Biodiversity Informatics, 8, 2013, pp. 94-172

CONTENT ASSESSMENT OF THE PRIMARY BIODIVERSITY DATA PUBLISHED THROUGH GBIF NETWORK: STATUS, CHALLENGES AND POTENTIALS

SAMY GAIJI [(1)]*, VISHWAS CHAVAN [(1)], ARTURO H. ARIÑO [(2)], JAVIER OTEGUI [(2)], DONALD HOBERN [(1)], RAJESH SOOD [(1),] ESTRELLA ROBLES [(2)]

(1) Global Biodiversity Information Facility Secretariat, Universitetsparken 15, DK-2100, Copenhagen, Denmark

(2) University of Navarra, Pamplona, Spain

**Corresponding author, Email: sgaiji@gbif.org*

Abstract —With the establishment of the Global Biodiversity Information Facility (GBIF) in 2001 as an inter-governmental coordinating body, concerted efforts have been made during the past decade to establish a global research infrastructure to facilitate the publishing, discovery, and access to primary biodiversity data. The participants in GBIF have enabled the access to over 377 million records of such data as of August 2012. This is a remarkable achievement involving efforts at national, regional and global levels in multiple areas such as data digitization, standardization and exchange protocols. However concerns about the quality and 'fitness for use' of the data mobilized in particular for the scientific communities have grown over the years and must now be carefully considered in future developments. This paper is the first comprehensive assessment of the content mobilised so far through GBIF, as well as a reflexion on possible strategies to improve its 'fitness for use'. The methodology builds on complementary approaches adopted by the GBIF Secretariat and the University of Navarra for the development of comprehensive content assessment methodologies. The outcome of this collaborative research demonstrates the immense value of the GBIF mobilized data and its potential for the scientific communities. Recommendations are provided to the GBIF community to improve the quality of the data published as well as priorities for future data mobilization.

Keywords— Primary Biodiversity Data, Content Assessment, and Gap Analysis.

INTRODUCTION

Free and open access to primary biodiversity data is essential both to enable effective decision-making and to empower those concerned with the conservation of biodiversity and the natural world (Bisby, 2000; Gaikwad and Chavan, 2005; GBIF, 2008). However, the history of publishing of primary biodiversity data is very recent. With the establishment of the Global Biodiversity Information Facility (GBIF) in 2001, concerted efforts to publish primary biodiversity data using community driven and agreed standards and tools gained momentum. GBIF was created to facilitate free and open access to biodiversity data worldwide, via the Internet, to underpin scientific research, conservation and sustainable development. The GBIF network, through its data portal (http://data.gbif.org), already facilitates access to over 377 million records from more than 400 data publishers[1]. The progress achieved in GBIF's first decade indicates that the development of a global informatics infrastructure, facilitating free and open access to biodiversity data, is indeed a realistic aspiration. One of the key future challenges for GBIF is now to ensure that such volume of knowledge about biodiversity on earth is indeed of high relevance for the scientific communities.

[1] As of August 2012.

Why assess the content of GBIF-mobilised data?

Despite GBIF's achievements, questions are frequently raised about whether it can yet be considered a global facility (Yesson et al., 2007), and about the usefulness of the data mobilised. GBIF has been criticised for the taxonomic, thematic, geospatial as well as temporal biases in the data mobilised by its network of data publishers (Johnson, 2007). There have been isolated studies to assess gaps, quality and fitness for use of GBIF-mobilised data (e.g. Guralnick et al., 2007; Collen et al., 2008; GBIF, 2010a). In 2010, an initial overview of the data published through the GBIF network (GBIF, 2010b) provided a first set of indicators on the content mobilized so far as well as major bias such as in the taxonomy and temporal areas. Recognising this, the GBIF-constituted Content Needs Assessment Task Group (CNATG) recommended that assessment of GBIF-mobilised content at various levels (global, regional, national and thematic) is crucial for determining the demand-driven approach for data mobilisation (Faith et. al., 2013, 2013). In 2011, in response to these recommendations, a series of improvements to the GBIF infrastructure were made such as the rework of the GBIF 'backbone taxonomy' with up-to-date checklists and taxonomic catalogues such as the Catalogue of Life 2011[2]. Other improvements such as the automated interpretation of the coordinates, country location and scientific names used in published records have been improved to screen out inaccuracies – for example, ensuring that records identified as coming from a particular country are shown as occurring within the borders and territorial waters of that country. The current study attempts to assess the gaps and fitness for use of the GBIF-mobilised data. It aims to provide a comprehensive overview of the 'state of the network' for data published through the GBIF network in 2012. Such assessment is aimed at demonstrating the value of the content mobilised and how it can contribute to our improved understanding of biodiversity in particular by the scientific community.

To achieve this objective and taking into account the large volume of information to be analysed, the authors of this study have adopted two complementary methodologies. One approach led by the GBIF Secretariat (GBIFS) focused on two temporal complete studies (December 2010 and February 2012) while the Department of Zoology and Ecology at the University of Navarra (UNZYEC) focused on processing random samples of the full content. The research outputs of these two studies were compared and complemented each other.

The outcomes of these two complementary exercises are presented in three categories: (a) data quality assessment, (b) trends/patterns assessment, and (c) fitness-for-use assessment.

Data flow of the GBIF network

As of August 2012, the GBIF network is comprised of 419 data publishers from 44 countries and 15 international organisations. Together they publish through GBIF 10,028 occurrence based data resources (or datasets).

Figure 1 depicts the typical flow of the data publishing processes through the GBIF network. Data publishers can use a variety of tools and protocols (e.g. DiGIR[3], BioCASE[4], Tapir[5], GBIF Integrated Publishing Toolkit[6]) and data standards (e.g. DwC[7] and ABCD[8]) in order to publish primary occurrence records to GBIF. After successful registration of their resources through the central registry, GBIF centrally indexes a limited but essential number of core data elements

[2] Ruggiero M., Gordon D., Bailly N., Kirk P., Nicolson D. (2009). The Catalogue of Life Taxonomic Classification, Edition 2, Part A. In: Species 2000 & ITIS Catalogue of Life, 3rd February 2012 (Bisby F.A., Roskov Y.R., Culham A., Orrell T.M., Nicolson D., Paglinawan L.E., Bailly N., Appeltans W., Kirk P.M., Bourgoin T., Baillargeon G., Ouvrard D., eds). DVD; Species 2000: Reading, UK.

[3] http://www.digir.net/

[4] http://www.biocase.org/

[5] http://wiki.tdwg.org/TAPIR/

[6] http://www.gbif.org/orc/?doc_id=2935

[7] http://rs.tdwg.org/dwc/index.htm

[8] http://www.tdwg.org/standards/115/

detailing the ‘what’ (species), ‘when’ (date/time), ‘where’ (location), “with what evidence” (basis of record) and ‘by whom’ (collector/observer) of the primary biodiversity data published by the GBIF network (also called GBIF-mediated data). The list of core data elements (Table 1) follows a common data standard: the Darwin Core standard[9]. This data standard has been used for the discovery of the vast majority of specimen occurrence and observational records published through the GBIF network. The Darwin Core standard was originally conceived to facilitate the discovery, retrieval, and integration of information about modern biological specimens, their spatio-temporal occurrence, and their supporting evidence housed in collections (physical or digital). These elements are compiled into a central database (also called GBIF Index) and their discovery and access is enabled through the GBIF data portal (http://data.gbif.org) as well as through web services (http://data.gbif.org/tutorial/services). Such a global discovery system is aimed at promoting access to the original information sources owned by each single publisher participating in the GBIF network, where more information can be found (*e.g.* media, richer data etc.).

While all data publishers are expected to follow common standards (*e.g.* DwC), their data resources discoverable through the GBIF infrastructure have varying precision and quality. This could be explained by incomplete information at the publisher level, errors during the publishing processes (*e.g.* formatting of date information) as well as errors during the central GBIF harvesting and indexing procedures.

In order to assess the content mobilised through the GBIF network, this study will focus on using the content of the GBIF Index as a proxy to the information published by the contributing publishers.

CONTENT ASSESSMENT OF GBIF-MOBILISED DATA

Methodology

In the last two decades, the informatics field has evolved to a stage where the handling of very large volume of data is becoming the central component of data discovery[10]. The capacity to store, manage and analyse a large volume of data is becoming a fundamental requirements in the field of Biodiversity Informatics and in particular for infrastructures like GBIF[11]. Today, technologies like Hadoop[12] and Hive[13] offer the ability to process such huge volumes of information on certain kinds of distributable problems using a large number of computers.

The assessment carried out by GBIFS used this new technology to process and analyse the full GBIF Index is depicted in Figure 2. The full GBIF Index was extracted in the form of Hive tables in December 2010 and February 2012. All outputs of the data-mining processes were stored in MySQL tables for easy processing and visualisation. The results of these analyses were kept so that in the future similar experiments could be repeated and compared temporally.

The Hadoop/Hive technology allowed the processing and analysis of the full GBIF Index in a reasonable amount of time compared to conventional technologies like relational database using known database management systems like MySQL. However such methodology requires a dedicated infrastructure with sufficient IT expertise and understanding of the processes involved in manipulating such large volume of information at once.

UNZYEC used two separate approaches in their assessment (Figure 3). In one, a random sample of the GBIF Index was obtained by issuing

[9] http://rs.tdwg.org/dwc/index.htm

[10] Jiawei Han and Jing Gao, “Research Challenges for Data Mining in Science and Engineering", in H. Kargupta, et al., (eds.), Next Generation of Data Mining, Chapman & Hall/CRC, 2009, pp. 3-28.

[11] http://www.gbif.org/communications/news-and-events/showsingle/article/important-quality-boost-for-gbif-data-portal/

[12] http://hadoop.apache.org/

[13] http://hive.apache.org/

an automated set of queries through the portal's web services[14]. This approach mimics an ecological sampling where a vast amount of data is represented by a subset, thus greatly reducing the data processing requirements. In another approach, mirrors of both the GBIF Index and the raw data harvested from the participants were queried using standard SQL statements and scripts. Although much more taxing in terms of resources, this approach enabled the authors to finely track the flow of information (not just data) from the publishers to the index. In this way, gaps caused by the data processing flow can be detected.

The UNZYEC team made queries and samplings during a three-year period, over ten versions of the GBIF Index. However, for the purpose of this assessment, analyses were made mostly on the November, 2010-released mirror, in order to provide an independent comparison of GBIFS-obtained results.

Limitations of the methodologies

The methodology used in this article enables the fast data mining of the GBIF data index but does not address issues such as:

-The level of accuracy of the data (e.g. precision in geospatial coordinates).

-The risk of misidentification of taxa.

-Duplicate records that can arise from:

i. Datasets being unwittingly published repeatedly,

ii. Duplicate records within and between datasets,

iii. Multiple digital records derived from the same physical specimen, such as a specimen being physically split and stored in multiple museums.

- Computing interpretation errors in the data harvesting and indexing routines.

For example, depending on the data schema used (Darwin Core or ABCD) and their versions, an occurrence date may be represented as a date-time stamp, an ISO-formatted date, a simple text string in varying formats, or composed of individual fields (day, month, year). The mapping of the data by the publisher may therefore introduce additional error or ambiguity, if for example month and day are swapped. In order to overcome this difficulty, we assumed the level of error of the year within a malformed date-time stamp as sufficiently low to be considered as a good proxy to assess the temporal dimension.

With regards to the conversion and validation of taxonomical information (e.g. genus, species, scientific names) the challenges are more complex. During the harvesting and indexing procedures, the taxonomical information is checked against the most up-to-date GBIF taxonomical backbone. Until end 2011, GBIF used the Catalogue of life (CoL) 2007 as its core taxonomical backbone and when unmatched names were identified during the harvesting/indexing procedures they were simply added to the backbone. In November 2011, GBIF has entirely refreshed its taxonomical backbone and uses now primarily the latest version of the Catalogue of life in addition to other resources (Table 2). Today, unmatched names are not added to the core backbone and whenever possible, expert taxonomists are consulted. Therefore the study undertaken in terms of taxonomical comparison (in 2010 and 2012) should be undertaken taking into account this particular bias due to the improvement of the GBIF taxonomical backbone and resolution services.

Material

For the purpose of this study, elements covering three dimensions ("what", "where" and "when") were extracted from the GBIF Index by GBIFS and UNZYEC in December 2010, and also from raw data as supplied by the providers by UNZYEC for some specific analysis. Further analyses using the February version of the GBIF Index were undertaken by GBIFS.

[14] http://data.gbif.org/tutorial/services

The elements covered in these analyses are:

Source of the data: The assessment has taken into account the identifiers of the data publisher and data resources. However, due to incompleteness and lack of accuracy of entries in the institution ID, collection ID and catalogue fields in the GBIF Index, we have decided to exclude these fields from the analysis.

Taxonomic data: Taxonomic ranks such as Kingdom, Phylum, Class, Family, Genus and Species are included. The assessments have also taken into account the synonyms as recorded in the GBIF Index, in order to provide the most accurate estimate of the number of species. Data from multiple synonyms get merged during the harvesting and indexing routines.

Geospatial data: Latitude and longitude information was used when available. However, due to scarce information provided by data publishers, it was not possible to consider precision. This is a serious limitation that will need to be addressed in future analysis.

Temporal data: Limited to the field year of observation/collection. The assessments ignored the day and month recorded in the date field, except for analysing possible causes of year mis-assignment.

Other data: The basis of records, a descriptive term indicating whether the record represents an object or observation, was included in the analysis. The basis of record actually contains useful information such as the level of evidence and other categories that may be considered enhanced subclasses of information.

Results of the content assessment of the GBIF-mobilised data

We present the salient outcomes of these two independent exercises in four categories, namely: (a) data quality, (b) trends/patterns and (c) fitness-for-use assessments. In most cases, both exercises reached similar conclusions and therefore validate each other. In some instances, significant differences arose and were assessed.

A. Data Quality Assessment:

Taxonomy:

Until November 2011, the processing of taxonomical references was made against some taxonomical references such as the checklist of Catalogue of Life 2007 (http://www.catalogueoflife.org/annual-checklist/2007/) or the International Plant Names Index (http://www.ipni.org). During the discovery of unmatched taxonomical references against the accumulated GBIF taxonomical backbone, these are automatically added. Therefore, the 2010 GBIF taxonomical backbone contained accepted names (e.g. from CoL 2007) and new names discovered during the indexing process. This also means that in our December 2010 assessment, we had limited capacity to distinguish between authoritative names (e.g. referring to Catalogue of Life 2007 version) and added names, which had no validation against any taxonomical reference. In November 2011, the GBIF taxonomical backbone was rebuilt using primarily the latest version of the Catalogue of Life as well as many new taxonomical authoritative references (Table 2). Therefore the February 2012 assessment on taxonomical names can be considered as much more accurate.

Matching against the Catalogue of Life

Using a less advanced interpretation techniques developed in 2006 by the GBIFS, the backbone taxonomy that covers the occurrence records has 1,946,429 concepts at species or lower ranks, of which 458,716 (24%) is provided by the Catalogue of Life 2007 Annual Checklist[15]. A more recent study made in December 2010[16] showed that 52 per cent of the distinct canonical names found in the GBIF Index matched to a name in the CoL 2010 using straight, case insensitive matches. This can be slightly increased to 54% if a ‘fuzzy’ matching with a maximum difference of 10% in characters is used. In February 2012, a

[15] GBIFS personal communication (March 2011)

[16] http://code.google.com/p/gbif-occurrencestore/wiki/TaxonomicIntegration

similar study (Table 3) showed than 53.47% of names were straight, case insensitive matched of the canonical names in the Catalogue of Life 2011 Annual Checklist.

Completeness of the taxonomical classification

In order to study the completeness of taxonomical classification in the GBIF Index, we assessed for each rank (kingdom, phylum, class, order, family, genus and species) the valid references generated after the harvesting and indexing routines. The level of completeness is therefore based on valid taxonomical references within the GBIF taxonomical backbone. In cases where for example a family name wasn't mapped correctly, a 'null' value is assigned to this field in the published occurrence record. For each rank, we evaluated the number of occurrences and species (or lower taxa) having incomplete or unknown taxonomical status – or 'null' values (e.g. counting all occurrences having an `unknown` status for the kingdom rank). Table 4.a provides a summary of our findings in December 2010 and Table 4.b the summary for February 2012.

In 2010, a total of 114,721 species or lower taxa corresponding to 15 million occurrences representing 5.6% of the GBIF Index were not 'mapped' against the GBIF taxonomical backbone at the kingdom level. Similar trends are observed for other taxonomical ranks with somehow a variation in amplitude of incompleteness (e.g. 14.5% for species and lower taxa at the family level and 7.4% at the species level). This analysis confirmed similar results obtained in 2008 and 2010 (GBIF, 2010b and Ariño and Otegui, 2008). However some of the correctly matched names against the GBIF taxonomy backbone may not be valid names if referred to authoritative references such as Catalogue of Life. The reasons being that some of these names if not matched to the existing GBIF taxonomy backbone during the harvesting and indexing processes were simply added as valid references. The mixing of valid taxonomical references with new unverified references with limited capacity to track such changes over time caused serious difficulties to our study. The assessment summarized in Table 4.a provides therefore more a status of incompleteness of the taxonomical backbone rather than a real comparison to any authoritative taxonomical references.

In December 2010, our preliminary findings suggested the need for an urgent review of the GBIF taxonomical backbone in particular against the most critical taxonomical authorities such as the annual checklist Catalogue of Life 2010 (http://www.catalogueoflife.org/) and other sources such as the Interim Register of Marine and Nonmarine Genera (IRMNG). The decision not to mix unverified names with existing authoritative names was critical. In November 2011, GBIFS successfully upgraded its taxonomical backbone against the latest version of the Catalogue of Life (2011) and other authoritative references.

This resulted in our February 2012 study in a more accurate assessment of the taxonomical gaps within the GBIF Index. The results of this analysis are presented in Table 4.b. The percentages of incompleteness observed in 2012 were significantly lower (i.e. 0,35%, 1.81%, 2.82%, 2.17% respectively at the Kingdom, Class, Family and Genus levels) than the once observed in December 2010 (i.e. 7.0%, 14,5%, 14,5% and 4.7% respectively at the Kingdom, Class, Family and Genus levels) with the exception of the species rank. Similar trends are observed taking into account occurrences. Therefore a high number of unmapped taxonomical ranks from Kingdom to Genus levels were resolved using the upgraded GBIF taxonomical backbone. The higher number of taxonomical references used to construct the GBIF taxonomic backbone largely explains this. The observed percentages of unresolved names at the species level represents 9.15% in 2012 while in 2010 this percentage was of 7.4%. However these numbers can't be compared because of the changes in the taxonomical backbone between these dates.

Taking into account these improvements in taxonomical name resolution, we have tried to assess the additional data quality improvements that could be undertaken. To achieve this, we have looked at the top 10 possible misidentifications (at the kingdom level) by number of occurrences (Table 5). The three species within the genus *Zonotrichia* listed as within the plantae kingdom are wrongly assigned. These species belong to the American sparrows group of the family Emberizidae[17]. This misidentification is due to the generic homonym *Zonotrichia* being both present in the Plantae and Animalia Kingdom. This misidentification is being resolved in the GBIF taxonomical backbone and these obvious misidentifications progressively corrected[18]. For the other cases listed in Table 5, the discrepancy with CoL 2011 version is resolved in the latest version of the CoL (February 2012) or other taxonomical authorities (i.e. Marine Species Identification Portal). Once these changes are implemented we estimate that 1,808,488 occurrences would be correctly mapped and the total of occurrences with ‘unknown’ status at the species level would decrease from 25,343,834 to 23,535,346. This shows that while the GBIF Index has grown from 267 to 324 million occurrences (+21.3%) from December 2010 to February 2012, corrections on the top 10 species misidentifications in February 2012 would have resolved a substantive volume of the GBIF Index: the growth in occurrences with ‘unknown’ status at the species rank would have grown of only 2.3% (from 23,015,905 to 23,535,346).

It is therefore reasonable to extrapolate that: a large portion of the gaps identified in Table 4.b will in the future be resolved with newest versions of the taxonomical authorities used to build the GBIF taxonomic backbone. The rate of resolved names should in principle directly be correlated with the growth in volume of the taxonomic authoritative references used by GBIF.

Table 6.a provides a summary of the taxonomical misidentification at the Kingdom level and an indication of the total number of associated occurrences affected. For example, correcting the wrong assignment of 90 species from the Kingdom Plantae to Animalia will impact more than 1.3 million occurrences within the GBIF Index as of February 2012. On the other hand correction of the wrong assignments to Animalia of 26 species will only affect 1,536 occurrences. Similar breakdowns are provided for Phylum (Table 6.b) and Class (Table 6.c). This table shows that the effort in correcting misidentifications at a high taxonomical rank (e.g. Kingdom) will impact a limited number of occurrences (1.3 million representing less than 0.5% of the GBIF Index)

Only 9.15% of the discovered scientific names in the GBIF network have not been mapped to a taxonomic reference at the species level. Such volume of unknown references includes for example species not yet endorsed by existing authoritative references used to construct the GBIF taxonomy backbone, as well as misidentified or wrongly spelled names. This represents 7.82% of the GBIF Index in terms of volume of occurrences (i.e. 25.3 million occurrences). We have also demonstrated that this volume of unmapped scientific names has grown less than the growth of the GBIF Index: +9.9% (25.3 million in 2012 against 23 million in 2010) while the GBIF Index has grown in the same period of +21% (267 million occurrences in 2010 and 323 million in 2012). The study also demonstrated that compared to the largest authoritative reference - the Catalogue of Life (CoL) – only half (53.47%) of the species names known to GBIF would have been recognized. The other half are mostly names known to other taxonomical references but unknown to CoL.

Geospatial:

During the harvesting and indexing routines, these geo-referenced occurrences are checked in particular for wrong assignments (e.g. when the latitude and longitude information is not corresponding to the country where the occurrence was observed/collected). In the context of this study, we considered geo-referenced occurrences as a record in the GBIF Index with the latitude and

[17] http://en.wikipedia.org/wiki/Zonotrichia

[18] http://dev.gbif.org/issues/browse/CLB-119

longitude within the earth-bounding box (i.e. -90<=latitude<=90 and -180<=longitude<=180). This amounted to 99.67% of all occurrence records where geo-spatial information was provided in the GBIF Index in December 2010- the remainder being 'extra-terrestrial' (Otegui et al., 2009). This includes a substantial number of records being reported as 0.0N, 0.0E and therefore suspicious. This could happen for example when the publisher maps a zero value to the latitude or longitude fields instead of a 'null' value. In order to solve such problems, publishers should in addition to ensuring that the mapping of the fields is appropriate, provide for example the country in which the observation/collection has occurred. This would greatly facilitate the validation of geo-referenced occurrences during the harvesting and indexing routines.

In end of 2010, 18.45% of the mobilised data were not geo-referenced. This percentage was lower (14.1%) in our assessment of February 2012. As shown in Figure 4, the rate of geo-referenced records is increasing over time. Such rate is higher for recent years of collection/observation (e.g. from 1973 the rate is constantly greater than 80% for the February 2012 assessment). For older occurrences the rate of geo-referencing is decreasing substantially. For example, before 1930, the rate of geo-referencing was largely lower than 50% and this can be explained by technology limitations (e.g. absence of GPS), absence or limited data collection standards covering geo-location (e.g. latitude/longitude, or location fields in collection forms) or simply due to the absence of such information in the collection forms.

During the harvesting and indexing procedures, a series of verifications on geospatial fields (e.g. latitude, longitude, country boundaries etc.) are performed enabling for example identification of potential latitude/longitude incorrect assignments. This can be the case for occurrences where longitude and latitude values were swapped; or simply when the longitude value was incorrectly assigned causing for example occurrences originally collected/observed in North America to appear on the Asian continent. In February 2012, we estimated that less than 3.6% of the total georeferenced occurrences are falling in this category. In addition we estimated that occurrences without latitude and longitude information but with information for the locality represent 11.1% of the total GBIF Index. Taking into consideration that GBIFS is not mandated to apply corrections to the original published occurrence records, these records with possibly wrong coordinates are therefore only flagged during the harvesting and indexing routines. These occurrences aren't displayed on maps through the GBIF data portal but the original occurrences records are kept intact. Users of the GBIF Index (e.g. scientists) should be aware of this limitation and ensure that they consult the 'geospatial issue' flag provided by the GBIFS. While this addresses partly the problem, it is important to note that the verification and correction of the original occurrences records lies with the publishers. The availability of better guidelines[19] and practices in recording biodiversity observations/specimens should support publishers in this effort. In addition, the use of tools like BioGeomancer and GEOLocate should be recommended and more widely promoted. Such tools parse place name descriptions in multiple languages and provide in return a set of longitude/latitude coordinates associated with that description. The data curators can therefore enrich their database content, increase the quality and accuracy of the content mobilised through GBIF and thus makes it suitable for wider uses.

The high percentage of georeferenced records within the GBIF Index as well as the observed positive improvements in our two assessments is an important quality stamp of the GBIF mobilised data.

The study showed that the rate of geo-referencing in the GBIF-Index is increasing over time due mostly to better data quality checking

[19] Principles of Data Quality - Arthur Chapman http://www.gbif.org/orc/?doc_id=1229

activities both at the publisher and central levels (Figure 4). The improvements are observed for all decades since 1900. However, the variability is very high and older occurrences are expected to have a lower probability of valid geospatial information (e.g. prior to 1930: <50%, 1930 – 1960: <70%, 1980-2010: >80%). More importantly, our study shows that the percentage of potential wrong geo-spatial records is very low (<3.6%). In most cases, such situation can be explained by wrong latitude or longitude sign assignments and these can be easily resolved by swapping coordinates.

Temporal:

As detailed in Table 7, 30.8% of the GBIF Index contains records with null or not valid year in the date time stamp field. The UNZYEC analysis (Table 8) estimated a similar percentage (31%) although it distinguished invalid years (i.e. before 1750 or in the future). The breakdown provided in this analysis shows that 4.3% had not valid date stamp data and 26.7% had no data or null values. However, the comparison between raw data and processed data uncovered some issues on date processing, such as mismatches between the published and interpreted date stamp. For example, 8.6% of the records with a value in the date stamp field were nullified during the harvesting and indexing procedures. In addition, 5.0% of the null values in the publisher data were converted to valid date stamp values after harvesting and indexing. More details about this mismatch can be found in Otegui & al., 2013 (this volume). Thus, according to the UNZYEC study (Table 8), 36.1% of the records would be either undated or doubtfully dated at the year level. These analyses show that a large volume of date stamp information does not convert to a valid date stamp, or at least lacks information about the year of collection/observation. In 2011, these preliminary findings were taken into account by the GBIFS and existing processes to interpret date stamp at the publisher level were reviewed and improved. Table 7 shows also the comparison between the assessment made in December 2010 and February 2012. While during this period of time, the GBIF Index has grown by 21% in total (respectively 267 in December 2010 to 324 million occurrences in February 2012), the total number of occurrences with no year provided in the final GBIF Index has decreased by 47.9%. This amounted to 13.3% of the total GBIF Index compared to 30.8% in December 2010. Most of these improvements relate to improved interpretation of malformed data stamp information in the published resources during the harvesting and indexing routines.

Temporal information is useful for two classes of questions: 1) biogeographic changes over time and 2) phenological. The year information is the most important element within temporal date stamp information to note long-term changes. However, the month and day elements provide additional accuracy in particular when looking at migratory species moving for example from feeding to reproduction areas during the same year. Partial date, as found on many older specimens may be useful for one or the other of these purposes even if they cannot serve all needs. Such gaps in the temporal attributes are a limitation for certain types of analysis, such as population cycles or changes in migration patterns related to climate change.

Alone, the low percentage of occurrence records without temporal information (13.2%) is not considered as a major limitation. However, combined with other parameters like geo-referencing, it could become a serious limitation for scientists in particular when dealing with analyses requiring the combination of these (e.g. ecological niche modelling). As shown in Table 9, if we consider only presence of valid temporal and geospatial information as determinants of ‘fitness for use’ in the context of ecological niche modelling analysis, 78.8% of the GBIF Index is meeting these criteria. This total represents 484,963 (48.6%) species from the total identified in the full GBIF Index of 995.974 species as of February 2012. But this also indicates that 51.4% of the species recorded in the GBIF Index don’t have a single information on the temporal*geospatial dimensions.

Background 'noise':

In December 2010, we estimated that 121.7 million records had missing, doubtful or wrong information in at least one of the three key attributes (i.e. taxonomy, georeferencing and temporal). This represented 45.6% of the GBIF-mobilised data records (267 million records). Although this was an improved figure compared to the 50.1% calculated in May 2008, it calls for concerted efforts firstly to sensitise data publishers of the need to ensure that all available temporal, taxonomical and geospatial information are correctly mapped during the publishing process to GBIF. In 2011, GBIFS had greatly improved its harvesting and indexing processes in order to optimize its ability to interpret as accurately as possible the information of publishers. In February 2012, the taxonomical backbone was greatly improved and the indexing processes fine-tuned. This has led to a lower percentage (21.3%) of the GBIF Index with absence of information in at least one of the three variables: temporal, taxonomical and geospatial. While these data quality trends are promising (Figure 5), they are mostly due to technical improvements in the GBIF IT infrastructure and much more efforts are required at the level of the data publishers within the GBIF community.

Collection curators should be encouraged to explore ways to improve the quality of the published information in particular for three dimensions, namely: taxonomical, temporal and geospatial. Many tools are aimed at helping curators to identify possible errors and to standardise data in accordance with authoritative references. Some key examples are:

- SpeciesLink developed by CRIA (Centro de Referência em Informação Ambiental) available at: http://splink.cria.org.br/
- BioGeomancer coordinated by the University of California at Berkeley (http://www.biogeomancer.org)
- DIVA-GIS developed by Robert Hijmans (http://www.diva-gis.org/)
- BIDDSAT developed at UNZYEC (Otegui & Ariño, 2012) available at: http://www.unav.es/unzyec/mzna/biddsat/

Duplicates

Concerns about the amount of record duplicates in the GBIF Index were also raised over recent years (Hobern, 2003; Page, 2012). Such situation could happen for example when the same dataset is published more than one time through GBIF. Comparing datasets on criteria like taxonomy, temporal and geospatial information can easily identify these cases. To assess these cases, we assumed that a duplicate record would be identified when the values respectively for taxonomical (species id), temporal (timestamp date e.g. YYYYMMDD) and geospatial (latitude and longitude) are identical. Based on this assumption, we calculated in February 2012 the total amount of potential duplicates between resources. The results are summarized in Table 10. We have identified 42 combinations of datasets with at least 100,000 potential duplicate occurrences representing a total of more than 30 million occurrences. This represents more than 9.5% of the GBIF Index. The top 20 potential duplicate combinations are listed in Table 11. In all cases (e.g. INBio, CNIN/Lepidoptera, Pelagic Fish Observations 1968-1999, Birds (KIEE-BI)) it appeared that the resources were republished twice to GBIF but with a different name (e.g. 'Pelagic Fish Observations 1968-1999' and 'Pelagic Fish Observations 1968-1999 (Australian Antarctic Data Centre)'). What appears very surprising is that most of these potentially duplicated resources were registered with very similar names (e.g. 'CNIM/Lepidoptera' and 'Colección de Referencia de Lepidópteros Diurnos Mexicanos de la CNIN'). When a new resource is registered, a simple text comparison between the title of the new resource with existing published ones would have enabled rapid identification of obvious possible duplication. This has never been implemented up to now in GBIF but efforts are underway to automate this process as well as to resolve the already identified potential duplicates in close communication with the respective GBIF publishers.

An improved monitoring of the resource at the time of registration is indeed an immediate solution but ultimately the adoption of persistent identifiers for each resource published (e.g. DOI), with proper metadata, would have been a much more robust solution.

B. Trends and patterns assessment

Taxonomy:

In December 2010, of the 267 million occurrences records accessible through the GBIF network 62% belonged to Kingdom Animalia, followed by Kingdom Plantae (23%), Fungi (1.55%), Protozoa (0.67%), and Bacteria (0.59%) (Figure 6.a). A similar assessment in February 2012 (Figure 6.b) showed that the major variation was the increase for the Plantae from 23% to 30%. Between these two assessments the GBIF taxonomical backbone was reviewed with the latest version of the Catalogue of Life. Monitoring of the taxonomical name resolution during the GBIF harvesting and indexing procedures has shown that a large proportion of names previously classified as 'Unranked/Unknown' were now reclassified in particular within the Kingdom Plantae (GBIF, personal communication).

In December 2010, as depicted in Figure 6.a, 52% of the occurrences belonged to Phylum Chordata (Kingdom Animalia) followed by 17.7% belonging to Phylum Magnoliophyta (Kingdom: Plantae), and 9.8% to Phylum Arthropoda (Kingdom Animalia). A breakdown at the Class rank (Figure 8.a, Figure 9) shows that the largest Class in the GBIF mobilised data is Aves (43%). This is mostly due to field observation from the ornithological community as depicted in Figure 10. In December 2010, the Bird Observation Checklist database represented 42.21 million occurrences or 15.7% of the total GBIF Index at that time. Within this top 5, four resources are related to bird watching activities (e.g. Bird Observation Checklist database, Project FeederWatch, Great Backyard Bird Count, Southern African Bird Atlas Project). While these figures clearly indicate the dominance of bird observations among the data accessible through the GBIF network, it also demonstrate the effectiveness of a given specialized network to leverage on the existence of GBIF to enable the publishing, discovery and access to such type of biodiversity observations.

These figures also show that the spread of occurrences across various taxonomical levels is also rather heterogeneous (Figure 7.a). Some phyla are extremely underrepresented, while specific classes such as Aves dominate, or even orders within the class Hexapoda (insects) (Figure 8.a and b, Figure 9). The hierarchy of the most represented groups (irrespective of taxonomic level) shows Classes Aves, Actinopterygii (bony fishes), Poales (grasses), Mammals and Asterales as the largest groups, followed by Order Lepidoptera within the Class Hexapoda. On the other extreme, for example Phyla Zygomycota, Nemata or Platyhelminthes, or Kingdom Bacteria, have marginal occurrence despite their natural abundance. However it is important to note that many of these apparently overrepresented taxa are species-rich, and have greater biomass and greater visibility, but also a higher number of competent specialists and observers. Having such large amount of data for a relatively small number of taxa should also be considered as a positive asset in particular when looking at temporal species distribution, provided that these taxa are ecologically diverse as well as representative. The availability of such high-density information for fewer taxa should not be under-estimated.

The analysis of the temporal spread of GBIF-mediated data for the two dominant Kingdoms (Animalia and Plantae, Figure 11) shows that the exponential increase observed from 1960 is mostly explained by the abundance of occurrences for the Kingdom Animalia. This increase of bird observation data records exceeds the mobilised data from all other classes from year 2000 onwards. In the same period (1960-2010), we also observed that beside a peak in 1999, the trend for Plantae is stable varying from 1.5 to 2.1 million occurrences observed/collected per year. As shown in Figure 13, the exponential increase of data records in the GBIF Index in recent years is largely explained by the growth of occurrences in the class Aves. Figure 14 provides a breakdown of occurrence records by basis of record within the class Aves. Since 1960, bird observation data have been growing almost exponentially while the trend remains stable for specimen and other types of data. If these trends are confirmed in upcoming years, it is expected that the growth of data records in Aves will be the main driver behind the growth of GBIF Index in terms of volume.

This phenomenon is even more revealing when listing the top 15 species by the number of data records. Tables 12.a and 12.b show that all of the top 15 species are birds, mostly published through networks like ‘eBird Bird Observation Checklist’, or other similar resources (e.g. Project FeederWatch, Great Backyard Bird Count).

In order to demonstrate the difference between the Kingdom Animalia and Plantae, we have generated two sub-indexes for each Kingdom from the February, 2012 version of the GBIF Index. Each sub-index was subdivided in new subsets based on the range of occurrence numbers for each species. Table 13 provides the summary of the results. For example, from the total of 457,340 Animalia species in the GBIF Index, 400,088 are species with less than 100 occurrences each, and represent 2.4% of the total number of occurrences in the Kingdom Animalia. The breakdown of species by occurrences did not show any major differences between the two Kingdoms except for 20 species in the Kingdom Animalia holding more than 1 million occurrences each, while no species had as many occurrences within the Kingdom Plantae. However, occurrences themselves diverged between Kingdoms. The set of 20 species having more than 1 million occurrences each identified in the Animalia Kingdom accounted for 15.9% of all Animalia occurrences (zero for Plantae), and for the species in the range 100,000-1 million occurrences a higher percentage was also observed for Animalia (39.8%). Plantae occurrences concentrated around species represented each by less than 100,000 occurrences. We conclude that the abundance of occurrences records in the GBIF Index for a few Animalia species is representing a significant portion of the full GBIF Index. These records are mostly represented by bird observation data. However this trend shouldn’t under-estimate the amount of species from all Kingdoms having less than 1 million and more than 1,000 occurrences, since these do represent a large portion of the GBIF Index (74.5% of Animalia and 78.7% of Plantae).

We have compared the distributions of the year of collection/observation of occurrences for both Plantae and Animalia Kingdoms (see Figure 12.a. and 12.b) taking into consideration the December 2010 and February 2012 versions of the GBIF Index. Both figures show that over time the rate of data mobilised per year tend to increase in both cases. However, for Plantae (Figure 12.a) we observed that, aside a few artefacts (e.g. year 1999) the rate of mobilisation is increasing at a slower rate to even stagnate from year 2000 compared to the Animalia Kingdom (Figure 12.b). On the other hand, we observed that the evolution for the Animalia Kingdom was approximately exponential in both versions of the GBIF Index. As indicated previously, this is attributed to the increased proportion of bird data in the GBIF Index in particular in the last decade, as shown in Figure 13. This confirms the fact that the rapid growth of the volume of occurrences in the GBIF Index is mostly driven by the bird observation data. The spread of other large publishers is perhaps wider, the main difference being the concentration of bird data towards recent years and few publishers (see Otegui & Ariño, 2013).

The value of observational data in comparison to voucher specimen in museums or accessions stored in genebanks is a subject for another discussion. However this study (Figure 14) demonstrates the over-representation of observational occurrences in the GBIF Index. The ratio between observation and specimen was very close to 1:1 until 1975. Thenceforth, the amount of observation occurrences has grown exponentially while the trend for specimen data was very much stagnating until 2000, where we observed a decline. In the last decade, the proportion of observation occurrences represented more than 90% of the yearly collected/observed occurrences.

The dominance of bird observational data in the last decades, as well as the drop for data mobilized in recent years for other classes during the last decades, is cause of concern. While on one hand the availability of such large volume of bird data will enable advanced research in temporal trends of bird populations, it also reveal the difficulty to undertake such valuable research in other classes. Part of this can be explained by rapid data mobilisation of the “low-hanging fruits” (or relatively easy to digitise and publish) vouchered specimen data (Berendsohn et. al., 2010). Many of the large natural history museums have digitised their main historical records (Ariño, 2010) and published them through GBIF. It is therefore expected if this situation of dichotomy between bird observation data and the other classes will increase in the next years.

Taking into consideration the existing major threats to biodiversity, the GBIF community needs to greatly strengthen its capacity to assess trends also for all non-bird biodiversity records. For example, GBIF could evaluate the opportunity to develop a list of priority species based on known references, for example the IUCN Red List; gather information about their distribution, and evaluate for each the availability of rich yet still undigitized or electronically unavailable occurrence data in the GBIF community. This approach would lead to a series of strategic data mobilisation strategies for each priority species.

***Geospatial*:**

In the December 2010 assessment we observed (Table 14) that the majority of the occurrences present in the GBIF Index were located in Northern America (28.19%), Northern (30.06%) and Western (11.48%) Europe. This represents a total of 69.73% of the GBIF Index. In February 2012, we observed the same trend where these three regions represented 70.8% of the GBIF Index, with minor variations in the order (*e.g.* Northern America was classified as the second region in 2010 while it became first in the 2012 assessment). There are multiple reasons that can explain this distribution.

The comparison between existing financial contributions to the GBIF Secretariat (Table 15, as of year 2010) on a regional basis shows that the sum of the contributions of these three regions equals 64.9% of the total GBIF operational budget, which is very similar to the percentage of occurrences collected/observed in these regions (69.73%). The major discrepancy observed in this table is the financial contribution of Eastern Asia countries (22.1%) for only 2% of the occurrences in the GBIF Index. This can be explained by the contribution of Japan within a region where the rate of data mobilisation is still low.

In 1999, The OECD Biodiversity Informatics Subgroup in its Working Group on Biological Informatics report made major recommendations for the establishment of GBIF. It is therefore not surprising today to observe (Table 16) that the majority of the occurrences in the GBIF Index are located in OECD countries (84.45%). Taking into consideration megadiverse countries, large countries like Brazil, China, Democratic Republic of Congo, Indonesia, Malaysia, Papua New Guinea or Venezuela are not yet members of GBIF in 2012 making it difficult for information from these countries to be published through GBIF. Thus, GBIF mobilised data are very much biased towards it original founders, mostly OECD countries. A clear example can be found in Otegui et al. (2009), where the geospatially-explicit provenance of data contributed by European publishers in the 2008 sample nicely matches the publisher' country (Figure 19).

As shown on Table 17, not surprisingly the majority (85%) of the occurrences were located in high-income countries, 11.7% in upper-middle income countries and less than 4% in lower-middle and low income countries.

The distribution of occurrences along latitude (Figure 16) confirmed also the large proportion of occurrences located in northern hemisphere, where the three regions contributing most records are located (Table 14). The peak observed in the southern hemisphere is mostly explained by the recent publication of a large volume of occurrences from South Africa, Australia, and in particular through the Atlas of Living Australia. However, the species richness, as measured by density of species per half-degree of latitude (Figure 17), showed a slightly different trend. We did not observe the large dichotomy between the two hemispheres that appeared in the density of occurrences, and the species richness ranged from 10,000 to 40,000 species per half-degree. Figure 18 provides a justification for these trends. The average number of occurrences in the southern hemisphere did not exceed 35 occurrences per species at that latitude range resolution, while this rate exceeded 50 for much of the latitudes north of 50ºN, and even reached peaks higher than 160 occurrences per species per half-degree. We therefore conclude that despite a bias of occurrences towards northern hemisphere, the species richness observed is equally distributed between hemispheres. We also conclude that species in northern hemisphere had a higher rate of occurrences/species than in southern hemisphere. This can suggest a wider distribution of temporal data for these species in the northern hemisphere, and therefore the availability of information more suitable for studying the temporal trends of species distribution in these regions. For the southern hemisphere, we also conclude that many species may not have sufficient occurrences to perform such analysis.

Temporal:

The temporal evolution in the GBIF Index is summarized in Figure 20. With the exception of a few artefacts (1950, and 1987 for the December 2010 curve), we observed that the availability of occurrence data over time grew almost exponentially. A striking feature in this trend was the presence of large peaks in specific years. These peaks seemed to respond to a combination of a provider effect and a possible mismatch between published data and indexed data arising from the date processing algorithms, that is explained in detail in Otegui et al., 2013 (this volume). The drop observed in the last period (between 2007 and 2010) for the December 2010 assessment can be attributed to the lag time required between the data collection/observation, digitization and publishing through GBIF. The same lag time (3 years) was later confirmed for the February assessment. We conclude that the amount of biodiversity data collected or observed tends to be greater for more recent years than for any older period (e.g. prior to 1970-1980). We also analysed the evolution of such trend by comparing the December 2010 and February 2012 assessments (Figure 21). The two horizontal lines represent the average growth in the GBIF Index for all occurrences and for occurrences having temporal information. The difference can be explained by two factors: (1) the improvement of the GBIF indexing processes in 2011, which enabled greater recovery of malformed date-stamp fields; and (2) the greater percentage of well-formed temporal fields (*e.g.* date of collection/observation) in the recently published data. The graph shows that for more recent decades (*e.g.* 1971-1980 onward) the growth of data in the GBIF Index is faster than for older data. More remarkably, we observe that for the latest decade (2000-2010) the variation is of 89.6%, which is the highest growth rate ever observed. The exponential growth of recent data in the GBIF network content is particularly driven by the availability of bird observational data during the last decade; this growth in recent content is sometimes termed a 'data deluge'.

The trends for the number of species collected/observed every year since 1900 (Figure 22) for both Plantae and Animalia were very similar. We observed an increase until the 1990's, and then stagnation followed by a drop from year 200 (with the exception of few artefacts). The drops observed in 1914 to 1917 as well as from 1939 to 1942 can be easily explained by the effect of the two world wars. What is troublesome, however, is the drop in both curves from 2000 onwards. The drop for the Animalia is even more severe than for Plantae. While the volume of occurrences mobilized every year is increasing until 2009, we note that at the same time these occurrences belonged to fewer species across both Kingdoms. One possible explanation could have been related to a lower number of data resources publishing since year 2000 (Figure 23), but it should be noted that the decline in species richness started more than one decade earlier. We have also calculated for each year the rate of geospatial occupancy in a grid with a resolution of half-degree (Figure 24). We observed that in all cases the grid occupancy for Animalia species was higher than for Plants. From 1963 to 1993, grid occupancy for Animalia was stable followed by a peak in 2000. For Plantae, grid occupancy was stable from 1970 until 2000. We also noted that in both cases grid occupancy started to decrease in 2000.

A detailed analysis of these trends will be further presented in a separate study.

C- 'Fitness-for-use' assessment:

Assessing the value of the GBIF mobilised data for a variety of usages is challenging. In this study, we decided to focus on the most common uses for GBIF-mobilised data reported in the scientific literature: ecological niche modelling (ENM) (Grinnell, 1917; Fernández et al., 2009; Peterson and Vieglais, 2001) and related analyses. The compilation of scientific literature using or citing GBIF is available since 2011 on-line at: http://www.mendeley.com/groups/1068301/gbif-public-library/. Such modelling techniques (*e.g.* using Maxent) required occurrence records with proper temporal attributes, correct geo-referencing attributes as well as sufficient volume of well-distributed data-points. The minimum number of distinct data-points for a niche modelling analysis is in the range of 10 to 20 (Pearson et al., 2007; Grantham et al. 2008). Recent studies on GBIF-mediated data using more than 19,000 plant species showed that a preferred threshold of 20 to 40 points is recommended (Jarvis, Personal communication). Maxent models generated for species meeting these criteria have an area under the curve (AUC) greater than 0.75 in more than 95% of the cases.

If time series are part of the models, then the requirements on number of data points can be an order of magnitude higher. For example, Ariño and Pimm (1995) showed that successful modelling of the evolution of population extremes require a minimum of 15 distinct time-dependent population estimates. In terms of ENM, it could be argued that if using cell frequencies in the ENM as an indicator of potential population estimates, at least 15 independent models, each time-constrained, should be needed to adequately characterize any time-dependent changes in the model. This may hold for both terrestrial and marine models, despite their intrinsic differences (Warner et al., 1995).

In this study we have decided to use the threshold of "presence in at least 20 distinct cells in a 1/10 degree grid" to define whether a species has sufficient occurrences in the GBIF Index to be used for ecological niche modelling. We used this threshold for temporal/spatial requirements to assess the number of species suitable for such ecological niche modelling analysis, but make no attempt yet to assess whether each selected species can be adequately modelled over time.

Tables 18.a, b and c provide a distribution of the number of species falling in various categories of grid occupancy. Our analysis was based on the February 2012 version of the GBIF Index due to the improved accuracy of the taxonomical matching, the greater resolution of date stamp as well as for geo-referencing attributes. For the full GBIF Index, more than 995,975 species (Table 18.a) were recorded in the GBIF Index (with at least one occurrence record). However only 603,532 species had at least one occurrence present in the GBIF Index with at least one presence in a distinct 1/10-degree grid. This means that 39.4% of the species recorded had no geo-referenced attributes. 747,988 species had at least one occurrence with a valid temporal attribute. This number dropped to 485,105 species if we added the condition of at least one geo-referenced attribute within a 1/10-degree grid. If we consider the ENM threshold of 20 presences in 1/10-degree grid with valid temporal attributes, the total number of species that were suitable for ENM analysis fell to 81,057. This represents 8.1% of the species recorded in the GBIF Index. While this percentage could be interpreted as a low percentage, the number of species falling in this category is already very high for many scientists and researchers interested in estimating the actual species distribution as well a projections in the future taking into account future climatic scenario. The use of such information is extremely valuable already for advanced scientific research and in particular in support of global biodiversity assessments such as the Strategic Plan on Biodiversity of the MEAs (also called Aichi Targets). Taking into account the constant growth of the GBIF Index with the addition of new datasets for example, it is logical to expect that this amount of ‘eligible’ species will increase over time.

What was also remarkable was the number of species with a presence in at least 100 1/10-degree cells with valid temporal attribute: 14,041. This rich reservoir of species with high quality occurrences is already an important message to the research community seeking to assess the species distribution evolution over time as well as future predictions. As shown in Table 18.b and c, this number is somehow equally distributed between species within the two dominant Kingdoms: Plantae (6,100) and Animalia (6,756). Figure 24 also shows the temporal trends between these two Kingdoms. Even if the grid occupancy is constantly higher on a yearly basis for Animalia, the trends between these two Kingdoms are very similar. The same trends are also observed for a low threshold of 20 1/10-degree grid presences. We also noted no major differences between Kingdoms in the breakdown assessment of the Animalia (Table 18.b) and Plantae (Table 18.c): 36,462 Animalia species were suitable for ENM (7.9% of the total number of Animalia species recorded in the GBIF Index) against 37,730 Plantae species (8.7%). Therefore the concerns about the over-representation of bird observation data within the Kingdom Animalia are contradicted here in terms of ‘fitness for use’ since a large number of plant species were already meeting the ENM suitability criteria. Figure 25 proves how much the two Kingdoms can’t be distinguished when looking at a presence higher than 20 in 1/10-degree grids.

In this study, we have also tried to assess the grid occupancy at Class (Table 19), and Family (Table 20) levels. Our objective was to assess the percentage of species within each rank suitable for ENM. For example, within the class Aves 45.2% of the recorded species in the GBIF Index were already suitable for ENM studies. In this particular case, we can conclude that the GBIF Index as of today can be used to estimate the biodiversity of most species within the class Aves. Taking into account the actual trend in terms of bird observation data, it is therefore expected that this percentage will grow in the future. Such volume of information can now open new opportunities such as studies on the over-sampled areas (e.g. North America or Western Europe) and recommendations for new areas where collection of new specimen/observation is required. Studies on the estimated number of species (Chao2) at a regional and global level could now be performed. Table 19 also shows that other classes are eligible for such analysis: cartilaginous fish (Elasmobranchii, Holocephali), bryophyte plants (Marchantiophyta). For many, the limited number of species within that Class can explain this. However, for some larger classes like Actinopterygii (ray-finned fishes), Elasmobranchii (cartilaginous fish) or Pinophyta (conifers) the GBIF Index holds sufficient information for more advanced ENM or other biodiversity assessment analysis. Figure 26 provides a visual representation of these trends. With the exception of the Class Aves, the other Classes shared a similar trend.

Such analysis may thus include a bias: the number of species representing each taxon group at a given level. Large families or classes, *e.g.* Hexapoda (insects), may not be listed in the Top 10 or 20 lists. Figure 27 shows the distribution of Families taking into account the number of species within each Family. While we observed that classes Aves and Actinopterygii were listed as the ones with the highest suitability for large-scale ENM, families within other classes, such as some insects (*e.g.* Cryptophagidae - beetles), or mammals (Phyllostomidae - New World leaf-nosed bats) are also to be considered. For even larger families (*e.g.* with a number of species greater than 1,000) it was not surprising to observe that the percentage of species suitable for ENM was lower. However for such large families, the suitability for ENM of 5-10% of their known species is probably a good proxy to initiate an assessment of the full class. In this category, in addition to families of insects we observe some large families of reptiles (*e.g.* Scincidae – lizards, Colubridae – snakes). For plants (Figure 28), the distribution of families is somehow distorted due to the high to very high number of species found within each class.

Discussion

The idea that birthed GBIF ten years ago remains as simple and powerful now as it was then: to make the world's biodiversity information freely and universally available for science, society and a sustainable future (OECD, 1999). After 10 years of existence, the GBIF network represents the largest resource of primary biodiversity data that is freely accessible to all. With over 377 million occurrence records about nearly one million species (as of August 2012), the GBIF mobilised data provides a data-driven window to the state of the world's biodiversity. Access to such large volume of data opens for example new research avenues from assessing the state of biodiversity, identifying the potential threats up to monitoring trends and predicting future evolution and composition of biodiversity and ecosystems (Rödder & Lötters, 2010; Ramírez-Villegas *et al.*, 2010; Ready *et al.*,2010). Since 2008 till June 2012, over 600 scientific peer reviewed papers have been published which are based on analysis and interpretations of GBIF mediated data (GBIF, 2012a).

To become such a truly 'global biodiversity information facility', GBIF needs now to take into consideration the primary applications it originally intended to offer to the public such as in policy formulation, economic development, environmental protection, education, and scientific research. In order to ensure its relevance for such applications, the GBIF community needs to warrant that the information it delivers is of relevance to address the major science, societal and policy challenges.

While these needs are very diverse and difficult to categorize they do have in common essential pre-requisites that can be summarized as follows:

- "Can I trust the information provided?"
- "Is the information representative of biodiversity on earth?"
- "Can I use the data to model biodiversity over time?"

The present study was therefore aimed at assessing the data quality, bias and 'fitness-for-use' of the GBIF mobilized content. These challenging questions were addressed by tasking two separate teams to evaluate the content using different methodologies. The results from the GBIFS and UNZYEC teams were similar and they both demonstrated the validity of the conclusions presented hereby.

Are the GBIF-mediated data scientifically credible/reliable?

GBIF mobilised data is often criticized for errors (Yesson et al., 2007; Otegui et al., 2009). However, these errors are reflections of the data as collected, collated, and published by the heterogeneous data publishers across the globe. The role of GBIF is to provide a discovery window on the published data. Such a role requires reconciling, interpreting and publishing the essential key attributes: taxonomic, temporal and geospatial. In assessing the state of data quality in the GBIF Index over time, inevitably such study will combine data quality improvements at the level of the data publishers as well as at the central discovery point.

The recent improvements made by GBIF in the re-building of its taxonomical backbone and data quality checking routines have positively impacted on the level of data quality in the GBIF Index. However these improvements are explained by the improvements of the informatics infrastructure and processing algorithms, but these are not addressing the most critical underlying causes of poor data quality: accuracy and gaps. Informatics routines alone can eventually spot but cannot recover missing attributes,(if anything, perhaps hint or guess), in particular when these attributes were not mapped correctly at the publisher level or if they weren't even digitized from the original voucher specimen.

Taxonomy:

A majority of the scientific names published by the GBIF network are now recognized as valid references against a collection of authoritative taxonomic catalogues. The current GBIF taxonomy backbone provides an appropriate resolution service to the large majority of the scientific names discovered by GBIF. Given the fact that GBIF taxonomic backbone is a combination of multiple authoritative taxonomic catalogues (e.g. CoL, WORMS, IPNI, NCBI, and ITIS etc.), it has potential to serve larger systematicians communities than any specific taxonomic group alone. While doing so, informatics approaches are proven to be effective; however, questions about future improvements can be raised. Linkages with more authoritative taxonomic catalogues (Recommendation 4 in Faith *et al*., 2013) and involvement of taxonomic expertise will soon be required to resolve taxonomic discrepancies.

In order to continuously assess the effectiveness of its taxonomic backbone, GBIF Secretariat should perform regular estimation of completeness at all taxonomic ranks as described in Table 4.b. Such an analysis should in particular assess the amount of mis-identifications (*e.g.* species within genus *Zonotrichia)*. GBIF should also improve its reporting services to the original publishers so that potential taxonomic mis-identifications are reported (Recommendation 6 in Faith *et al.*, 2013). GBIF should also monitor over time the taxonomic data quality improvements made in the GBIF Index (*e.g.* indicators of taxonomic completeness at the class, order or family levels). In addition, GBIF should provide means to assess the effectiveness of its taxonomic names resolution services used during the harvesting and indexing processes. All taxa mis-identifications should be documented and calls to expert groups (*e.g.* marine biologists, crop wild relatives experts) should be considered in order to tap into taxonomist expertise and increase their engagements (Chavan *et al.*, 2005) in improving the quality of such valuable global resource.

Temporal and Geospatial:

Setting an ideal target for the rate of geo-referenced occurrences within the GBIF index is a difficult task. While the ideal scenario would be that all occurrences are georeferenced, the reality is that in many cases the original records for example specimens in zoological or botanical collections itself won't have such information. Some voucher specimens (especially older ones) have in general a lower percentage of geo-referenced records compared to recent field observation records. There is a high variability between data resources within the GBIF Index. However and as shown in

Figure 4, the average percentage of georeferenced records has increased between 1-5% in average during the period 1990-2010 and is consistently higher in February 2012 than in December 2010. The data publisher community is therefore addressing this challenge in particular for recent records. While GBIF's role is to enable the discovery of primary biodiversity data from a network of publishers (Recommendation 14, Faith *et al.,* 2013), it is not mandated to undertake or correct the content published. However, this can be questioned in particular when a simple correction, such as a sign correction on a longitude or latitude field, could be undertaken within the GBIF Index and therefore immediately improve the quality of data published. Taking into consideration the growing difficulties in communicating with a large network of publishers, such option may be considered for the most obvious data corrections. While this can be seen as a limitation, one way forward would be to set targets by periods where we observe low variation of the geo-referencing average (e.g. 1900-1930, 1930-1960, 1960-1990 and 1990-today). Within each period, a georeferencing target could be set based on a subset of data resources (*e.g.* comparing all datasets publishing insect occurrences against the top 10% best georeferenced datasets). However, any decision on such baselines would need to be discussed and agreed with the community of publishers. Experts could investigate datasets falling well under these baselines and reports with recommendations on possible corrections should be sent to the original publishers. However, this approach would require engagement from expert groups as well as willingness and availability of data owners to undertake more accurate verifications such as getting back to the original voucher specimen (Recommendation 3, Faith *et al.,* 2013).

As shown in Figure 5, we demonstrated that in December 2010 approximately 50% of the occurrences in GBIF Index had at least one of the taxonomic, temporal and/or geospatial attributes missing; this percentage dropped to less than 22% by February 2012. This means that occurrence records with essential attributes represent now more than three quarters of the GBIF Index. Taking into consideration that more recent occurrences tend to be of such quality, it is expected that over time this percentage will continue to increase.

The most critical priority for the GBIF network in this field is now to engage the data publisher community at large (including data curators and original collectors) to be (1) aware of the importance of data quality and accuracy; (2) alerted of the possible data gaps and/or quality issues identified centrally; and (3) investigate and fix these whenever possible (*e.g.* by checking the data publishing process up to involving the original curators and specimen) (Recommendation 6, Faith *et al.,* 2013).

To achieve this, a distributed annotation service will be required whereby reports on possible data quality issues are communicated to the original publishers. However such service would in turn require the promotion of effective identification of data objects such as persistent identifiers and sustainable resolution services (Recommendation 13, Faith *et al.,* 2013). GBIF should therefore place the use and re-use of persistent identifiers as a high priority activity and possibly as mandatory for all datasets (GBIF, 2009).

Are the GBIF-mediated data increasingly representative of "some" biodiversity on earth?

The GBIF Index has recorded information about 995,974 species, which is a remarkable amount compared for example with the Catalogue of Life, which contains, as of June 2012, more than 1.3 million species. However, the GBIF Index is facing a bias toward the Kingdom Animalia and to a less extent towards Plants. Other Kingdoms like Fungi, Protozoa and even Bacteria are under-represented within the data mobilized so far. Therefore, the GBIF Index is not representative of all Kingdoms and can't be used yet as a proxy to all biodiversity on earth (Recommendations 1 & 2, Faith *et al*., 2013).

The over-representation of bird observation data should not be considered as problematic, as the “over-“ bit means just by comparison to other groups. Our study shows that the bird observation community has managed in particular over the last two decades to mobilize a vast amount of information on a number of species. Such volume is remarkable and of great value to understand not only the distribution of species at a given time but also on a temporal basis. This is, for example, of immense value when dealing with monitoring potentially invasive alien species or reaction to climate change over time (Sullivan et al., 2009). Citizen scientists are using such bird observation network (*e.g.* eBird) data to monitor the biological patterns and the environmental and anthropogenic factors that influence them. These networks are providing today a near real-time observational network, a model to be followed by many other networks.

However, what is of greater concern to the authors is the flat data mobilization rate since the 1990’s for classes other than Aves (Figure 13), a phenomenon masked by the approximately exponential growth of bird observation data, or even more critically (Figure 14) by the exponential growth of the observation/specimen ratio. Taking into consideration the greater intrinsic value of specimen versus observation data (*e.g.* accuracy, taxonomic validation, validation by experts, and availability of voucher specimen for verification), the stagnation of such valuable resources over the last 2 decades is a priority that needs to be addressed by GBIF (Recommendation 3, Faith *et al*., 2013). Instead of focusing on a volume target, such as the one- and two-billion-records goals as set by the GBIF Governing Board in 2007 and 2009 respectively GBIF (GBIF, 2008, 2012b), GBIF should instead focus on an optimal distribution of such volume across Kingdoms, Classes, Orders, Families and Genera (Recommendations 1, 2 & 8, Faith *et al*., 2013). It is unrealistic to hope that the GBIF network will manage within the next decade to mobilize as many data for all classes as what has been mobilized so far for birds.

The volume of records has always been an easy and tempting target. We argue here that this is not an appropriate indicator of the success of GBIF as being a window on earth biodiversity. We demonstrated in our study that even though the volume of occurrences has a clear bias towards the northern hemisphere (Figure 18: 70% for Northern America, Western and Northern Europe - Table 15), in terms of species richness (Figures 17) we do not observe such bias. Therefore, even with fewer occurrences per species, the southern hemisphere shows to have a similar amount of species richness in the GBIF Index than the northern hemisphere. This can be explained by the recent addition of species-rich datasets from South Africa (SANBI), Australia (Atlas of Living Australia), and Costa Rica (InBIO).

Therefore if GBIF needs to become a window on earth biodiversity, and taking into account the stagnation of specimen records and the exponential growth of observational data, one way forward would be to engage non-bird observation networks (*e.g.* flowering plants, snails, fish, butterflies etc.) to participate as actively as the bird observation community. Because, the additions of biodiversity-rich datasets with fewer records per species can make a large difference in the representativeness of GBIF Index of the earth biodiversity distribution (Recommendations 1, 2, 3 & 8, Faith *et al*., 2013].

Another complementary strategy would be to focus on priority species derived from key scientific-policy priorities (e.g. reducing threats caused by invasive alien species, reducing the loss of threatened species etc.) and priority regions/areas (e.g. biodiversity hotspots, protected areas, high biodiversity regions) (Recommendation 8, Faith *et al*., 2013).

Therefore we conclude that if GBIF wants to become the main window on earth biodiversity it needs to articulate data mobilization strategies engaging the full GBIF network on a list of priority species and regions (Recommendation 2, Faith *et al*., 2013). To decide on these two major priorities, GBIF needs to undertake a more advanced data gap analysis looking at what biodiversity needs to be monitored and in which locations (Recommendations 1 & 10, Faith *et al.,* 2013). This would be somehow a radical shift from the former 'opportunistic' to a more pragmatic demand-driven approach (Berents *et al*., 2010). This shift in strategy may however have some serious financial implications since most of the high biodiversity regions in the world (*e.g.* Amazonia, Tropical Africa, etc.) are difficult of access or even dangerous (*e.g*. war zones or unstable regions).

Are the GBIF-mediated data opening new opportunities for the scientific communities to assess the state of biodiversity as well as the pressure it faces and its response?

Our study on the 'fitness-for-use' of GBIF-mediated data can be summarized in Figure 29. From a large volume of 323 million occurrences and 995 thousand species, the GBIF Index can be synthetized to a smaller volume of information (71.4 million occurrences) covering less than 50% of the known species in GBIF. This first filter of the GBIF Index is based primarily on the availability of valid taxonomical references, temporal and geospatial elements.

Less than 200,000 species have sufficient occurrences (requiring a minimum presence in at least 10 distinct 1/10 degree grids) to be used to assess their distribution through ecological niche modelling (ENM) analysis. Still, this represents a large volume of valuable information that can be immediately used to assess for example the status of biodiversity for a group of species within a given ecosystem. Our study also showed that many classes and families already have many species meeting these ENM requirements. The growth of scientific literature using GBIF mediated data in recent years is also an additional indicator that the GBIF mediated data is a valuable resource.

Therefore we concluded that the GBIF Index is a valuable resource that can already be used by the scientific community to assess the status of biodiversity at least for major groups such as for birds, fish, plants and insects (Rödder & Lötters, 2010; Ramírez-Villegas et al., 2010; Ready et al., 2010). Enhancing the fitness-for-use and trustworthiness of GBIF mobilised data is a natural course of action in this direction, that needs be attained urgently at all levels of data management chain (Recommendation 5 & 6, Faith *et al.,* 2013). We further opine that GBIF as a community needs to proactively advocate the use of GBIF mediated data in scientific analysis, which may result into sound decision making and effective conservation and sustainable uses of biological resources (Recommendation 9, Faith *et al*., 2013). In order to encourage the cross-sectional scientific and naturalist communities in publishing primary biodiversity data through the GBIF network, a comprehensive 'data publishing framework' (Chavan & Ingwersen, 2009; Moritz *et al.,* 2011) needs to be promoted and implemented (Recommendation 11, Faith *et al.,* 2013).

CONCLUSION

The Third Global Biodiversity Outlook (GBO3) published by the CBD in 2010 concluded that the target agreed by the world's governments in 2002, "*to achieve by 2010 a significant reduction of the current rate of biodiversity loss at the global, regional and national level as a contribution to poverty alleviation and to the benefit of all life on Earth*", was not met (SCBD, 2010). The loss of biodiversity is an issue of profound concern for its own sake, but biodiversity also underpins the functioning of ecosystems, which provide a wide range of services to human societies (SCBD, 2010). Its continued loss, therefore, has major implications for current and future human well-being.

The lack of a consistent baseline data and ongoing monitoring of biodiversity has been often cited as a major obstacle towards improving the scientific evidence of the consequences of biodiversity loss. GBIF through its mission is providing a mean to achieve greater improvements in this evidence base in particular through its global network and the discovery, access and use of the largest resources of primary biodiversity data.

This study has clearly demonstrated the great value of the GBIF mediated data in various aspects from improved data quality and accuracy, progressive reduction of gaps in the content and increased fitness-for-use. More importantly, the GBIF mediated data is offering today opportunities to undertake scientific research that has never been possible before. Access to three fundamental and essential biodiversity variables (i.e. taxonomy, geospatial and temporal) opens opportunities for example in assessing today's distribution of species as well as predicting their future distribution. Taking ecological niche modelling as the model application of GBIF-mediated data already opens a myriad of research opportunities, such as assessing the state of biodiversity (*e.g.* threatened species assessments, genetic diversity, etc.) or the pressures (*e.g.* invasive alien species, effect of climate change or land cover change) as well as the responses.

GBIF is therefore uniquely positioned today to become the 'data to science' interface in support of major scientific research trends such as in support of the Strategic Plan on Biodiversity as agreed in Nagoya in 2010 (SCBD, 2012). GBIF must therefore take concrete steps toward effective monitoring of current trends in science and policy, such that it is maximally responsive and effective as a mega-science data infrastructure.

This is how GBIF's focus can remain on being the single most important infrastructure for primary biodiversity data at the organism level, accompanied by strong, effective, and targeted links to data at the genetic, genomic, and ecosystem levels.

ACKNOWLEDGEMENTS

Authors are grateful to their colleagues in the Global Biodiversity Information Facility Secretariat, and University of Navarra for their support in this study and providing access to the GBIF Index.

REFERENCES

Ariño A.H. (2010). Approaches To Estimating The Universe Of Natural History Collections Data. Biodiversity Informatics, 7: 81-92.

Ariño A.H., Otegui J. (2008). Sampling Biodiversity Sampling. Proceedings of TDWG, 2008: 77-78.

Ariño A.H., Pimm S. (1995). On the nature of population extremes. Evolutionary Ecology, 9: 429-443.

Berendsohn W., Chavan V., Macklin J.A. (2010). Recommendations of the GBIF Task Group on the Global Strategy and Action Plan for the Mobilization of Natural History Collections Data. Biodiversity Informatics, 7: 67-71.

Berents P., Hamer M., Chavan V. (2010). Towards demand-driven publishing: approaches to the prioritization of digitisation of natural history collections data. Biodiversity Informatics 7, 113-119.

Chavan V., Ingwersen P. (2009). Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. BMC Bioinformatics, 10 (Suppl. 14): S2.

Chavan V., Rane N., Watve A., Ruggiero M. (2005). Resolving taxonomic discrepancies: role of electronic catalogues of known organisms. Biodiversity Informatics 2: 70-78.

Collen B., Ram M., Zamin T., Louise M. (2008). The tropical biodiversity data gap: addressing disparity in global monitoring. Tropical Conservation Science, 1(2): 75-88.

Faith D., Collen B., Ariño A.H., Koleff P., Guinotte J., Kerr J., Chavan V. (2012). Bridging the biodiversity data gaps: recommendations of the GBIF Content Needs Assessment task Group. Biodiversity Informatics, 2013.

Fernández M.A., Blum S.D., Reichle S., Guo Q., Holzman B., Hamilton H. (2009). Locality uncertainty and the differential performance of four common niche-based modeling techniques. Biodiversity Informatics 6: 36-52.

Bisby. F.A. (2000). The quiet revolution: biodiversity informatics and the Internet. Science 289: 2309-2312.

Gaikwad J. Chavan V. (2005). Open access and biodiversity conservation: challenges and potentials for the developing world. Data Science Journal, 5: 1-17.

GBIF (2008). GBIF Work Programme 2009-2010. Copenhagen: Global Biodiversity Information Facility, 59pp. Accessible at http://www2.gbif.org/WP2009-10.pdf.

GBIF (2009) Adoption of Persistent Identifiers for Biodiversity Informatics -

Recommendations of the GBIF LSID GUID Task Group, 6 November 2009 (http://imsgbif.gbif.org/CMS_ORC/?doc_id=2956)

GBIF (2010a). GBIF position paper on future directions and recommendations for enhancing fitness-for-use across the GBIF network, version 1.0. Authored by Hill, AW, Otegui J, Ariño AH, and RP Guralnick. 2010, Copenhagen: Global Biodiversity Information Facility, 25 pp. ISBN: 87-92020-11-9. Accessible online at http://www2.gbif.org/GPP-Final.pdf.

GBIF. (2010b). 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, xx pp. ISBN: 87-92020-13-5. Accessible online at http://www.gbif.org.

GBIF (2010b). Best practice guide for ‘Data discovery and publishing strategy and action plans’ version 1.0. Authored by Chavan, VS, Sood RK, and AH Arino. 2010. Copenhagen: Global Biodiversity Information Facility, 29 pp. ISBN: 87-92020-12-7. Accessible online at http://www.gbif.org/BestPracticeGuide-final.pdf.

GBIF (2012.a). Monthly statistics - Accessible online at: http://www.gbif.org/communications/resources/monthly-statistics/ as of 21 June 2012

GBIF (2012.b). History oif GBIF –m Accessible online at: http://www.gbif.org/communications/press/history-of-gbif/

Grantham H.S., Moilanen A., Wilson K.A., Pressey R.L., Rebelo T.G., Possingham H.P. (2008). Diminishing returns on investment for biodiversity data conservation planning. Conservation Letters, 1(4): 190-198.

Grinnell J. (1917). Field tests of theories concerning distributional control. American Naturalist 51:115.

Guralnick R.P., Hill A.H., Lane M. (2007). Towards a collaborative, global infrastructure for biodiversity assessment. Ecology Letters, 10(8): 663-672.

Hanson T., Brooks T.M, Da Fonseca G.A.B, Hoffmann M., Lamoreux J.F, Machlis G, Mittermeier C.G, Mittermeier R.A, Pilgrim J.D. (2009). Warfare in Biodiversity Hotspots. Conservation Biology, Volume 23, No. 3, 578–587

Hobern D. (2003). An Introduction to GBIF Biodiversity Informatics (version 1.0 – Final). GBIF, 25 pp, accessible at: http:// http://www.gbif.es/ficheros/GBIFBiodiversityInformaticsIntroduction-v1.0-final.pdf

IUCN and UNEP (2009). The World Database of Protected Areas (WDPA). 2010 BIP Indicator: Coverage of Protected Areas. UNEP-WCMC. Cambridge, UK. http://www.wdpa.org/Statistics.aspx

Johnson N.F. (2007). Biodiversity informatics. Annual Review of Entomology, 52: 421-438.

OECD (1999). Final report of the OECD Megascience Forum Working Group on Biological Informatics, January 1999, pp. 74, accessible at http://www.oecd.org/dataoecd/24/32/2105199.pdf.

Otegui J., Ariño A.H. (2013). On the time series of records mobilised through GBIF

Otegui J, Ariño A.H. (2012b). BIDDSAT: visualizing the content of biodiversity data publishers in the GBIF network. Bioinformatics, 2012: doi: 10.1093/bioinformatics/bts359.

Otegui J., Robles E., Ariño A.H. (2009). Noise in Biodiversity Data. Poster presented at e-Biosphere 09. *e-Biosphere Conference 2009, Conference Abstracts*, London; McLeod N. & J. Edwards, eds. P. 190. http://www.e-biosphere09.org/assets/files/e-Biosphere%20Abstracts%20Volume%20-%20FINAL.pdf

Mittermeier R.A., Robles-Gil P., Hoffmann M., Pilgrim J., Brooks T., Mittermeier C.G., Lamoreux J., da

Fonseca G.A.B. (2004). Hotspots revisited. CEMEX, Mexico.

Page R. (2012). How many specimens does GBIF really have? Retrieved August 2, 2012, from http://iphylo.blogspot.com.es/2012/02/how-many-specimens-does-gbif-really.html

Pearson R.G., Raxworthy C.J., Nakamura M., Townsend Peterson A. (2007). Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. Journal of Biogeography, 34, 102-117.

Peterson A., Vieglais D. (2001). Predicting species invasions using ecological niche modelling: new approaches from bioinformatics attsack a pressing problem. Bioscience 51: 363-371.

Pino-del-Carpio A., Villarroya A., Ariño A.H., Puig J., Miranda R. (2011).

Communication gaps of knowledge of freshwater fish biodiversity: implications for the management and conservation of Mexican Biosphere Reserves. Journal of Fish Biology, 79(6): 1563-1591.

Ramírez-Villegas J., Khoury C., Jarvis A., Debouck D. G., Guarino L. (2010). A Gap Analysis Methodology for Collecting Crop Genepools: A Case Study with Phaseolus Beans. PLoS ONE, 5(10), e13497. Public Library of Science. Retrieved from http://dx.doi.org/10.1371/journal.pone.0013497

Ready J., Kaschner K., South A.B., Eastwood P.D., Rees T., Rius J., Agbayani E., et al. (2010). Predicting the distributions of marine organisms at the global scale. Ecological Modelling, 221(3), 467-478. doi:DOI: 10.1016/j.ecolmodel.2009.10.025

Rödder D., Lötters S. (2010). Potential distribution of the alien invasive brown tree snake, Boiga irregularis (Reptilia: Colubridae). Pacific Science, 64(1), 11-22. University of Hawai'i Press. doi:10.2984/64.1.011

Secretariat of the Convention on Biological Diversity (2010). Global Biodiversity Outlook 3. Montreal, 94 Pages.

Secretariat of the Convention on Biological Diversity (2012). Strategic Plan for Biodiversity 2011-2020, including Aichi Biodiversity Targets accessible at: http://www.cbd.int/sp/

Sullivan B.L., Wood C.L., Iliff M.J., Bonney R.E., Fink D., Kelling S. (2009). eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation, 142(10), 2282-2292. doi:DOI: 10.1016/j.biocon.2009.05.006

Warner S., Linburg K., Ariño A.H., Dushoff J., Dodd M., Stergiou K., Potts J. (1995). Time Series Compared Across the Land-Sea Gradient. Pp. 242-273 en: Powell T.M. & Steele J.H. (eds.:) Ecological Time Series. Chapman & Hall, New York etc.

Yesson C., Brewer P.W., Sutton T., Caithness N., Pahwa J.S., et al. (2007) How Global Is the Global Biodiversity Information Facility? PLoS ONE 2(11): e1124. doi:10.1371/journal.pone.0001124

LIST OF TABLES

Table 1. Essential core data elements (in the GBIF-Index occurrence table).

Table 2. Resources currently available through GBIF ChecklistBank used to build the GBIF taxonomical backbone.

Table 3. Taxonomical rank matching with Catalogue of Life 2011 (February 2012)

Table 4.a. Scientific names and occurrences summary for each 'unknown' taxonomic rank (as of December 2010).

Table 4.b. Scientific names and occurrences summary for each 'unknown' taxonomic rank (as of February 2012).

Table 5: Potential misidentification at the Kingdom rank and tentative resolution through CoL 2011 and more recent version (February 2012)

Table 6.a: Estimation of the taxonomical misidentification at the Kingdom level (February 2012)
Table 6.b: Estimation of the taxonomical misidentification at the Phylum level (February 2012)
Table 6.c: Estimation of the taxonomical misidentification at the Class level (February 2012)
Table 7. Percentage (%) temporal quality of the GBIF mobilised data records according to GBIFS methodology.
Table 8. Breakdown of the year provided in the GBIF mobilised data records.
Table 9. Breakdown of the temporal and geospatial data availability.
Table 10. Summary of potential resource duplicates.
Table 11. Top 20 potential resources duplicates.
Table 12.a: Top 15 species with highest number of data records (December 2010).
Table 12.b: Top 15 species with highest number of data records (February 2012).
Table 13: Breakdown of the species and occurrences richness for Kingdom Animalia and Plantae (February 2012).
Table 14. Breakdown of occurrences by continents.
Table 15. Proportion of financial contribution to GBIF (2010) by continents.
Table 16. Distribution of occurrences per members to the Organisation for Economic Co-operation and Development (OECD).
Table 17. Distribution of occurrences per country income status.
Table 18.a Grid (1/10 degree) occupancy (February 2012).
Table 18.b Grid (1/10 degree) occupancy for Animalia.
Table 18.c Grid (1/10 degree) occupancy for Plantae.
Table 19. Grid (1/10 degree) occupancy by Classes (Animalia and Plantae).
Table 20. Grid (1/10 degree) occupancy by Families (Animalia and Plantae).

LIST OF FIGURES

Figure 1. Typical flow of data discovered and published through the GBIF network.
Figure 2. Data mining methodology employed during content assessment exercise carried out by the GBIF Secretariat.
Figure 3. Data mining methodologies employed during content assessment exercise carried out by the University of Navarra.
Figure 4. Evolution of the percentage of geo-referenced records (1800-2010).
Figure 5. Evolution of incomplete records (taxonomical*temporal*geospatial) in the GBIF-Index
Figure 6.a. Data records by Kingdom (December 2010).
Figure 6.b. Data records by Kingdom (February 2012).
Figure 7.a. Data records by Phylum (December 2010).
Figure 7.b. Data records by Phylum (February 2012).
Figure 8.a. Data records by Class (December 2010).
Figure 8.b. Data records by Class (February 2012).
Figure 9. Treemap of occurrences according to taxon group, down to Class. Cell surface proportional to number of occurrences. Blue: Invertebrates; purple: Vertebrates; green: Higher plants; yellow: Algae and Ferns; brown: Fungi; red: unicellular organisms.
Figure 10. Top 10 data resources publishing maximum number of data records (December 2010).

Figure 11. Kingdom wise distribution of data records by Year.
Figure 12.a. Plantae Kingdom wise distribution of occurrences by year (February 2012).
Figure 12.b. Animalia Kingdom wise distribution of occurrences by year (February 2012).
Figure 13. Aves wise distribution of occurrences by year (February 2012)
Figure 14. Basis of records distribution of occurrences by year (February 2012)
Figure 15. Distribution of voting and associate country participants (February 2012)
Figure 16. Distribution of occurrences by latitude (February 2012)
Figure 17. Distribution of species richness by latitude (February 2012)
Figure 18. Distribution of the average number of occurrences per species by latitude (February 2012)
Figure 19. Georeferenced records (dots) published by the largest data providers in Europe. Colors represent distinct providers. (From Otegui et al., 2009).
Figure 20. Records available in the GBIF-Index by date
Figure 21. Variation of the records available in the GBIF-Index by date (December 2010 – February 2012)
Figure 22. Distribution of distinct species collected/observed over time (February 2012)
Figure 23. Distribution of distinct resources contributing to the GBIF-Index over time (February 2012)
Figure 24. Grid occupancy (1/2 degree grid) of the GBIF-Index over time (February 2012)
Figure 25. Grid occupancy (1/10 degree grid) of the GBIF-Index by Kingdom (February 2012)
Figure 26. Grid occupancy (1/10 degree grid) of the GBIF-Index by Classes (February 2012)
Figure 27. Distribution of families for FFU-ENM – Kingdom Animalia (February 2012)
Figure 28. Distribution of families for FFU-ENM – Kingdom Plantae (February 2012)
Figure 29. Selection of fit-for-use records in the GBIF Index for ecological niche modelling (ENM).

Table 1. Essential core data elements (in the GBIF-Index occurrence table).

Title	**Description**
Publisher	Publisher of the resource/dataset
Dataset	Resource/Dataset
Institution	The name (or acronym) in use by the institution having custody of the object(s) or information referred to in the record.
Collection	The name, acronym, code, or initials identifying the collection or data set from which the record was derived.
Catalogue number	An identifier (preferably unique) for the record within the data set or collection.
Scientific name	The full scientific name, with authorship and date information if known. When forming part of identification, this should be the name in lowest level taxonomic rank that can be determined.
Taxon author	The authorship information for the Scientific name.
Taxon rank	The taxonomic rank of the most specific name in the Scientific name. Recommended best practice is to use a controlled vocabulary.
Kingdom	The full scientific name of the kingdom in which the taxon is classified.
Phylum	The full scientific name of the phylum or division in which the taxon is classified.
Class	The full scientific name of the class in which the taxon is classified.
Order	The full scientific name of the order in which the taxon is classified.
Family	The full scientific name of the family in which the taxon is classified.
Genus	The full scientific name of the genus in which the taxon is classified.
Species epithet	The name of the first or species epithet of the Scientific name.

Infraspecific epithet	The name of the lowest or terminal infraspecific epithet of the Scientific name, excluding any rank designation.
Latitude	The geographic latitude (in decimal degrees) of the geographic center of a Location. Positive values are north of the Equator; negative values are south of it. Legal values lie between -90 and 90, inclusive.
Longitude	The geographic longitude (in decimal degrees) of the geographic center of a Location. Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive.
Coordinate precision	A decimal representation of the precision of the coordinates given in the Latitude and Longitude.
Maximum altitude	The upper limit of the range of elevation (altitude, usually above sea level), in meters.
Minimum altitude	The lower limit of the range of elevation (altitude, usually above sea level), in meters.
Altitude precision	A decimal representation of the precision of the altitude.
Minimum depth	The lesser depth of a range of depth below the local surface, in meters.
Maximum depth	The lesser depth of a range of depth below the local surface, in meters.
Depth precision	A decimal representation of the precision of the depth.
Continent or ocean	The name of the continent in which the Location occurs. Recommended best practice is to use a controlled vocabulary such as the Getty Thesaurus of Geographic Names or the ISO 3166 Continent code. Recommended best practice is to use a controlled vocabulary such as the Getty Thesaurus of Geographic Names.
Country	The name of the country or major administrative unit in which the Location occurs. Recommended best practice is to use a controlled vocabulary such as the Getty Thesaurus of Geographic Names.
State or province	The name of the next smaller administrative region than country (state, province, canton, department, region, etc.) in which the Location occurs.
County	The full, unabbreviated name of the next smaller administrative region than State or Province (county, shire, department, etc.) in which the location occurs.

Name of collector/observer	A list (concatenated and separated) of names of people, groups, or organizations responsible for recording the original occurrence.
Locality	The specific description of the place. Less specific geographic information can be provided in other geographic terms. This term may contain information modified from the original to correct perceived errors or standardize the description.
Year of collection	The four-digit year in which the collection or observation event occurred, according to the Common Era Calendar.
Month of collection	The ordinal month in which the collection or observation event occurred.
Day of collection	The integer day of the month on which the collection or observation event occurred.
Basis of record	The specific nature of the data record. Recommended best practice is to use a controlled vocabulary such as the Darwin Core Type Vocabulary (http://rs.tdwg.org/dwc/terms/type-vocabulary/index.htm).
Name of identifier	A list (concatenated and separated) of names of people, groups, or organizations that assigned the taxon to the subject.
Identification date	The date on which the subject was identified as representing the taxon. Recommended best practice is to use an encoding scheme, such as ISO 8601:2004(E).
date of creation	Timestamp of creation of this raw occurrence record in the index.
date of modification	Timestamp of last update of this raw occurrence record in the index.
date of deletion	Timestamp of deletion of this raw occurrence record in the index (obsolete).

Table 2. Top 10 resources currently available through GBIF 'ChecklistBank' used to build the GBIF taxonomical backbone.

Title	Version	Families	Genera	Species
The Catalogue of Life	2012-01-14	8,149	129,461	1,379,178
Register of Marine and Nonmarine Genera (IRMNG)	2012-01-13	34,119	790,025	1,017,851
International Plant Names Index	2011-07-13	791	59,766	1,317,317
NCBI Taxonomy	2012-01-13	7,223	59,404	668,915
The Integrated Taxonomic Information System (ITIS)	2012-01-14	6,972	45,531	306,358
World Register of Marine Species	2012-05-02	6,370	41,293	233,811
Index Fungorum	2011-07-13	2,926	10,569	267,553
Fauna Europapaea	2011-07-13	-	37,214	131,671
Wikipedia Species Pages - English	2011-09-04	-	-	-
GRIN Taxonomy for Plants	2012-01-14	492	12,909	58,773

A full up-to-date list can be accessed at: http://ecat-dev.gbif.org/

Table 3. Taxonomical rank matching with Catalogue of Life 2011 (February 2012)

Taxonomical rank matching with Catalogue of Life 2011							Percentage of the GBIF-Index	Percentage of the total number of species
Kingdom	Phylum	Class	Order	Family	Genus	Species	(324,247,283 occurrences)	(995,974 species in total)
							0.05%	0.27%
✔							1.38%	0.29%
✔	✔						0.53%	0.89%
✔	✔	✔					0.77%	1.35%
✔	✔	✔	✔				0.76%	2.40%
✔	✔	✔	✔	✔			4.54%	13.36%
✔	✔	✔	✔	✔	✔		9.13%	27.98%
✔	✔	✔	✔	✔	✔	✔	82.83%	53.47%
							100.00%	100.00%

Table 4.a. Scientific names and occurrences summary for each ‘unknown’ taxonomic rank (as of December 2010).

Taxonomy	**Scientific name with ‘unknown’ status**	**% of total species recorded in GBIF**	**Occurrences with ‘unknown’ status**	**% of total occurrences recorded in GBIF Index**
Kingdom	114,721	7.0%	15,030,014	5.6%
Phylum	223,433	13.8%	22,180,639	8.3%
Class	235,857	14.5%	23,071,180	8.6%
Order	261,706	16.1%	24,605,925	9.2%
Family	235,089	14.5%	21,508,688	8.1%
Genus	76,416	4.7%	8,665,178	3.2%
Species	120,362	7.4%	23,015,905	8.6%

Table 4.b. Scientific names and occurrences summary for each ‘unknown’ taxonomic rank (as of February 2012).

Taxonomy	Scientific name with ‘unknown’ status	% of total species recorded in GBIF	Occurrences with ‘unknown’ status	% of total occurrences recorded in GBIF Index
Kingdom	5,153	0.35%	167,208	0.05%
Phylum	11,305	0.77%	4,640,252	1.43%
Class	26,266	1.81%	3,963,750	1.22%
Order	52,007	3.58%	6,304,444	1.94%
Family	41,932	2.82%	6,015,636	1.86%
Genus	31,565	2.17%	8,959,016	2.76%
Species	133,086	9.15%	25,343,834	7.82%

Table 5: Potential misidentification at the Kingdom rank and tentative resolution through CoL 2011 and more recent version (February 2012)

Kingdom	Species	Occurrences	CoL 2011
Plantae	*Zonotrichia albicollis*	775,671	Accepted name in CoL 2012 in Animalia kingdom
Plantae	*Zonotrichia leucophrys*	362,767	Accepted name in CoL 2012 in Animalia kingdom
Protozoa	*Neogloboquadrina pachyderma*	141,720	Not in CoL 2011 Accepted in CoL 2012
Plantae	*Zonotrichia atricapilla*	106,804	Accepted name in CoL 2012 in Animalia kingdom
Protozoa	*Globigerinoides ruber*	86,563	Not in CoL 2011 Accepted name in CoL 2012
Protozoa	*Globigerina bulloides*	82,643	Not in CoL 2011 Accepted name in CoL 2012
Protozoa	*Globigerinita glutinata*	74,617	Not in CoL 2011 Accepted in CoL 2012

Protozoa	*Globorotalia truncatulinoides*	64,707	Not in CoL 2011 Not CoL 2012 Identified in Marine Species Identification Portal (as of Feb 2012)[20]
Protozoa	*Globorotalia inflata*	57,706	Not in CoL 2011 Not in CoL 2012 Identified in Marine Species Identification Portal (as of Feb 2012)[21]
Protozoa	*Orbulina universa*	55,290	Not in CoL 2011 Not in CoL 2012 Identification Portal (as of Feb 2012)[22]

Table 6.a: Estimation of the taxonomical misidentification at the Kingdom level (February 2012)

Incorrect Kingdom assignment	**Correct Kingdom in CoL 2011**	**Occurrences**	**Species**
Plantae	Animalia	1,308,111	90
Animalia	Plantae	1,536	26
Chromista	Animalia	1,504	1
Chromista	Plantae	310	3
Animalia	Fungi	190	23
Fungi	Animalia	186	10
Protozoa	Chromista	100	8
Plantae	Fungi	98	11
Plantae	Protozoa	61	2
Plantae	Chromista	43	6

[20] http://species-identification.org/species.php?species_group=zsao&id=1387

[21] http://species-identification.org/species.php?species_group=zsao&id=1384

[22] http://species-identification.org/species.php?species_group=zsao&id=1397

Fungi	Plantae	41	2
Animalia	Chromista	26	3
Bacteria	Protozoa	22	1
Plantae	Bacteria	13	5
Protozoa	Plantae	9	5
Protozoa	Fungi	6	1
Fungi	Protozoa	2	1
	Total	**1,312,258**	**198**

Table 6.b: Estimation of the taxonomical misidentification at the Phylum level (February 2012)

Incorrect Phylum assignment	**Correct Phylum in CoL 2011**	**Occurrences**	**Species**
Bryophyta	Magnoliophyta	17,488	24
Magnoliophyta	Arthropoda	2,788	36
Cnidaria	Chordata	2,213	12
Ochrophyta	Arthropoda	1,504	1
Chordata	Magnoliophyta	833	5
Cyanobacteria	Proteobacteria	312	5
Ochrophyta	Rhodophyta	309	2
Arthropoda	Magnoliophyta	297	10
Arthropoda	Chlorophyta	244	1
Magnoliophyta	Chordata	201	5
Magnoliophyta	Cnidaria	176	2
Labyrinthista	Sarcomastigophora	116	4
Ascomycota	Chordata	115	1
Marchantiophyta	Bryozoa	114	1
Annelida	Tardigrada	111	1
Arthropoda	Ascomycota	93	6
Arthropoda	Rhodophyta	82	2
Magnoliophyta	Ascomycota	80	2

Mollusca	Ascomycota	48	10
Bryozoa	Magnoliophyta	47	4
Sarcomastigophora	Ochrophyta	46	4
Ascomycota	Magnoliophyta	40	1
Ascomycota	Arthropoda	39	3
Chlorophyta	Magnoliophyta	34	2
Brachiopoda	Ascomycota	32	2
Mollusca	Arthropoda	30	2
Arthropoda	Nematoda	26	1
Arthropoda	Ochrophyta	25	2
Annelida	Magnoliophyta	24	1
Pinophyta	Arthropoda	23	1
Platyhelminthes	Arthropoda	22	9
Rhodophyta	Arthropoda	21	1
Basidiomycota	Arthropoda	19	2
Arthropoda	Bacillariophyta	17	5
Chlorophyta	Rhodophyta	16	2
Chlorophyta	Cyanobacteria	13	5
Echinodermata	Arthropoda	12	1
Bacteroidetes	Proteobacteria	11	3
Magnoliophyta	Ochrophyta	8	5
Bryophyta	Rhodophyta	8	1
Ascomycota	Bryozoa	7	2
Echinodermata	Cnidaria	6	1
Magnoliophyta	Rotifera	6	1
Ciliophora	Chlorophyta	4	1
Arthropoda	Mollusca	4	1
Arthropoda	Pinophyta	4	1
Ascomycota	Bacillariophyta	3	2
Euglenozoa	Rhodophyta	3	2

Annelida	Bacillariophyta	3	1
Ascomycota	Cnidaria	3	1
Ascomycota	Echinodermata	3	1
Chlorophyta	Arthropoda	2	1
Platyhelminthes	Bacillariophyta	2	1
Magnoliophyta	Bacillariophyta	2	1
Platyhelminthes	Acanthocephala	1	1
Pteridophyta	Arthropoda	1	1
Ascomycota	Chlorophyta	1	1
Cnidaria	Ochrophyta	1	1
Arthropoda	Platyhelminthes	1	1
Dinophyta	Rhodophyta	1	1
	Total	**27,695**	**210**

Table 6.c: Estimation of the taxonomical misidentification at the Class level (February 2012)

Incorrect Class assignment	**Correct Class in CoL 2011**	**Occurrences**	**Species**
Bryopsida	Andreaeopsida	21,966	82
Bryopsida	Liliopsida	17,488	24
Magnoliopsida	Insecta	2,429	12
Hydrozoa	Arachnida	2,213	12
Phaeophyceae	Insecta	1,504	1
Actinopterygii	Magnoliopsida	808	2
Insecta	Malacostraca	664	11
Malacostraca	Insecta	359	1
Phaeophyceae	Florideophyceae	309	2
Insecta	Trebouxiophyceae	244	1
Liliopsida	Insecta	239	15
Lecanoromycetes	Dothideomycetes	220	1

Magnoliopsida	Hydrozoa	176	2
Liliopsida	Andreaeopsida	175	1
Insecta	Magnoliopsida	167	3
Insecta	Liliopsida	116	6
Labyrinthulea	Polycystina	116	4
Jungermanniopsida	Gymnolaemata	114	1
Polychaeta	Eutardigrada	111	1
Insecta	Florideophyceae	82	2
Magnoliopsida	Lecanoromycetes	78	1
Liliopsida	Maxillopoda	59	1
Insecta	Lecanoromycetes	58	1
Magnoliopsida	Arachnida	57	7
Lobosa	Coscinodiscophyceae	54	4
Stenolaemata	Magnoliopsida	47	4
Zoomastigophora	Craspedophyceae	45	3
Lecanoromycetes	Magnoliopsida	40	1
Lecanoromycetes	Insecta	35	1
Insecta	Leotiomycetes	35	5
Chlorophyceae	Liliopsida	34	2
Rhynchonellata	Lecanoromycetes	32	2
Ostracoda	Secernentea	26	1
Actinopterygii	Liliopsida	24	2
Polychaeta	Liliopsida	24	1
Pinopsida	Insecta	23	1
Turbellaria	Arachnida	22	9
Florideophyceae	Insecta	21	1
Agaricomycetes	Insecta	19	2
Magnoliopsida	Actinopterygii	18	2
Insecta	Agaricomycetes	17	5
Chlorophyceae	Florideophyceae	16	2

Maxillopoda	Liliopsida	14	1
Asteroidea	Entognatha	12	1
Sphingobacteria	Alphaproteobacteria	8	2
Bryopsida	Florideophyceae	8	1
Insecta	Phaeophyceae	8	1
Lecanoromycetes	Gymnolaemata	7	2
Magnoliopsida	Eurotatoria	6	1
Granuloreticulosea	Lecanoromycetes	6	1
Magnoliopsida	Reptilia	6	1
Magnoliopsida	Coscinodiscophyceae	5	3
Ciliatea	Chlorophyceae	4	1
Magnoliopsida	Entognatha	4	1
Ostracoda	Gastropoda	4	1
Dothideomycetes	Insecta	4	2
Magnoliopsida	Liliopsida	4	2
Insecta	Pinopsida	4	1
Rhabditophora	Turbellaria	4	1
Dothideomycetes	Asteroidea	3	1
Polychaeta	Bacillariophyceae	3	1
Dothideomycetes	Eurotiomycetes	3	1
Euglenida	Florideophyceae	3	2
Flavobacteria	Gammaproteobacteria	3	1
Leotiomycetes	Hydrozoa	3	1
Liliopsida	Actinopterygii	2	1
Magnoliopsida	Agaricomycetes	2	1
Turbellaria	Bacillariophyceae	2	1
Lecanoromycetes	Granuloreticulosea	2	1
Ulvophyceae	Insecta	2	1
Pezizomycetes	Leotiomycetes	2	2
Magnoliopsida	Phaeophyceae	2	1

Liliopsida	Coscinodiscophyceae	1	1
Eurotiomycetes	Dothideomycetes	1	1
Dinophyceae	Florideophyceae	1	1
Filicopsida	Insecta	1	1
Appendicularia	Liliopsida	1	1
Gastropoda	Orbiliomycetes	1	1
Neoophora	Palaeacanthocephala	1	1
Anthozoa	Phaeophyceae	1	1
Zoomastigophora	Synurophyceae	1	1
Arachnida	Turbellaria	1	1
		50,434	**290**

Table 7. Percentage (%) temporal quality of the GBIF mobilised data records according to GBIFS methodology.

	December 2010	**February 2012**	**Difference**
Occurrences with no year provided	82,300,746	42,890,654	-47.9%
Percentage of the GBIF Index	30.8 %	13,2%	

Table 8. Breakdown of the year provided in the GBIF mobilised data records.

RAW refers to records as supplied by the publisher, whereas OCC indicates records available through the portal after processing. According to UNZYEC methodology. "Not valid" year includes years supplied as <1750 (including explicit zero) or in the future. "Null" includes records with year provided as null value but do not include years explicitly stated as a numerical zero value. "Matching/not matching" indicates whether the value for year in a record matches between the raw data collected from providers (RAW) and the processed data made available through the portal (OCC).

			Valid OCC	Not valid OCC	Null value	%
RAW	Valid	Matching	63,9%	--	--	63,9%
		Not matching	--	0,1%	8,6%	8,7%
	Not valid	Matching	--	0,2%	--	0,2%
		Not matching	0,1%	--	0,4%	0,5%
	Null	Matching	--	--	17,7%	17,7%
		Not matching	5,0%	3,9%	--	8,9%
		Total	69,1%	4,3%	26,7%	100,0%
				31%		

Table 9. Breakdown of the temporal and geospatial data availability

	With year	**Without year**	**Total**
Georeferenced	255,4 (78.8%)	24.4 (7.5%)	279.8 (86.3%)
Not georeferenced	25.9 (8%)	18.5 (5.7%)	44.4 (13.7%)
Total	281.3 (86.8%)	42.9 (13.2%)	324.4 (100%)

(in million occurrences)

Table 10. Summary of potential resource duplicates

Estimated potential duplicates	Number of resources combinations	Total number of 'potential' duplicates'
>100.000	42	30.905.772
10.000 – 100.000	215	5.460.179

1.000 – 10.000	560	1.989.482
100 – 1.000	820	309.663
	1.637	**38.665.096**

Table 11. Top 20 potential resources duplicates

Resource Name (1)	**Resource Name (2)**	**'Potential' duplicates**
Biodiversidad de Costa Rica	Especímenes INBio	12,993,467
CNIN/Lepidoptera	Colección de Referencia de Lepidópteros Diurnos Mexicanos de la CNIN (Lepidoptera: Papilionoidea) (IBUNAM)	1,779,872
(Appendix 1) Planktonic foraminifera abundances in ODP Site 181-1123	(Appendix 1) Census data of planktic foraminiferal faunas together with estimates of mean annual SST for ODP Site 181-1123	1,620,028
Planktic foraminifera counts of sediment core MD95-2040	Planktonic foraminifera, stable isotope record and temperature reconstruction of sediment core MD95-2040	1,514,214
Pelagic Fish Observations 1968-1999	Pelagic Fish Observations 1968-1999 (Australian Antarctic Data Centre)	1,288,625
(Fig. 2) Abundance of Neogloboquadrina pachyderma sinistral in sediment core MD95-2040	Planktic foraminifera counts of sediment core MD95-2040	1,132,425
(Appendix B5) Distribution of planktic foraminifera in DSDP Site 90-594 east of New Zealand	(Appendix 1) Census data of planktic foraminiferal faunas together with estimates of mean annual SST for DSDP Site 90-594	770,149
Planktic foraminifera counts of sediment core MD95-2040	(Appendix 4) Stable oxygen isotope record of Globigerina bulloides and abundances of Neogloboquadrina pachyderma and ice-rafted debris in sediment core MD95-2040	761,421
Birds (KIEE-BI)	Birds (MNHM-BI)	593,674

Benthic foraminifera abundance in counts of Hole PRAD1-2	Benthic foraminifera abundance in per cent of Hole PRAD1-2	585,766
(Table 3) Distribution and abundance of selected planktonic foraminifera of the Pliocene DSDP Hole 41-366A	(Table 4) Distribution and abundance of selected planktonic foraminifera of the Pliocene DSDP Hole 41-366A	515,705
(Appendix 3) Assemblage of benthic foraminifera in sediment core M5/2_KL15	Relative abundance of benthic foraminifera in sediment core M5/2_KL15	480,479
Birds (UWEP-BI)	Birds (KIEE-BI)	479,080
Planktic foraminifera counts of sediment core MD95-2040	(Fig. 8g-h, 11) Abundance of planktonic foraminifera and estimation of sea surface temperature and export production of sediment core MD95-2040	420,615
Planktic foraminifera abundance in counts of Hole PRAD1-2	Planktic foraminifera abundance in per cent of Hole PRAD1-2	395,733
(Fig. 2) Abundance of Neogloboquadrina pachyderma sinistral in sediment core MD95-2040	Planktonic foraminifera, stable isotope record and temperature reconstruction of sediment core MD95-2040	368,550
(Table 3) Occurrences of planktonic foraminifers in samples from ODP Hole 105-647A	(Table 2) Occurrences of planktonic foraminifers in samples from ODP Hole 105-647A	321,750
Planktic foraminifera abundance of Hole 41-369A	(Table 2) Distribution and abundance of selected planktonic foraminifera of the Pliocene DSDP Hole 41-369A	286,390
Hatikka Observation Data Gateway	Tiira information service	281,119
(Appendix A) Stable carbon and oxygen isotopes and paleoproductivity reconstructions for the last 550 kyr of ODP Hole 130-807A from the Ontong Java Plateau, Pacific Ocean	(Appendix A) Benthic foraminiferal assemblages in sediments of the last 550 kyr of ODP Hole 130-807A from the Ontong Java Plateau, Pacific Ocean	249,300

Table 12.a: Top 15 species with highest number of data records (December 2010).

	Occurrences	**(% of total)**	**Georeferenced (%)**
Zenaida macroura	2,163,341	0.8%	99.8%
Cardinalis cardinalis	1,953,522	0.7%	99.8%
Passer domesticus	1,892,301	0.7%	94.9%
Sturnus vulgaris	1,852,357	0.7%	98.0%
Junco hyemalis	1,735,767	0.6%	99.4%
Cyanocitta cristata	1,697,922	0.6%	99.8%
Picoides pubescens	1,673,374	0.6%	99.8%
Carduelis tristis	1,671,365	0.6%	99.8%
Carpodacus mexicanus	1,611,847	0.6%	99.7%
Poecile atricapillus	1,518,874	0.6%	99.9%
Corvus brachyrhynchos	1,369,418	0.5%	99.8%
Baeolophus bicolor	1,259,308	0.5%	99.9%
Sitta carolinensis	1,244,000	0.5%	99.8%
Turdus migratorius	1,220,210	0.5%	99.5%
Anas platyrhynchos	1,202,864	0.4%	99.0%
	24,066,470	**9.0%**	

Table 12.b: Top 15 species with highest number of data records (February 2012).

	Occurrences (difference with 2010)	**(% of total)**	**Georeferenced (%)**
Zenaida macroura	2,270,891 (+4,9%)	0.7%	99.9%
Sturnus vulgaris	2,171,136 (+17,2%)	0.7%	99.7%
Passer domesticus	2,029,427 (+7.2%)	0.6%	99.0%
Cardinalis cardinalis	1,779,316 (-8.9%)	0.5%	99.9%
Picoides pubescens	1,776,269 (+6.2%)	0.5%	99.9%
Junco hyemalis	1,731,413 (-0.2%)	0.5%	99.3%
Cyanocitta cristata	1,695,019 (-0.1%)	0.5%	99.8%
Carduelis tristis	1,666,477 (-0.3%)	0.5%	99.8%
Poecile atricapillus	1,612,214 (+6.2%)	0.5%	99.9%
Carpodacus mexicanus	1,609,953 (-0.1%)	0.5%	99.8%

Table 13: Breakdown of the species and occurrences richness for Kingdom Animalia and Plantae (February 2012)

	Species		**Occurrences**	
Number of occurrences/species	**Animalia**	**Plantae**	**Animalia**	**Plantae**
<100	400.088 (87.5%)	374.524 (86.5%)	4.608.908 (2.4%)	5.655.807 (6%)
100-1.000	46.264 (10.1%)	49.643 (11.5%)	14.226.733 (7.4%)	14.455.647 (15.3%)
1.000-10.000	9.178 (2%)	7.428 (1%)	25.228.569 (13.1%)	18.437.768 (19.5%)
10.000-100.000	1.483 (<0.1%)	1.126 (<1%)	41.730.120 (21.6%)	33.901.539 (35.9%)
100.000-1.000.000	306 (<0.1%)	129 (<0.1%)	76.851.930 (39.8%)	22.025.055 (23.3%)
>1.000.000	20 (<0.01%)	-	30.641.798 (15.9%)	-
Total	**457.340** **(100%)**	**432.851** **(100%)**	**193.288.058** **(100%)**	**99.475.816** **(100%)**

Table 14. Breakdown of occurrences by continents

	% of GBIF Index	
Region	**(December 2010)**	**(February 2012)**
Northern America	32.26%	28.19%
Northern Europe	30.24%	30.06%
Western Europe	8.34%	11.48%
Southern Africa	4.12%	3.75%
Central America	3.61%	4.43%
South America	2.72%	2.63%
Southern Europe	2.35%	2.21%
Australia and New Zealand	2.09%	7.03%
Eastern Asia	1.72%	2.00%
Eastern Africa	1.00%	0.66%
Eastern Europe	0.87%	0.90%
South-Eastern Asia	0.70%	0.64%
Caribbean	0.47%	0.44%
Melanesia	0.47%	0.42%
Antartica	0.37%	0.30%
Western Africa	0.31%	0.25%
Western Asia	0.31%	0.30%
Middle Africa	0.30%	0.28%
Southern Asia	0.30%	0.33%
Northern Africa	0.14%	0.15%
Micronesia	0.06%	0.06%
Polynesia	0.05%	0.05%
Central Asia	0.04%	0.04%
Unknown	7.16%	1.08%
Total	**100%**	**100%**

Table 15. Proportion of financial contribution to GBIF (2010) by continents

Region	**% of contribution (2011)**	**OECD countries**	**% of GBIF Index (February 2012)**
Northern America	23.0%	2	28.19%
Northern Europe	17.8%	8	30.06%
Western Europe	24.1%	7	11.48%
Southern Africa	1.0%	-	3.75%
Central America	1.4%	1	4.43%
South America	0.5%	1	2.63%
Southern Europe	6.3%	5	2.21%
Australia and New Zealand	3.6%	2	7.03%
Eastern Asia	22.1%	2	2.00%
Eastern Africa	<0.1%	-	0.66%
Eastern Europe	<0.1%	3	0.90%
South-Eastern Asia	<0.1%	-	0.64%
Caribbean	<0.1%	-	0.44%
Melanesia	<0.1%	-	0.42%
Antartica	<0.1%	-	0.30%
Western Africa	<0.1%	-	0.25%
Western Asia	<0.1%	2	0.30%
Middle Africa	<0.1%	-	0.28%
Southern Asia	<0.1%	-	0.33%
Northern Africa	<0.1%	-	0.15%
Micronesia	<0.1%	-	0.06%
Polynesia	<0.1%	-	0.05%
Central Asia	<0.1%	-	0.04%
Unknown	-		1.08%
Total	**100%**		**100%**

Table 16. Distribution of occurrences per members to the Organisation for Economic Co-operation and Development (OECD).

	Occurrences	**%**
OECD countries	261,377,957	84.45%
Non-OECD countries	48,126,509	15.55%

(based on occurrences where geospatial information is provided)

Table 17. Distribution of occurrences per country income status.

Country Income status	**Occurrences**	**%**
High	263,073,917	85.00%
Upper middle	36,197,975	11.70%
Lower middle	7,292,802	2.36%
Low	2,939,563	0.94%

Table 18.a Grid (1/10 degree) occupancy (February 2012)

Presence in grid	**All**		**With valid date**	**%**
In at least 1 grid	603,532	(60.6%)	485,105	(48.7%)
In at least 10 grids	150,771	(15.1%)	127,408	(17.0%)
In at least 20 grids	95,783	(9.6%)	81,057	(10.8%)
In at least 50 grids	38,520	(3.9%)	31,596	(3.2%)
In at least 100 grids	18,072	(1.8%)	14,041	(4.2%)
Total number of species	***995,975***	***(100%)***	***747,988***	***(100%)***
				[75.1%]

Table 18.b Grid (1/10 degree) occupancy for Animalia.

Presence in grid	**All**		**With valid date**	**%**
In at least 1 grid	307,871	(67.3%)	213,498	(72.3%)
In at least 10 grids	68,475	(15.0%)	55,561	(18.8%)
In at least 20 grids	43,761	(9.6%)	36,462	(12.3%)
In at least 50 grids	17,747	(3.9%)	15,071	(5.1%)
In at least 100 grids	7,893	(1.7%)	6,756	(2.3%)
Total number of species	***457,600***	***(100%)***	***295,380***	***(100%)***
				[64.6%]

Table 18.c Grid (1/10 degree) occupancy for Plantae.

Presence in grid	**All**		**With valid date**	**%**
In at least 1 grid	253,477	(58.5%)	233,720	(62.5%)
In at least 10 grids	71,332	(16.5%)	61,771	(16.5%)
In at least 20 grids	44,572	(10.3%)	37,730	(10.1%)
In at least 50 grids	17,325	(4.0%)	13,454	(3.6%)
In at least 100 grids	8,690	(2.0%)	6,100	(1.6%)
Total number of species	***433,174***	***(100%)***	***373,885***	***(100%)*** ***[86%]***

Table 19. Grid (1/10 degree) occupancy by Classes (Animalia and Plantae).

Class	**Total number of species in GBIF-Index**	**Number of species with sufficient grid presence (>=20)**	**%**
Aves	12,065	5,452	45.2%
Holocephali	54	19	35.2%

Marchantiopsida	156	54	34.6%
Cephalaspidomorphi	47	14	29.8%
Elasmobranchii	1,275	286	22.4%
Sphagnopsida	293	64	21.8%
Jungermanniopsida	1,750	381	21.8%
Actinopterygii	26,417	5,047	19.1%
Pinopsida	1,064	199	18.7%
Phascolosomatidea	43	8	18.6%
Sipunculidea	95	16	16.8%
Nuda	7	1	14.3%
Bryopsidophyceae	431	52	12.1%
Asteroidea	1,664	199	12.0%
Ulvophyceae	716	69	9.6%
Liliopsida	65,196	6,237	9.6%
Thaliacea	75	7	9.3%
Bangiophyceae	108	9	8.3%
Holothuroidea	1,178	98	8.3%
Echinoidea	1,290	106	8.2%
Appendicularia	51	4	7.8%
Hydrozoa	2,704	210	7.8%
Amphibia	5,665	409	7.2%
Haplomitriopsida	14	1	7.1%
Magnoliopsida	324,251	22,677	7.0%
Filicopsida	21,260	1,477	6.9%
Anthocerotopsida	58	4	6.9%
Bivalvia	11,956	801	6.7%
Cephalopoda	2,871	174	6.1%
Insecta	223,933	13,146	5.9%
Malacostraca	23,167	1,203	5.2%
Gastropoda	33,760	1,701	5.0%
Myxini	64	3	4.7%

Crinoidea	552	25	4.5%
Maxillopoda	7,864	345	4.4%
Cubozoa	25	1	4.0%
Anthozoa	7,747	278	3.6%
NULL	5,548	159	2.9%
Aplacophora	223	4	1.8%
Stenolaemata	859	14	1.6%
Demospongiae	4,589	63	1.4%
Neoophora	201	2	1.0%
Rhynchonellata	2,033	12	0.6%
Cycadopsida	352	2	0.6%
Hexactinellida	504	1	0.2%
Leiosporocerotopsida	1	-	-
Sarcopterygii	36	-	-
Remipedia	17	-	-
Cephalocarida	9	-	-
Somasteroidea	2	-	-
Eucycliophora	2	-	-
Lingulata	163	-	-
Pleurastrophyceae	7	-	-
Pedinophyceae	2	-	-
Pararotatoria	2	-	-
Cestoda	150	-	-
Pauropoda	185	-	-
Eoacanthocephala	25	-	-
Phylactolaemata	32	-	-
Myxosporea	72	-	-
Eutardigrada	107	-	-
Collembola	1	-	-
Archoophora	5	-	-

Table 20. Grid (1/10 degree) occupancy by Families Top 10 (Animalia and Plantae).

Taxonomical ranks (Kingdom –Phylum – Class – Order)	**Families with sufficient grid presence (>=20) for a number of species greater than 100**	**Number of species**	**% of total**	**Total number of species in the family**
Animalia - Chordata - Aves - Anseriformes	Anatidae	133	**(73.1%)**	182
Animalia - Chordata - Actinopterygii - Myctophiformes	Myctophidae	173	**(68.4%)**	253
Animalia - Chordata - Actinopterygii - Perciformes	Carangidae	103	**(66.9%)**	154
Animalia - Chordata - Aves - Ciconiiformes	Laridae	102	**(66.2%)**	154
Animalia - Chordata - Aves - Passeriformes	Thraupidae	139	**(48.9%)**	284
Animalia - Chordata - Aves - Passeriformes	Tyrannidae	212	**(48.0%)**	442
Animalia - Chordata - Aves - Passeriformes	Emberizidae	159	**(46.8%)**	340
Animalia - Chordata - Aves - Piciformes	Picidae	115	**(46.4%)**	248
Animalia - Chordata - Aves - Ciconiiformes	Accipitridae	132	**(43.7%)**	302
Animalia - Chordata - Aves - Passeriformes	Furnariidae	128	**(40.3%)**	318

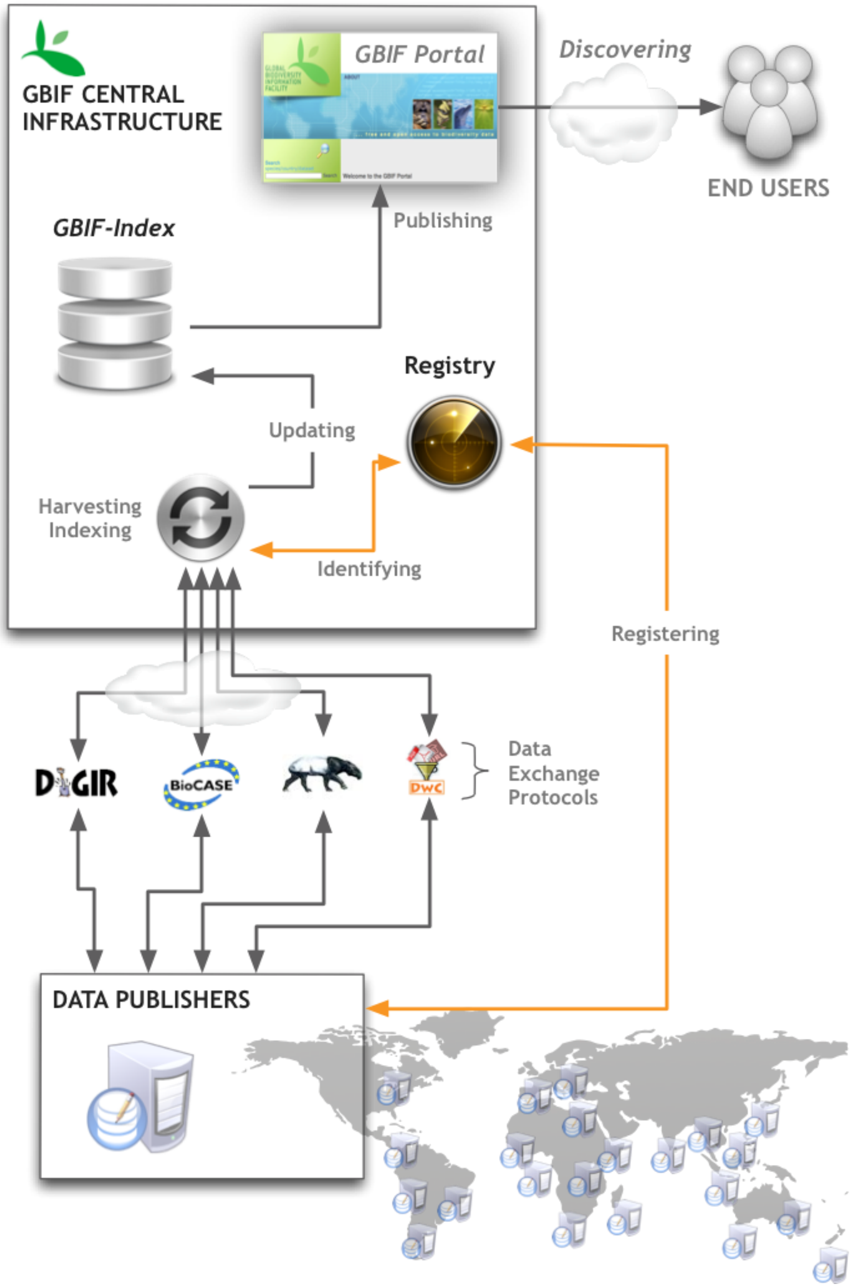


Figure 1. Typical flow of data discovered and published through the GBIF network.

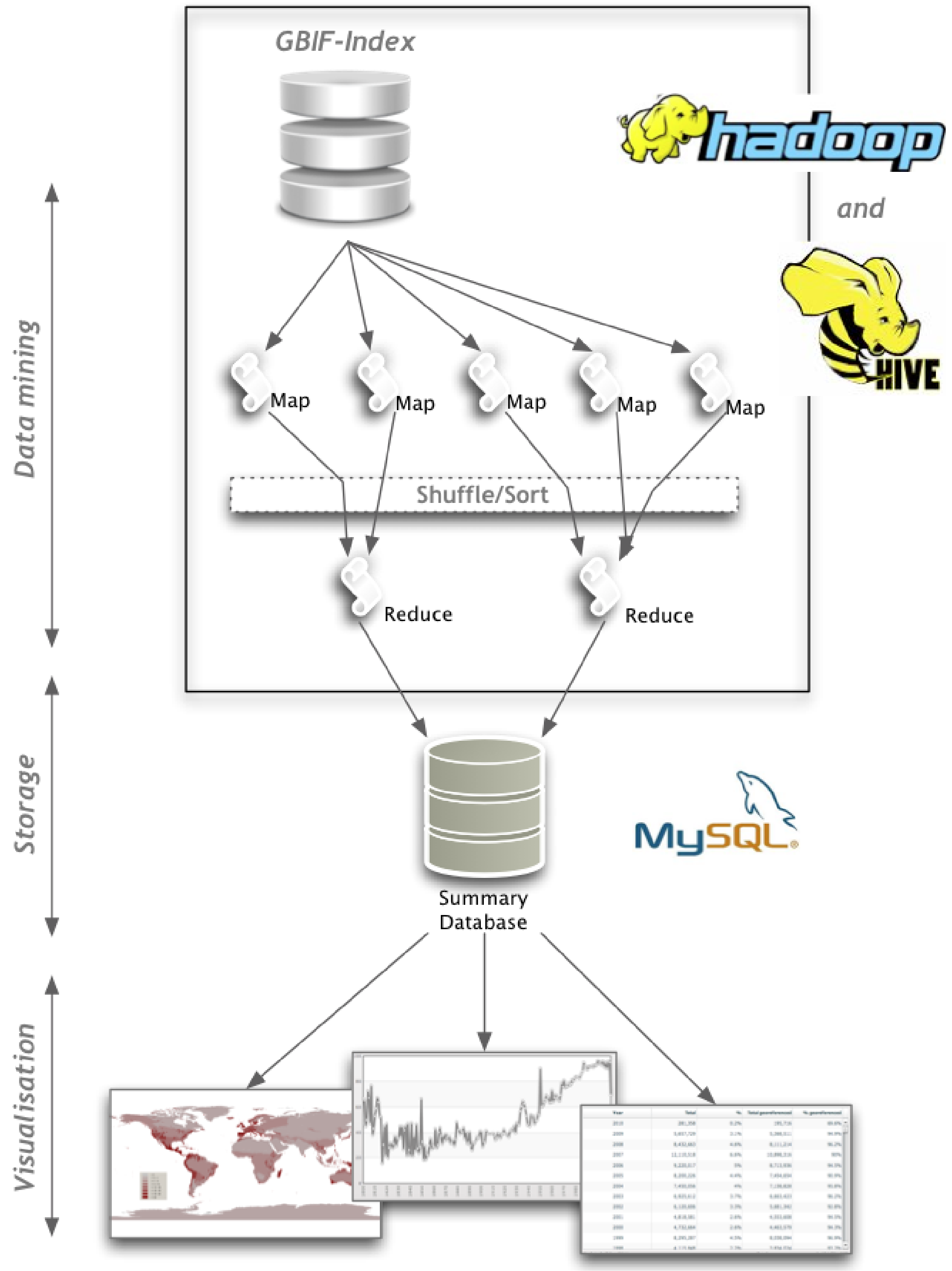
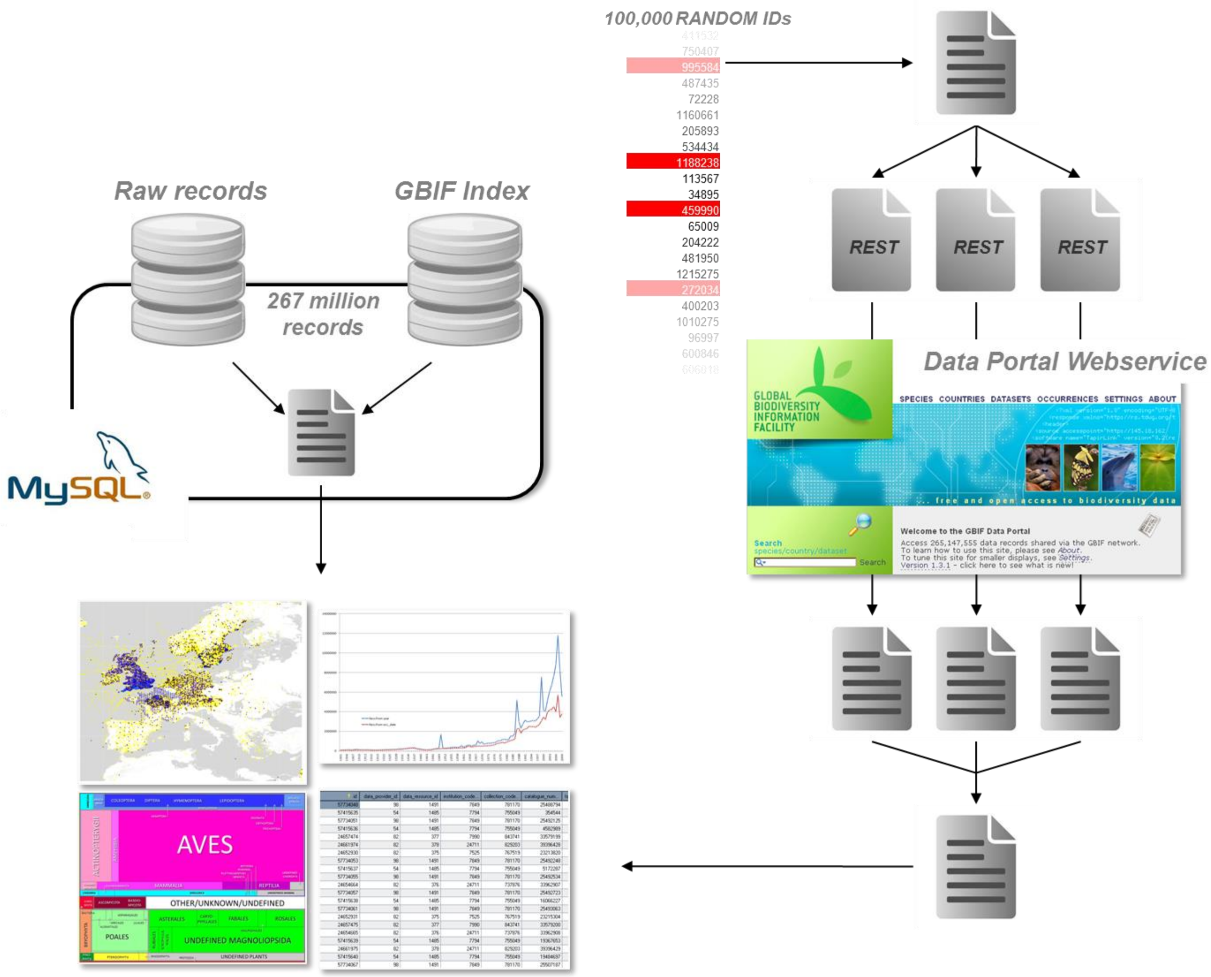


Figure 2. Data mining methodology employed during content assessment exercise carried out by the GBIF Secretariat.



Welcome to the GBIF Data Portal

Access 265,147,555 data records shared via the GBIF network.
To learn how to use this site, please see *About*.
To tune this site for smaller displays, see *Settings*.
Version 1.3.1 - click here to see what is new!

Figure 3. Data mining methodologies employed during content assessment exercise carried out by the University of Navarra.

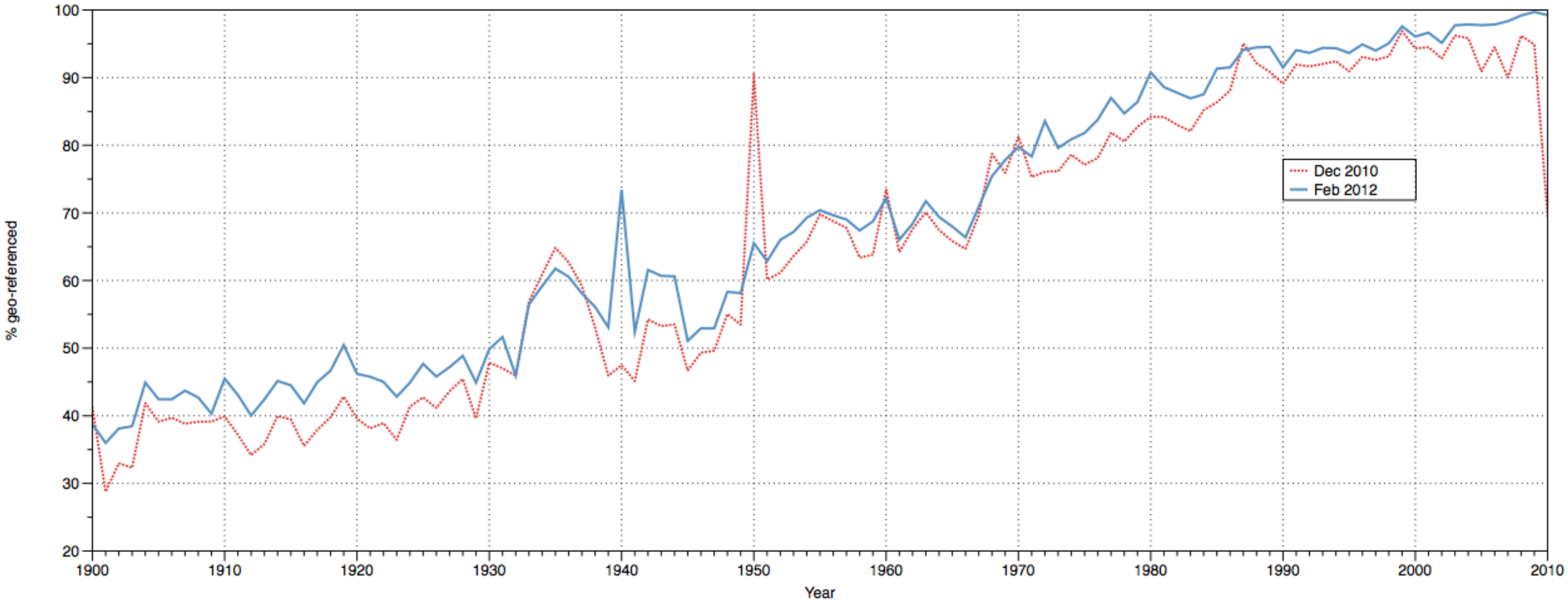


Figure 4. Evolution of the percentage of geo-referenced records (1800-2010).

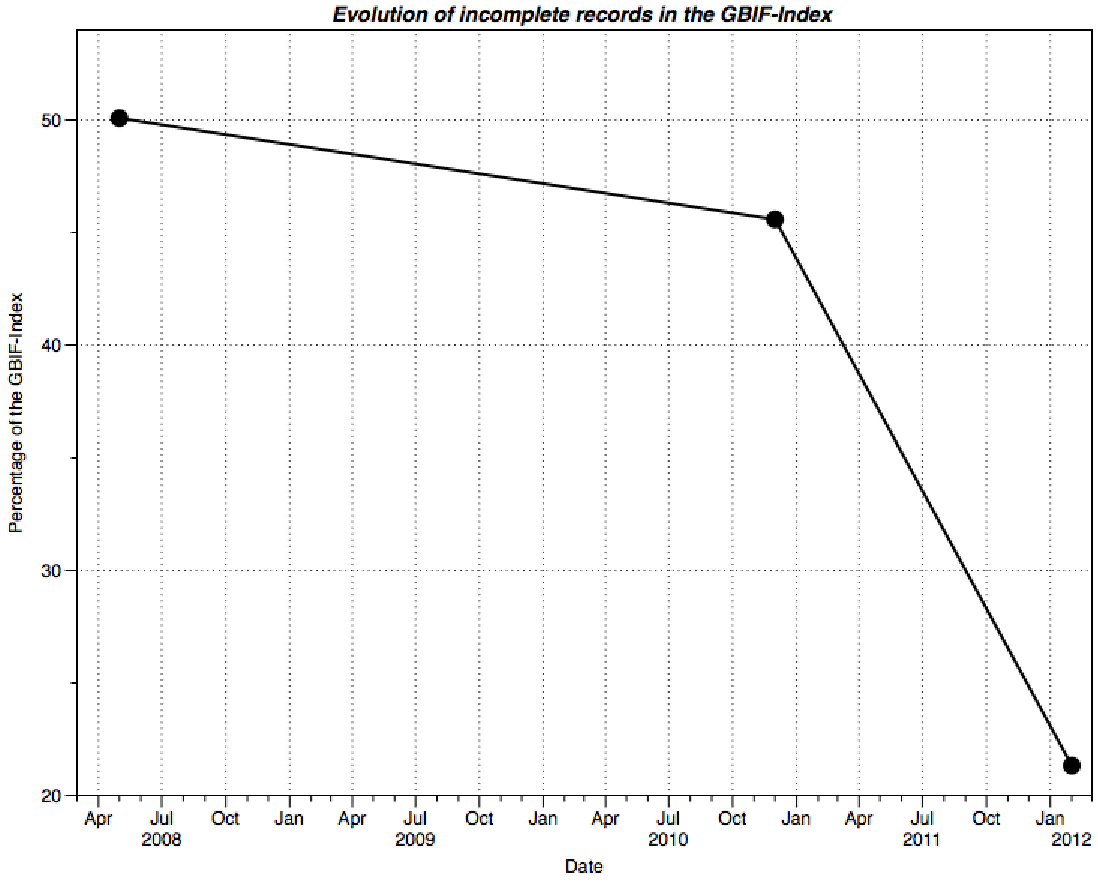


Figure 5. Evolution of incomplete records (taxonomical*temporal*geospatial) in the GBIF-Index

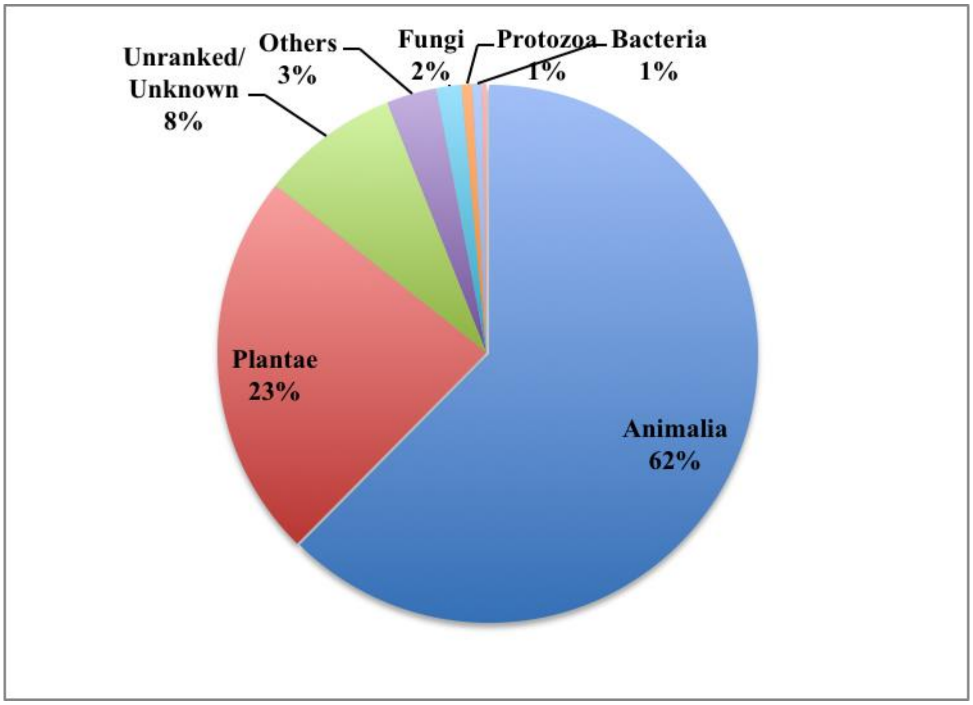


Figure 6.a. Data records by Kingdom (Dec 2010).

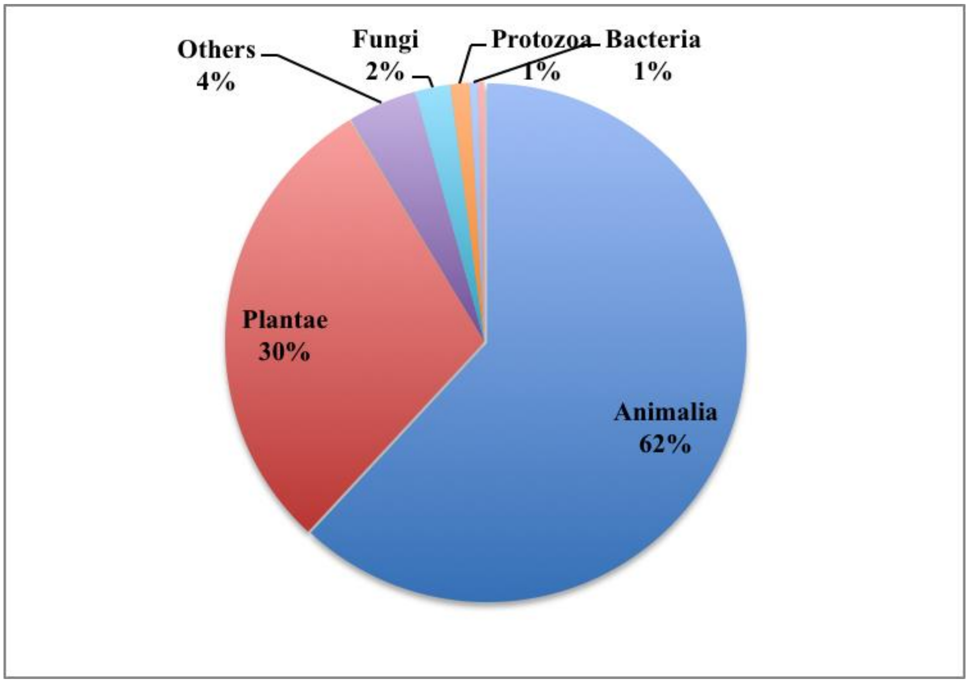


Figure 6.b. Data records by Kingdom (February 2012).

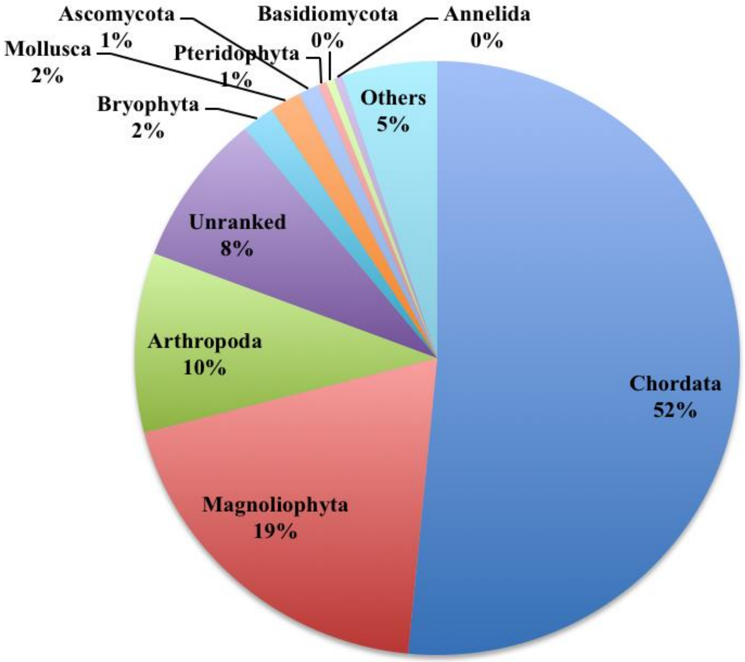


Figure 7.a. Data records by Phylum (December 2010).

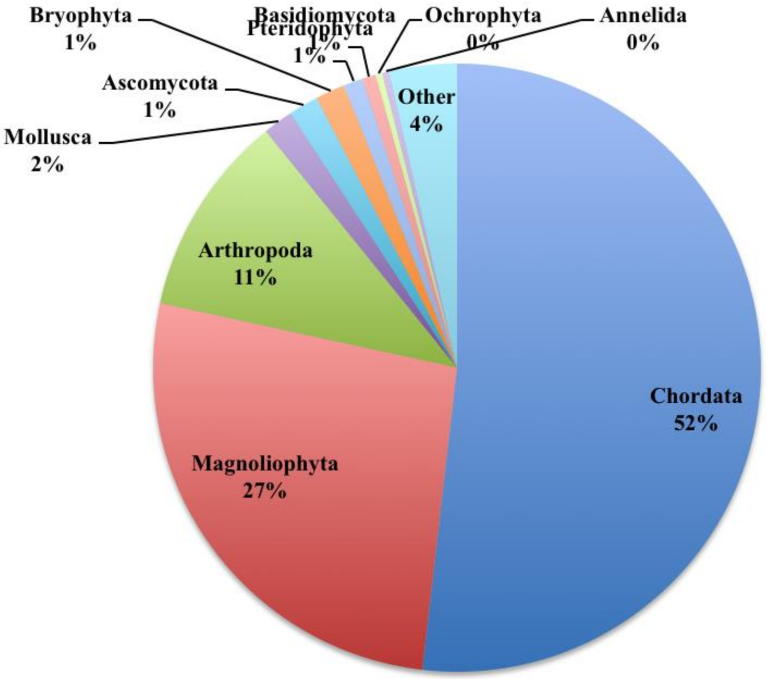


Figure 7.a. Data records by Phylum (February 2012).

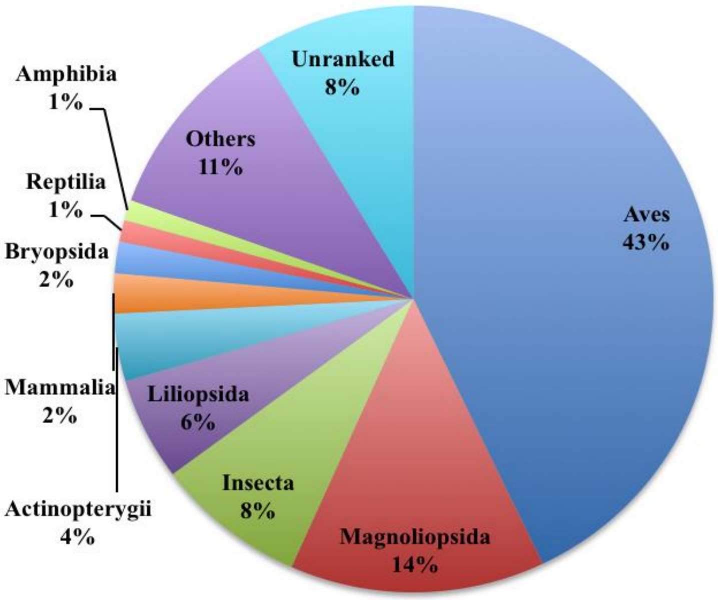


Figure 8.a. Data records by Class (December 2010).

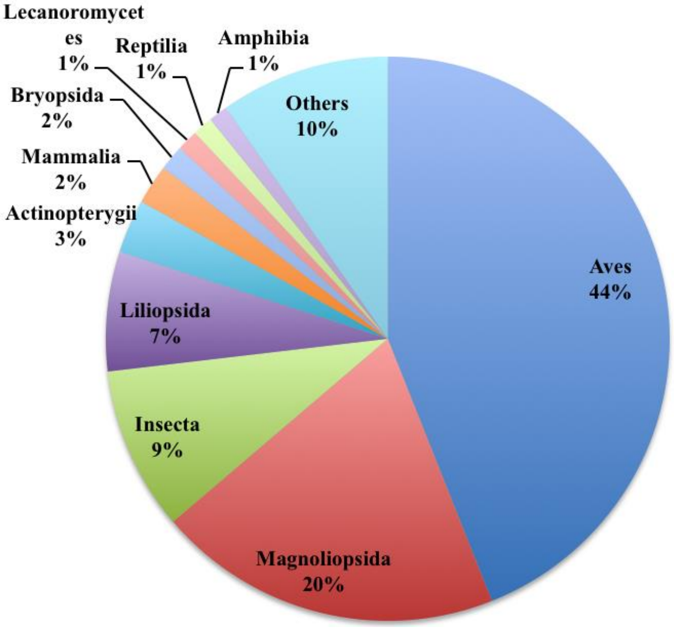


Figure 8.b. Data records by Class (February 2012).

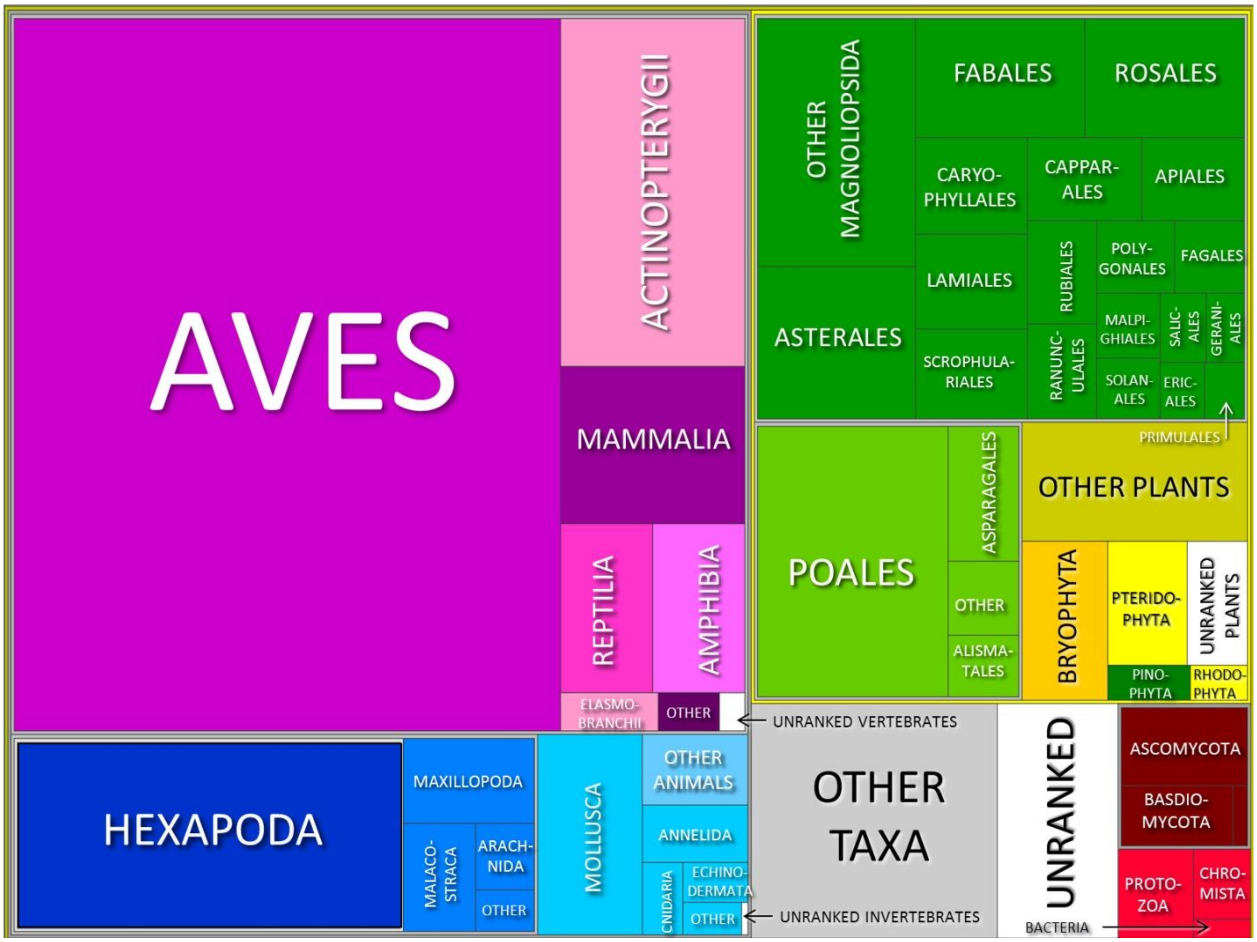


Figure 9. Treemap of occurrences according to taxon group, down to Class. Cell surface proportional to number of occurrences. Blue: Invertebrates; purple: Vertebrates; green: Higher plants; yellow: Algae and Ferns; brown: Fungi; red: unicellular organisms.

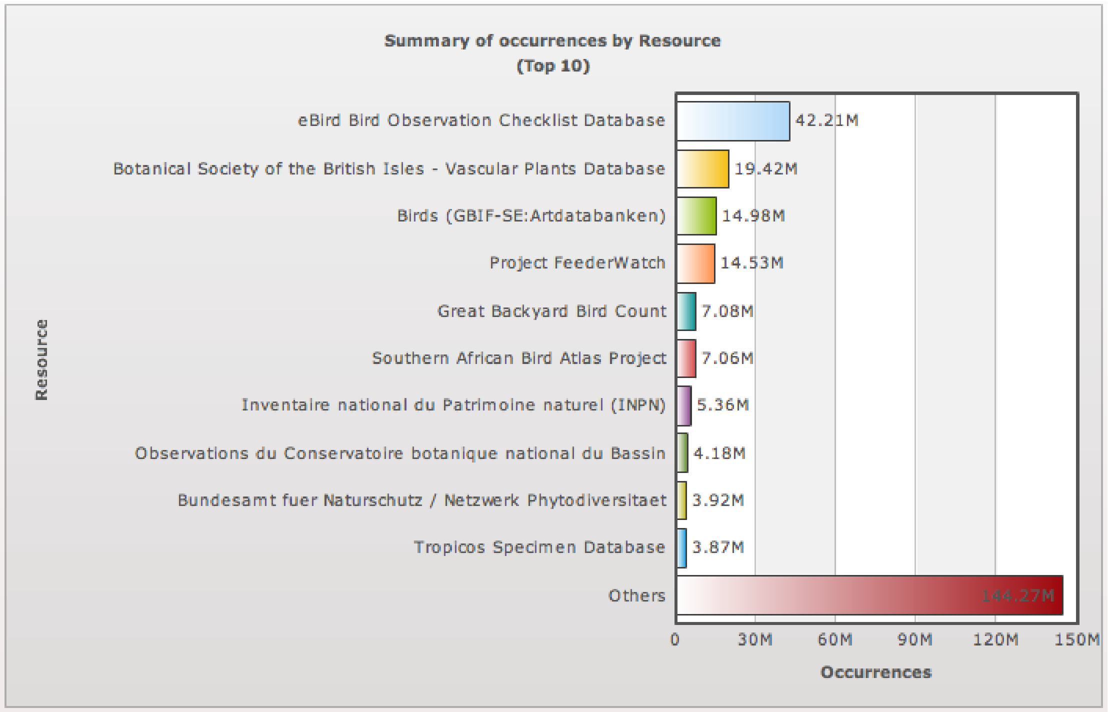


Figure 10. Top 10 data resources publishing maximum number of data records (December 2010).

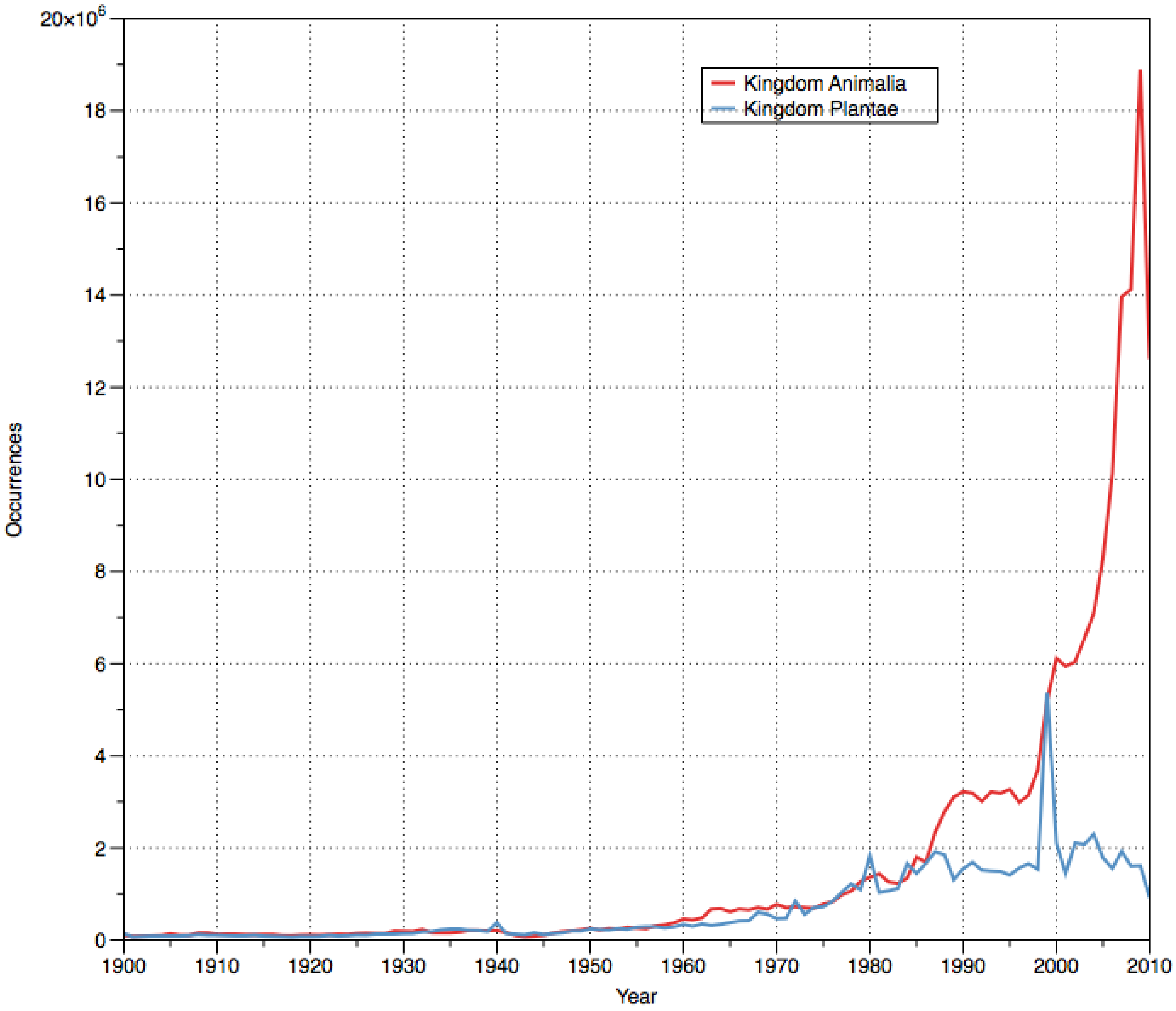


Figure 11. Kingdom wise distribution of data records by Year (Feb 2012).

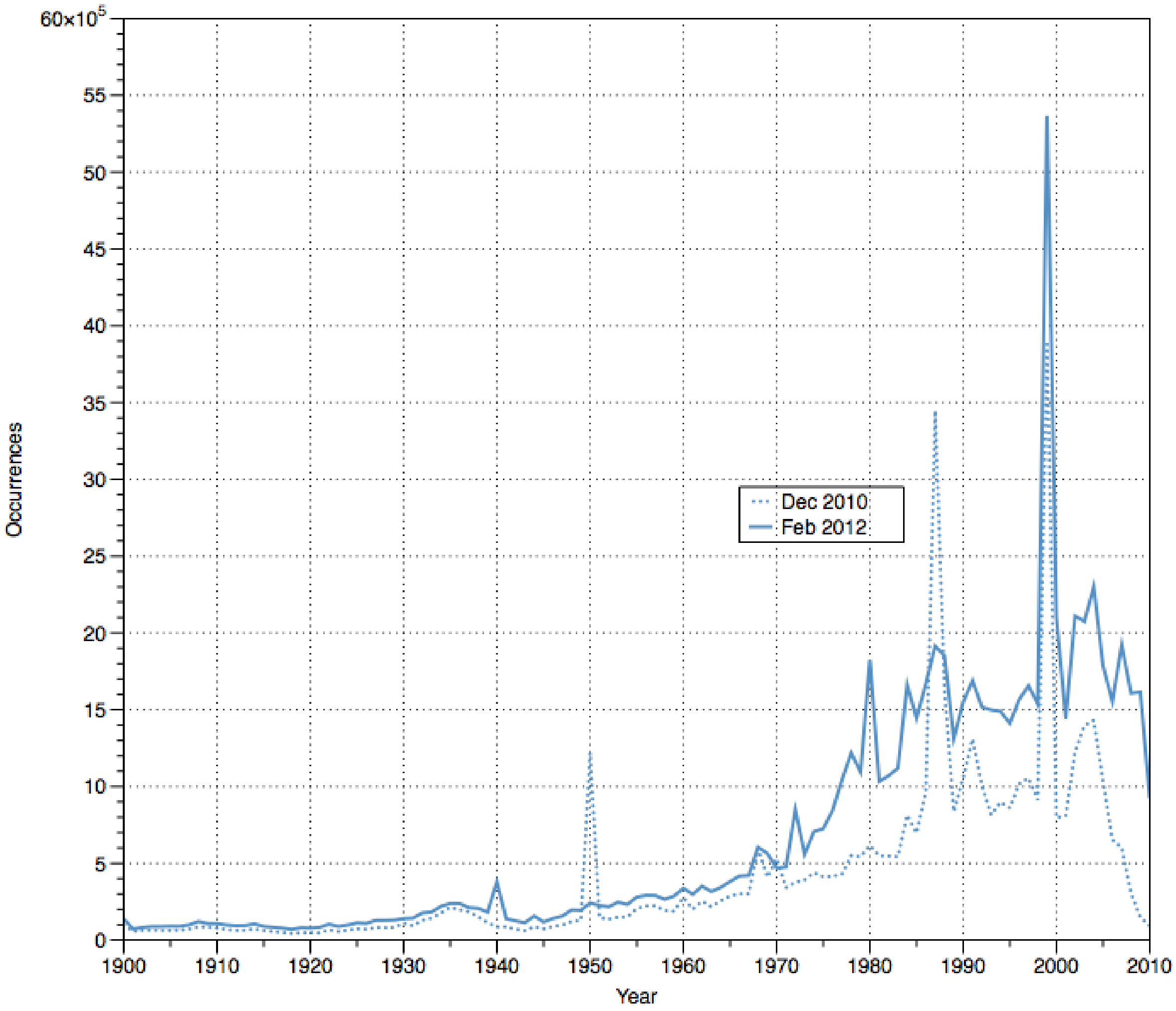


Figure 12.a. Plantae kingdom wise distribution of occurrences by year.

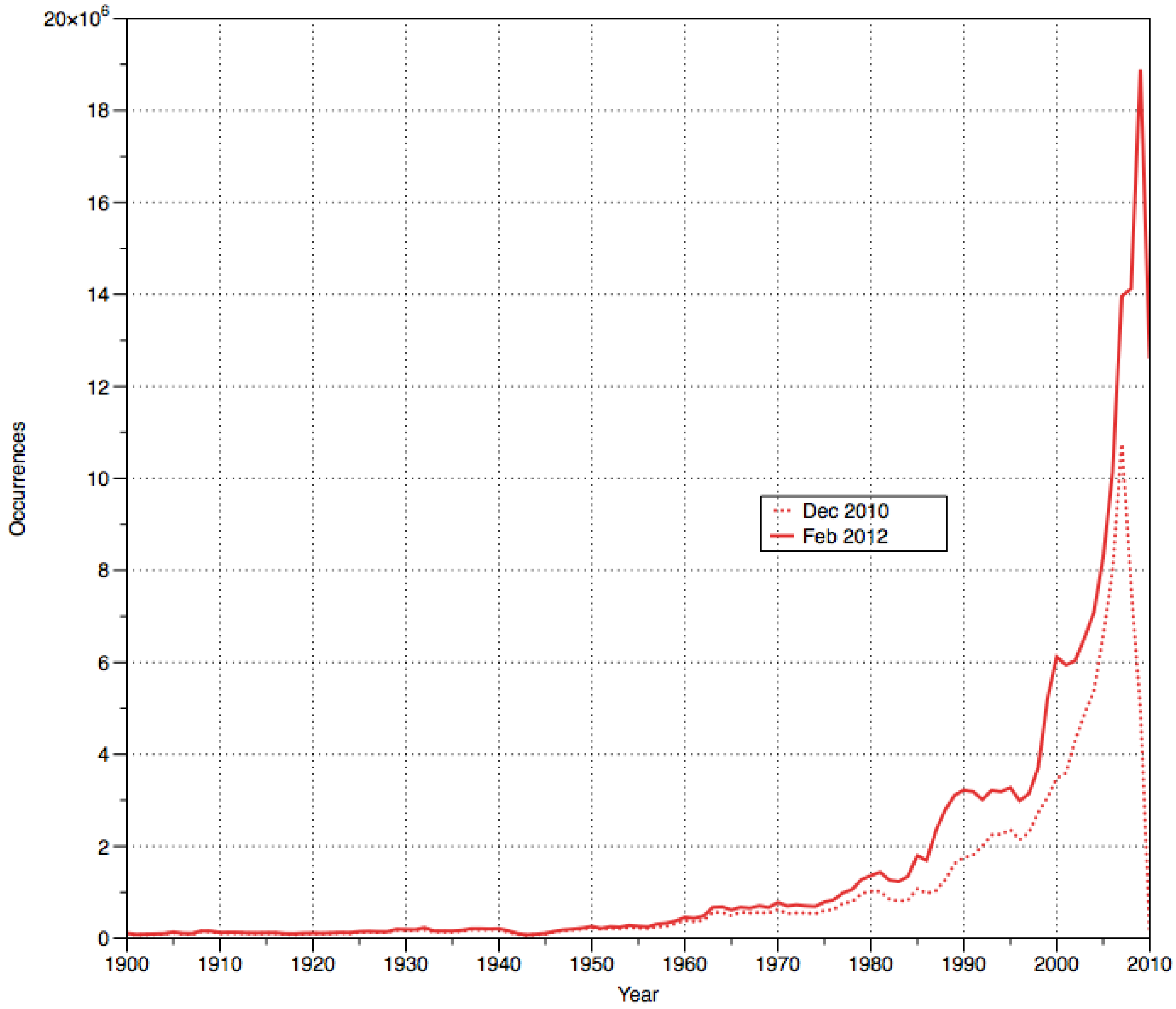


Figure 12.b. Animalia kingdom wise distribution of occurrences by year (February 2012).

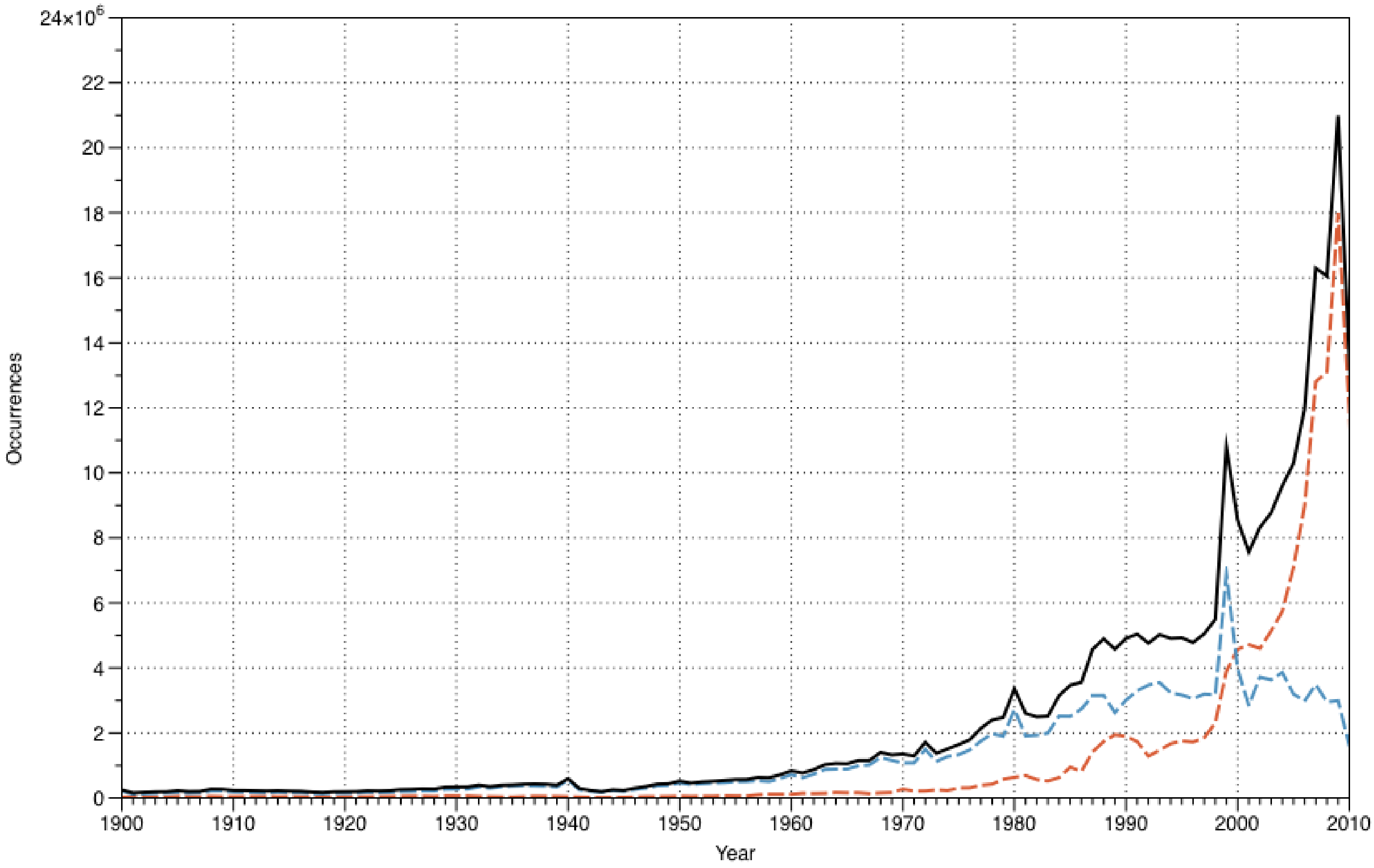


Figure 13. Aves wise distribution of occurrences by year (February 2012)

(red: Aves, blue: other classes, black: all classes)

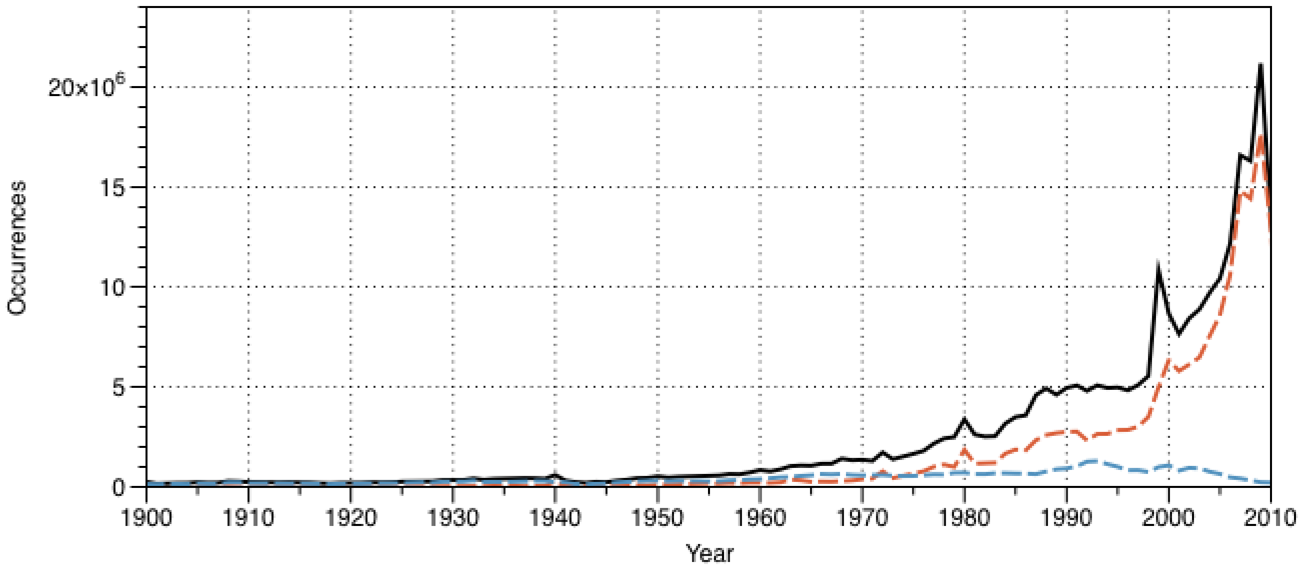


Figure 14. Basis of records distribution of occurrences by year (February 2012)

(red: observation, blue: specimen, black: all types)

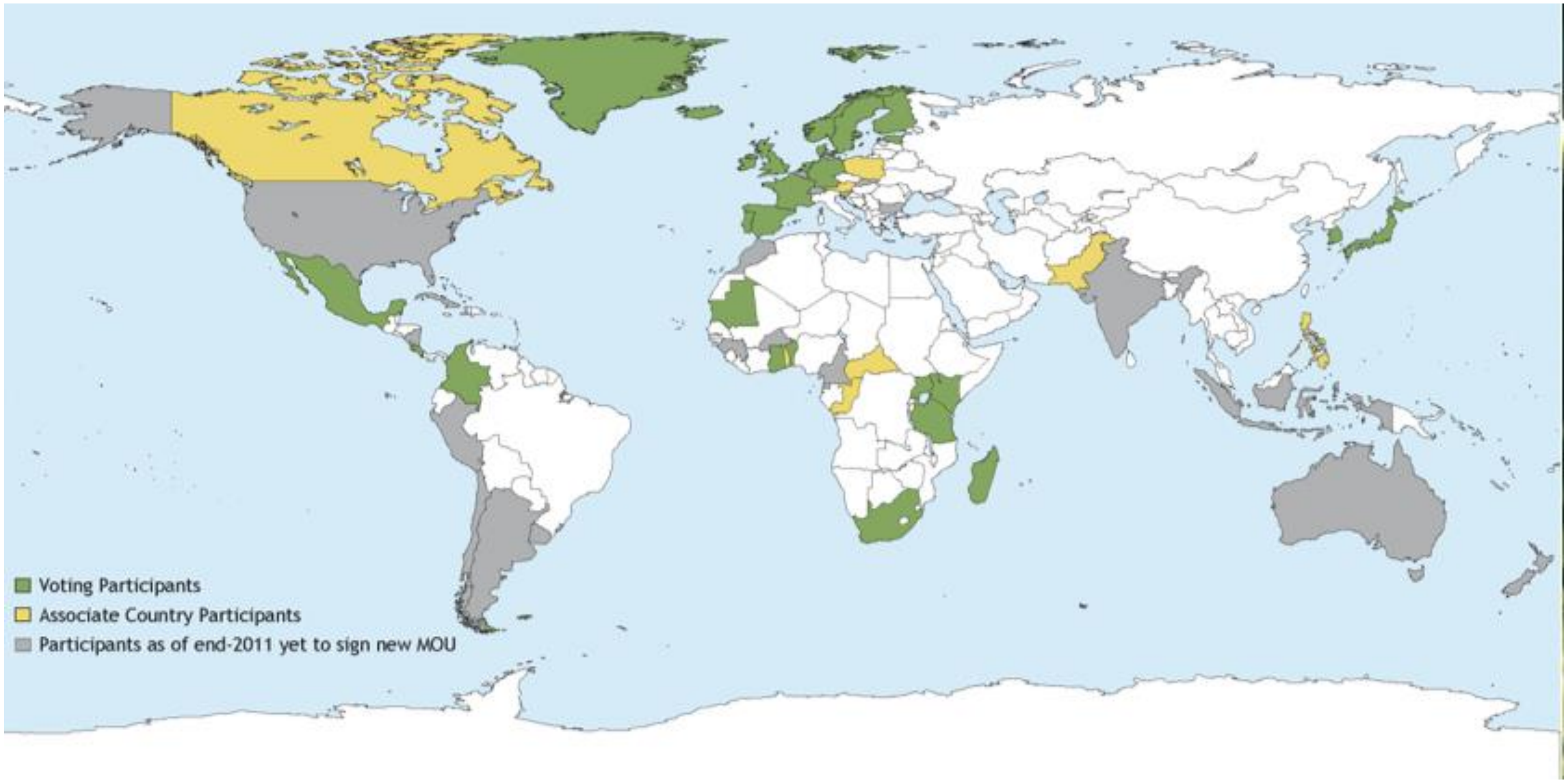

🟩 Voting Participants
🟨 Associate Country Participants
⬜ Participants as of end-2011 yet to sign new MOU

Figure 15. Distribution of voting and associate country participants (February 2012)

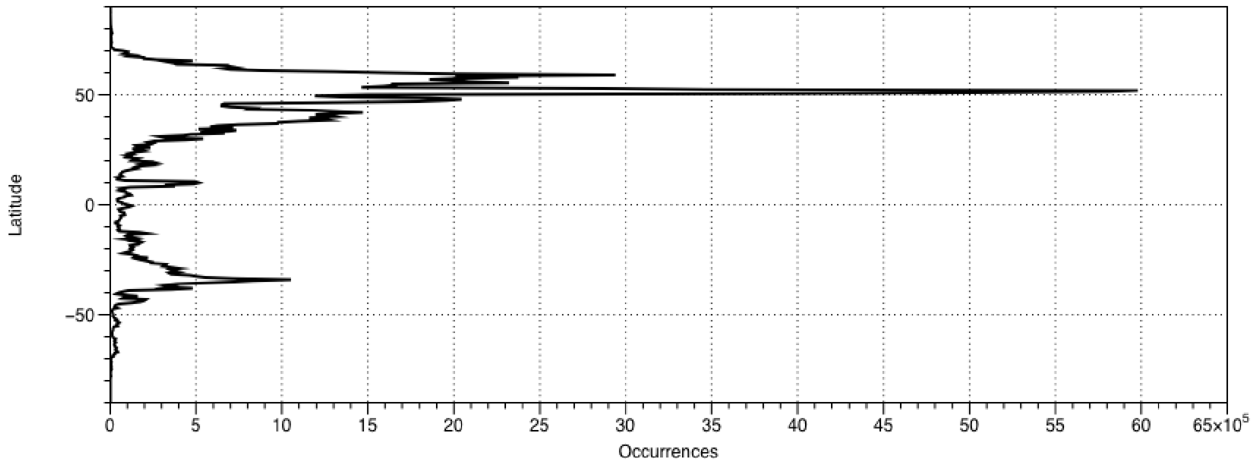


Figure 16. Distribution of occurrences by latitude (February 2012)

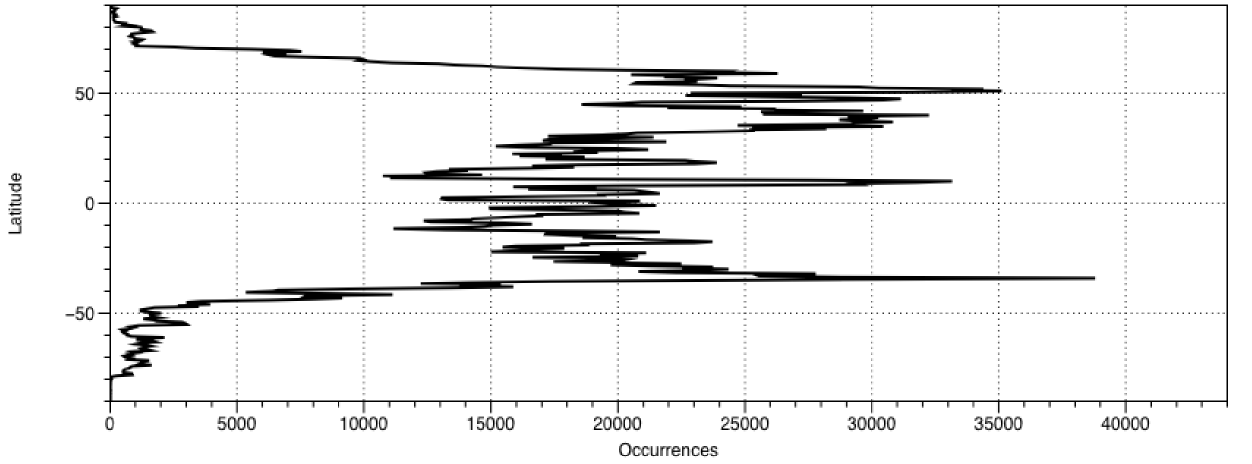


Figure 17. Distribution of species richness by latitude (February 2012)

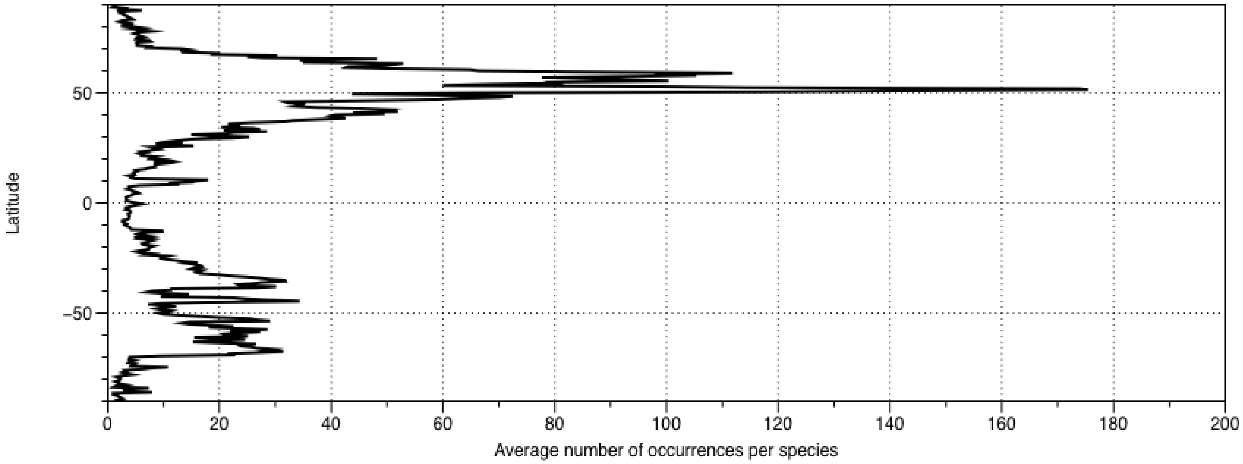


Figure 18. Distribution of the average number of occurrences per species by latitude (February 2012)

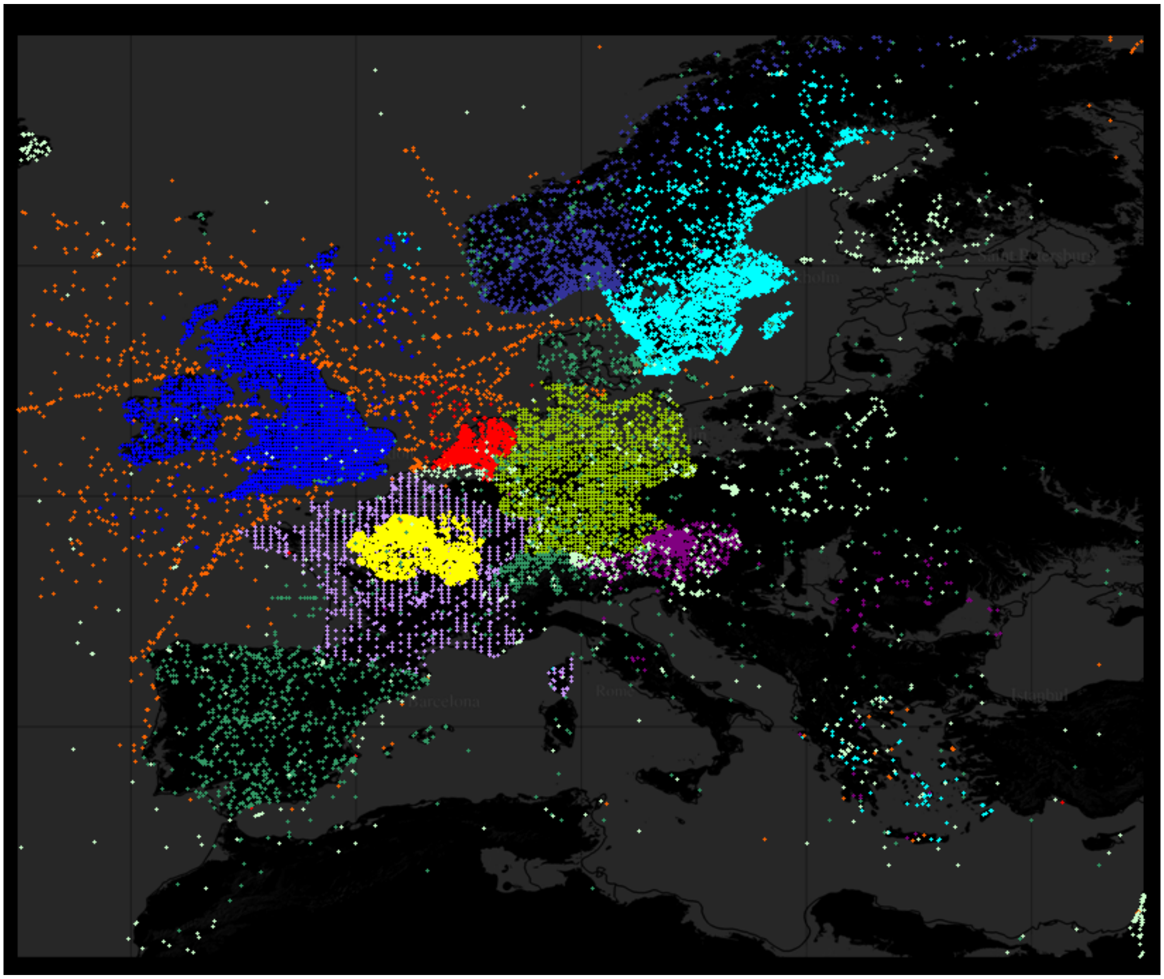


Figure 19. Georeferenced records (dots) published by the largest data providers in Europe. Colors represent distinct providers. (From Otegui et al., 2009).

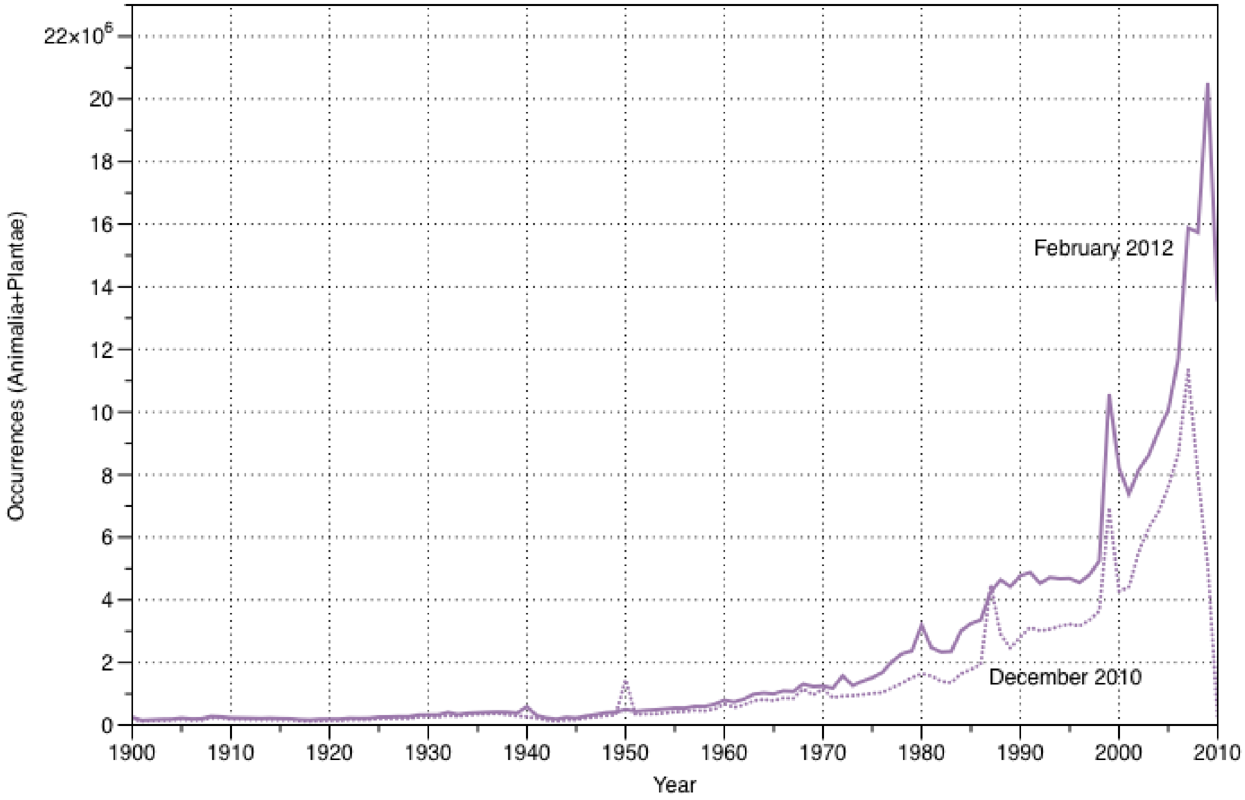


Figure 20. Records available in the GBIF-Index by date

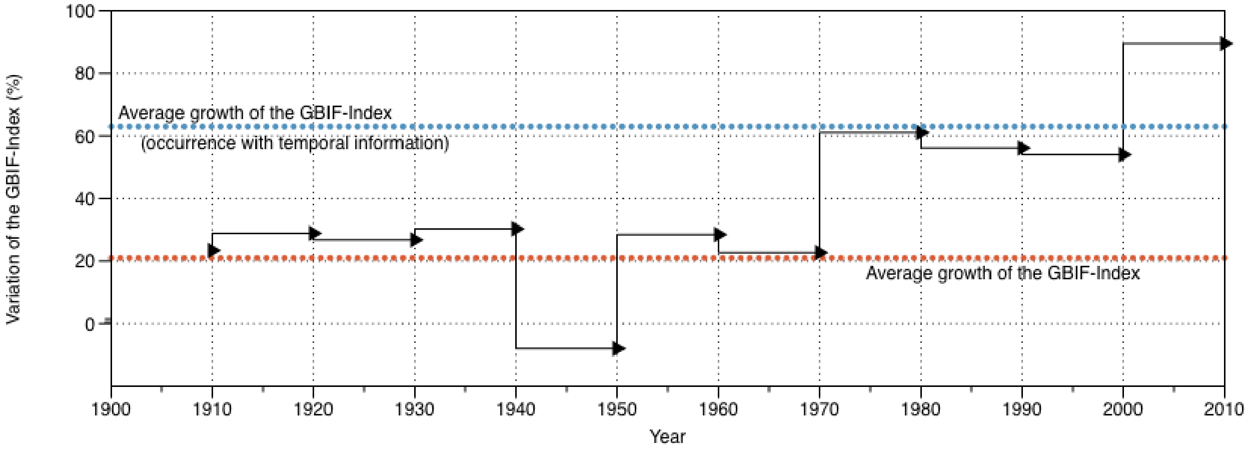


Figure 21. Variation of the records available in the GBIF-Index by date (December 2010 – February

2012)

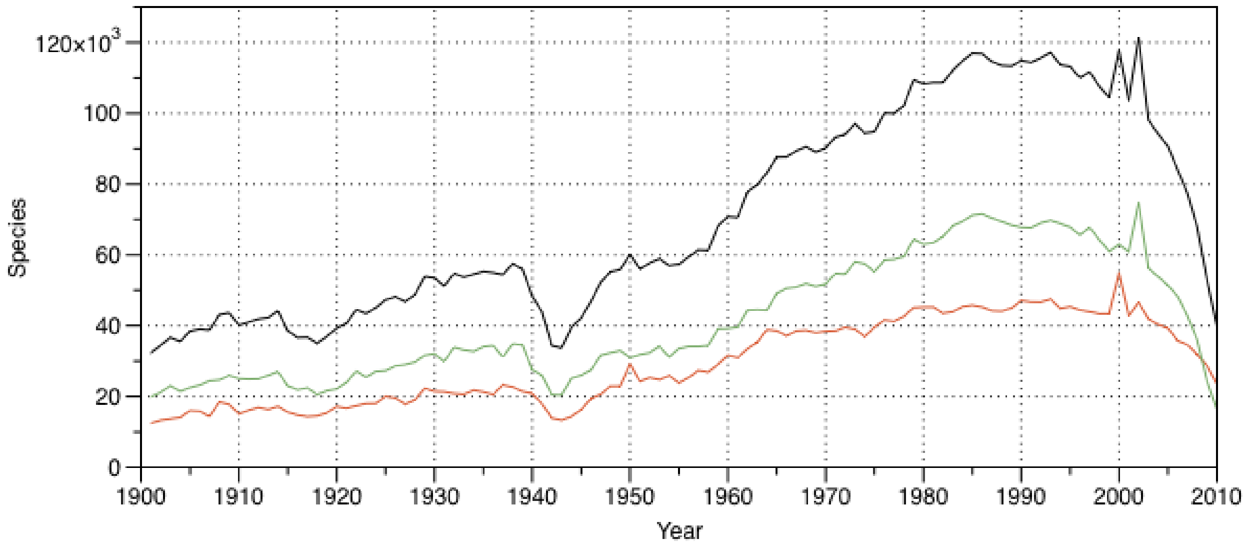


Figure 22. Distribution of distinct species collected/observed over time (February 2012)

(red: Plantae, green: Animalia, black: both)

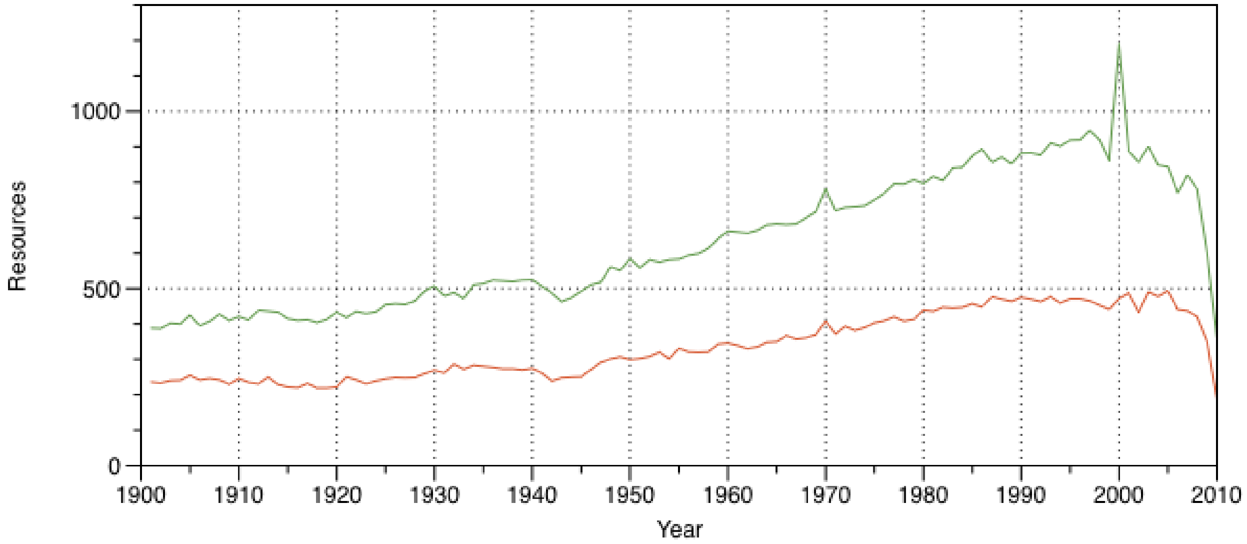


Figure 23. Distribution of distinct resources contributing to the GBIF-Index over time (February 2012)

(red: Plantae, green: Animalia)

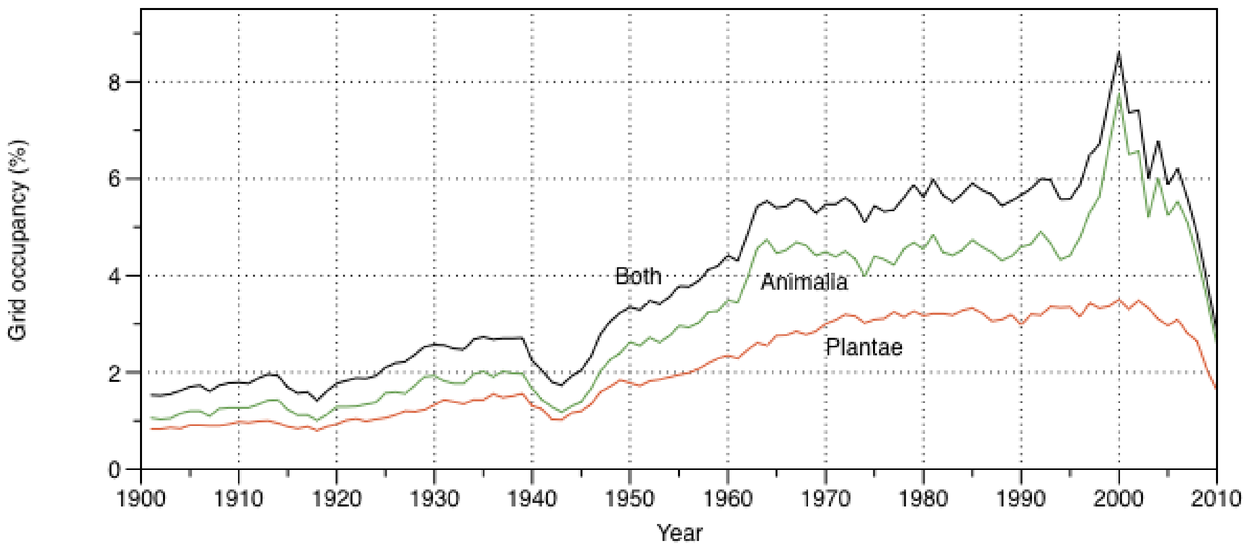


Figure 24. Grid occupancy (1/2 degree grid) of the GBIF-Index over time (February 2012)

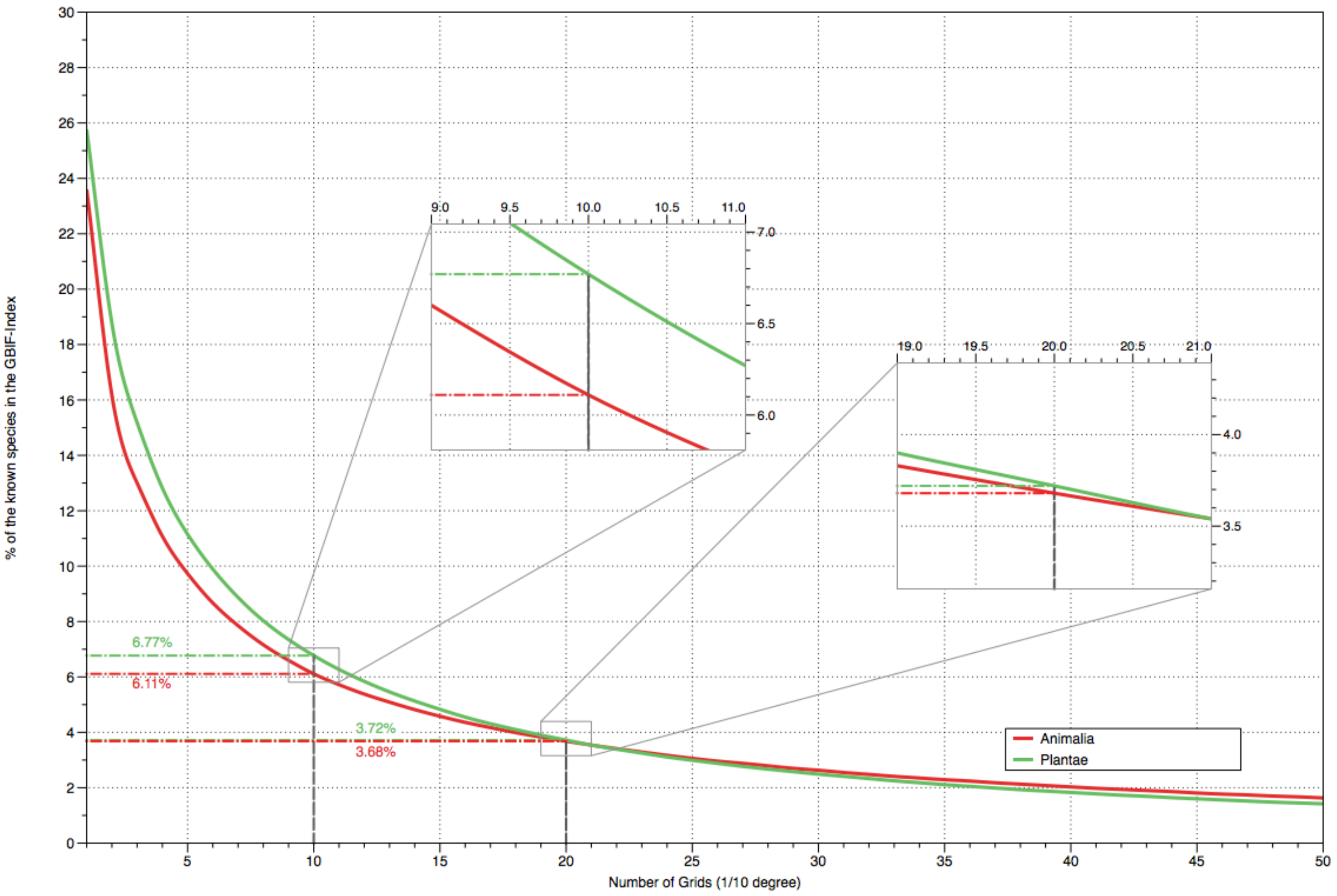


Figure 25. Grid occupancy (1/10 degree grid) of the GBIF-Index by Kingdom (February 2012)

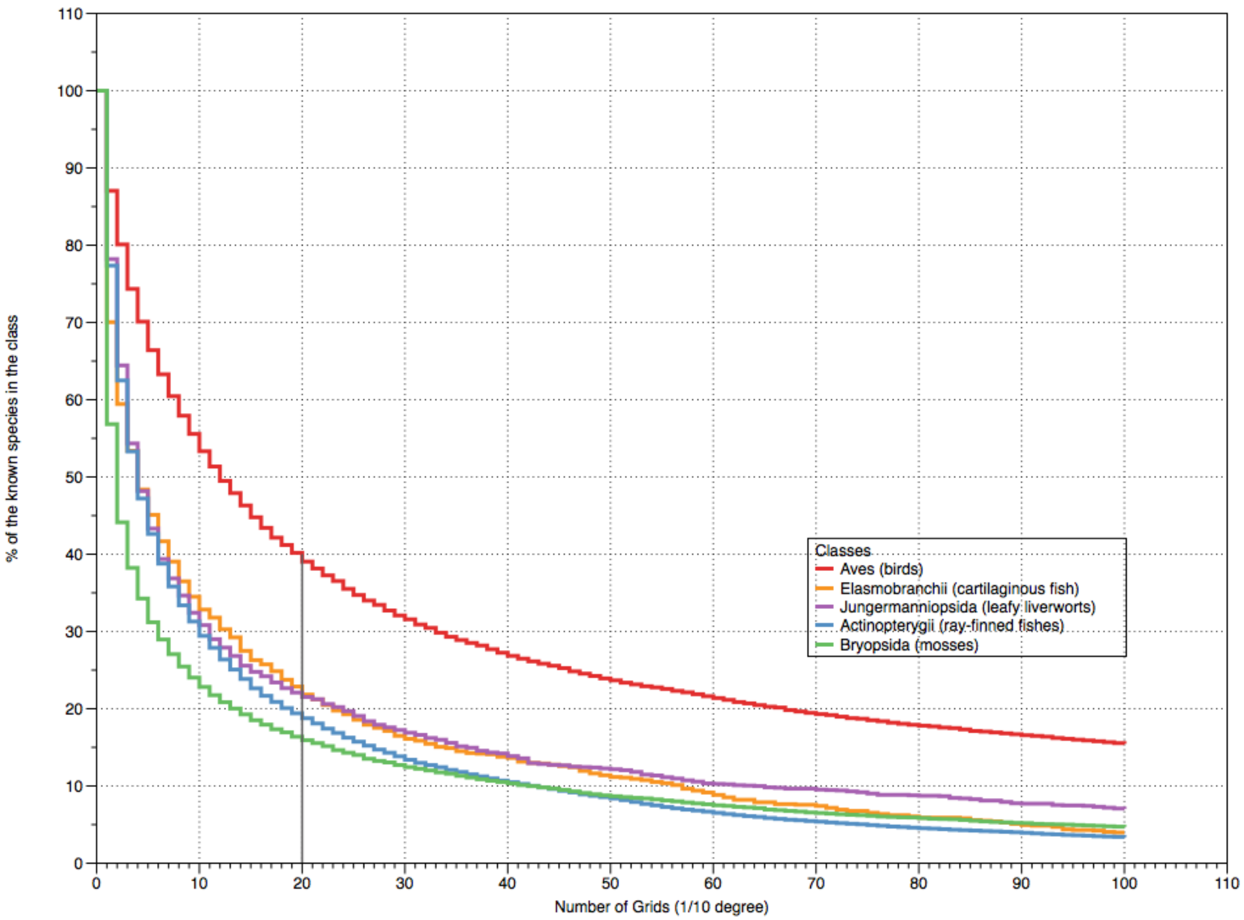


Figure 26. Grid occupancy (1/10 degree grid) of the GBIF-Index by Classes (February 2012)

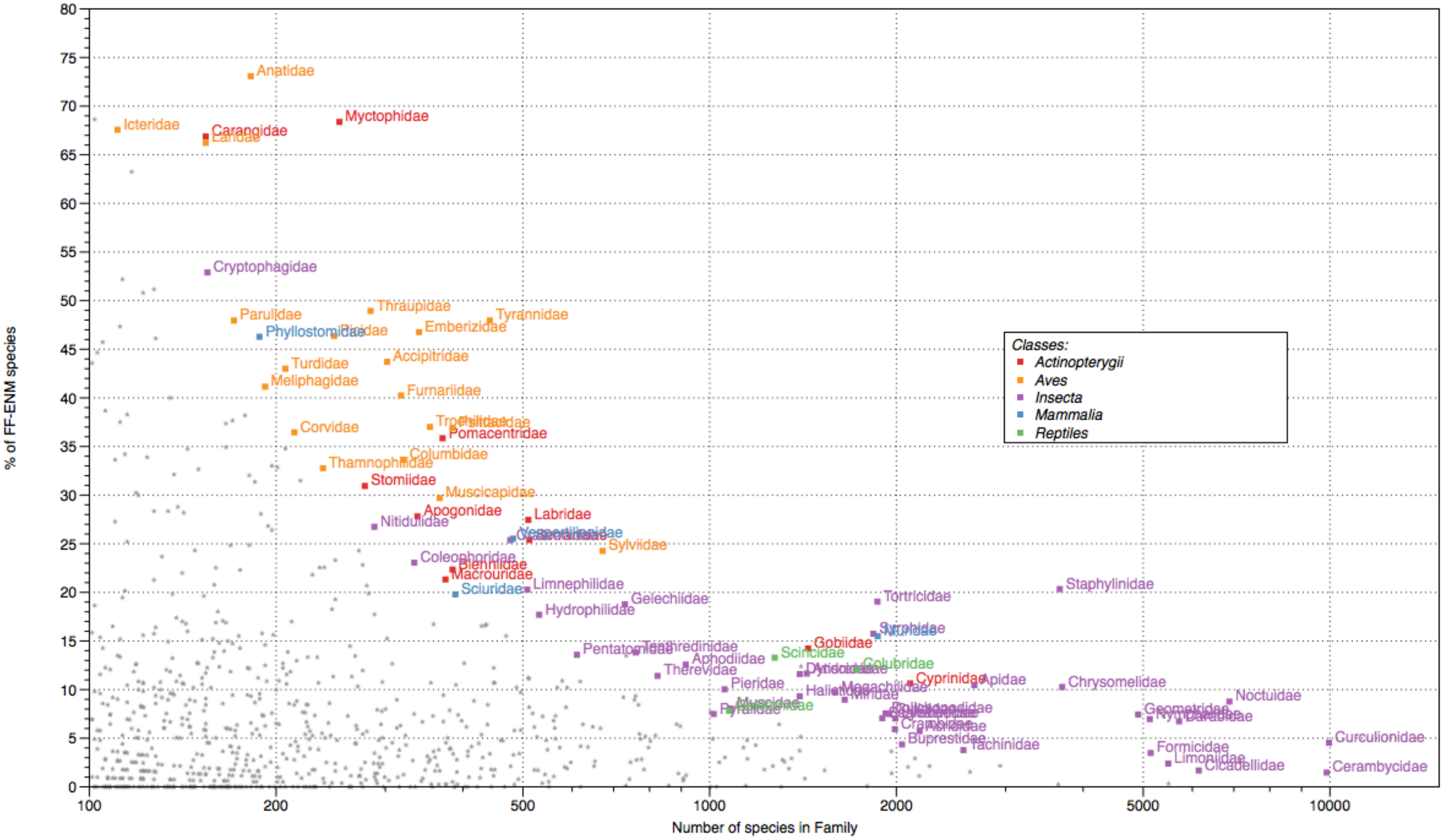


Figure 27. Distribution of families for FFU-ENM – Kingdom Animalia (February 2012)

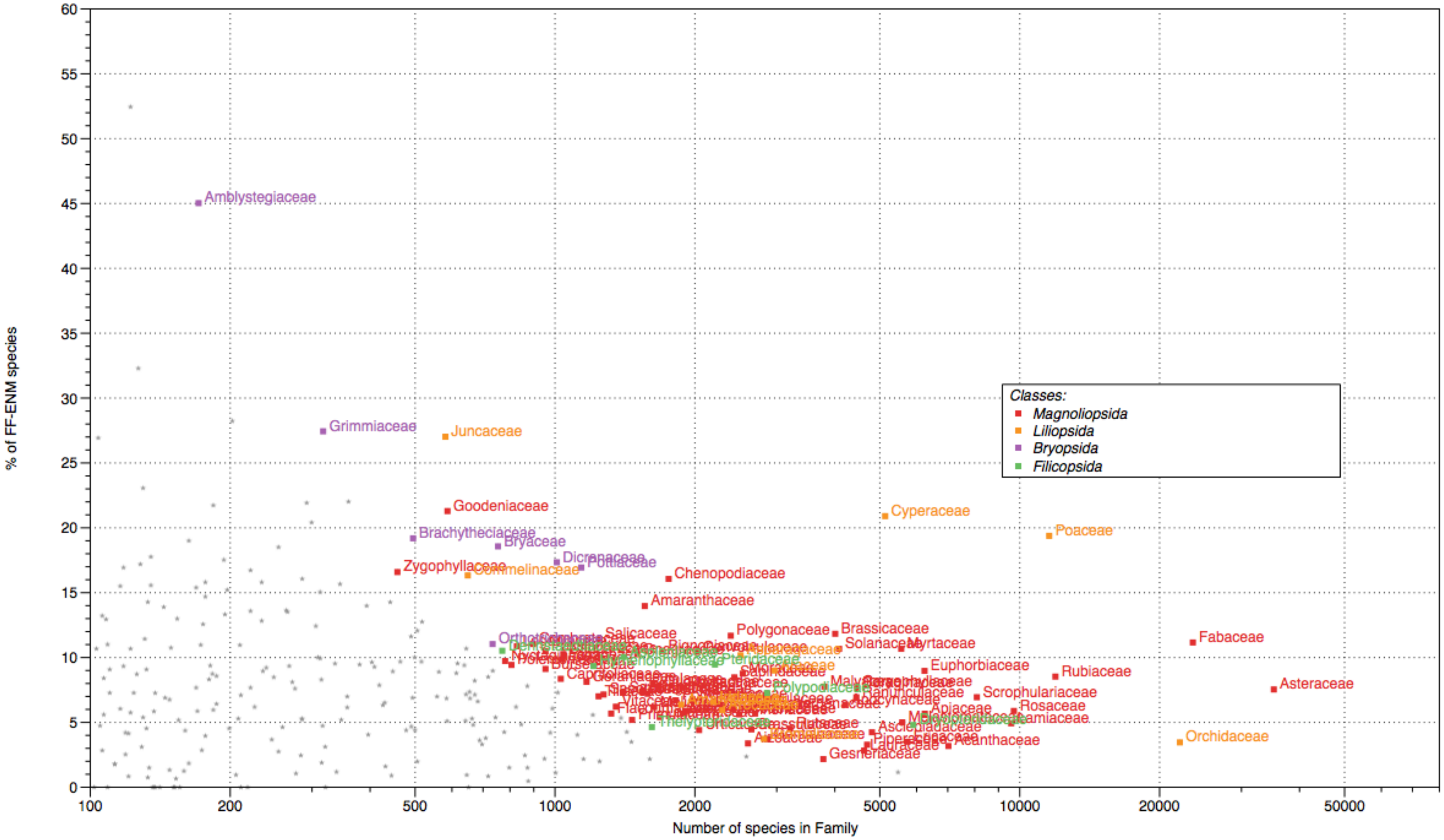


Figure 28. Distribution of families for FFU-ENM – Kingdom Plantae (February 2012)

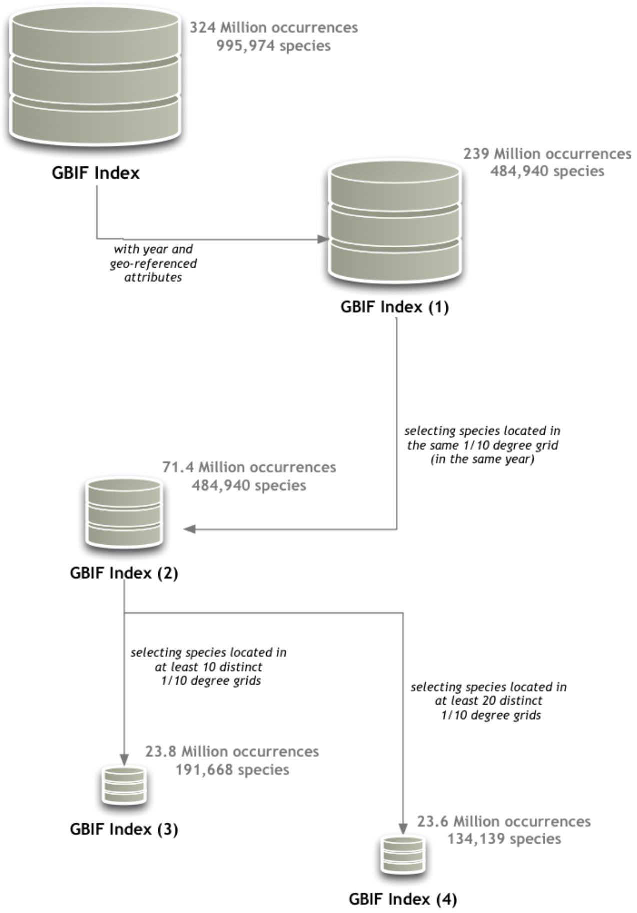


Figure 29. Selection of fit-for-use records in the GBIF Index for ecological niche modelling (ENM).