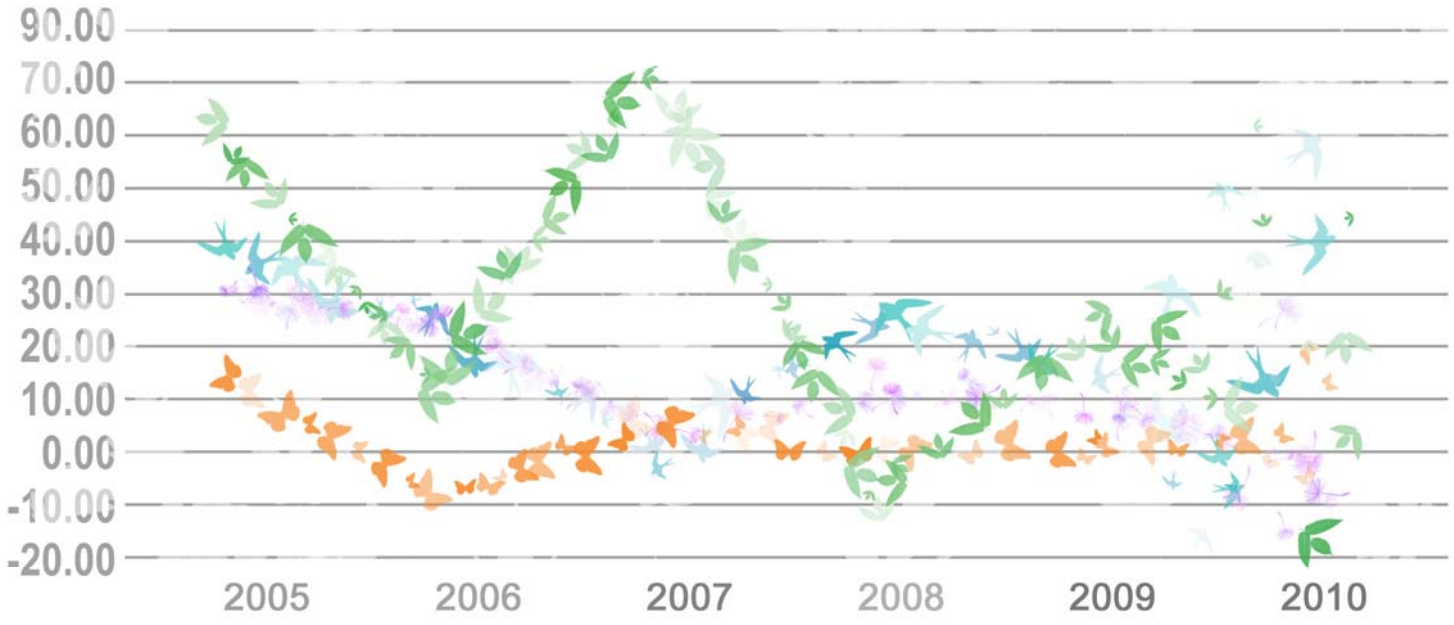
GLOBAL
BIODIVERSITY
INFORMATION
FACILITY

State-of-the Network 2010: Discovery and Publishing of Primary Biodiversity Data through the GBIF Network



October 2010

www.gbif.org

Suggested citation:

GBIF. 2010. State-of-the-Network 2010: Discovery and Publishing of the Primary Biodiversity Data through the GBIF Network. Authored by Chavan, V. S., Gaiji, S., Hahn, A., Sood, R. K., Raymond, M., and N. King. 2010. Copenhagen: Global Biodiversity Information Facility, 36 pp. ISBN: 87-92020-13-5. Accessible online at http://www.gbif.org.

Copyright 2010 © Global Biodiversity Information Facility
Cover design: GBIF Secretariat

ISBN: 87-92020-13-5

CONTENTS

1.0 Executive Summary

The Global Biodiversity Information Facility (GBIF) has a mandate to facilitate free and open access to primary biodiversity data worldwide. GBIF currently (October 2010) facilitates access to over 216 million primary biodiversity data[1] records. This report is an effort to present a comprehensive overview of (a) the data discovery and publishing potential, (b) the status of data publishing, and (c) a content assessment of accessible data, in order to assess the preparedness of the GBIF network to meet its ambitious targets. The following are the salient observations / outcomes of this exercise.

Data Publishing Potential of the GBIF network

- The GBIF network has huge potential for both discovery and publishing of primary biodiversity data. For instance, in 2009, 27 of the 95 Participants estimated their holdings of primary biodiversity data records as 2.453 billion, of which 818 million data records were reported to be digital.

Status of Data Publishing through the GBIF network

- The GBIF network currently (October 2010) facilitates access to over 216 million primary biodiversity data records contributed by 318 publishers from 43 countries covering 10371 data resources[2].
- The percentage of annual increase in number of data records, data publishers and data resources is declining with each passing year.
- Participant Nodes from developed regions are the largest contributors of the data records accessible through the GBIF network.
- The Avian Knowledge Network[3] is the largest single data publisher within the GBIF network.
- Of the 10 135 data resources registered till July 2010, 7353 resources are indexed, leaving the remaining 2782 data resources to be indexed.

[1] Primary Biodiversity Data is defined as: 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 recording.
[2] Data resources are also referred as datasets.
[3] http://www.avianknowledge.net

- The DiGIR protocol[4] is used by over 60% of data publishers.
- Data resources published by USA-based publishers receive the maximum number of search hits and return maximum of data records, download events, and downloaded records through the GBIF network.

Content Assessment of the GBIF mobilised data

- The majority of the data records accessible through the GBIF network belong to the Kingdom Animalia, most of which are bird observations.
- Over 74% (148+ million) of the data records currently accessible through the GBIF network pertain spatially to the Northern hemisphere, mainly in Europe and North America.
- Over 25% of the data records published through the GBIF network were collected during the period 2001-2010.
- Over 60% of the data records published through the GBIF network are observation-based, and a little over 25% are specimen-based.

Preparedness of the GBIF network

- Data discovery is in an extremely nascent stage across the GBIF network. Metadata catalogues are limited in scope and numbers across the network. There is a lack of national policy on metadata cataloguing across the GBIF network.
- Data publishing is opportunistic rather than deterministic and demand-supply driven in nature across the GBIF network.
- The majority of the Participants of the GBIF network have no plans for conducting Content Needs Assessment. The majority of the Participants do not consider having systematic Content Needs Assessment as an essential activity.
- Data discovery and publishing activities across the GBIF network are unplanned.
- The majority of the GBIF Participants do not consider that 'data discovery and publishing strategy and action plans' are an essential component of data publishing activities.

[4] http://digir.sourceforge.net

- Very few Participants have taken specific actions to expedite activities to meet the 2 billion target goal approved by GB15[5].
- One third of the GBIF Participants (22 country + 5 Associate organisations) estimated that 2.453 billion data records are available within their domain, of which 818 million are digital data records. However, these Participants committed to publish only 23% (195+ million) records by the end of 2010.

Overcoming current impediments

- Data discovery and publishing strategies and action plans by the Participants are essential for expediting progress in data publishing to meet growing demands for data.
- Our progress in data discovery and publishing is directly proportional to the status, mandate, capacity, vision and resources of the Participant BIFs[6].

The outcome as provided by this assessment of the network's preparedness and progress highlights the need for significant changes at the Participant level. In particular a more strategic, multi-agency approach to biodiversity data needs and investment at the national level is imperative if the individual participants and wider GBIF network is to benefit from the establishment of GBIF.

[5] http://www2.gbif.org/GB15_ExecutiveSummary.pdf
[6] A Participant BIF refers to a network of data holders, users and other stakeholders established by a GBIF Participant to promote, facilitate, and coordinate the biodiversity data sharing activities within its domain.

2.0 Introduction

Discovery and accessibility of Primary Biodiversity Data[7] (GBIF, 2008a)[8] is critical to make sound and informed decisions for sustainable use of biological resources and conservation of biodiversity at all levels, local to global. Since its inception in 2001, the Global Biodiversity Information Facility (GBIF) has made steady progress to ensure access to over 216 million primary biodiversity data records (as at October, 2010). However, the progress made by the GBIF network is far from the needs expressed by the stakeholder communities. This calls for frequent monitoring, and gap analyses, leading to development of demand-driven strategies and action plans towards expedited discovery and publishing of primary biodiversity data.

Since the launch of the GBIF Data Portal (http://data.gbif.org) in 2004, the progress by the GBIF network in data publishing is reported through Annual Reports (GBIF, 2004[9]; GBIF, 2005[10]; GBIF, 2006[11]; GBIF, 2007[12]; GBIF, 2008b[13]; GBIF, 2009[14]), the Nodes Surveys in 2007 and 2008, and the Participant Report in 2009. Additionally, gap analysis exercises are also commissioned at regular intervals either directly by the Secretariat or through commissioned studies.

This report is an effort to present a comprehensive overview of (a) the data discovery and publishing potential, (b) the status of data publishing, and (c) a content assessment of accessible data, in order to assess the preparedness of the GBIF network to meet its ambitious targets.

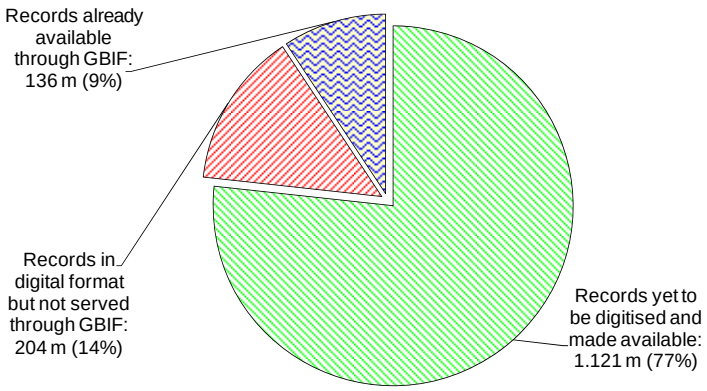
[7] Primary Biodiversity Data is defined as: Digital text or multimedia data record detailing facts about the instance of occurrence of an organism, i.e. on the what, where, when, how and by whom of the occurrence and the recording.
[8] http://www2.gbif.org/WP2009-10.pdf
[9] http://www2.gbif.org/annual_report_2004.pdf
[10] http://www2.gbif.org/annual_report_2005.pdf
[11] http://www2.gbif.org/annual_report_2006.pdf
[12] http://www2.gbif.org/annual_report_2007.pdf
[13] http://www2.gbif.org/annual_report_2008.pdf
[14] http://www.gbif.org/fileadmin/AR2009/index.html

The report is structured in five sections, viz., (a) Data publishing potential of the GBIF network, (b) Status of data publishing through the GBIF network, (c) Content assessment of the GBIF mobilised data[15], (d) preparedness of the GBIF network, and (e) lessons learnt and future plans for similar exercises.

3.0 Data Publishing Potential of the GBIF Network

Since the beginning of the data publishing activities across the GBIF network (in 2004) the approach has been largely opportunistic to tap into low hanging fruits. Since 2007, planned efforts are made to understand the data publishing potential of GBIF network through survey exercises[16].

In 2007, in response to specific questions about their data mobilisation capacity, 33 Nodes reported that they hold the potential of publishing over 1.325 billion data records (Figure A1). Of these, 204 million (14%) are already in digital format, but are not published through the GBIF network. The remaining 1121 million (77%) records are known to these Participants, but need to be digitised. At this point in time only 136 million records were accessible through the GBIF network.



[15] GBIF mobilised data: also called as 'GBIF mediated data'. Data records discovered and published through GBIF Network. In the context of this report accessible through GBIF data portal, http://data.gbif.org.
[16] Survey exercises carried out since 2007 includes Nodes Surveys in 2007 and 2008, and Participants Report in 2009 respectively.

Figure A1. Estimate of universe of primary biodiversity data by 33 Nodes (October 2007) (Source: GBIF Executive Secretary Report to GB15, October 2007).

In 2008, the Participant Nodes were further queried about their data publishing potential. 29 Participant Nodes responded to questions and identified 2.065 billion data records, of which 735 million were reported to be in digital form (Figure A2). Of these 2.065 billion records 1.6 billion records were reported to be observational in nature in comparison to 407 million specimen records. At this point in time only 147 million records were accessible through the GBIF network.

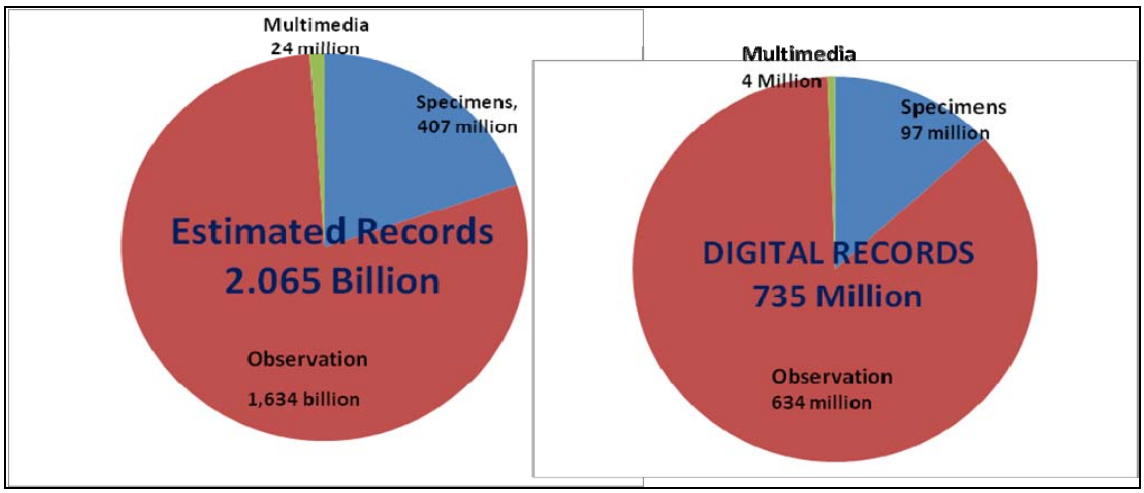


Figure A2. Estimate of universe of primary biodiversity data by 29 Nodes (October 2008) (Source: Executive Secretary Report to the GB15 (November 2008).

In 2009, only 27 Participants answered questions specific to their data publishing potential. Together, these 27 Participants (22 countries and 5 Organisations) estimated their universe of primary biodiversity records as 2.453 billion, of which 818 million data records were reported to be digital, whereas the remaining 1.635 billion data records still needed to be digitised (Figure A3).

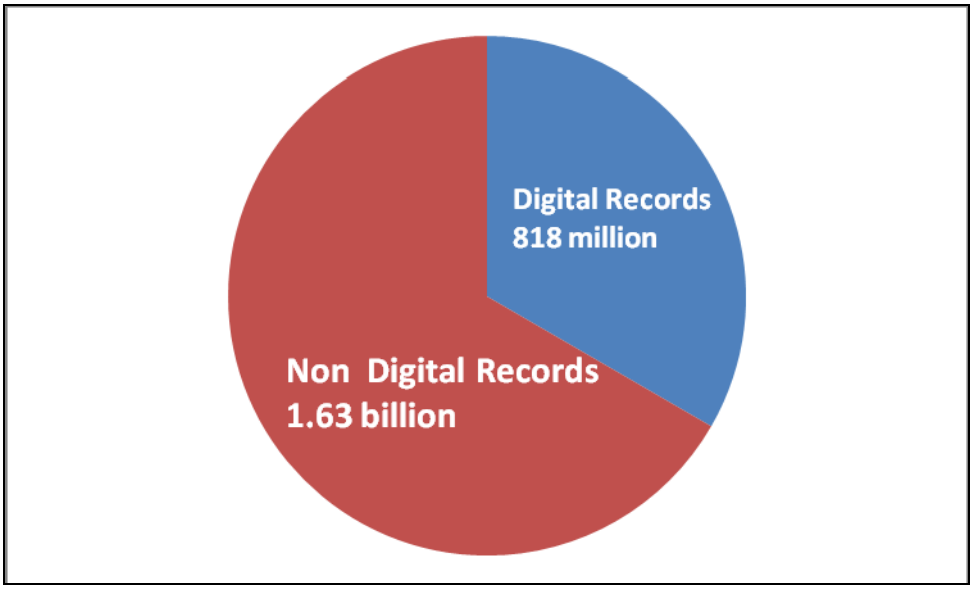


Figure A3. Participant Report 2009 estimated universe of primary biodiversity data records as 2.45 billion of which 818 million were digital records (Source: GBIF Participants Report, 2009).

The 22 country Participants[17] estimated that a total of approximately 2.428 billion primary biodiversity data records were accessible within their countries, of which nearly 33% (794+ million) were currently in digital form (Table A1). In addition to this, 5 Associate Organisation Participants estimated that approximately 25.114 million records were available within their organisations, of which 94% (23.667+ million) were already in digital form (Table A2).

Please provide the best estimate of the total amount of primary biodiversity data currently available within your country.			
Record type	**Total no. records**	**No. digital records**	**No. nondigital records**
Specimen based occurrence data	1,392,186,463	99,542,463	1,292,644,000
Observation based occurrence records	863,916,645	524,216,645	339,700,000
Multimedia data linked to primary biodiversity data	2,190,000	690,000	1,500,000
Population / ecological monitoring records	70,003,750	70,003,750	0
Impact Assessment associated data records	30,000	30,000	0
Other types of primary biodiversity data	100,005,000	100,005,000	0
TOTAL	**2,428,331,858**	**794,487,858**	**1,633,844,000**

Table A1. Estimate of available primary biodiversity data records as reported by 22 country Participants in 2009 (Source: Participants Report 2009).

[17]Country Participants responded to data publishing query in Participant Reporting System, 2009: Argentina, Australia, Austria, Belgium, Burkina Faso, Canada, Colombia, Costa Rica, Denmark, Finland, France, Ireland, Japan, Republic of Korea, The Netherlands, Norway, Peru, Poland, Solvakia, South Africa, Switzerland, and United Kingdom.

Q66: Please provide the best estimate of the total amount of primary biodiversity data currently available within your organisation as of 2009.			
Record type	**Total no. records**	**No. digital records**	**No. nondigital records**
Specimen based occurrence data	1,458,287	1,458,287	0
Observation based occurrence records	15,989,096	15,969,096	20,000
Multimedia data linked to primary biodiversity data	279,484	279,484	0
Population / ecological monitoring records	2,500	2,500	0
Impact Assessment associated data records	800	800	0
Other types of primary biodiversity data	7,384,513	5,957,708	0
TOTAL	**25,114,680**	**23,667,875**	**20,000**

Table A2. Estimate of available primary biodiversity data records as reported by 5 organisation Participants (Source: Participants Report 2009).

It is to be noted that only one-third of the total Participants responded to data publishing related queries. In 2007, only 33 of the 79 Participants responded to data publishing related questions. Whereas in 2008, 29 of 88 and in 2009, 27 of 95 Participants responded to data publishing related queries. Thus, on average only one-third of the total Participants answered the queries.

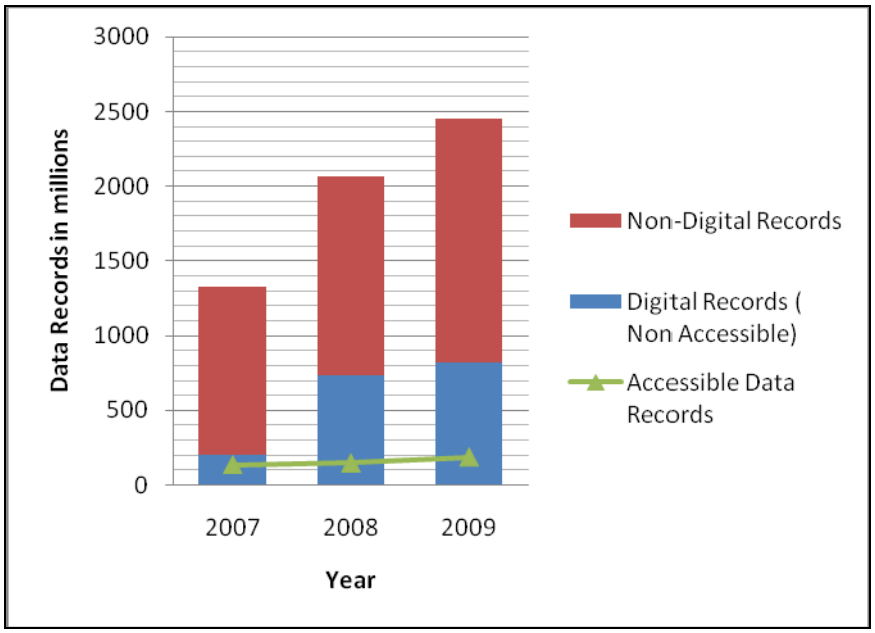


Figure A4. Trend of Data discovery v/s Data Publishing across GBIF Network (Source: GBIF Nodes Survey 2007, 2008, and Participants Reporting System, 2009)

As depicted in Figure A4, discovery of both digital and non-digital data records is rising; however, the data accessibility through the network remains to be linear. This indicates that the GBIF network has huge potential for both discovery and publishing of primary biodiversity data. This

has been the basis for the GBIF Governing Board to approve ambitious targets of discovery of upto 5 billion data records, and publishing of upto 2 billion data records by December 2010 (GBIF, 2008a)[18].

4.0 Status of Data publishing through the GBIF network

4.1 Data Publishing: Pattern of Growth

Since the launch of the GBIF Data Portal Prototype in February 2004[19] (http://data.gbif.org/), there has been a steady increase in both data resources and data records published through the GBIF Network. With a modest beginning of 46 million records in 2004, as of June 2010, the GBIF Data Portal facilitates access to over 203 million records, contributed by 309 Publishers from 43 countries through 10186 data resources (Table B1). In 2009, the number of accessible primary biodiversity records grew by 20% from 163 million to 196 million.

Year	Records (in millions)	Publishers	Country of Publishers	Data Resources
2004	46	95	30	400
2005	85	150	37	1214
2006	114	215	35	1480
2007	119	234	38	7586
2008	163	274	39	7237
2009	196	311	40	9834
June 2010	203	319	43	10186

Table B1. Increase in primary species occurrence records published through the GBIF Network since 2004 till June 2010.

From a 45.88% increase in data records in 2005 (in 2004 GBIF network published 46 million data records), the GBIF network's increase has been 16.84% in 2009, and 3.45% in 2010 (Figure B1). Similarly there is corresponding depletion in % of increase in numbers of Publishers joining GBIF network and numbers of data resources published.

[18] http://www2.gbif.org/WP2009-10.pdf
[19] http://www2.gbif.org/annual_report_2004.pdf

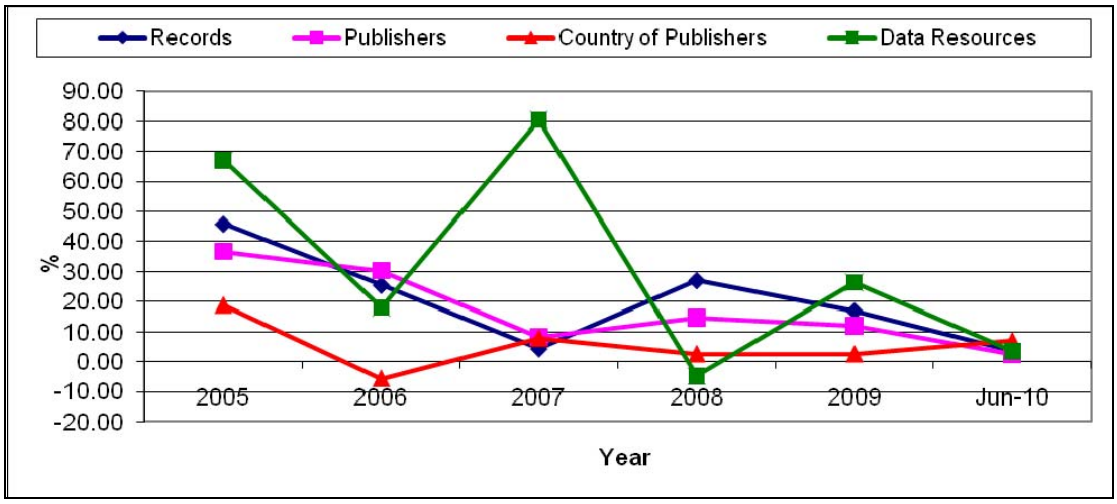


Figure B1: Percentage growth in GBIF Network mobilised data records, data resources, Publishers and Publisher countries.

Figure B2 depicts the growth pattern of primary biodiversity occurrence records captured at the time of publication of a new index database version (rollover). The percentage of georeferenced records has risen from 81.3% at the end of 2008 to 81.5% at the end of 2009.

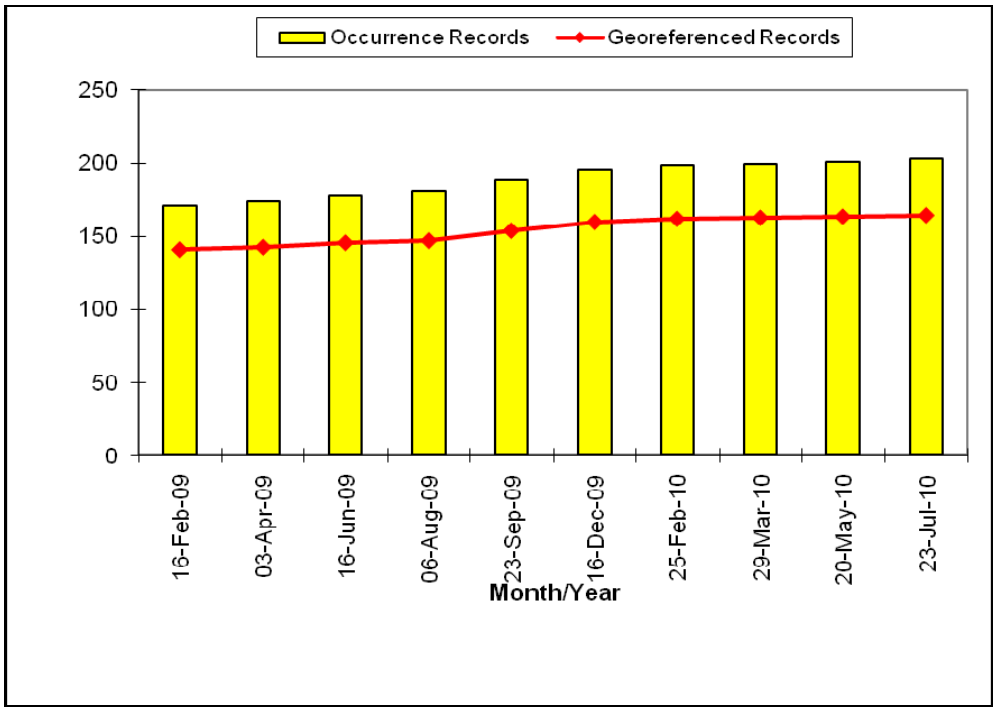


Figure B2. Pattern of primary biodiversity records (including georeferenced) published through the GBIF Network from February 2009 till June 2010.

4.2 Participant Nodes: Pattern of publishing

To date, Participant Nodes from the developed part of our globe are contributing the largest number of records through the GBIF network. Of these, the United States of America (USA) publishes over 75 million occurrence records (Table B2). Other leading data publishing Nodes include Sweden (over 24 million), United Kingdom (over 17 million), the Ocean Biogeographic Information System (OBIS; over 12 million) and France (over 10 million). Five Participant Nodes, namely Germany, The Netherlands, Denmark, Spain and Australia publish in the range of 4-8 million records, while thirteen Participant Nodes publish in the range of 1.5 to 4 million records, and 7 Participant Nodes publish in the range of 200,000 to 800,000 records (Figure B3).

Range of accessible data records (in millions)	Participant Nodes
> 50	United States of America
21-50	-
1**0**-20	Sweden, United Kingdom, OBIS, France
5-10	The Netherlands, Germany
1-5	New Zealand, Bioversity International, Republic of Korea, Canada, Poland, Mexico, EU, Japan, Finland, South Africa, Austria, Costa Rica, Norway, Australia, Spain, Denmark
< 1M	Belgium, NatureServe, Switzerland, EU-BioCASE, Iceland, Slovenia, Argentina

Table B2: Categorisation of Top 30 Participant Nodes on the basis of number of records published.

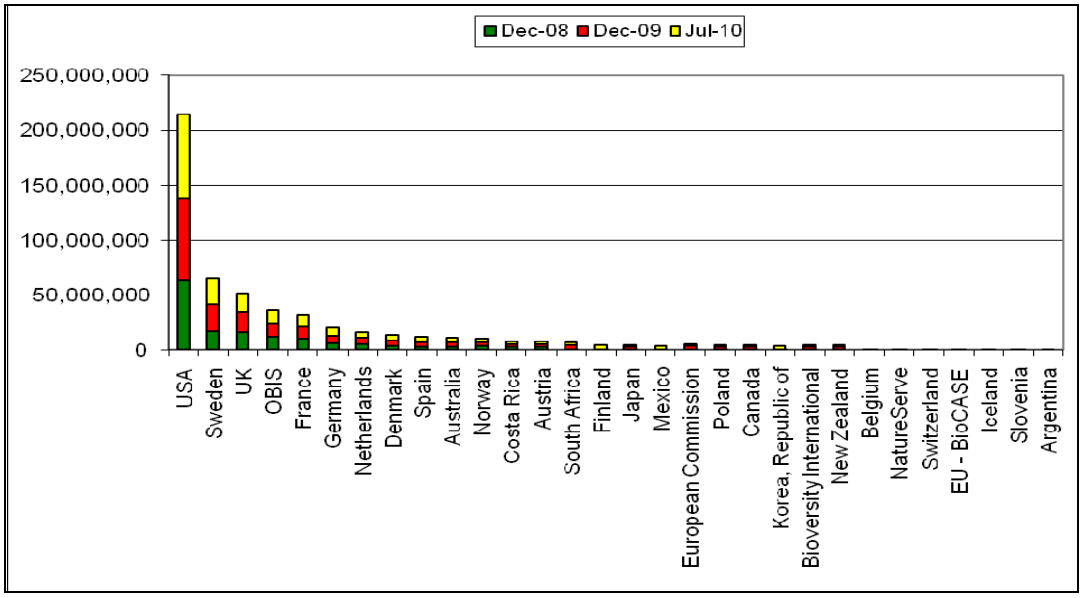


Figure B3. Top 30 Participant Nodes who publish data through GBIF network.

4.3 Data Publishers: Pattern of Publishing

Table B3 and Figure B4 depicts the 30 data publishers with the highest numbers of primary biodiversity data records published through the GBIF network. Of these, the top 3 data publishers, viz., the Avian Knowledge Network , GBIF-Sweden, and the UK National Biodiversity Network, each publish data records in the range of 16 to 44 million. The Ocean Biogeographic Information System (OBIS) publishes just over 10 million records; where as The Netherlands Biodiversity Information Facility (NLBIF), the MNHN and Danish Biodiversity Information Facility (DanBIF) publish little over 5 million records each. The remaining 23 of the 30 top data publishers publish less than 5 million data records each (Table B3).

Range of accessible data records (in millions)	Data Publishers
25-50	Avian Knowledge Newtork
10-25	Oceab Biogeographic Information System, UK National Biodiversity Network, GBIF-Sweden
5 -10	DanBIF, Service du Patr. Nat., MNHN, NLBIF
1-5	New York Botanical Garden, Field Musuem, PANGAEA, Yale University Peabody Museum, National Museum of Nature & Science, Japan; CONABIO, GBIF-New Zealand, Bioversity International; OZCAM; European Environment Agency; USDA Plants; Finnish Musuem of Natural History; Biologiezentrum Liz; KU Biodiversity Research Center; SANBI; InBIO, Costa Rica; Natural History Museum, University of Oslo; National Musuem of Natural History, Smithsonian Institution; Missouri Botanical Garden; BfN; NetPhyD; Cons. Bot. Nat.; Bassin Parisien; GBIF-Spain

Table B3. Categorisation of Top 30 Data Publishers on the basis of number of records published.

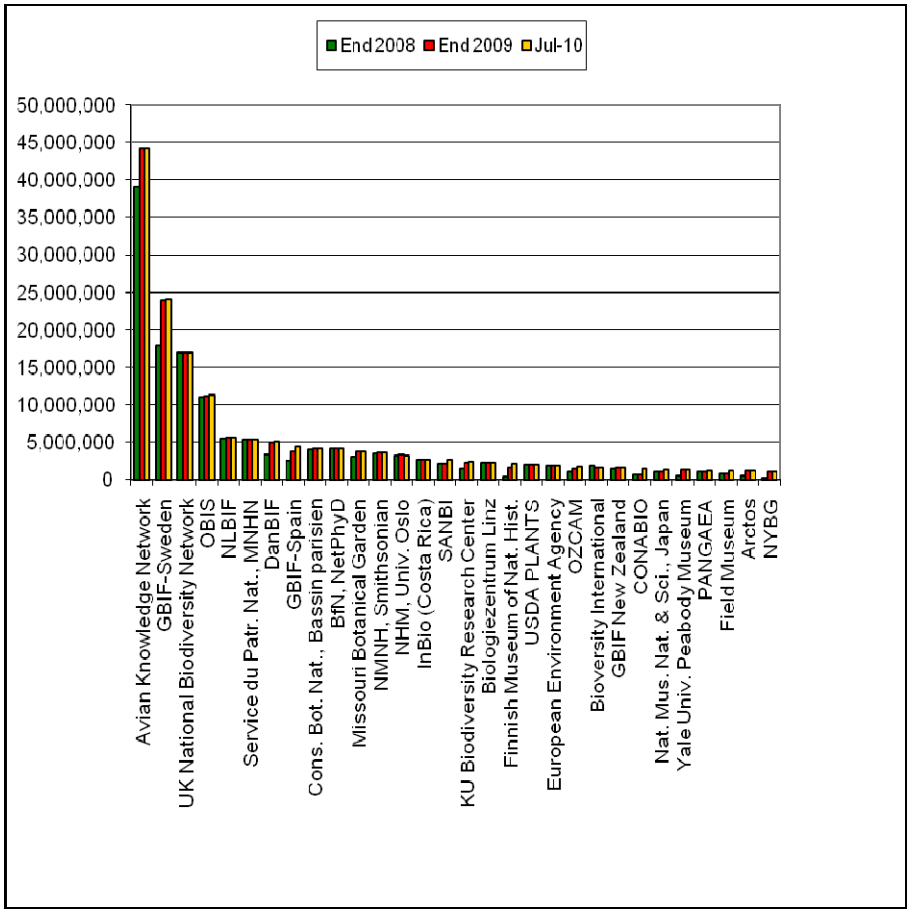
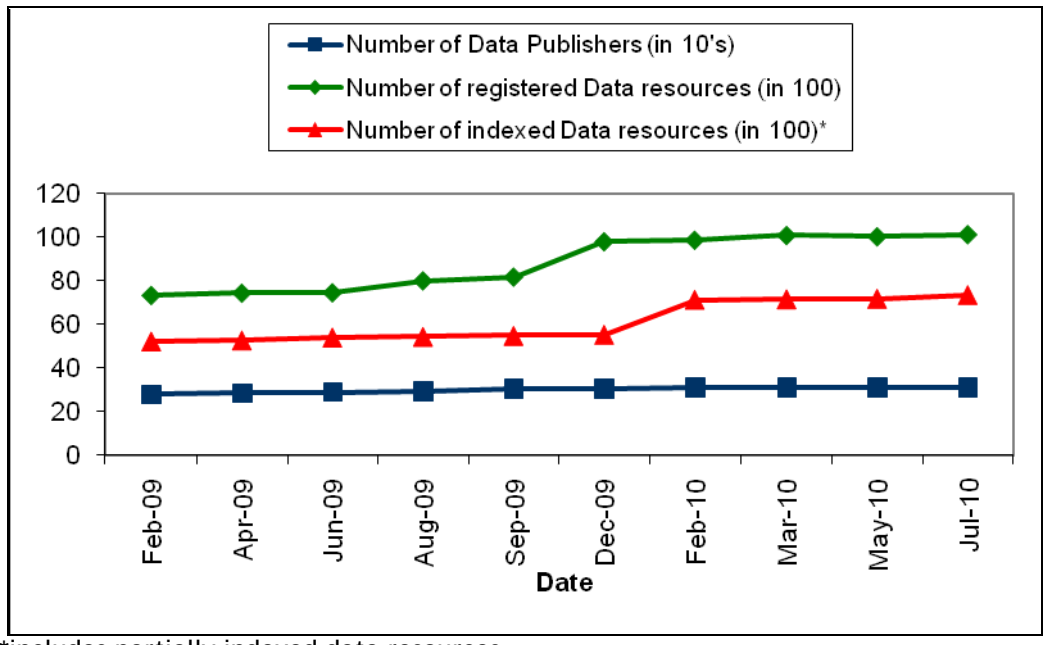


Figure B4. Top 30 Data Publishers during 2008 - June 2010.

4.4 Registered and Indexed Data Resources: Pattern of Growth

Figure B5, depicts the Data Publisher, registered and indexed data resources growth pattern at each rollover since February 2009. Numbers of Data Publishers increased by 33 from 280 in February 2009 to 313 in July 2010. During 2010, seven (7) new Data Publishers joined the publishing activities through the GBIF network (from 306 in December 2009 to 313 in July 2010).

During the same period the numbers of registered data resources increased by 2794 (from 7300 in December 2009 to 10135 in July 2010). To date, 7353 of these data resources are indexed, leaving a remaining 2782 data resources to be indexed.



*includes partially indexed data resources.

Figure B5: Growth pattern of (a) data publishers and (b) registered data resources and (c) indexed data resources at each rollover- during February 2009 until July 2010.

4.5 Access Point Protocols

To date, the DiGIR protocol (197 instances) is still in use by a large number of Data Publishers (Figure B6), followed by BioCASE (75 instances), and TAPIR (31 instances). From the representation it appears that DiGIR is the most used protocol. However, in 2004, when the GBIF network began to publish data, it was the only protocol available, explaining the high number

of DiGIR installations. In addition, the protocol had been adopted by several thematic networks early on, and is still in active use within those.

There are also data publishers who employ more than one protocol publishing data. Usually this reflects either a part-way migration from one protocol to another, or differing requirements or preferences in different parts of an institution. The most frequently found combination is that of DiGIR and TAPIR (7), closely followed by TAPIR and BioCASE (6) as well as DiGIR, TAPIR, and BioCASE (6). Four publishers employ a combination of DiGIR and BioCASE.

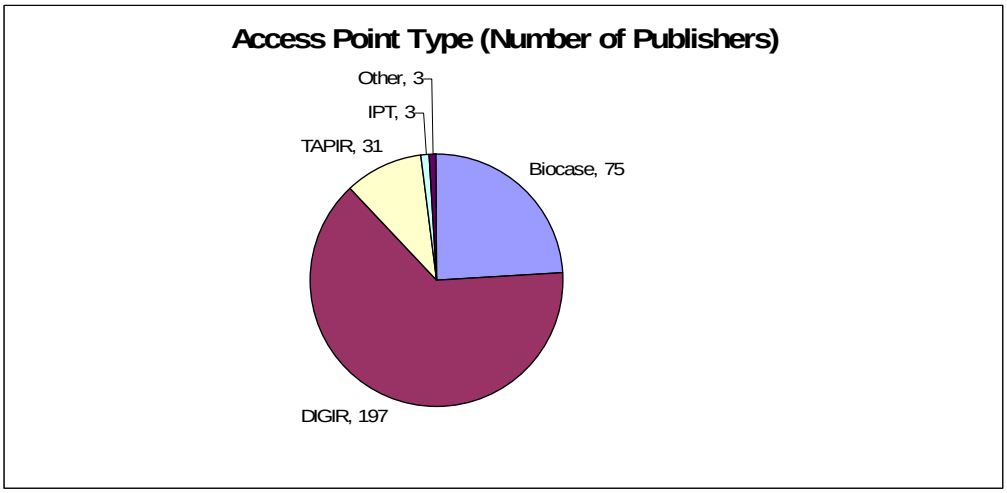


Figure B6. Access Point Protocols used by GBIF Data Publishers.

4.6 Data Usage Pattern

4.6.1 Pattern of search hits and records returned

Figure B7 shows the top 18 Participant data resources that receive the maximum number of search and return maximum of data records through the GBIF Data Portal (http://data.gbif.org). Leading amongst these are data resources published by the United States of America, followed by Germany, the Ocean Biogeographic Information System, Spain and Australia.

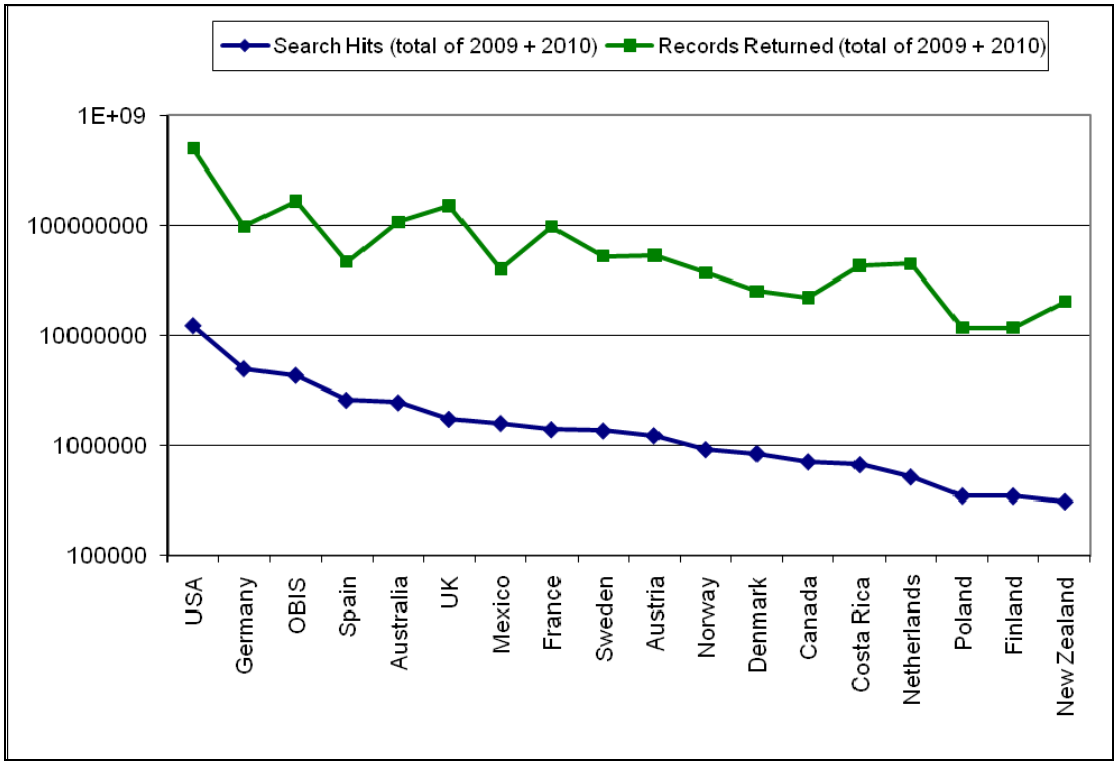


Figure B7. Top 18 Participant Nodes receiving maximum search hits and returing maximum data records from January 2009 until July 2010.

4.6.2 Pattern of download events and number of records downloaded

Figure B8, depicts the pattern of data download events and downloaded records from January 2009 to July 2010. Data resources published by the USA, Germany, Spain, Mexico, Sweden, United Kingdom, Japan, France, The Netherlands, Austria, Australia, Canada, OBIS, Norway, Poland, Denmark, Costa Rica and Belgium receive maximum number of download events and number of records downloaded in descending order.

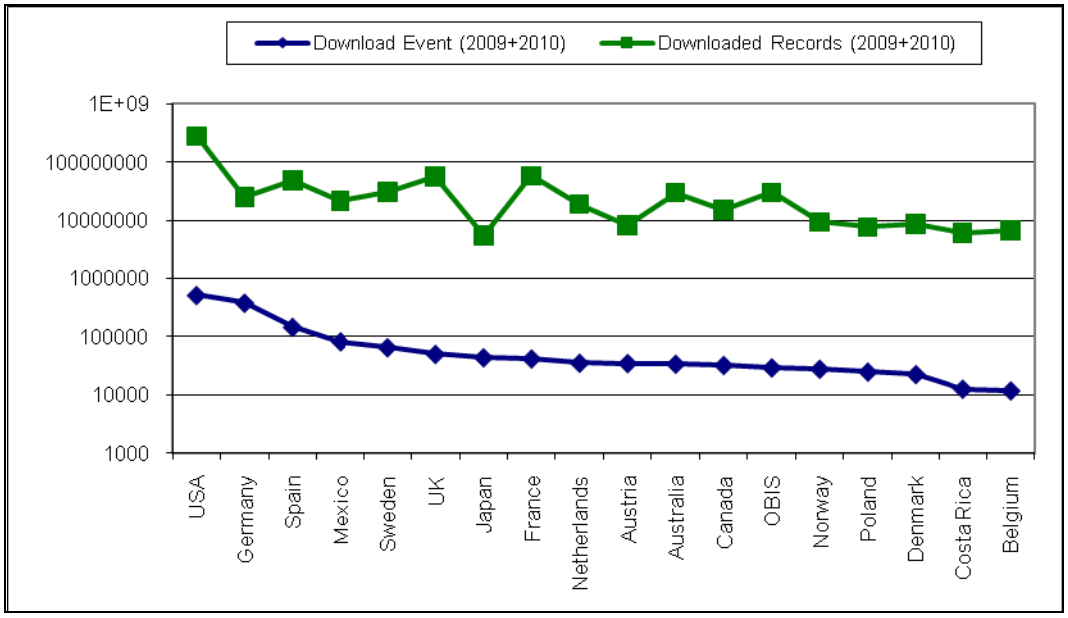


Figure B8. Participants with maximum number of download events and maximum number of records downloaded from January 2009 to July 2010.

5.0 Content Assessment of the GBIF mobilised data

For the purpose of performing a Content Assessment of GBIF mobilised data, the 201,275,468 data records accessible in June 2010 were used. The objective of this assessment was to understand the taxonomic, geographic, temporal coverage of GBIF mobilised data. It further aimed at identifying the gaps (taxonomic, geographic and temporal gaps) in currently accessible data. It was felt that interpretation of such an analysis would highlight both the potential usability of accessible data to address ecological as well as social issues. It would further help in channelling efforts towards demand-driven discovery and publishing of primary biodiversity data.

Four types of assessment, namely, (a) taxonomic, (b) geographic, (c) temporal, and (d) basis of records were carried out. The following sections highlight the salient observations and preliminary interpretations.

5.1 Taxonomic Assessment

Of the 201 million primary biodiversity data records accessible through GBIF, the majority of the data records belongs to Kingdom Animalia (64.2%), followed by Plantae (25.7%), Fungi (1.5%), Protozoa (0.8%) and bacteria (0.1%). As shown in Figure C1, 4.2% of data records are unclassified[20], where 3.2% are unknown[21].

About 85% of the records belonging to Kingdom Animalia are georeferenced, compared with 76% of the data records belonging to Kingdom Plantae (Table C1).

Kingdom	Records	Georeferenced (%)	% of total
Animalia	129,223,621	85	64.2
Archaea	906	0.4	0
Bacteria	186,348	20.8	0.1
Chromista	554,878	85.6	0.3
Fungi	3,085,909	53.8	1.5
Plantae	51,721,703	76	25.7
Protozoa	1,677,433	91.7	0.8
Unclassified...	8,414,605	72.6	4.2
Unknown	6,409,214	74.9	3.2
Viruses	851	0	0

Table C1. GBIF mobilised data by Kingdoms.

[20] Unclassified data: Data returned by Portal index is null.

[21] Unknown data: Data returned by Portal index is flagged as unknown.

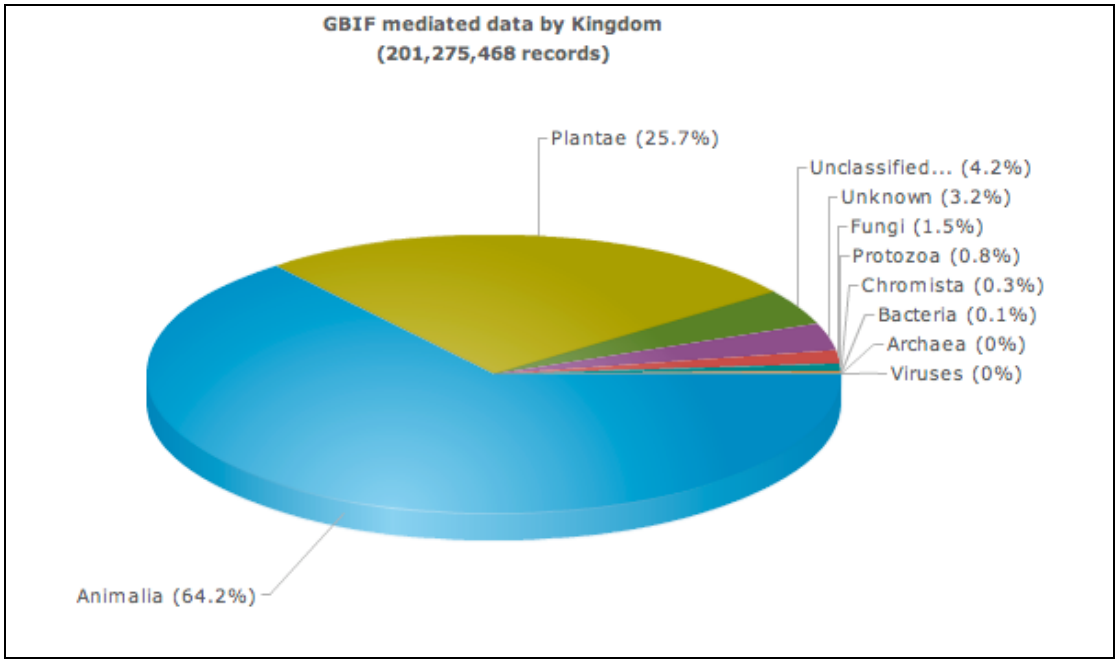


Figure C1. GBIF mobilised data by Kingdoms.

Of the 129+ million data records belonging to Kingdom Animalia, 76.8% are those belonging to Phylum Chordata followed by 17.7% belonging to Phylum Arthropoda (Figure C2). There are 158,953 records termed as 'unclassified' as they lack sufficient taxonomic information to classify them in any Phylum. The majority of the data records belonging to Kingdom Animalia are georeferenced (Table C2).

Phylum	Records ▼	Georeferenced (%)	% of total
Chordata	99,258,748	87.4	76.8
Arthropoda	22,872,946	79.5	17.7
Mollusca	3,990,870	64.1	3.1
Annelida	953,188	86.7	0.7
Cnidaria	499,697	78.2	0.4
Brachiopoda	365,106	40.3	0.3
Echinodermata	362,266	80.9	0.3
Ectoprocta	205,200	90	0.2
Unclassified	158,953	88.9	0.1
Porifera	140,337	66.7	0.1

Table C2: GBIF mobilised data records for Kingdom Animalia.

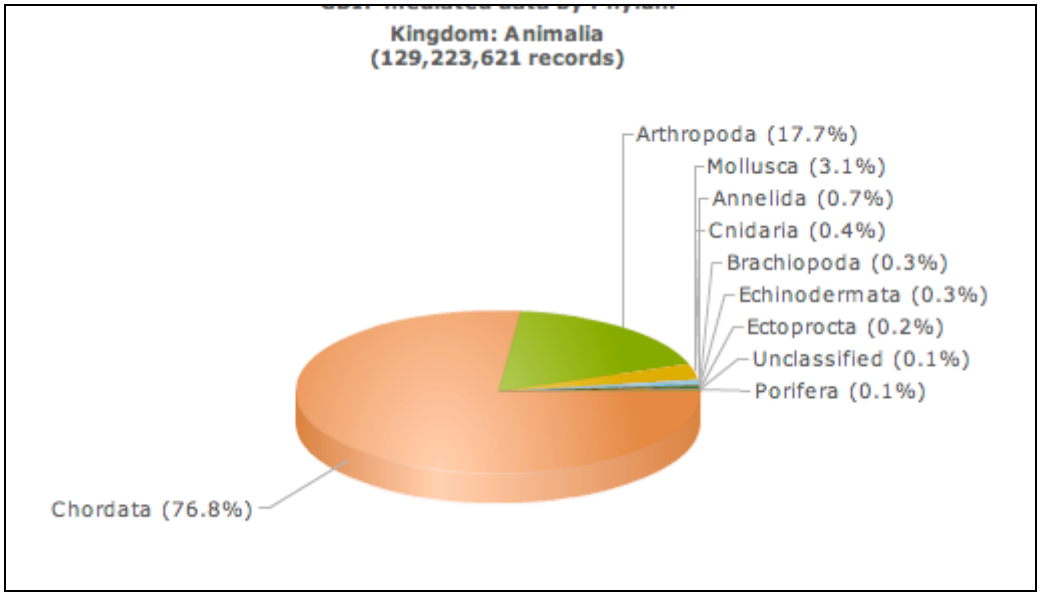


Figure C2: GBIF mobilised data for Kingdom Animalia. Phylum Chordata constitute over 76% of the Animala related primary biodiversity data records.

Of the 99+ million data records belonging to Phylum Chordata 79.1% are avian observations, followed by 9.4% belonging to Class Actinopterygii (Figure C3). Again the majority of these records belonging to Phylum Chordata are georeferenced (Table C3).

Class	Records ▼	Georeferenced (%)	% of total
Aves	78,533,672	93.1	79.1
Actinopterygii	9,356,550	79.3	9.4
Mammalia	4,774,556	63.3	4.8
Reptilia	2,898,518	40.4	2.9
Amphibia	2,728,641	40.5	2.7
Elasmobranchii	445,425	91.1	0.4
Unclassified	151,971	80.9	0.2
Chondrichthyes	94,284	98.9	0.1
Ascidiacea	61,369	89.3	0.1
Appendicularia	60,284	99.9	0.1

Table C3: GBIF mediated data records belonging to Phylum Chordata. The majority of the records are georeferenced.

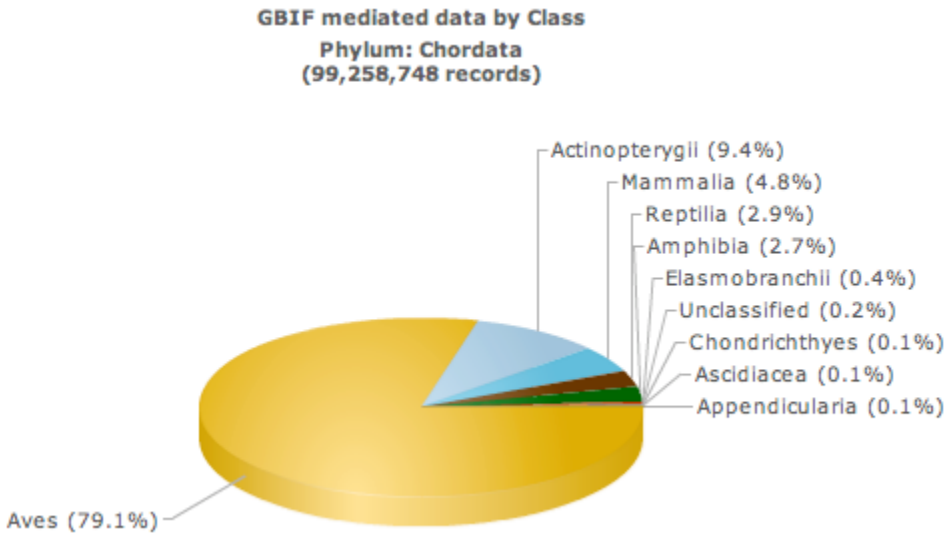


Figure C3: GBIF mediated data for Phylum Chordata. Class Aves constitute over 79% of the primary biodiversity data records belonging to Phylum Chordata.

This clearly indicates the dominance of bird observations among the data accessible through GBIF network.

Of the 51+ million data records belonging to Kingdom Plantae, data records related to flowering plants (Phylum Magnoliophyta) constitute 81.8% (Figure C4), of which 77% are georeferenced (Table C4).

Phylum	Records ▼	Georeferenced (%)	% of total
Magnoliophyta	42,296,509	77	81.8
Bryophyta	2,923,562	67.7	5.7
Pteridophyta	1,855,660	69	3.6
Unclassified	779,803	78.5	1.5
Bacillariophyta	617,566	97.5	1.2
Pinophyta	590,743	68.7	1.1
Hepatophyta	558,955	72	1.1
Anthophyta	395,807	80.3	0.8
Rhodophyta	379,257	57.1	0.7
Spermatophyta	364,713	69.8	0.7

Table C4: GBIF mobilised data records for Kingdom Plantae. Over 81% of these records belong to Phylum Magnoliophyata of which nearly 77% are georeferenced.

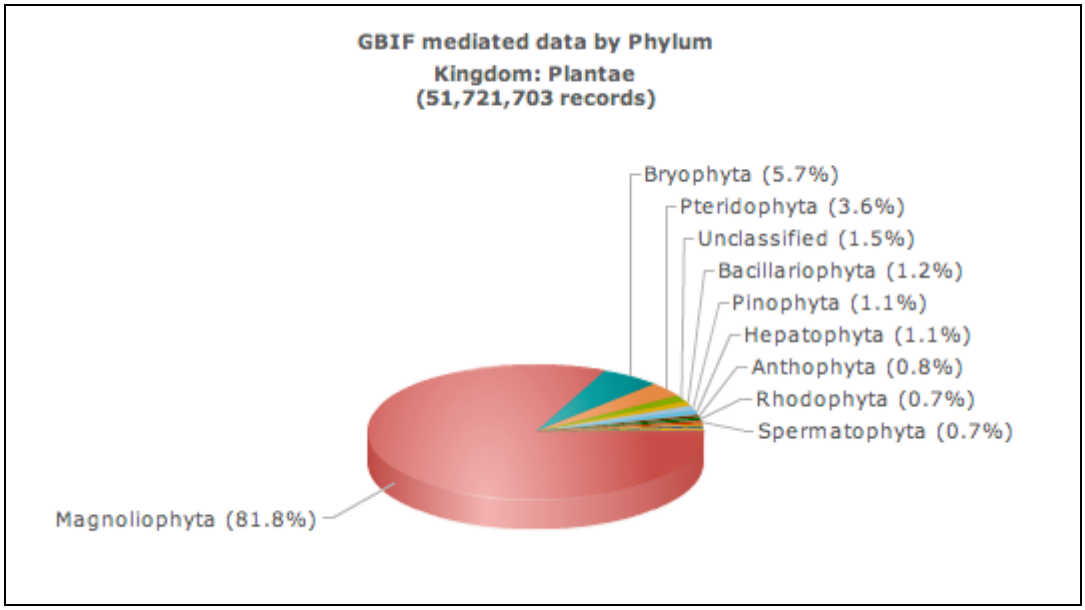


Figure C4: GBIF mobilised data records for Kingdom Plantae. Phylum Magnoliophyta constitute over 81% of the Plantae records mobilised by GBIF.

Of the 42+ millon data records belonging to Phylum Magnoliophyta, over 30+ million (71%) are those related to Class Magnoliopsida (Figure C5), of which 77.5% are georeferenced (Table C5).

Class	Records ▼	Georeferenced (%)	% of total
Magnoliopsida	30,137,196	77.5	71.3
Liliopsida	12,133,881	75.7	28.7
Monocotyledona	15,969	66.4	0
Unclassified	8,780	53.4	0
Classindet	635	23.5	0
Dicotyledon	48	45.8	0

Table C5: GBIF mobilised data records by Class for Phylum Magnoliophyta.

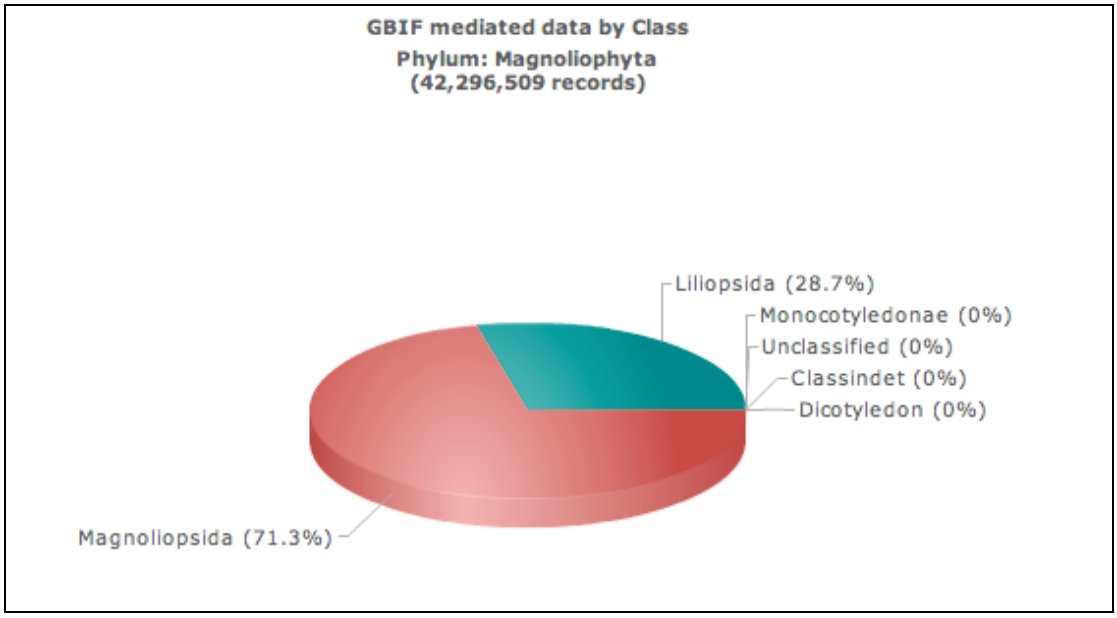


Figure C5. GBIF mobilised data for Phylum Magnoliophyta. Class Magnoliopsida constitute over 71% of the data records.

5.2 Geographic Assessment

Geographic assessment of the GBIF mobilised data reveals that 40.6% (or 81+ million) of the currently accessible data records are observed or collected from Europe (Figure c6, Table 6c). This is followed by 33.6% (67+ million) data records observed or collected in Northern America. Therefore, the total number of GBIF mobilised records observed or collected in the Northern hemisphere constitutes 72.4% of the total (148+ million).

Consequently, data records observed or collected from Southern hemisphere (which is also the biodiversity rich region) constitute only 27.6%. of the total. Records observed or collected in South America constitute 6.1% (12+ millions), followed by Africa, Asia and Oceania respectively. The numbers of data records collected from Africa amount to 9+ million, followed by Asia (7+ millions) and Oceania (6+ millions). Over 789+ thousand records are observed or collected from Antarctica, which are not represented in Figure C6. Over 15+ million records are classified as unknown.

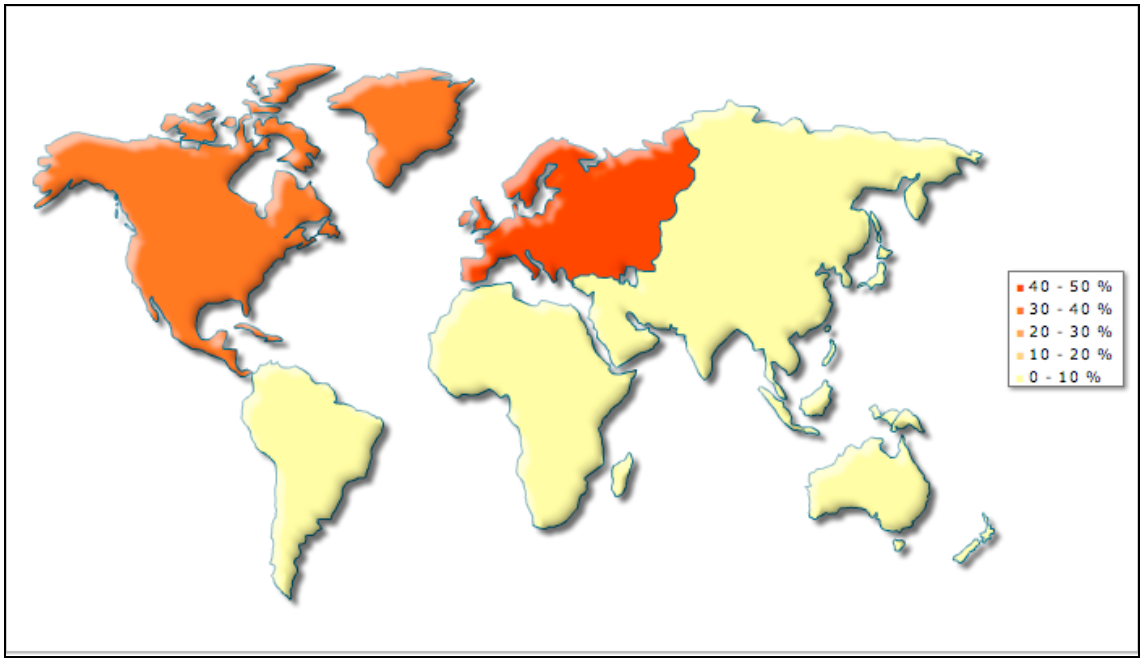


Figure C6. GBIF mobilised data by continents.

Continent	Records	Georeferenced (%)	% of total
Europe	81,718,041	90.9%	40.6%
North America	67,632,690	84.1%	33.6%
Unknown	15,891,727	54.1%	7.9%
South America	12,223,884	63.4%	6.1%
Africa	9,352,126	79.3%	4.6%
Asia	7,033,171	41.2%	3.5%
Oceania	6,634,505	77.8%	3.3%
Antartica	789,324	98.3%	0.4%

Table C6. GBIF mobilised data by continents.

Geographic assessment by country provides a greater level of detail to the observed regional patterns. The top 15 countries where data records are observed or collected include, the United States of America, Sweden, United Kingdom, France, Canada, South Africa, Denmark, The Netherlands, Spain, Norway, Mexico, Costa Rica, Australia, Germany, Austria, Ireland. Thus, this pattern clearly indicates that the majority of the GBIF mobilised data is either observed or collected in the data-rich part of the globe, thereby hindering any global analysis and interpretation at this stage.

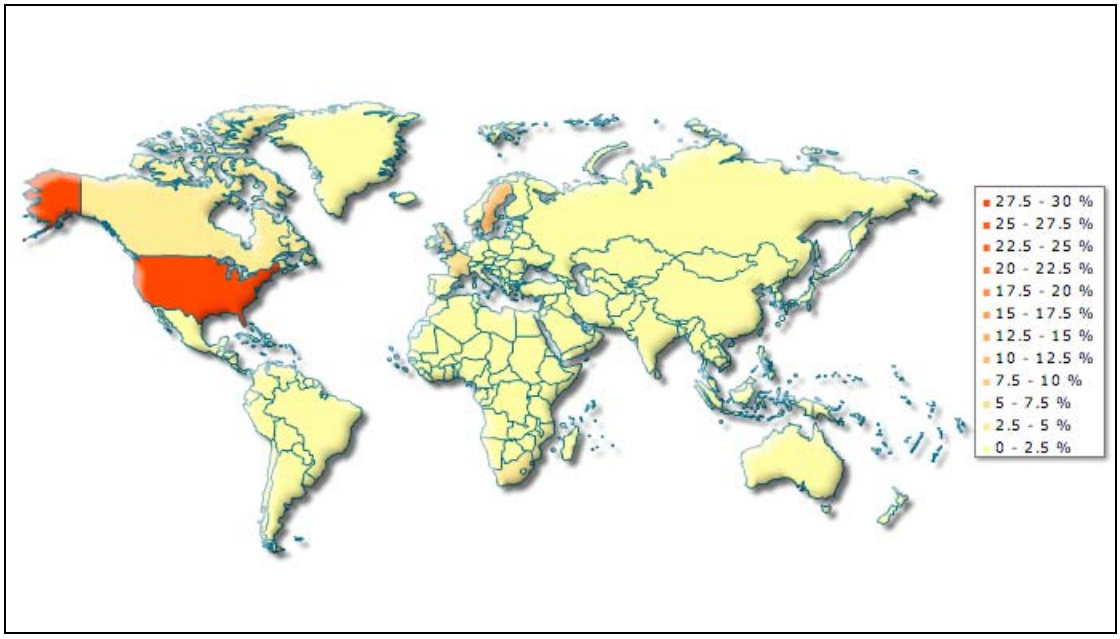


Figure C7: GBIF mobilised data by Country.

Country	Records	Georeferenced (%)	% of total
United States	57,581,449	84.9%	28.6%
Sweden	23,097,566	88.9%	11.5%
Unknown	15,891,727	54.1%	7.9%
United Kingdom	15,719,382	98.6%	7.8%
France	10,378,224	94.4%	5.2%
Canada	6,361,574	90.9%	3.2%
South Africa	5,809,191	95.2%	2.9%
Denmark	4,918,476	97.7%	2.4%
Netherlands	4,726,379	97.1%	2.3%
Spain	4,489,590	68.7%	2.2%
Norway	3,695,401	95.2%	1.8%
Mexico	3,689,667	60.2%	1.8%
Costa Rica	3,652,704	92.2%	1.8%
Australia	3,502,135	90.3%	1.7%
Germany	3,393,103	87.4%	1.7%
Austria	2,409,151	95.1%	1.2%
Ireland	2,085,814	97.9%	1%

Table C7. GBIF mobilised data by country.

5.3 Temporal Assessment

Analysis of the temporal spread of GBIF mediated data records shows that records date back to pre-1800. There are 86,030 records that are dated

before 1800. However, a large volume of GBIF mobilised data is dated after 1950. Over 55 million data records published through the GBIF network are sampled during 2001-2010 (Figure C8). Temporal attributes are missing for a quarter (26%) of the GBIF mobilised data.

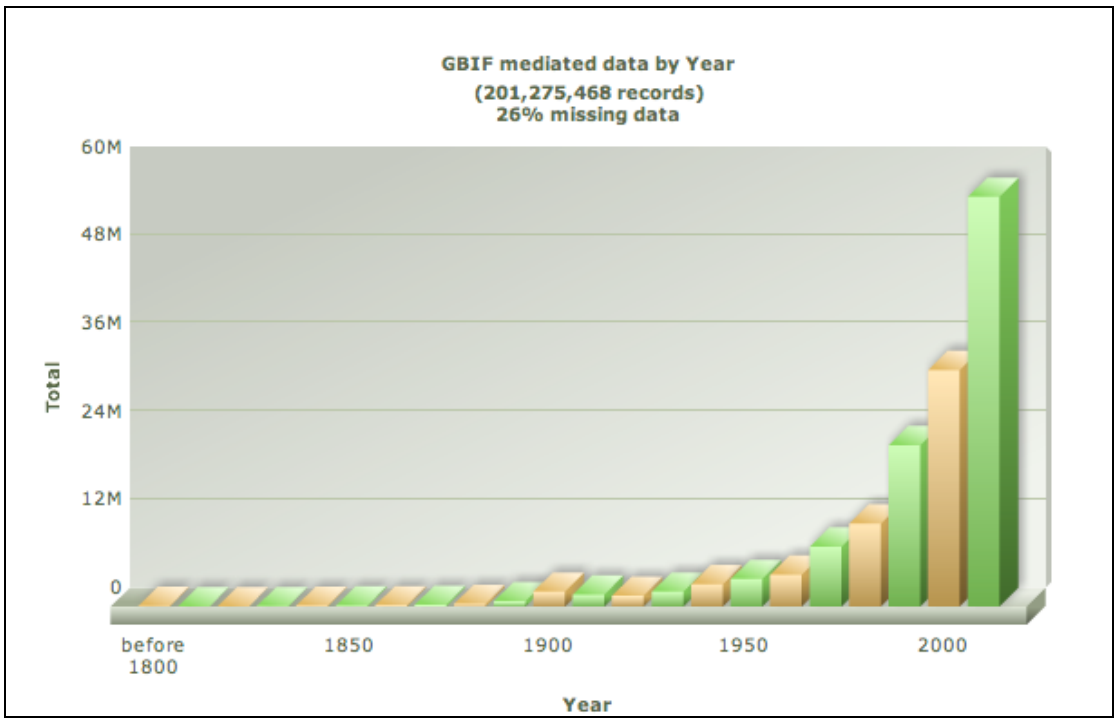


Figure C8: GBIF mobilised data by decade of collection/observation.

A taxon based temporal analysis indicates that our progress in publishing data records related to Kingdom Animalia is gradual (Figure C9). However, the GBIF network is far more efficient in publishing data collected and observed during this decade than in the previous decades. For the 1991-2000 period, over 20 million records are accessible; however records collected or observed since 2001 amounts to 48+ million. It is worth mentioning that 22% of records for Animalia do not supply a collection date, so they cannot be evaluated here.

With regards to the temporal coverage of data records related to Kingdom Plantae, the GBIF network publishes maximum records observed or collected during 1981-90 (Figure C10), followed by 1991-2000, and 2001- to date. This depleting trend of temporal coverage from 1991-to date is an issue of concern that needs to be investigated. Both Kingdom Animalia and Kingdom Plantae have data records collected or observed prior to1800.

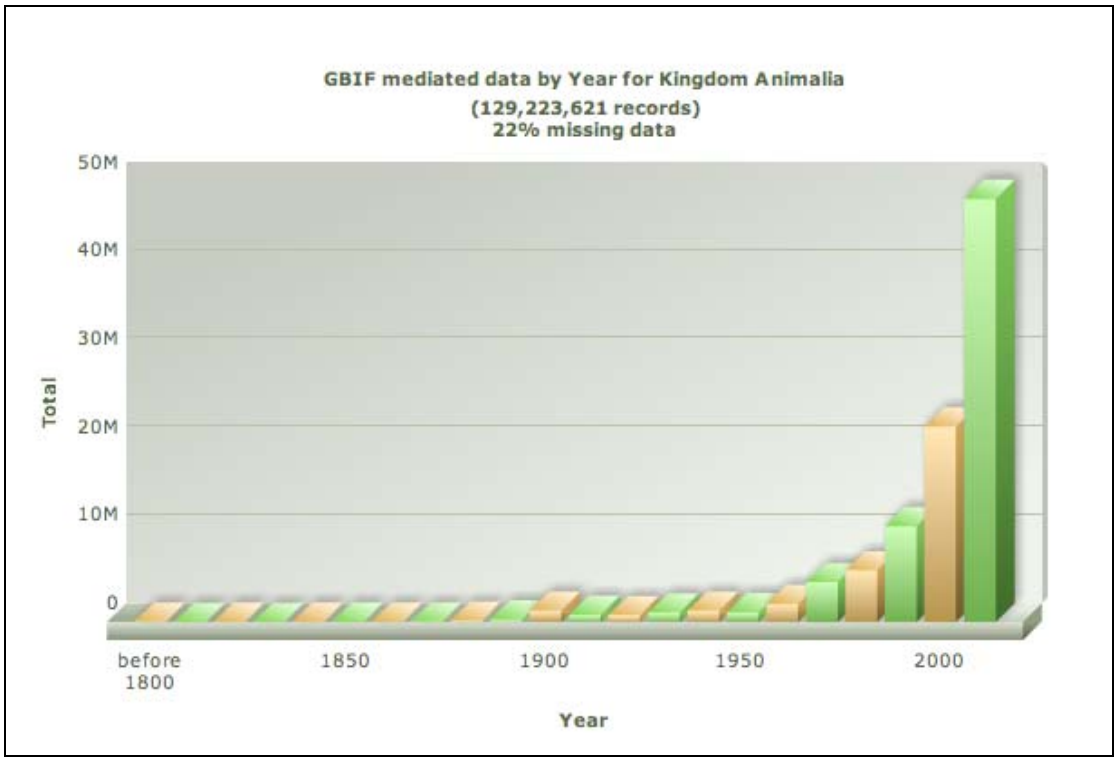


Figure C9: GBIF mobilised data by decade for Kingdom Animalia.

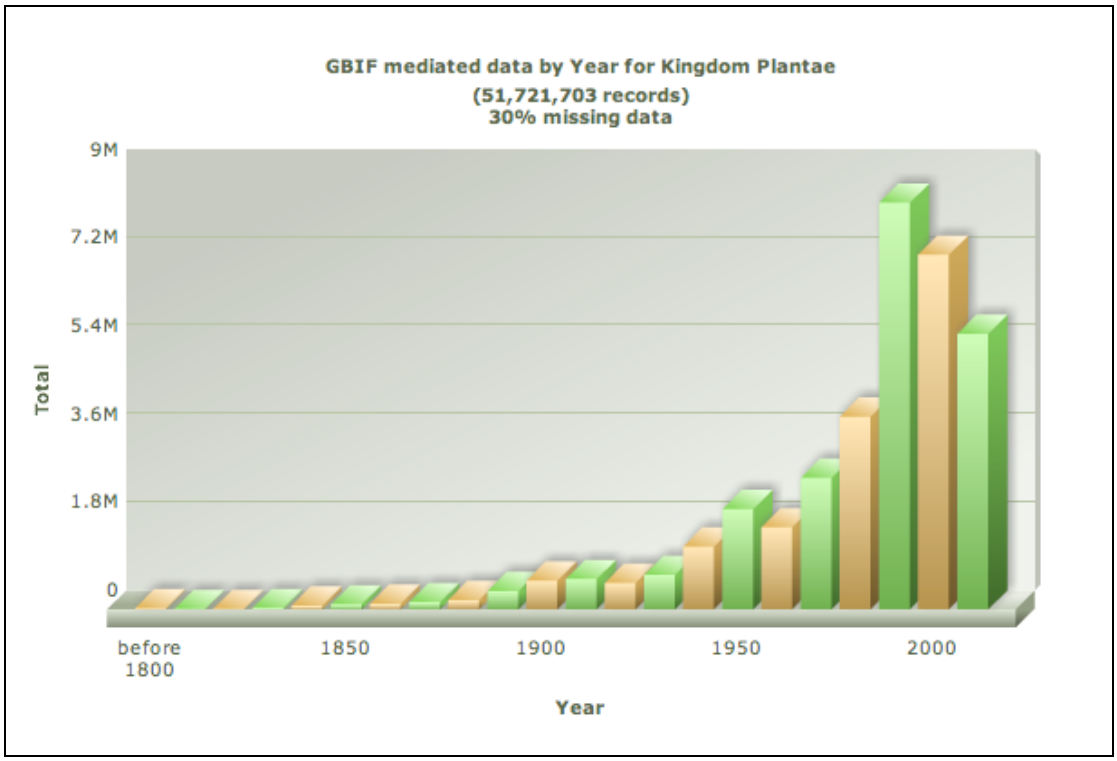


Figure C10: GBIF mobilised data by decade for Kingdom Plantae.

5.4 Basis of Records

Currently 62.8% of the GBIF mobilised data records are observation based of which 97% are georeferenced (Figure C11). Specimen based data records constitute 25.7% of the GBIF mobilised data of which only 52.9% are georeferenced. Surprisingly, the basis of record- is unknown or undocumented for 10.7% of the data records, of which 59.5% are georeferenced.

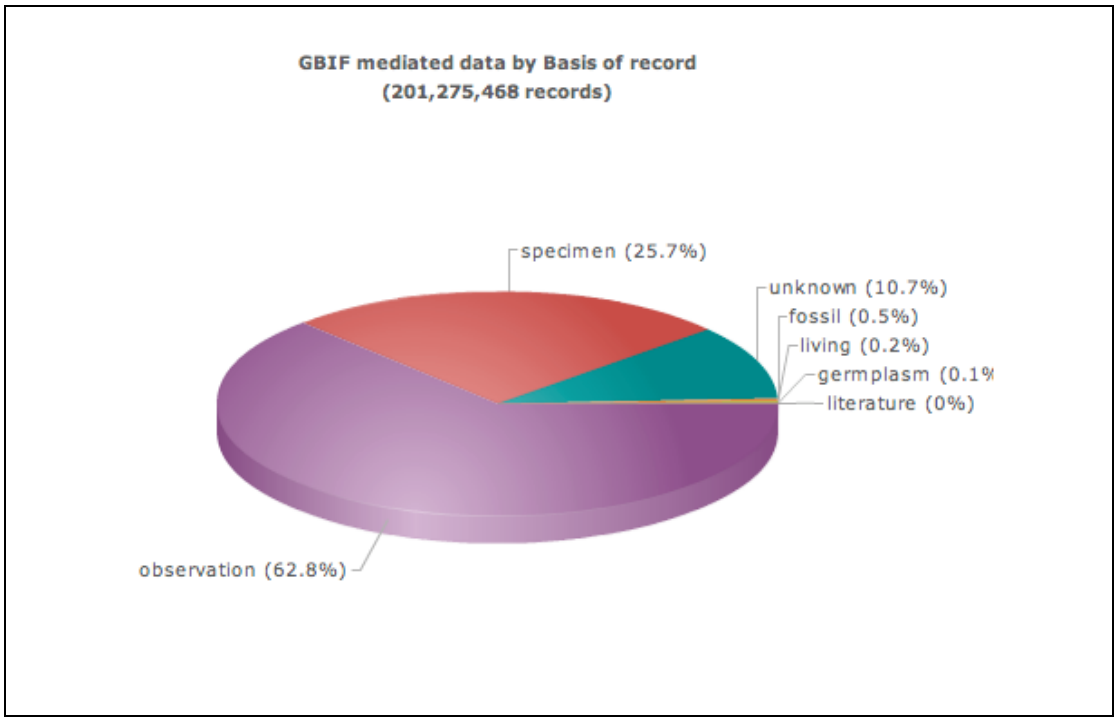


Figure C11: GBIF mobilised data by basis of records.

The decadal distribution of data records with specimen as basis of records indicates gradual increase in coverage till 2000 (Figure C12), which seems to be decreasing in the current decade (2001 to date).

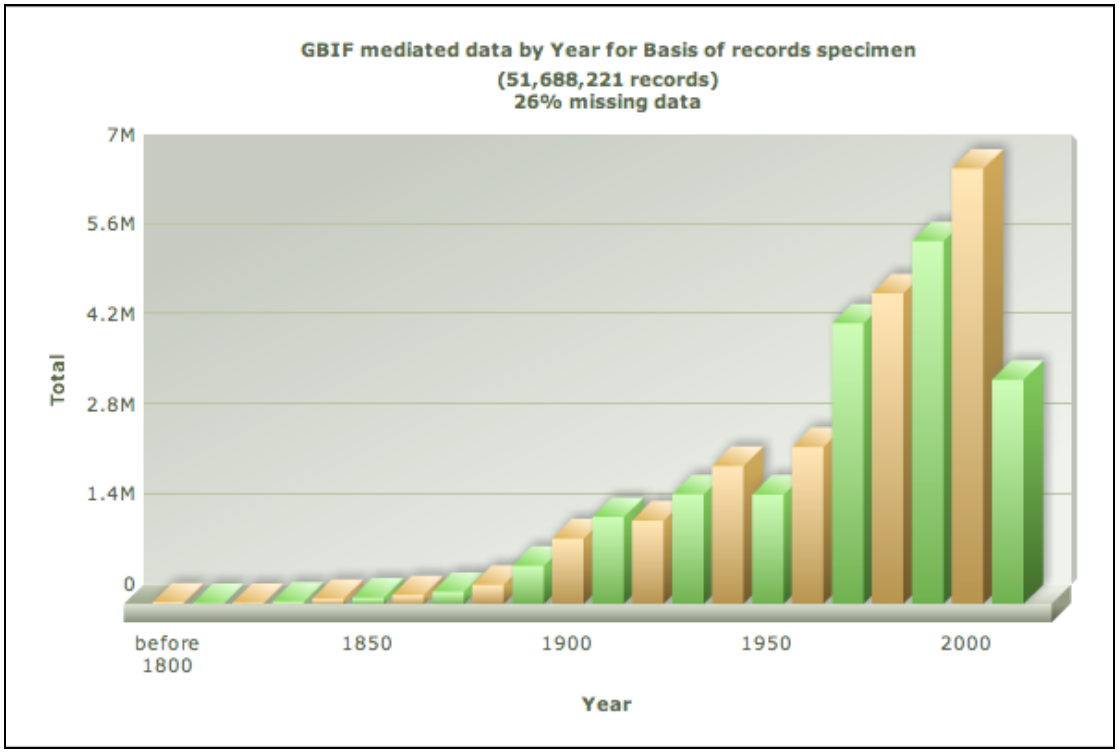


Figure C12. GBIF mobilised data by Year for basis of records 'Specimen'.

On the contrary the decadal coverage of data records with 'observation' as basis of records is rising. In fact, in the current decade (2001-2010 to date) temporal coverage of data records with 'observation' as basis of records has doubled (Figure C13).

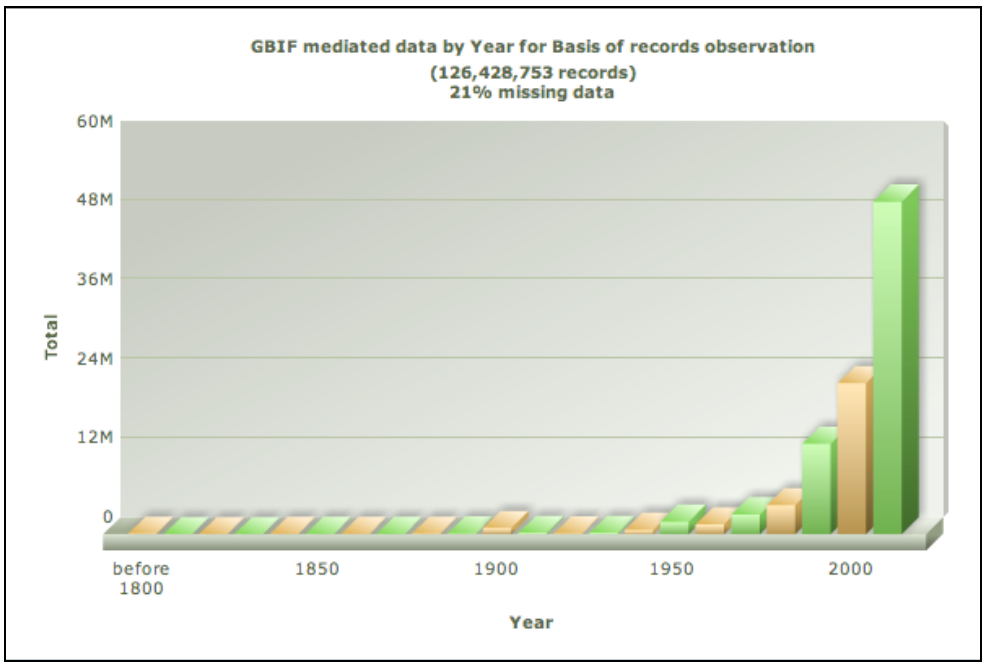


Figure C13. GBIF mobilised data by year for basis of records 'observation'.

6.0 Preparedness of the GBIF Network

This State-of-the-Art and Potentials report of the discovery and publishing through the GBIF network would remain incomplete without a mention of the preparedness of the network for meeting the ambitious targets set at the 15th Meeting of the Governing Board (GB15) in November 2008[22]. The analysis presented in this section is based on two key activities, namely, (a) the Participant Report 2009, and (b) the 'Data Discovery and Mobilisation Strategy' discussions held with 25 Participants in 2009. Table D1, lists 25 Participants who participated in 'Data Mobilisation Strategy Discussions' during 2009, and Table D2 lists the Participants who responded to the Participant Report, 2009.

Argentina, Austria, Belgium, Canada, Chinese Taipei, Columbia, Costa Rica, Denmark, Discover Life, Finland, Iceland, Japan, Mexico, NatureServe, New Zealand, Norway, Ocean Biogeographic Information System (OBIS), Poland, Republic of Korea, Scientific Committee on Antarctic Research (SCAR), Sweden, The Netherlands, and United Kingdom

Table D1. 25 Participants participated in 'Data mobilisation Strategy discussions' during 2009.

Country or Economy Participants
Argentina, Australia, Austria, Belgium, Burkina Faso, Cameroon, Canada, Chinese Taipei, Colombia, Costa Rica, Cuba, Denmark, Finland, France, Germany, Ghana, Guinea, Iceland, Indonesia, Ireland, Japan, Kenya, Republic of Korea, Madagascar, Mexico, The Netherlands, New Zealand, Norway, Pakistan, Peru, Philippines, Poland, Slovakia, South Africa, Spain, Sweden, Switzerland, United Kingdom, United States of America
Associate Organisation Participants
AdinoNet, BioNET-International, Bioversity International, CABI Bioscience, Consortium of European Taxonomic Facilities, Encyclopedia of Life, ETI Bioinformatics, Endangered Wildlife Trust, Inter-American Biodiversity Information Network, International Centre for Insect Physiology and Ecology, International Long Term Ecological Research, Integrated Taxonomic Information System, MORENA, NORDGEN, Ocean Biogeographic Information System, Pacific Biodiversity Information Forum, Scientific Committee on Antarctic Research, Society for Management of Electronic Biodiversity Data, Society for the Preservation of Natural History Collections, Taxonomic Database Working Group, United Nations Environment Programme – WCMC, World Data Center for Biodiversity and Ecology, World Federation for Culture Collections, Wildscreen

Table D2. Participants responded to the Participants Reporting System, 2009. However, not all responded to questions on data discovery and publishing.

[22] http://www2.gbif.org/GB15_ExecutiveSummary.pdf

6.1 Data Discovery

It is agreed that better discovery of data resources is a first step towards hassle-free and efficient publishing of biodiversity data. However, in the early years data discovery was not the focus of the GBIF network, which was rather 'data publishing' of low-hanging fruits (i.e. readily available digital data). Recognising that data discovery is critical to developing demand-driven data publishing strategy and action plans, in 2009, GBIF engaged in scoping requirements and planning for the construction of a distributed metadata system. Therefore, data discovery is in an extremely nascent stage across the GBIF network.

Less than 50% of our Participants responded to a question on discovery and metadata activities in the Participant Report 2009. Of these 38% (15) country Participants indicated that they have a metadata catalogue to document and describe biodiversity data resources, compared with 41% (16) that reported that they plan to establish national metadata catalogues, and 21% (8) that reported that no national metadata catalogue was available in their country.

However, further discussions with Participants who responded affirmatively to this question reveal that the majority of these metadata cataloging systems are limited in scope, and are certainly not the priority activity for the Nodes. Furthermore, there is a lack of national policy on metadata cataloging for easier discovery of biodiversity data with the exception of France, Guinea, Ireland and United States of America.

6.2 Data Publishing

Data Publishing preparedness of the GBIF Network is assessed on the basis of Participant responses to questions related to: (a) content needs assessment; (b) data discovery and publishing strategies; (c) actions to meet 2 billion targets, and (d) data publishing by 2010.

6.2.1 Content Needs Assessment: Comprehensive 'Content Needs Assessment' is the first step towards demand-supply driven 'data discovery and publishing strategies', which fulfills the needs of the stakeholder communities. Only six Participants, namely Argentina, Australia, Burkina Faso, Costa Rica, Cuba and South Africa had completed a systematic content needs assessment. A few others are in the process of conducting such an assessment, or had carried out an assessment but not systematically. However, a large part of our network has no plans to conduct 'Content Needs Assessment'. Furthermore, the investigation during the 'data mobilisation strategy discussions' reveals that (a) there is lack of understanding of the significance of CAN in developing strategies and action plans for demand-driven data discovery and publishing, and (b) the majority of the Participants do not consider having systematic content needs assessment as an essential activity.

6.2.2 Data Discovery and Publishing Strategies: The GBIF network status on this issue is similar to that for content needs assessment. Only 9 of the 54 country Participants had such a national data discovery and publishing strategy in place. These include Australia, Colombia, Denmark, France, Guinea, Mexico, South Africa, United Kingdom and the United States of America.

Further discussion on this topic with both the country and organisation Participants revealed that: (a) data discovery and publishing activities across the network are mostly unplanned; (b) there is a lack of understanding of the significance of strategies and action plans for data discovery and mobilisation, and (c) that the majority of the Participants do not consider this as an essential step forward.

6.2.3 Actions to meet the 2 billion targets: Despite the setting of ambitious targets for data discovery and publishing[23], to date our progress has been linear. Only 9 country Participants (Colombia, Denmark, Ghana,

[23] http://www2.gbif.org/WP2009-10.pdf

Ireland, The Netherlands, Peru, South Africa, Switzerland, and the United States of America) and 5 Associate organisation Participants (SCAR, OBIS, EWT, SMEBD and ITIS) Participants have taken specific actions to expedite the progress. However, closer scrutiny of the measures taken indicates that these actions are not sufficient to meet the target.

6.2.4 Data publishing by 2010: Responses by the Participants on specific queries as to how much data will be published by December 2010 are discouraging with regards to the targets set. For instance, the Participant Report 2009 identified 818 million digital records of the 2.45 billion data records identified. However, as shown in Table D2 and D3, Participants committed to publishing 195+ million records by the end of 2010. This is approximately 23% of the 818 million digital records reported by these 27 Participants in 2009 (Participant Report, 2009).

Q26: What are the estimates of numbers of records that the data holders within the the domain of your national node plan to mobilise and publish via the GBIF data portal (http://data.gbif.org) by end 2010?		
Record type	**2009**	**2010**
Specimen based occurrence data	11,167,563	18,565,962
Observation based occurrence records	40,141,645	105,130,000
Multimedia data linked to primary biodiversity data	81,700	277,100
Population / ecological monitoring records	55,000	4,422,000
Impact Assessment associated data records	0	50,000
Other types of primary biodiversity data	12,100	5,532,000
TOTAL	**51,458,008**	**133,977,062**

Table D2. Estimate of data publishing by 22 country Participants by end 2010.

Q64: What are the estimates of numbers of records that the data holders within the the domain of your organisation plan to mobilise and publish via the GBIF data portal (http://data.gbif.org) by end 2010?		
Record type	**2009**	**2010**
Specimen based occurrence data	330,000	330,000
Observation based occurrence records	4,700,000	4,185,000
Multimedia data linked to primary biodiversity data	7,100	7,100
Population / ecological monitoring records	500	200
Impact Assessment associated data records	700	200
Other types of primary biodiversity data	475,610	475,720
TOTAL	**5,513,910**	**4,998,220**

Table D3. Estimate of data publishing by 5 organisation Participants by end 2010.

Engagement with Potential Data Publishers: National and thematic level engagement with potential data publishers is also not yet optimal. For instance, 33 countries reported that they have a contact list of 2464 institutions and 9174 individuals, which is an under-estimation of the total number of potential data Publishers in these countries.

6.3 Summarising the state of preparedness:

- Data discovery and publishing activities across the GBIF network are opportunistic, and aimed at only tapping low hanging fruits.
- Data discovery activities are at an early stage, and there is a lack of policy for biodiversity data discovery at the Participant level.
- Many Participants have not yet undergone a systematic content needs assessment, and lack data discovery and publishing strategies.
- Actions to meet 2 billion data publishing target by the end of 2010 are not sufficient.
- In spite of estimating that data publishers in their domain have nearly 818 million data records, Participants only estimate that they will be able to publish 23% of these data records by the end of 2010.
- Engagement with potential data publishers is far from what is required to achieve the targets.

7.0 Recommendations to overcome current impediments

As mentioned in several key strategic documents, the discovery and publishing of primary biodiversity data is the major reason for GBIFs' establishment (GBIF 2008). If GBIF has to be scientifically and socially relevant and ensure its sustainable progress, we must evolve quickly from the existing modes of discovery and publishing of data, which are producing linear growth in accessible records. One approach is to adopt strategies and action plans that will expedite our progress to meet the growing demands for data by our stakeholder communities. To date, it is evident that our progress in data discovery and publishing is directly proportional to the status, mandate, capacity, vision and resources (human, infrastructural and

financial) of the Participant BIFs[24]. Thus, the strengthening of our Participant BIFs is crucial to our success.

As depicted in Figure E1, this calls for local to global scale, coordinated activities by ALL the Participants at an early date. These include systematic content needs assessment, data discovery strategy and action plans, fundable business plans, increased investment by stakeholders, and uptake of a comprehensive data publishing framework[25]. To address the need expressed by the Participants for guidelines to develop needs-driven data discovery and publishing strategies, a 'Best Practice Guide for Data Discovery and Publishing Strategy and Action Plans' is released for community uptake[26].

Figure E1. Local to global scale investment by ALL Participants in data discovery and publishing activities is essential.

[24] Participant BIF refers to a network of data holders, users and other stakeholders established by a GBIF Participant to promote, facilitate, and coordinate the biodiversity data sharing activities within its domain.
[25] http://www.biomedcentral.com/1471-2105/10/S14/S2.
[26] http://www2.gbif.org/BestPracticeGuide-final.pdf

8.0 Lessons Learnt and Future Plans

8.1 Lessons learnt

Monitoring of the network's progress in discovery and publishing of the data is not a onetime exercise. It needs to be carried out at frequent intervals. Furthermore, comprehensive understanding of the network's progress can only be understood if such an exercise is carried out on a local-to-global scale. Active involvement of all network Participants is essential for the success of such exercises. Future data discovery and publishing efforts should be based on the scientific and societal relevance towards the mission of GBIF. Participants should develop their 'data discovery and publishing strategies and action plans' accordingly taking into consideration the existing in-country resources – financial, human, infrastructure, and expertise.

8.2 Future Plans

It was felt that network Participants will greatly benefit from having a dynamic progress monitoring facility. Therefore, it is planned to develop features that can provide dynamic and up-to-date assessment of the progress made either by a particular Participant or a network as a whole in data discovery and publishing.