* Team to visit GBIF Secretariat to understand the scope and nature of the organization, and resolve any concerns over the scope of the Review Report output (as primary baseline for the Forward Look). *(Sept/Oct)*
- *Develop and present draft plan to develop Forward Look (future Strategic Plan):* Draft plan for development and scope of Forward Look presented to GB 16. *(6 October)*
- *Develop Draft FL/Strategic Plan:* Brainstorming of Team, sourcing other inputs/expertise as required. Telephonic interviews with selected stakeholders if required *(selected following consultation with the Executive Committee).* Development of a draft Forward Look/Strategic Plan with appropriate commentary.*(1 February)*
- *Draft FL/Strategic Plan circulated to ExCo:* Draft SP circulated to the Executive Committee, comments incorporated/addressed. *(1 April 2010)*
- *Presentation to the ExCo:* Presentation by the full FL Team to the Executive Committee at the 2010 mid-term of the draft Strategic Plan. *(30 April 2010)*
- *Revise draft FL/Strategic Plan:* Following input from Executive Committee members, revise the draft Strategic Plan. *(31 May 2010)*
- *Circulate draft final Strategic Plan to GB:* Circulate for Governing Board comment, comments incorporated. *(end June 2010 )*
- Final Report circulated August 2010 and presented to GB at GB17. (+-October 2010)

## REPORT DEVELOPMENT PROCESS

The assignment is scheduled to start in August 2009. The assignment should be divided into the following phases, in accordance with the timeline above:

- Kickoff meeting with GBIFS in Copenhagen
- Design of forward look and draft plan
- Plan agreed/amended at GB16
- Biodiversity informatics landscaping and future scenarios assessment
- Review of GBIF documents, gathering of wider information (interviews, etc.), including preliminary findings of Review Team
- Preliminary analysis of findings
- Source wide/broad external comment on preliminary findings; research gap analysis
- Preparation of draft report (including recommendations on broad/high-level components of a possible Strategic Plan 2012-2017 or beyond)
- Draft report circulated to ExCo for comments, presented at mid-year ExCo April 2010
- Preparation and submission of draft final report to GB, June 2010
- Revisions incorporated, Final Report circulated to GB, August 2010
- Final report presented at GB17, October 2010

## REPORT FORMAT

The primary output of this assignment will be a 'Forward Look' Report regarding a third phase of GBIF post-2011. The main Report is expected to be not more than 20 pages, with additional annexes as required.

The contents of the report should include (but not necessarily be limited to):

- Executive Summary
- Introduction
- Forward Look approach description
- Findings and Analyses
- Recommendations
- Priority Components of a GBIF Strategic Plan post-2011
- Annexes, including References

# XVI. Appendix 3 – Letters of Commentary from External "Visionaries"

**From Jesse Ausubel (Program Director, Sloan Foundation; Director, Program for the Human Environment, Rockefeller University):**

This is thoughtful and well-edited.

I have only one significant comment. The draft underplays the very basic fact that GBIF is an Intergovernmental Organization and the special leverage that can flow from that power. It is also what distinguishes GBIF from CoL, EOL, DIVERSITAS and the dozens of other e-Biosphere players. In most countries the museums and universities are almost entirely controlled by governments – through ministries of education or culture or science. I would put at the top of my own Forward Look a much harder press to get GBIF to use its governmental power to get governmental entities to deposit data and take other relevant actions. Even in the USA, NIH's requirements for GENBANK and PubMed have transformed the practice of science. GBIF should be starting to get governmental muscle behind its actions and mandates. While the USA situation may be uniquely complex because of the big private sector in the USA, in most other GBIF members, the government is in a position to mandate many needed actions for the growth of the substance of GBIF. If, within 5 more years, the agencies in GBIF member countries concerned with management of and research on biodiversity are not routinely instructing relevant national agencies to contribute to and use GBIF, then that is a big failure. It means GBIF could just have been an interacademy program under ICSU or some other nongovernmental entity.

I urge your committee to think much more aggressively about what it means for GBIF to be intergovernmental, and how it can use that power.

**From Sahotra Sarkar, Professor, University of Texas**

I agree with almost everything in the report and want only to emphasize one major shortcoming: among the drivers (Section IV) that should be considered--and where GBIF would be particularly relevant--is that computer-based decision support tools are being used for biodiversity conservation, management, and restoration planning in all continents and at least 158 countries and all major transnational NGOS (including WWF, CI, TNC). Data from GBIF

can be critical for this purpose and one major problem has been a lack of freely available georeferenced data at the species level. The problem with the use of all these tools has been the poor quality of data. And, yet, plans based on these tools are being implemented in many countries already, including Australia, Colombia, and Indonesia—beside, possibly, a variety of countries such as Suriname in which CI is working.

I would also have like to have seen some mention of interaction with various agencies and organizations providing remote-sensed data which would complement the species-level data. (I would add this as a separate category besides genomic and ecosystem data.)

I would emphasize that to achieve its objectives, GBIF would do well to establish much stronger links with the scientific community including the academic community and selected NGOs. The governance system for GBIF should have adequate representation from the scientific community. While I appreciate the emphasis on capacity-building (for instance, the sabbatical idea [Section VI, which is in general a very strong section]) even this could be better done through outsourcing to universities, etc. Having a diverse and competent scientific advisory board, not chosen on political grounds, would do a lot to help.

(A lesser worry: there's too much emphasis on national governments, not enough on big NGOs, and certainly not remotely enough on small local NGOs which are the main drivers of change in many countries, including India—the mention of citizen science in Section IV seems like pious lip service and nothing else.)

About reduction in current practices (p. 5): I fully agree with the recommendations, especially taking the focus away from developing range maps. (Technology, especially computational tool, development should be outsourced in particular to academic institutions which are relatively inexpensive sources of expertise). In fact, I think that there is still too much emphasis on technique development in Section VIII when these could also be optimally outsourced. But, then, I don't think that something necessarily as politically governed as GBIF will ever be the source or serious technical innovation and should not even try to do so. Following the pattern of IPCC may be a better way to organize its goals.

I couldn't follow the discussion of gap analysis (Section VIII): a lot more clarity is required on this point. The last point is related to a nagging worry that I have: given resource constraints (and I am thinking more about trained personnel than money) it is clear, as you realize, that not all species can or should be equally covered. What rules do we use to prioritize species? (You simply suggest birds and other vertebrates, etc. Why? What's the rationale other than historical bias.) Perhaps this is where there should be feedback from users.

A. Townsend Peterson
University Distinguished Professor
Biodiversity Institute
The University of Kansas
town@ku.edu

Dear Townsend,

Thank you for inviting me to give my opinion about, and express my thoughts on, the Global Biodiversity Information Facility. First of all, let me tell you that I have been a stout supporter of the GBIF idea right from its beginnings. The idea of free and open access to biodiversity data online has been extremely appealing for me.

However, confronted with the question you asked me, namely, why have I, or have not, incorporated GBIF's data infrastructure into my own work and thinking, I realized that I have used the GBIF data facility really very little.

Before analyzing the reasons, let me analyze GBIF's mission. According to GBIF's own website, the institution is an international government-initiated and funded initiative focused on making biodiversity data available to all and anyone, for scientific research, conservation and sustainable development. The GBIF model hinges around three core services and products:

1. An information infrastructure – an Internet-based index of a globally distributed network of interoperable databases that contain primary biodiversity data – information on museum specimens, field observations of plants and animals in nature, and results from experiments – so that data holders across the world can access and share them
2. Community-developed tools, standards and protocols – the tools data providers need to format and share their data
3. Capacity-building – the training, access to international experts and mentoring programs that national and regional institutions need to become part of a decentralised network of biodiversity information facilities.

Stated like that, it seems great. However, most of my colleagues do not use it, and I certainly do not either. Why? The reason seems to be that, at the scale many ecologists and biogeographers like me work, the GBIF data facility is simply not sufficiently accurate. Let me give a few examples, which I developed after doing a couple of "quick and dirty" tests to see how well the GBIF data portal could work to help me with my own research. IN order to do this, I chose off the top of my head three species I am very familiar with, and that are, geographically speaking, very common and should be well represented in the collections.

1. I first checked the distribution of the North American creosote bush *Larrea tridentata* Coville, and much to my dismay found that the database contains a series of records in South America. After searching in detail, I found that the GBIF the database is considering the creosote bush as a synonym of the South American "jarilla" *Larrea divaricata* Cavanilles. And indeed, one author (R. Felger) believes that the two plants belong to the same species, but apart from him, massively the literature recognizes the two as distinct species. The database does not warn you

about synonyms, and it is up to the user to find them in the digitized herbarium labels. A trained taxonomist can figure these things out with some work, but the GBIF database is aimed to a broad audience, such as conservationists, students, or simply persons interested in sustainable development. How can they figure this out?

2. I then tried the "algarrobo blanco" *Prosopis chilensis* Stuntz, a mesquite species strictly endemic to South America. Again, it was distressing to see that many specimens were recorded in North America, all of them under cultivation (including a specimen collected at Walt Disney World nurseries in Florida!). However, as with the creosote bush, it takes going to the herbarium labels (that are recorded in the database) to get the true information and figure out what is going on. The question is the same as with the example above: an ecologist can probably figure this out with some work, but not the broader audience to which GBIF is aimed.
3. I then tried the Monk Parakeet *Myiopsitta monacha*, a very common South American species that is frequently seen as an urban invader in the Northern hemisphere. I found some 76 records in the GBIF database, all of them in European cities! In this case, the collection labels of the European Environmental Agency do not even indicate that it is a recently introduced species that survives as a peri-urban invasive in Europe. A student working with the data could get the impression that it is, indeed, a native European bird, and would never be able to figure out the species true origin and distribution.

In short, what I mean with all this is that

1. there is a certain level of taxonomic confusion in the GBIF database that forces the user to be really cautious about the way the data may be interpreted,
2. it is often not easy to separate records of species growing in the wild from specimens obtained under artificial conditions such as nurseries, urban environments, or under cultivation, and
3. some areas are very poorly covered while others seem to be covered quite exhaustively.

I also have a comment on the niche modeling algorithm. I tried to analyze the distribution of the rigid spineflower (*Chorizanthe rigida* Torr. & Gray), a little desert plant in the Polygonaceae (Buckwheat Family) that grows in the northern Sonoran Desert, in the Mojave, in the southern part of the Great Basin, and in the driest parts of the California Floristic Province. The Polygonaceae are to this region what the Proteas are to the Cape Region of South Africa, they have radiated into myriad species that are endemic to these drylands. So its distribution is limited by climatic and edaphic factors, but also and very importantly by the evolutionary history of the region.

However, if you run a niche model analysis for the rigid spineflower using all the existing predictor variables, the model will predict a suitable habitat in almost any desert in the world. And indeed, this may be true from an almost agronomical perspective, but suitable habitat is quite different from actual distribution that defined by is the intersection of all the suitable habitats and the evolutionary / biogeographical history of the lineage.

Any researcher knows this, of course, but the GBIF algorithm does not explain it in any detail. What the page says is that the Generated Niche Model generates a probability distribution using the Envelope Score Algorithm. But what is that probability? Is it is the estimated probability that if I plant the plant it will actually grow? If this is the case, the probability is possibly quite accurate. Or, alternatively, is it the estimated probability of finding the plant there? If any user interprets "probability" as the latter, they

can make a very serious overestimation of the species' true distribution. Unfortunately, the difference between suitable habitat and true distribution (similar to the difference between potential and realized niche in Hutchinson's 1950's lingo) may be quite clear for a trained ecologist, but a non-expert might end looking for *Chorizanthe* in the Namibian Desert, where, according to the niche model, the probability is very, very high!

At the same time, there are myriad local distributed databases that are growing like mushrooms in many parts of the world. A favorite of my students is the SEINET portal, that connects you to all the herbaria in the American Southwest. It is very accurate, and works great! Other local online databases such as Baja Flora (SDNHM) or CalFlora are also excellent and are very much used by my students.

In conclusion, this is what I believe as a user of online biological databases:

1. The GBIF data system has a lot of potential, but is still not there in terms of completeness and accuracy for many research problems. There is obviously lack of expertise and field knowledge in the management of species names and distributions, and many local databases are yet to be incorporated.
2. Biodiversity databases are growing exponentially in many regions, and distributed use is also growing, but the GBIF has not been there at the time of inception to interact with these initiatives in a collaborative manner. SEINET, for example, is one of the best distributed databases I know, because it is a grassroots initiative, well rooted in the herbaria and the botanical community, and very well curated. A sort of "mom and pop" initiative that has grown away from, and ignored by, many of the larger biodiversity databasing initiatives. IF GBIF's explicit goal is "to promote community-developed tools, standards and protocols," then shouldn't GBIF be interacting with these rich local initiatives?
3. Finally, and related to the previous point, GBIF has as one its main goals helping national and regional institutions to become part of a decentralised network of biodiversity information facilities. There is a lot of room for improvement here, I believe. Many of the errors I detected with the few species I checked would not be there if regional collections and local leadership were more involved

I hope my comments may be of some use for you. Don't take them as overly critical, I am still a very staunch supporter of the GBIF model, but I also believe that a lot can be donde to improve the way the model operates.

Best wishes,

Exequiel Ezcurra