GLOBAL BIODIVERSITY INFORMATION FACILITY

Recommendations of the GBIF Observational Data Task Group

September 2008

Suggested citation:

Kelling, Steve; Baban Ingole, Brenda Daly, Bruce Stein, Denis Lepage, Eamonn O'Tuama, Jerry Cooper, Matthew Jones, Tapani Lahti, and Vishwas Chavan (2008). **Recommendations of the GBIF Observational Data Task Group**. September 2008, Pp. 21.

GBIF Observational Data Task Group members at its face-to-face meeting during 16-18 June 2008 at Copenhagen. Sitting (L to R): Bruce Stein, Steve Kelling. Standing (L to R): Tapani Lahti, Vishwas Chavan, Matthew Jones, Denis Lepage, Brenda Daly, Baban Ingole, Éamonn Ó Tuama. Absent: Jerry Cooper who could not attend the meeting.

For comments/suggestions contact –

Vishwas Chavan
Senior Program Officer for Digitisation and Mobilisation of Primary Biodiversity Data
Global Biodiversity Information Facility
Universitetsparken 15, Copenhagen 2100
DENMARK

Tel. +45 35 32 14 75
E-mail: vchavan@gbif.org

Cover page credit: Ciprian Marius Vizitiu, GBIF Secretariat, Copenhagen, Denmark

CONTENT

1. Background

The Global Biodiversity Information Facility (GBIF) recognises the need for Primary Biodiversity Data[1] and Information to extend beyond natural history collections. To this end, GBIF convened an Observational Data Task Group (ODTG) to develop a strategy to expand the types of species' occurrence (observation) data it can make available. In particular, the charge of this working group was to recommend the requirements needed to improve the existing GBIF infrastructure to facilitate the discovery, access, and analysis of observations of the occurrence of species. Specifically, the questions posed to the ODTG included:

1. What constitutes an observation of species' occurrence?
2. What additions are required to GBIF's data sharing infrastructure to more accurately represent species' occurrence data?
3. How could GBIF improve its function as a resource discovery service for species' occurrence data?
4. Who are the potential publishers[2] of species' occurrence data and how can they be encouraged to contribute to GBIF?

2. Objectives

The mission of the Global Biodiversity Information Facility (GBIF) is to facilitate free and open access to biodiversity data worldwide via the Internet. To best serve this mission a comprehensive resource discovery service must be implemented for all organism observation data. This service must go beyond indexing distributed databases via an interchange schema into a central data cache[3], but include the following:

[1] Primary Biodiversity Data: Definition

- Digital text or multimedia data record detailing facts about the instance of occurrence of an organism, i.e. on the What, Where, When, How and By Whom of the occurrence and the recordings (as per GBIF Work Programme 2009 – 2010)
- All observational data including multimedia detailing facts about the instance of occurrence of an organism including WHO, WHAT, WHERE, WHEN, and HOW an observation was gathered (as defined by the Observational Data Task Group)

[2] Publishers: Throughout this report term "Publishers" or "Data Publishers" has been used instead of "Data Providers" as being used in previous GBIF reports and communications. GBIF facilitate discovery, and access to data.

[3] Central Data Cache: Data are cached through DiGIR, TAPIR or BioCASE provider tools using one of two data exchange schemas – Darwin Core (DwC), and Access to Biological Collections Data (ABCD). Then the data are indexed using a GBIF defined data model that is primarily in DwC, but includes specific ABCD elements.

1. GBIF must provide the functionality for resource discovery that enhances access and scientific analysis. This requires metadata that explicitly describes the datasets and allows users to discover what data resources exist, how to access those resources, and how to properly use those resources in analysis.
2. The data within each dataset must be organized into an interoperable format. This requires a data federation architecture that is sufficiently comprehensive for all observation data and includes all required information needed to serve overall analysis goals.
3. GBIF would like to increase participation, not only through increasing the volume of federated data, but also by increasing the actual use of the data. This requires enhancing the data processing, analysis, and visualization tools, which can be interactively run, freely shared, combined in unique ways, and incorporated in publications.

3. Participants and Affiliations

Task Group was chaired by Steve Kelling, Director, Information Science, Cornell Laboratory of Ornithology, Cornell University, USA.

Members of the group include;

- Baban Ingole, Scientist, National Institute of Oceanography, Goa, INDIA..
- Matthew Jones, Lead Development Engineer, National Centre for Ecological Analysis and Synthesis, UC Santa Barbara, USA.
- Brenda Daly, Endangered Wildlife Trust, SOUTH AFRICA.
- Tapani Lahti, IT Architect, Finnish Museum of Natural History, Helsinki, FINLAND.
- Denis Lepage, Senior Scientist, Bird Studies Canada, Ontario, CANADA.
- Bruce Stein, Vice President and Chief Scientist, Nature Serve, Arlington, Virginia, USA (current address: National Wildlife Federation, Washington, DC)
- Jerry Cooper, Science Leader, Informatics, Landcare Research, NEW ZEALAND (did not attend meeting but participated in conference calls and commented on drafts).

- Éamonn Ó Tuama, Senior Program Officer for IDA[4], Global Biodiversity Information Facility, Copenhagen, DENMARK.
- Vishwas Chavan, Senior Program Officer for DIGIT[5], Global Biodiversity Information Facility, Copenhagen, DENMARK.

Vishwas Chavan, Senior Program Officer for DIGIT was the GBIF Secretariat coordinating officer for the Task Group. The Task Group was constituted on January 25, 2008.

4. Modus Operandi

Given the geographical spread of the Task Group **members**, most of the business was conducted through email mailing list, wiki and teleconferences. Email discussions are archived in GBIF Secretariat's LiveLink system. The ODTG wiki can be accessed at http://wiki.gbif.org/gbif/wikka.php?wakka=ObservationalDataTaskGroup. Three tele-conferences were held on 6th February 2008, 22nd April 2008, and 27th May 2008. In addition to these, ODTG also carried out a SurveyMonkey survey of potential observational data providers. The survey was commissioned in early May 2008 and concluded on 2nd June 2008. Results of the survey are detailed in *Annexure 1*.

[4] IDA – Inventory, Discovery and Access is one of the work areas of GBIF Informatics thematic area.
[5] DIGIT: Digitisation and mobilisation of primary biodiversity data is one of the work areas of GBIF Informatics thematic area.

5. Outcomes

A. What is a primary species occurrence record?

Observations can be gathered through a myriad of mechanisms — as diverse as measuring the physical processes of the biosphere, recording birds visiting a backyard feeder, and differentiating genotypes of virus strains — and describe some entity (e.g., presence of a species), a trait of that entity (e.g., age), or a process involving the entity (e.g., behaviour). For all their diversity, observations share common thematic components, laying the foundation for organization of massive datasets. GBIF indexes observation data of organisms from an ever growing multitude of sources through the GBIF Participants[6] into a central data cache. Data are cached through DiGIR, TAPIR or BioCASE provider tools using one of two data exchange schemas – Darwin Core (DwC), and Access to Biological Collections Data (ABCD). Data are then indexed using a GBIF defined data model that is primarily in DwC, but includes specific ABCD elements. Currently, about 80% of data publishers contributing to the GBIF network use Darwin Core, with the rest being provided in ABCD format. GBIF facilitate access to almost 145 million observations using these schemas, but new challenges arise as new, increasingly diverse data sources are added.

The intent of the ODTG is to provide recommendations to GBIF on how to properly manage and organize increasingly diverse species' observation data. Observations are documented in various ways, with physical collections of the organisms representing just one type of documentation. Because different forms of documentation (e.g., collections, physical evidence, photos, recordings, sight observations) can have varying degrees of confidence associated with them, special consideration must be given to metadata enabling users to evaluate the quality. (It should be noted, however, that although collections can be re-examined and identifications verified, data quality is also an issue for collections-based observations). Nonetheless, most of the new observation data being

[6] GBIF Participants: This term is used in this report to represent wider community of GBIF national and thematic functionaries including NODES, data publishers (formerly known as data providers) as well users who often provide feedback on the quality, and fitness-for-use of GBIF mobilised data.

provided to GBIF is coming from major projects where much care was used in assuring quality in the data collection process.

The ODTG must have a clear understanding of how GBIF defines the primary occurrence of an organism. GBIF defines primary occurrence data as the presence of an organism at a particular place and a particular time. The data model employed by GBIF does a good job in handling such presence-only data associated with serendipitous observations such as those associated with museum specimen collections. But as more and varied observations of species occurrence that fall outside the realm of museum specimen collections are added to the GBIF data model, improvements are needed to express the primary occurrence of an organism. This is because much of GBIF's new data come from protocol-driven projects where not only the occurrence, but also the data collection process, must be described.

The addition of information on the data collection process has the potential to substantially increase the types and quality of analyses that can be conducted. For instance, more accurate predictions of species occurrence can be made by providing the context in which the observations were made, including information on how data were collected, sampling design, and potentially numbers of individuals observed. Finally, data from even more rigorously designed hypothesis-driven experiments, or even simulations, which have the potential to be considered observations, may become part of GBIF. The result is that more information must be provided through descriptive metadata (see section B), and the GBIF data model should be expanded (see section C).

B. How can GBIF increase its resource discovery functionality to enhance access and scientific analysis?

The current GBIF informatics infrastructure is comprised of a distributed network of data publishers who are linked to a central data caching system. The data publishers provide data in Darwin Core or Access to Biological collections data schemas. Data are harvested using DiGIR, TAPIR, or BioCase and are cached in the GBIF defined data model. This current GBIF infrastructure is strongly tied to organizing primary species' occurrence records) and making these records directly available.

To improve GBIF's resource discovery functionality GBIF should adopt a more comprehensive infrastructure that not only organises individual records, but also provides better resource discovery mechanisms. This requires not only organising the data but also documenting the projects that collect the data.

To properly use biodiversity datasets, users must first know what is available, and how they can be accessed. Metadata describes the projects and collections that contribute data and in so doing provide information about what is available and how it can be properly used in analysis and visualisations. Descriptive metadata provide information on the provenance, project description, protocols used to collect the data, spatial attributes, identification, quality, spatial context, data structure, taxonomy, distribution of collections, and other project features using a common framework that prevents loss of the original meaning and value of the dataset, and increases use. Thus, without this descriptive metadata, discovering that a resource exists, what data were collected, and how to properly use the data in analysis would not be possible.

IDA Program Officer presented to the ODTG a vision for enhanced descriptive metadata for datasets delivered via the GBIF network. Specifically, GBIF intends to provide through its data portal a search and browse interface to a metadata catalogue populated with enhanced descriptive metadata (potentially based on Ecological Metadata Language) that accurately identifies the datasets available through GBIF.

The ODTG concurs that GBIF needs better tools for dataset discovery, and that the addition to the GBIF portal of a descriptive metadata catalogue with intuitive search and browse functions would enhance use of GBIF resources.

The ODTG concurs that improving GBIF's descriptive metadata is essential.

Specific ODTG Recommendations for descriptive metadata for datasets:

1. GBIF is encouraged to be proactive in the improvement of descriptive metadata for datasets and the development of mechanisms facilitating metadata harvesting and data discovery as well as being a clearinghouse for data access.

2. Spatial attributes of the dataset are essential in the proper use of the data. Attention should be paid to the spatial context of how data were gathered which includes: limits of geographic coverage, and a description of how the spatial data are stored.
3. A description of the taxonomic system used, with references to methods employed for identification (this might also be necessary at the record level).
4. As the volume of datasets available through GBIF increases, the protocols used for data collection will be varied. This requires that for each dataset, details describing sampling protocol, process of vetting, basis of record, documentation are essential. This information should be part of the minimum requirements for the descriptive metadata.
5. Proper attribution for the dataset contributor that includes branding and data use agreements should be part of the minimum requirements for the descriptive metadata.

C. What additional fields are required for GBIF data federation to be sufficiently comprehensive for all observation data?

The GBIF 'Global Data Portal[7]' maintains a central indexed cache of taxon occurrence data (primary occurrence data) and names served by GBIF Participants who make their data available. As previously stated, the GBIF data cache uses its own data model, and the data attributes follow, in a large part, the DwC core elements, their equivalents in ABCD, and some specific ABCD elements (e.g., multiple image URLs). This creates a data model that serves as a "common denominator" between the schemas and focusing on elements that are critical to ensuring GBIF can successfully provide access to primary occurrence data. The number of elements stored in the central index has to be limited for performance reasons, which is why it would be difficult to index full ABCD records.

A dichotomy exists for comprehensive data access initiatives between data caching and data discovery services. Two general approaches are used to

[7] Global Data Portal: Global Data Portal refers to http://data.gbif.org/. This to prevent any confusion with several other national, regional and thematic portals hosted and maintained by the GBIF Participants.

provide access to biodiversity and other ecological databases. The first is to provide sufficient metadata documentation to allow an analyst to understand the structure and interpret the contents of disparate datasets and in so doing provide the opportunity to download and construct customised datasets for specific analyses. The second mechanism is to make data interoperable through exchange schemas, which transform disparately structured source data into a standardised target schema. The drawbacks to these approaches are that either the analyst must generate the dataset or the federation schemas almost always lose some domain specific content.

The primary goal of GBIF is to provide the free and open access to biodiversity data. GBIF makes disparate data interoperable through its data model, and makes this openly available through an Internet portal. To successfully meet this goal GBIF must provide the ability for users to discover data, assess the fitness of the data, and retrieve the data in a format appropriate for the identified purpose. This must be accomplished through an interface that limits the need for a user to understand specific data format protocols for individual datasets, but instead, provides seamless discovery and access.

GBIF is bringing together observational data on species' occurrences from a global network of contributors. . The challenge in federating data from disparate projects is that data collection techniques are varied, the data are widely dispersed, and data formats and organization are varied. While the Darwin Core has been successful in organizing museum specimen records, additional fields must be added to enable GBIF to describe characteristics of the data-collection process and other features not found in data from museum specimens. For example, sufficient variables must be included in a data schema to identify data-gathering protocol(s), to incorporate both presence and absence data, to deal with multiple organisms observed during single data-collecting events, and other features. The ODTG considers that there are two options open to GBIF to enrich the index: 1) developing extensions to DwC or 2) specifying a richer subset of ABCD.

The ODTG feels strongly that GBIF needs to make a clear statement about how it intends to proceed in its data clearinghouse mechanisms. Should it focus on descriptive metadata for datasets and provide a data organizational strategy that is sufficient to provide simple species' occurrence records, or should it strive to improve the GBIF data model to provide more robust species' occurrence information? It is recognised by the ODTG that additions to GBIF's version of Darwin Core should be fairly modest. But, GBIF should increase its usability by adding fields relevant to observational data in its data model. Additionally, the ODTG recognises that efforts are underway to develop new unified models for observation data.

One specific challenge with observation data is how to capture the absence of organisms as well as their presence. In order to properly understand the spatiotemporal patterns of occurrence of organisms, it is critical to be able to separate the lack of data from absence data (that is, the fact that an organism was *not* observed, despite a sampling effort). Although one can almost never be entirely confident about the true absence of an organism, knowing that it was not detected under a set of specific conditions provides critical information that allows the modelling of variations in occurrence or in abundance across time and/or space. From a data modelling perspective, cataloguing all absence data as individual records in a flat structure like Darwin Core is almost always impractical, at best of times. The number of species not observed during a specific observation event is typically much larger than the number of species detected, therefore substantially inflating the number of records by as much as several orders of magnitude. A better solution is to have in place data fields that allow identifying unique sampling events from which absence can be inferred, in addition to information on sampling design that can affect detection probability (e.g., sampling effort and protocol).

Specific recommendations of the ODTG to improve observation-level data content:

6. Although the group concentrated on option 1 (developing extensions to DwC) the ODTG recommends that GBIF should explore a richer data model that includes extensions to Darwin Core and a richer subset of ABCD to be more comprehensive in describing the primary occurrence of an organism.

7. In the near term, GBIF should extend its existing Darwin Core schema and user a richer subset of ABCD to improve its ability to access biodiversity data. That being said, GBIF is not expected to organize deeply structured domain specific primary data, but 'facilitate' access to the deeply structured data. Those should be managed by associated thematic networks such as the OBIS[8] for serving marine data, Vegbank[9] for serving vegetation data, the Avian Knowledge Network[10] for serving bird data, or GISIN[11] for serving Alien Invasive Species data.
8. Additional fields are recommended for the GBIF version of Darwin Core. These focus on aspects of species' occurrence data that are general across all/most projects and include:
   a. Project Code: Allows linking of species' occurrence records to a "project" (typically recognized as set of observations gathered through a unified effort that adheres to a particular set of protocols, and covers a particular geography or time sequence. This allows the organisation of records within a spatio-temporal sampling unit, protocol, and effort.
   b. Sampling Event Identifier: Allows single observations to be grouped. The identifier must be unique within each project. A sampling event is typically defined as a series of observations made during a determined amount of time at a given location (i.e., a checklist of birds or other organisms, marine mammals counted along a transect).
   c. Survey Area Identifier: an identifier that uniquely represents a sampling area, to allow identify time-series data at individual points.
   d. Protocol Identifier: Allows the identification of the methods used to collect the species' occurrence data, using domain specific standards.
   e. Observation Count: Number of individuals detected for an individual record.
   f. All Species Reported: Identifies whether the Observation Count for a given taxon includes all individuals and taxa that have been detected within a higher taxonomic group, such as the class (eg, all Birds, all

[8] OBIS, Ocean Biogeographic Information System (http://www.iobis.org/)
[9] Vegbank (http://www.vegbank.org/)
[10] Avian Knowledge Network (http://www.avianknowledge.net)
[11] GISIN, Global Invasive Species Information Network

Fishes, or all Butterflies). "No" should be used to indicate that the sampling event should not be used to infer absence observations (the species was not reported). This approach is more heavily used in (and indeed may only be feasible) in certain groups (e.g., birds) than others.

9. GBIF should actively pursue the standardisation and ratification of versions of both the ABCD and Darwin Core schemas and the adoption of a framework to develop particular domain or thematic extensions
10. Darwin Core is currently available in many flavours, making it difficult for potential users and developers to know which version should be used. Official adoption by GBIF and proper versioning through TDWG is strongly recommended. Part of the framework for creating extensions should also address how updates to the core schema are rolled out to the extensions for instance.
11. Since it is impractical for GBIF to index all fields in a fully comprehensive and extended federation schema, then the GBIF data model should at least be able to expose the fact that richer data are available (possibly elsewhere, not served by GBIF). Those additional extensions could be describing multimedia, geographic extensions, or domain-specific ones such as Bird Monitoring (AKN), Marine Surveys (OBIS), etc.
12. In the long term, we recommend that GBIF keep abreast of developments of a unified model of observation data interoperability. GBIF representatives should participate in the Biodiversity Information Standards TDWG[12] Observation Data Task Group[13]. The intended outcome of this working group is to derive consensus on modelling strategies for achieving observational data interoperability. The task group will solicit and follow recommendations from a broad community of researchers to develop a unified model for scientific observation onto which current and future data models can be mapped.

[12] TDWG, Biodiversity Information Standards, also known as Taxonomic Database Working Group (http://www.tdwg.org)

[13] TDWG Observational Data Task Group (http://wiki.tdwg.org/Observational/)

D. How can GBIF increase participation?

The mission of GBIF is to facilitate the free and open access to biodiversity data worldwide via the Internet to underpin sustainable development and ensure biodiversity conservation. In section B of this document, the ODTG has made recommendations that will improve dataset discovery by enhancing the descriptive metadata for each project. In section C, the ODTG has made recommendations to extend the existing Darwin Core data schema used by GBIF that will improve the usefulness of the data. But increasing participation in GBIF will require more than enhanced metadata and a more comprehensive data schema. What is required are more tools to streamline data federation, and more visualization and analysis features that would entice more data holders to provide access to data.

The major recommendations of ODTG for increasing participation are to:

1. Enhance the data processing, analysis, and visualization tools, which can be interactively run, freely shared, and combined in unique ways, and incorporated in publications.
2. Improve data attribution and data sharing policies
3. Increase the number of languages in which GBIF materials are made available.
4. Become more proactive in gathering species occurrence data.
5. GBIF Participants (national, regional, thematic NODES, and data publishers) needs to be more proactive in discovery and publishing of observational data sets of all types.

In order to streamline data federation, GBIF is creating a Integrated Publishing Toolkit (IPT). At present GBIF provides tools that include:

- DiGIR data provider tools
- Data Cleaning Tools to assist in checking datasets for quality
- Data Repository Tools for hosting data, data validation, and sharing
- Access to a suite of tools developed by the community of GBIF participants.

A new Integrated Publishing Toolkit (IPT) developed with the intent of simplifying data harvesting and provide visualization tools for data providers is under development. While the ODTG did not directly address these tools, the general recommendation is that they continue to develop and streamline processes that encourage and support GBIF data providers.

Recently, GBIF released a suite of online analysis and visualization tools. These tools became available after the ODTG meeting, and were not reviewed by the committee. The GBIF 'Global Data Portal' now provides straightforward data search functionality based on species, country, occurrence, and dataset queries. Data visualisations were available via interactive maps. Data can be downloaded in a variety of formats. Finally, predictive models of species' occurrence based on a number of climate variables can be dynamically generated.

After reviewing these tools, the recommendation of the ODTG is to provide these as web services to regional nodes. Many nodes that do not have the resources to provide these services would greatly benefit from the access to interactive mapping and data that is available at the GBIF site.

Visualisation and analysis functionality will encourage participation and data mobilisation. Many ODTG participants presented web-enabled visualisations, analysis, and exploration tools and functionality for displaying species occurrence data. A general recommendation from the ODTG was to further increase these tools as they will increase participation. While GBIF has made inroads in providing more visualization, analysis and data exploration tools through its GBIF analysis portal, GBIF could also act as a clearinghouse for applications developed by GBIF Participant nodes. This could be accomplished by encouraging the GBIF user community to rely on GBIF adopted data standards. For example, facilitate development of tools that allow synthetic analysis of Darwin Core data (with or without extensions). Possible examples could be visualizing hotspots for biodiversity, niche modelling, or climate change analysis or other predictive, quantitative models of complex biological processes.

Most GBIF data contributors want the ability to track those who are using their data, how often they are used, how they are used, and will want to ensure that they are properly acknowledged. Members of the ODTG felt strongly that not having these features limited participation. Many domain specific biodiversity data publishers have developed detailed data sharing policies that could be used to develop a standard data sharing policy. For example, many potential data publishers may not be inclined to allow free and open access to their data resources directly, but may be willing to provide their data for specific analyses and visualization packages, or access to data only after specific permission was granted. Thus, the ODTG recommends that data could be provided to GBIF through a series of levels of access that could include:

- Data can be used in certain publicly available, predefined visualisations (i.e. maps and graphs), but direct access to the data is restricted.
- Datasets are available upon request from the original data publisher.
- Datasets are available for download directly via the Internet after a user agrees to the GBIF data use policy.

Many regions, very rich in biodiversity, exist in countries that do not use English as their official language. Since the GBIF website is only available in English, its usefulness is limited in these regions. Many new content management systems can now handle multiple languages. The ODTG strongly recommends that GBIF makes a special effort to begin to provide GBIF functionality (website, data provider tools, etc) in a diversity of languages. This can be accomplished through the data provider node community, where native speakers can translate materials that could then be made available.

Other clearinghouses of biodiversity data exist that are currently not GBIF partners, and potentially many more will be coming online. Often these resources provide open access to their data, but not in a standardized format. For example, the Knowledge Network for Biocomplexity[14] has access to over 1,500 studies, and has the potential of holding millions of species occurrence records. While GBIF should make an effort to encourage these groups to

[14] Knowledge Network for Biocomplexity, KNB (http://knb.ecoinformatics.org/index.jsp)

become partners, some may not be inclined to organize their data in the GBIF Darwin Core format. Consequently, in order for GBIF to access these data, a recommendation would be for GBIF to become more proactive in developing tools that harvest data from these resources in a format that could then be added to the GBIF data repository.

Specific recommendations of the ODTG to increase participation in GBIF:

13. Increase the variety of data visualization, analysis, and exploration tools. The ODTG strongly feels that providing value added functionality will increase participation in the GBIF initiative.
14. Provide GBIF data portal mapping and analysis functionality as web services that can be customised to regional data publishers. The services should allow the provider to select datasets and geospatial mapping extents for data visualisation.
15. Maintain a catalogue of data visualisation and analysis tools (i.e. graphs, histograms, charts) developed by GBIF Participants, and encourage those partners to make these resources generally available to the GBIF community through open source software development or as web services.

The final ODTG recommendations focused on addressing major impediments to data sharing (resources, cultural, etc.) to overcome challenges to participation in GBIF.

16. Improve attribution for data providers by creating data use procedures based on recommendations from data publishers, and the GBIF commissioned Data Citation Task Group[15].
17. Provide tools that allow data publishers to track usage, and data users to register applications that use those data, for example.
18. Enable data providers to specify dataset specific data use policies that can be stored as part of the metadata and easily be accessed online, particularly in conjunction with the data download functions.

[15] GBIF Data Citation Task Group (http://wiki.gbif.org/gbif/wikka.php?wakka=DataCitation).

19. Modify GBIF data access policies to include several levels of data access that range from data use only for web-based analysis and visualisations, to complete and open access.
20. GBIF must make special efforts to get data from countries that are rich in biodiversity and/or biodiversity data. Frequently, these efforts should include the translation of the GBIF website, provider tools, and other relevant materials from English.
21. Develop tools that actively index species occurrence data from data publishers who are not participants in GBIF and do not maintain their data in Darwin Core format. These tools should harvest data in their original format, and only after they are at GBIF be reformatted into Darwin Core.
22. GBIF Participants (national, regional, thematic NODES, and data publishers) needs to be more proactive in discovery and publishing of observational data sets of all types.
23. GBIF Participants needs to be proactive in rescue and hosting of orphaned or to be orphaned observational data sets.

6. Executive Summary

During the first seven years of GBIFs' existence its primary focus was to mobilise specimen based observations records through digitisation of natural history collections and to tap into the low-hanging fruits. However, 60% of the currently mobilised 145 million data records are non-specimen based observations. This clearly indicates the potential for encouraging mobilisation of a wide array of observation records. Furthermore, the GBIF Governing Board 14 endorsed the renewed target of mobilising 1 billion primary biodiversity records by December 2008. The ODTG also learned that the GBIF Work Programme 2009-2010 has set the target of discovery of datasets totalling 5 billion primary biodiversity records, and mobilisation of 2 billion records through its Participants, and non-participant networks.

Needless to say, an ambition of such magnitude can only be achieved if GBIF focuses on all types of primary biodiversity datasets in addition to specimen based ones. Therefore, commissioning this Task Group to investigate ways and means of mobilising in an exponential manner the desired quantity of “fit-for-use” observation records is a timely initiative. The Task Group with diverse global expertise debated on several aspects of encouraging digitisation of observation data and its mobilisation through the GBIF network. The ODTG realises the vast potential both within and outside the GBIF network to channel the heterogeneous and distributed observation datasets through multi-cultural data publishers and partners. The ODTG believes that, if implemented as early as possible, these 23 recommendations, as detailed in the preceding sections, will help GBIF to fulfil its aspirations of mobilising billions of primary biodiversity records in the next few years, making it a truly Global “Global Biodiversity Information Facility”!

Annexure 1

Results of the Survey of Observational Data Publishers

The major objective of this survey was to understand the extent of the universe of potentially useful, sharable observational biodiversity datasets and the data owners responsible for them.

The survey was also intended to

- discover the current barriers to the exchange/sharing these observational data sets, and
- determine the additions/modifications to GBIF's informatics infrastructure needed to facilitate this exchange/sharing through the GBIF network.

The survey also requested data publishers (providers) permission to publish the attributes and URL of any publicly available as well off-line data sets maintained by them. . This is to help GBIF develop a public directory of such observational data sets.

Survey asked 18 questions to each respondent. 146 attempts were made to undertake the survey of which only 53 completed the survey. Statistics in percentage of major questions based on 53 complete responses is detailed below.

1. Familiar with GBIF

Tell me more about GBIF	44.1%
I am familiar with GBIF	51.0%

2. Estimated total nos.of observational records you or your institutions hold in multiple datasest

Fewer than 999	7.5%
1000 – 9999	20.8%
10000 – 49999	11.3%
50000 – 99999	11.3%
100000 – 499999	18.9%
500000 or more	30.2%

3. Status of digitization and accessibility

	Fewer than 25%	26-50%	51-75%	76-100%	Unknown
Percentage of digital records	14.8%	9.3%	14.8%	57.4%	3.7%

Percentage of non-digital records	44.2%	16.3%	9.3%	18.6%	11.6%
Percentage of publicly accessible records	37.5%	12.5%	6.3%	39.6%	4.2%

4. Groups or types of organisms for which observational data records are collated in these data sets

Plants	73.6%
Insects	47.2%
Reptiles	45.3%
Amphibians	43.4%
Birds	64.2%
Mammals	49.1%
Fishes (freshwater)	39.6%
Fishes (marine)	30.2%
Other marine organisms	22.6%
Other organisms	32.1%

5. Basis of records

Organism focused field survey	71.7%
Vegetation / Community sampling	56.6%
Ecological field studies	58.5%
Site or area monitoring	60.4%
Migration studies	28.3%
Others	22.6%

6. Is the repository or data set dedicated to any particular themes (e.g. invasive species, native plants, arctic biota, etc.), to particular groups of organisms, or to particular locations

Themes	71.0%
Groups	61.3%
Locations	83.9%

7. Accurracy of attributes for data records

	None	100%	Up to 75%	Up to 50%	Up to 25%	Less than 25%	Unknown
Scientific Names	3.8%	57.7%	34.6%	0.0%	1.9%	0.0%	1.9%
Common Names	10.4%	33.3%	22.9%	4.2%	8.3%	8.3%	12.5%

Latitude – Longitude	8.0%	24.0%	38.0%	10.0%	4.0%	14.0%	2.0%
Place Name	4.1%	40.8%	32.7%	14.3%	2.0%	2.0%	4.1%

8. Do the data sets have associated metadata descriptions

Metadata for all data sets	45.1%
Metadata for few/some data sets	27.5%
No Metadata for any data set	27.5%

9. Authorisation to publish the repository description by GBIF

	YES	NO
Are you the manager of the repository described here	60.0%	40.0%
May the descriptive information for the repository be made public	76.5%	23.5%

10. Levels of access control for data sets

Used for data archive, only	15.9%
Available with prior consent, only	45.5%
Available for analysis and visualization, only	29.5%
Open access, like GBIF mobilised data	65.9%

11. Major barriers to sharing observational data sets through a GBIF type network

Data Quality: Need to improve it / Not sure of quality of data	28.9%
Funding	46.7%
Attribution / Acknowledgement/ Credits	48.9%
Concerns about misuse or other abuse of data	40.0%
Concerns about impact on sensitive resources	48.9%
IT resources (servers, internet connections, software)	31.1%
IT expertise	20.0%
Scientific expertise	11.1%
Management time	57.8%
Management of citizen scientists or other volunteers	11.1%
Authoritative determination of the species or ecosystem	8.9%
Scientific quality of the data (Data Quality)	22.2%
Other reasons	20.0%

12. Participation in GBIF promoted activities aiming towards mobilisation of observational data records

	YES	NO	Don't know
Participate in GBIF promoted observational data mobilisation exercises	56.8%	2.3%	40.9%
Participate in development /implementation if standards for the exchange of observational data records	65.9%	6.8%	27.3%
Already participate in GBIF's data mobilisation activities	53.3%	26.7%	20.0%

13. How important do you or your institution feel to share the observational data

Extremely important and essential to share all data	40.9%
Important to share data (except sensitive and economic ones)	54.5%
Not at all essential to share the data	4.5%