GBIF Global Biodiversity Information Facility

Final Report of the Task Group on GBIF Data Fitness for Use in Distribution Modelling

Version 1.1 published on 22 March 2016

Are species occurrence data in global online repositories fit for modeling species distributions? The case of the Global Biodiversity Information Facility (GBIF)

Authors (in alphabetical order)

Robert P. Anderson, City University of New York, USA, randerson@ccny.cuny.edu

Miguel Araújo, Museo Nacional de Ciencias Naturales, Spain maraujo@mncn.csic.es

Antoine Guisan, University of Lausanne, Switzerland, antoine.guisan@unil.ch

Jorge M. Lobo, Museo Nacional de Ciencias Naturales, Spain, mcnj117@mncn.csic.es

Enrique Martínez-Meyer, Universidad Nacional Autónoma de México, Mexico, emm@ib.unam.mx

A. Townsend Peterson, University of Kansas, USA, town@ku.edu

Jorge Soberón, University of Kansas, USA, jsoberon@ku.edu

with support from Dmitry Schigel, GBIF Secretariat, dschigel@gbif.org

Executive Summary

Primary Biodiversity Data (PBD) are defined as the basic attributes of observations or records of the occurrences of species. PBD is a fundamental concept of biodiversity informatics since it is substantial in quantity and provides the links to organize other large and independent bodies of data concerning species (= taxonomic information) and environments. In fact, PBD is at the core of the exploding field of biodiversity informatics, which in some sense now underlies biogeography, macroecology, landscape ecology and several other subdisciplines of biology.

A principal – and rapidly growing – class of research that can be performed using PBD is the estimation of a species' environmental requirements and the projection of these in both environmental and geographic spaces to estimate niches or distributional ranges, generally by using models of ecological niches and species' distributions (often called ENMs or SDMs, respectively).

The largest point of access to PBD in the world is the Global Biodiversity Information Facility (GBIF), and hundreds of papers have now used GBIF-mediated data to fit and apply ENM/SDM.

Experience has shown that GBIF, like other aggregated data research infrastructures, holds a number of potential problems related to incomplete or difficult access to all the fields in its schema, inconsistent information among fields, or simply erroneous or incomplete data. These drawbacks complicate ENM/SDM analyses considerably, and detract from the enormous scientific value of this information storehouse.

Three overlapping communities participate in GBIF's data process: providers (museums, herbaria, and observer's networks), users (scientists, analysts working for governments, NGOs or the private sector, the public) and the technical staff managing the huge databases, web services and servers at GBIF. Each can play a different role in fixing data issues of GBIF.

Our main recommendations for the GBIF Secretariat are the following:

- GBIF.org should serve indicators of precision, quality, and uncertainty of data that can be calculated practically, and preferably "on the fly", as well as summaries and metrics of completeness of inventories, at scales and for regions defined by the user. The summaries should display maps and graphs of completeness by region, time-period and taxa.
- The implementation of the GBIF information resource should go beyond unique identifiers of queries (DOIs for downloads, including the capability to re-run queries, http://www.gbif.org/publishing-data/summary#supporteddatasettypes), and to include identifiers of the individual data that make up the queried data.
- GBIF.org should include applications or functionalities enabling users to annotate errors or problems, and communicate those changes directly to providers, as it may be practical and appropriate. This point may need to be discussed with providers.
- A procedure enabling users to make accessible versions of their databases that have been improved and annotated should be supported, but this functionality should not lose the vital tie back to the original data records and the actual data provider.
- GBIF should partner with and/or support initiatives to do more for training and guiding users on the proper use of the data; such initiatives should incorporate actual expert uses in ENM/SDM to assure that current best practices are followed.

1 Introduction

1.1 Primary data on biodiversity

Information about different aspects of biodiversity has been accumulating for centuries, but most of it is preserved in heterogeneous formats, unreadable by machines, and largely unconnected and disorganized in textual form, images, or stand-alone databases (Scholes *et al.* 2008). The amount of biodiversity information accumulated in these heterogeneous formats is truly staggering, although it remains highly biased geographically and taxonomically.

Linking and making these knowledge domains interoperable is a major challenge and frontier for biodiversity informatics (Peterson *et al.* 2010). Links between domains that have proven to be both practical and powerful function via common fields, such as scientific name and geographic location, which can unite data records that have certain elements in common (e.g., place, species). Primary biodiversity data (PBD) consist of the basic attributes of individual specimens or observations, such as locality, date, scientific name, phenotypic measurements, images, voice recordings, etc. Links between domains, ideally, will be established via properties of individual organisms (i.e., sequence data of a specimen, body size, secondary chemistry), rather than linking via attributes associated with taxonomic names or polygons in maps, in other words, via secondary information. However, in practice, it is likely that such connections will not only take place through primary data, but using combinations (for instance, distribution of body mass on the basis of reported mean weight of taxa, rather than on weights of individual specimens).

Many applications can be built on the basis of such simple links, as the large number of papers on species distribution modeling (= linking data records by name to enumerate a large number of places, by which one links to environmental data) exemplifies (Guisan *et al.* 2013). However, to achieve the volume and reach necessary for a PBD database to become useful at geographic scales, data from various sources should be aggregated and made broadly available. Several databases have been providing this service since at least the last 20 years, including not only many in institutions of the developed world, but also several in the developing countries, such as Costa Rica's INBio[1], Colombia's Instituto von Humboldt[2], South Africa's SANBI[3], and Mexico's CONABIO[4]. More recently, still more comprehensive initiatives have begun, including the United States' iDigBio[5] and Australia's Atlas of Living Australia[6].

Among such initiatives, the Global Biodiversity Information Facility (GBIF)[7] is by far the largest point-of-access, indexing datasets provided by other publishers and/or initiatives, such as the Ocean Biogeographic Information System (OBIS)[8] and the Global Invasive Species Database (GISD)[9]. Compiling, maintaining and operating large PBD databases entails overcoming major technical and organizational obstacles, including challenges for both data providers and data users, particularly if the data have heterogeneous origins (Soberón *et al.* 2002a; Chapman 2005). When a heterogeneous-provenance database grows to the size of GBIF (more than 2.5 M names, including 1.6 M confirmed species in the Catalogue of Life[10], and more than 650 M occurrence records), issues related to the

[1] http://www.inbio.ac.cr/en/
[2] http://www.humboldt.org.co/en
[3] http://www.sanbi.org/
[4] http://www.biodiversidad.gob.mx/
[5] https://www.idigbio.org/
[6] http://www.ala.org.au/
[7] http://www.gbif.org
[8] http://www.iobis.org
[9] http://www.issg.org/
[10] http://www.gbif.org/dataset/7ddf754f-d193-4cc9-b351-99906754a03b

consistency, quality, and reliability of the data can become even more important. As a consequence, to assure full and appropriate uses, such a database should include:

- Fields describing error and uncertainty associated specifically with taxonomy, georeference, and collection date (and others);
- Ways and tools for users to display, visualize, and explore data, and highlight possible inconsistencies or errors (Soberón *et al.* 1996; Soberón *et al.* 2002b; Chapman 2005);
- Ways of providing feedback about inconsistencies and errors from users to the data aggregator, and from the aggregator to the original data providers;
- Stable unique identifiers for queries have been available in GBIF since Sept 2014. However, DOIs for the individual objects in the databases are crucial to full, individual-record-level linkage of data domains; lacking such individual identifiers, fields linking data records to non-PBD repositories can be included, like those related to environmental or species features (ethnobotany, DNA sequences, morphology, physiology – Peterson *et al.* 2010). It is clear to our panel that such DOIs are technically challenging and we encourage GBIF to keep working on them.

The actual activities related to the topics above are shared among three major (and often overlapping) classes of participants: (1) *data providers*, including natural history museums, herbaria, scientific projects, networks of amateurs, repositories of governmental reports, and others. (2) *Data aggregator* agencies or organizations, national or international, scientific or non-scientific, compile and serve data from different data providers. Finally, (3) data users, are mostly scientists and analysts working in academia, governmental agencies, NGOs, or consulting companies, but increasingly also the general public. These classes overlap, but the broad classification is useful to propose some solutions to the problems we identify.

Different types of errors/inconsistencies affect users in different ways, depending on the specific research question that a user intends to develop. For example, characterizing the flora or fauna of a site and developing a taxonomic revision will require high-quality information from different fields in the database. In this report we concentrate on questions related to the challenge of estimating associations between environmental datasets and the species-occurrences datasets; this is the modeling of ecological niches (ENM) which very often is used to estimate the distributions of species, called Species Distribution Modelling (SDM). Since here we do not make a distinction between ENM and SDM, we use the latter acronym throughout.

1.2 An overview of Primary Data Portals

Until the late 1980s, most data-sharing in biology was dependent on printed, hard-copy formats. For PBD, access to data was through consultation of published monographs, visits to museums or herbaria, or consultations with individual curators using conventional postal services. As a consequence there was great variability in speed and comprehensiveness of responses. In recent decades, PBD data-sharing has shifted to institution-wide, open-access-oriented policies using the Internet (Soberón & Peterson 2004), such that enormous volumes of PBD are now freely available to users. These initiatives come from individual institutions and other data holders that may or may not join worldwide consortia, and vary in scope among regions and taxonomic groups (Table 1).

Table 1. Some example global database initiatives for biodiversity-related information domains.

Worldwide/Global Biodiversity Data Initiatives		
Genbank	+188 million sequences	http://www.ncbi.nlm.nih.gov/genbank/
Global Genome Biodiversity Network	+100 thousand sequences from +23 thousand taxa	http://data.ggbn.org/
Catalogue of Life	+1.6 million species	http://www.catalogueoflife.org/
Bold Systems	+4 million barcode sequences	http://www.boldsystems.org/
Global Biodiversity Information Facility	+1.6 million species, ca. 650 million occurrences	http://www.gbif.org/
Freshwater Biodiversity Data Portal	+91 thousand species, +160 million occurrences	http://data.freshwaterbiodiversity.eu/
Ocean Biogeographic Information System	+148 thousand species, +34 million records	http://iobis.org/
Fossilworks	+330 thousand taxa, +1.2 million occurrences	http://fossilworks.org/

1.3 GBIF.org

GBIF is currently the largest-scale biodiversity data infrastructure in the world and is funded by governments. GBIF.org has provided free and open access to PBD since 2005, when it started with just over 40M records. Today, it provides access to almost 650M of records for >1.6M species; the data served via GBIF are aggregated from >15,000 datasets from >750 data holders (Fig. 1).

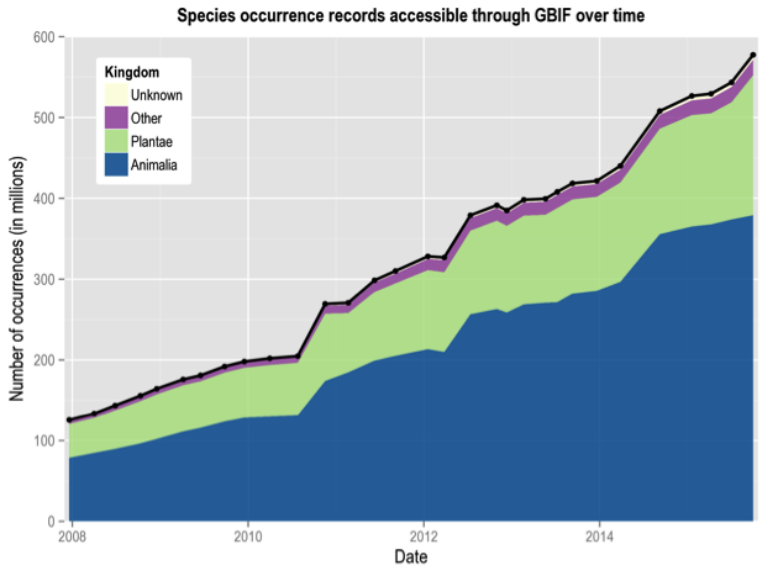

Figure 1. Accumulation of PBD records accessible via GBIF through time (2008-2015).

GBIF-mediated data provides information about biodiversity for almost all countries in the world, albeit with drastically different volumes (Soberón 2014). Amounts of data available via GBIF separated by world region are illustrated in Figure 2.

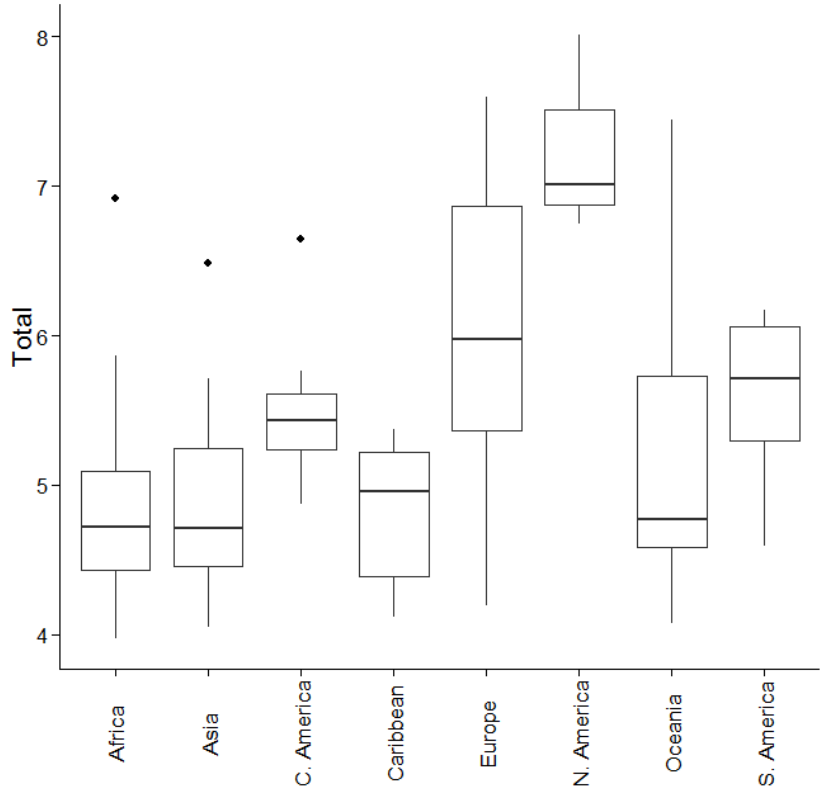

Figure 2. Decimal logarithm of total number of PBD records accessible via GBIF for various regions of the world (Soberón, 2014). The boxes represent the 25% and 75% quartiles, the bars the medians, and the whiskers represent the maximum to minimum interval without outliers (1.5 times the upper or lower quartiles)

Access to such massive amounts of PBD has catalyzed research in large-scale biodiversity science in many ways (Graham *et al.* 2004), with > 1750 peer-reviewed publications (Mendeley GBIF Public Library)[11] in which GBIF-mediated data have been applied to questions in a variety of fields, including macroecology (Beck *et al.* 2012), biodiversity responses to environmental change (Warren *et al.* 2013), public health (Peterson 2015), conservation (Guisan *et al.* 2013), agriculture (Lyal *et al.* 2008), ecosystem services (Allan *et al.* 2013), evolution (Antonelli *et al.* 2010), and invasive-species biology (Adhikari *et al.* 2015). The growth of usage and science dependence on GBIF as an information infrastructure (Figure 3) has been impressive.

[11] http://www.gbif.org/mendeley

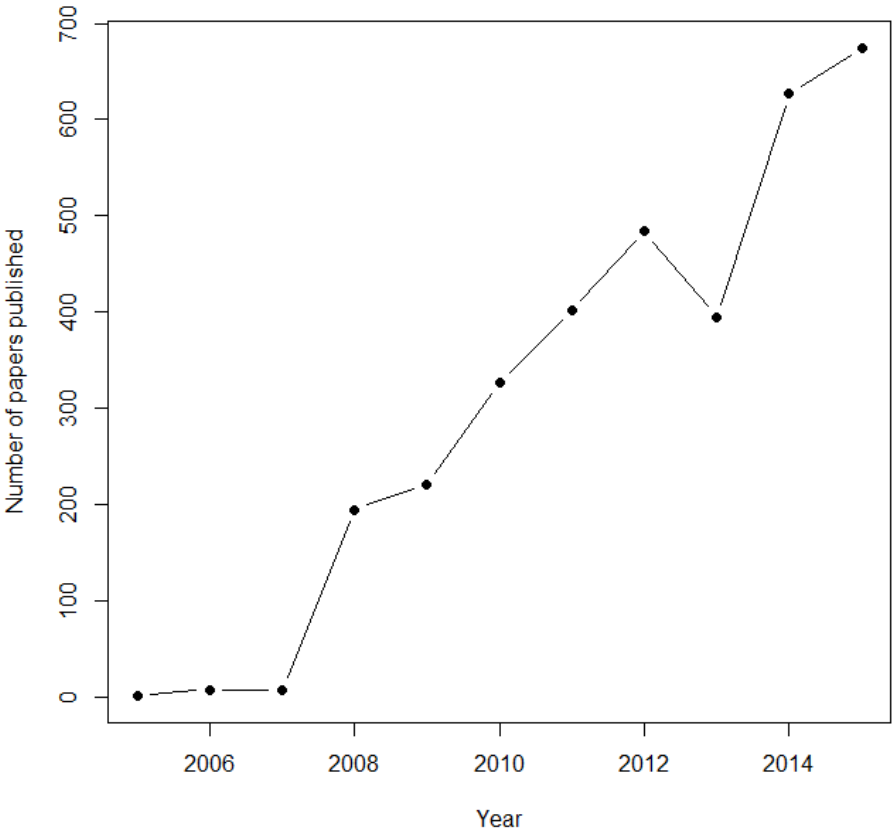

Figure 3. Per year number of publications based on data mediated by GBIF and discussing GBIF, since 2005. The accumulated growth is exponential.

1.4 Species Distribution Modelling

A large and diverse scientific literature has grown to study distributions of species on the basis of associations between geographic occurrences of species and the environmental characteristics of such occurrences, a field generally known as SDM (Franklin 2010; Peterson *et al.* 2011). SDM applications require

- PBD: data on sites where a species has been recorded (and ideally, places where sampling occurred but the species was not recorded), and
- Environmental data that characterize the landscape across which the species is distributed.

A variety of modelling algorithms can be used to integrate occurrence and environmental information to identify other localities (in the same or other time period) with similar environmental conditions (Guisan & Zimmermann 2000). Depending on the approach, outputs of the various methods can be interpreted in various ways, as formal probabilities of presence, given the environmental conditions, or as mere indices of relative environmental suitability with undefined scaling, or as various other probabilities (Peterson *et al.* 2011). One thing that all of these methods share, however, is the need for data on spatio-temporal occurrences of a species, or PBD.

One could argue that the exponential growth of SDM applications in the past 20 years (Guisan *et al*. 2013) results in large part from increases in PBD availability. Since, PBD available online is now approaching a billion records, such frequent and widespread use should not be surprising. However, experience shows that PBD cannot be used “as is” in

SDM applications without data preparation and cleaning. Indeed, some analyses have concluded that GBIF-mediated data in particular are not “fit for use” in analyses like SDM (Yesson *et al.* 2007, Beck *et al.* 2012; Beck *et al.* 2013; Otegui *et al*. 2013a; Otegui *et al*. 2013b). Although after appropriate preparation, a degree of fitness for use can almost always be achieved, there is no doubt that GBIF, as an information resource, is not used as extensively as it could be due to many of those drawbacks.

2. The challenges

Three classes of difficulties affect accessible digital PBD, such as those mediated by GBIF: (i) those affecting the data *per se*, (ii) those affecting access to the data, and (iii) those related to the use of the data. These limitations are detailed and discussed below.

Data issues

Problems with the primary data are generally the result of various inaccuracies, biases, and spatial and environmental data gaps, leading to issues of data incompleteness and uncertainty (Graham *et al*. 2004; Soberón *et al*. 2007; Newbold 2010; Sousa-Baena *et al*. 2013; Meyer *et al*. 2015a; 2015b; Hortal *et al*. 2015). Inaccuracies (errors, imprecision) that are most relevant in SDM are taxonomic (Graham *et al*. 2004) or locational (Hefley *et al*. 2013), although temporal problems have also been discovered (Otegui *et al*. 2013a). Biases originate from many sources: observer choices (e.g., sampling focused along roads, field stations, rivers, etc.; Bojórquez-Tapia *et al*. 1995; Kadmon *et al*. 2004), biased taxonomic focus (e.g., many more records of birds than for other taxa; Hortal *et al*. 2008), goals of surveys other than preparing an inventory of species (Edwards *et al*. 2005), biased of defective methodologies (Anderson 2003) or political barriers to access (Peterson *et al*. 2009).

Incompleteness can spring from many causes. It can be taxonomic (Yesson *et al*. 2007), known as the ‘Linnean shortfall’ (Whittaker *et al*. 2005), or spatial (Meyer *et al*. 2015a; 2015b), known as the “Wallacean shortfall” (Whittaker *et al*. 2005). Although many thousands of biodiversity records are ‘dark’ to use for lack of full taxonomic identification, perhaps the most significant information gap is that of georeferencing, which makes spatial views of the data difficult. SDM depends crucially on such spatial information. An important point that is often not appreciated is that modern and responsible SDM applications demand information on the precision and uncertainty associated with these georeferences.

Access problems

Problems of data accessibility include the point that full information is not always provided to users, and that functionalities that would allow users to query, extract, and manipulate data are lacking or difficult. Data are frequently served only partially either (1) to protect sensitive species against exploitation, or (2) to protect research interests of researchers. While the former may be a reasonable step for certain groups, in light of rampant exploitation, the latter should be kept within bounds, to prevent permanent information gaps. That is, keeping key information fields ‘dark’ for a few years while research reports are finalized, but if the process extends to years and decades, doubts arise over the progress of the research. For instance, Mexico’s CONABIO imposes a maximum delay for full disclosure of PBD records of 5 years on the data for which it provides funding.

Use issues

Further problems with the use of the data are that some PBD records may be used in analyses for which they were not suited (Beck *et al*. 2013; Joppa *et al*. 2013; Beck *et al*. 2014). The most frequent of such problems in SDM manifest when users either use data without any data cleaning or quality control steps, or when users wish to create “large-scale” implementations (e.g., developing SDMs for all of the plants of Asia), which necessarily reduces the care that can be applied to any particular record. Although these lessons are

stated and restated in synthesis after synthesis of SDM methods (Peterson, 2011; Peterson, 2014), many users nonetheless are not careful, and end up using PBD records inappropriately in SDM applications.

Generalities

We review these varied issues and propose solutions, based on literature, answers to an *ad hoc* survey, and our own experience. When looking at these problems, it is useful to consider the community to which each is most likely linked: by whom they were caused, by whom they are identified, and/or by whom they may be managed and fixed (Table 2). We distinguished three classes of actors earlier: data providers, data users, and data aggregators (in this case, GBIF). Some actors may play multiple roles (e.g., a curator of a collection who also conducts SDM analyses). Data issues and accessibility problems fall mostly with providers and aggregators, respectively, but both are usually identified and assessed by users. In contrast, weaknesses in usage are on the side of users, but are viewed, often with horror, by providers, aggregators, and especially users who read the resulting publications. It is thus particularly important to identify and discuss the roles of these different actors and their contributions, before attempting to solve these problems.

Table 2. The three major actors in the PBD world, and their main activities.

	Main activity	Interaction
Data Providers (DP)	Generate, clean, maintain, and update data. Review data based on internal processes and needs	Review data based on feedback by aggregators and users
Data Aggregators (DA)	Develop standards and protocols for data sharing and interoperability Provide agile, comprehensive, updated access to data. Provide tools to explore and visualize data	Use feedback from users to flag records; channel feedback from users to providers
Data Users (DU)	Use data responsibly, via three components: Know the data Know the algorithms Know the species Provide crucial use cases to demonstrate the importance of these initiatives	Communicate observations to aggregators or providers; publish use cases

The most pressing issues are those related to the reliability of the PBD records. In general, because these problems derive from the original databases made available via GBIF, in some sense, fixing them should be the responsibility of the providers. The reasons why PBD records are biased, incomplete, or include inaccuracies/errors, relate in largest part to the opportunistic data gathering process associated with these data. Generally speaking, the data records were accumulated with different aims and methodologies. Putting these data together in a PBD index like GBIF results in a melting pot of heterogeneous data that requires careful work to be fit for scientific uses (Soberón *et al*. 2002b; Graham *et al*. 2004). This process is certainly not optimized to meet objectives of taxonomic, environmental, and geographic completeness, not to mention other perspectives of biodiversity (Dawson *et al*. 2013, Hortal *et al*. 2015). Indeed, the process is so heterogeneous that one cannot trust that

each observer will have followed expected standards for collecting data, in terms of taxonomic or geographic accuracy, or in terms of the data types to be recorded (simple occurrence, abundance, population parameters, or other characteristics). Basic data-record-level errors, inconsistencies, and inaccuracies may be identified and corrected by relatively simple validity checks, but more systemic biases and incompleteness need more advanced methods and tools (see section 3.4).

Biases and incompleteness can be remedied, at least to a limited extent, by building their consideration directly into the use of the data (e.g. use of bias surfaces in species distribution models (Phillips *et al*. 2009; Hijmans 2012). Such activities are the province of users: given the magnitude of the challenge and extent of the gaps, they will require extensive collaboration, training, and funding (Costello *et al*. 2010; Wheeler *et al*. 2012). For instance, a detailed study of drivers of data completeness concluded that "completeness is mainly limited by distance to researchers, locally available research funding and participation in data-sharing networks, rather than transportation infrastructure, or size and funding of Western data contributors" (Sousa-Baena *et al*. 2013; Meyer *et al*. 2015a; Meyer *et al*. 2015b). More such studies are needed.

Difficulties of access are quite complex, and a combination of responsibilities of providers, users and aggregators. For instance, on the side of the aggregators, not all data necessary for certain analyses are available in the main, easily available views of GBIF.org; even if available, it may not be easy to access by users not capable of programming and scripting (e.g., accessing the data through an application programming interface, API). Moreover, currently, GBIF's database contains a very large number of fields, corresponding in large part to those in DarwinCore, but many of them are empty or nearly so. Sometimes this appears to be an issue with GBIF (i.e., fields that are populated in the provider version, and empty in GBIF's, as an example with the field "coordinateUncertaintyInMeters" and the provider VertNet shows e.g. for *Artibeus watsoni*), but sometimes this is an issue with the providers sending sparse datasets. Now, although conceivably some users may need the full Darwin Core dataset in a query, to access the total of the GBIF-mediated data, one needs to query the API directly, using programming tools, or download the database entirely, which is expensive in time and memory, as well as inefficient. This in turn becomes a challenge for the user. In the case of SDM, having access to uncertainty fields related to names, dates and georeferences is important. Ensuring that providers fully populate these fields may be impossible. That GBIF always imports them when available should be feasible, and that users are capable of accessing all the available datasets, probably via the API, or via an enhanced interface by GBIF, remains a thorny problem.

The final set of problems is the way in which data are misused. Clearly, this problem is mostly the responsibility of users. Examples of some misuses identified in our survey included use of coarse-resolution data to analyze phenomena at finer resolutions; lack of awareness of data inconsistencies; lack of experience with the taxonomy of the group in question leading to misidentification errors; or naïve use of complex software without appreciation of the caveats and assumptions (Joppa *et al*. 2013, Jamevich *et al*. 2015).

3. The solutions

In this section, we review briefly steps that could be taken towards enabling fuller and more effective use of existing biodiversity data. Key steps include adding the following elements:

- Encouraging providers to add geographic information as completely as possible (complying with the existing protocols, like the Darwin Core) to each biodiversity data record, accompanied by appropriate and complete metadata documentation of sources, methods, and associated uncertainty. GBIF then should ensure that the raw and verbatim versions of the databases, at least, contain such data.

- Make the users more aware of the existing 'flags' that individually identify records that include inconsistencies that may hallmark errors or problems in various dimensions[12], and develop new ones as necessary.
- Keep working on fully permanent, stable, individual-record-level identifiers that permit full cross-linking and enriching of PBD records with the information provided by other users.

To the extent that such steps towards improving PBD fitness-for-use can be implemented fully, the data will take on greatly amplified utility and importance as a scientific information resource.

3.1 Georeferencing

Ideally, PBD records would include both a description of the locality and geographic coordinates at fine spatial resolutions on the order of 10-10^3 m. In practice, a very large proportion of data in PBD databases include only verbal descriptions of a geographic reference, such as "USA, Kansas, Douglas County, Lawrence, 5 km NNW." While informative, such textual descriptions can be complex and difficult to use in large quantities, such that their translation into standardized, GIS-readable geographic coordinates is an enormous priority. This task, accomplished rigorously, can be performed ideally by providers or by users with specialized knowledge. This step is important—not just that it be taken, but that it be done *right*—as recent analyses indicate that the quality of georeferencing has major influences on the outcomes of SDMs (Engler *et al*. 2004; Graham *et al*. 2008; Lash *et al*. 2012).

In the fairly early history of biodiversity informatics, the Mammal Networked Information System initiative made important advances in georeferencing (Stein and Wieczorek 2004), centered around (1) establishment of methodologies that are well-documented and well-founded, (2) provision of detailed metadata documentation, and (3) close linkage between the georeferencing protocols and data architecture of the DarwinCore. The result, after several years of experimentation and experience gained in the course of several major projects, was a georeferencing protocol that is both practical and feasible to implement, and optimized and customized for biodiversity applications (Wieczorek *et al*. 2004; Chapman 2005; Chapman *et al*. 2006). This general protocol has also now seen quite a bit of investment of development effort towards automating and digitally enabling the process (Guralnick *et al*. 2006; Rios & Bart 2008).

As was demonstrated by the VertNet initiative in the USA (Constable *et al.* 2010), major data aggregators can play a crucial role in these data-improvement steps. A key point is that whenever available, GBIF should include the crucial summary of georeferencing uncertainty, for instance, the CoordinateUncertaintyInMeters of VertNet (Wieczorek *et al*. 2004). Although full automation of georeferencing is not yet feasible, parts of the process can be automated (Guralnick *et al*. 2006), which could be implemented centrally by the data aggregator. More importantly, however, data aggregators can use their networks to coordinate and enable georeferencing initiatives: the aggregator can create data packets consisting of the un-georeferenced records for certain taxa and/or regions, and facilitate their distribution to expert georeferencers (e.g., the scientific community of the region, the specialists on the taxon), as well as facilitating communication of the now-georeferenced records back to the original data providers, for addition to the original data record. This role can be key, as was demonstrated in the success of the VertNet initiative in this regard; GBIF could play this role at a global level, given its massive network of participants. Proper feedback by the original providers should be enabled as well.

[12] http://gbif.github.io/gbif-api/apidocs/org/gbif/api/vocabulary/OccurrenceIssue.html, and http://www.gbif.org/infrastructure/processing

3.2 Error flagging and data cleaning

PBD records can hold georeferences and still have errors and inconsistencies, to the point that they are not usable in analyses. Indeed, lack of care with respect to data consistency can have major and significant influences on the outcomes of SDMs (Peterson *et al*. 2014). Although at times it may be possible to identify actual errors, more commonly, the focus is on identifying data records that present inconsistencies and conflicts, either internally (e.g. geographic coordinates fall in a state other than that specified in the textual state field) or externally (e.g. taxon does not match an authority list or a list of taxa known from a particular country). Additional efforts may flag records with strong environmental departure from other records of the species, or may use collectors' itineraries to detect unlikely combinations of collection locality and date. In instances in which inconsistencies exist, the user should be extremely careful when using them. Search for consistency holds great promise in ensuring that data records are sufficiently free of errors for SDM performance not to be affected. If provided with reliable maps of subnational units GBIF could implement inconsistency analysis below the level of country, but this is contingent on the resolution and precision of subnational maps, and on criteria for inconsistency. Perhaps this type of inconsistency analysis should be left to the user, given the sheer magnitude of possibilities and details.

Several protocols and workflows for data cleaning have been proposed (Chapman 2005); efficient and highly visual online implementations have already been developed and applied to Brazilian biodiversity data (CRIA 2012). CONABIO in Mexico has developed handbooks used for data-cleaning, encompassing both georeferenced and taxonomic data cleaning (CONABIO 2012).

An efficient design for mass implementation of this step would involve centralized application of the data-cleaning tools, as several such tools will be more efficient when data are pooled as broadly as possible (e.g. Peterson *et al*. 2004). However, crowd-sourcing of this data-quality assurance step represents another option: inconsistencies and incongruences identified by users should be flagged and transmitted to the data providers for evaluation and potential incorporation into primary data records. Alternatively, interested communities or consortia can do this assessment step, to ensure data quality for their desired applications (Anderson 2012). The final stage of this process is that of either incorporating the corrections to the erroneous fields or flagging records as dubious or possibly problematic, in the original data sets curated and maintained by the data providers; this step has proven cumbersome at times, both as regards data quality and in terms of adding georeferences, as data providers have not always been particularly efficient in 'ingesting' data improvements back into original data sets.

3.3 Individual identifiers and cross-linking data realms

For two major reasons, one key innovation in biodiversity informatics is that of adding a unique, permanent, and stable identifier to individual records (GBIF 2011). The first reason is that such identifiers will greatly promote the repatriation, ingestion, and incorporation of changes and additions offered either by users or by aggregators (see Table 2), but that should be added to data sets by providers. The second reason is that these identifiers will allow linkages between data realms. This has been highlighted as key for this field for some time (Peterson *et al*. 2010), and is the focus of several of the suggestions and changes that are proposed later in this report as useful in next generations of SDMs.

The potential for producing and using such identifiers has already been incorporated in some of the data infrastructures (e.g. fields gbifID, occurrenceID at GBIF.org). However, full implementation in the sense of broad agreement regarding a preferred mechanism (e.g., the need of cross-linking among unique identifier systems[13]) has not yet occurred. We note that the current DOI implementation in the GBIF data portal creates unique identifiers for *query-*

[13] http://devpost.com/software/bioguid-org

level entities, not for individual records, and as such is not sufficient. This step of implementing full, individual-level identifiers is crucial to the error repatriation, data record improvement, and cross-linking among data realms that is called for in this review.

3.4 Data visualization and GAP analysis

As mentioned above, PBD records on biodiversity are full of biases and gaps. Existing data resources can be the basis for detailed analyses of gaps in coverage to guide strategic mobilization of resources (Soberón *et al*. 2004) or at least to identify them (Sousa-Baena *et al*. 2013, Pelayo-Villamil *et al*. 2015). Such analyses will identify well-sampled groups and localities as well as priorities to focus efforts to fill the last remaining gaps, so that coverage globally is as complete as is feasible (Sousa-Baena *et al*. 2013; Kouao *et al*. 2015).

Once gaps are identified, they can be either filled, as a more permanent remedy, or considered explicitly in analyses, for a given use. In taxonomic dimensions, poorly-known taxa can be prioritized for *de novo* data generation or for mobilizing data presently in analog formats. In temporal dimensions, detecting and filling gaps (and their converse, well-sampled sites) allows before-and-after comparisons of biotic community composition at different points in time (Peterson *et al*. 2015). Finally, in geographic dimensions, individual regions, countries, or states, can be the focus of concerted georeferencing efforts (or, if necessary, data capture efforts) that fill gaps (Navarro-Sigüenza *et al*. 2003). Different user communities will have different interests and priorities, such that consortia can assemble to fill gaps and improve data fitness for use towards those needs (Figure 4).

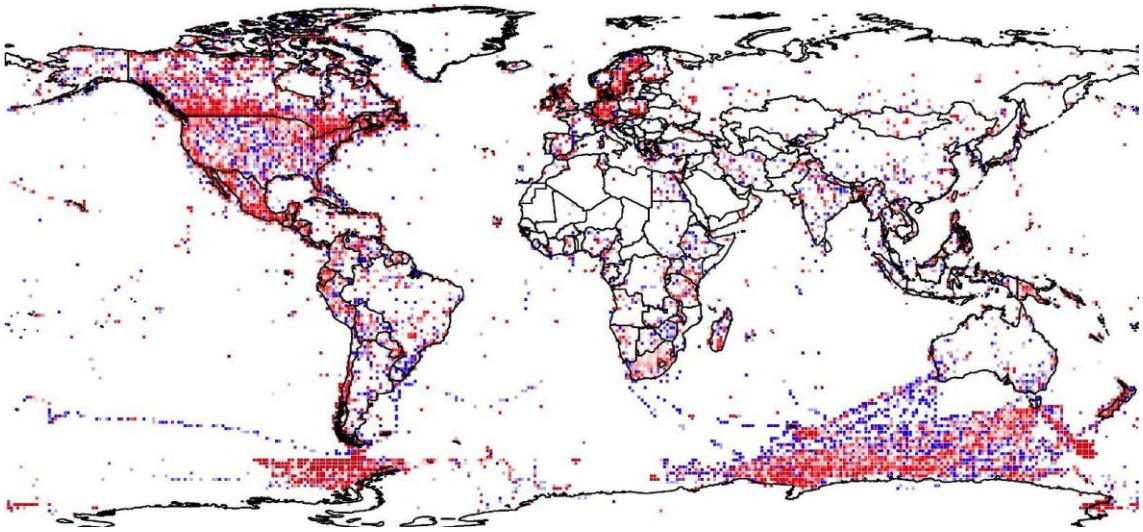

Figure 4. World bird geographic knowledge, expressed as degree of completeness (an index that goes from red when the estimated number of species is close to the observed, towards blue where the ratio of observed to expected is smaller than 0.5, to white where the database is empty or the ratio C is less than 0.1). Data as November of 2014, based on 167M records from GBIF. A. T. Peterson et al.*, unpublished data.*

Consider Figure 4, which summarizes completeness of PBD records for birds of the world, based on GBIF-mediated data from late 2014. Some major gaps that can be noted include (1) Russia, other portions of the former Soviet Union, and China; (2) the Sahara Desert and Middle East; (3) the Congo Basin; and (4) the Pacific islands. Interestingly, these gaps have different solutions: for instance, one gap can be resolved via political agreements with Russian and Chinese institutions that already have massive digital data resources that are not as yet broadly integrated into the global PBD resource. Gaps #2 and #4, on the other hand, depend much more on single institutions (Natural History Museum in London for #2, American Museum of Natural History for #4) digitizing key data resources, as the richest collections from those regions are housed in those single institutions. Finally, in the case of the Congo Basin, although significant collections are indeed housed in the Royal Museum

for Central Africa and smaller holdings exist in other institutions (e.g., American Museum of Natural History), a fullest sampling and dense coverage of this region may require additional field work. These examples of gaps are at a global scale, and are certainly over-simplified, but they illustrate the diversity of solutions to this problem.

Indeed, an intriguing possibility for a next generation of GBIF functionalities is that GBIF would provide summary data products that can be used as 'bias surfaces' for a variety of purposes that users might have. These maps would not just be counts of available points, but also of completeness indices and related indices. Such mapping of knowledge and ignorance should constitute one of the most important future challenges (as e.g. inSoberón *et al*. 2007, Meyer *et al.* 2015a; 2015b; Ruete *et al.* 2015).

3.5 Training of users

Many "problems" highlighted in our survey of PBD users were, in a strict sense, a manifestation of users not being aware of the caveats and complications underlying rigorous and proper SDM analysis. Although GBIF may not be the right organization to deal with the training of users, many training initiatives exist that provide access to such improved knowledge e.g., the Biodiversity Informatics Training Curriculum[14] (GBIF 2015). Moreover, other organizations, like the JRS Biodiversity Foundation[15] are supporting such initiatives, and for GBIF this is a great opportunity to partner and perhaps leverage resources specifically designed to train a larger community of sophisticated users, mostly in developing countries.

5. Future perspectives

This review, so far, has focused on well-known gaps and problems that have been pointed out several times in the past, and that should be remedied. In this section, we attempt to look into the future of the SDM field, and to anticipate needs and innovations that may arise. This set of considerations presents new challenges and new opportunities for the interaction between PBD providers and SDM.

5.1 Modelling and sampling

Models and spatial analyses can help design sampling strategies to complement data gaps for species (Guisan *et al*. 2006; Le Lay *et al*. 2010; Mokany *et al*. 2011, Meyer *et al*. 2015b), groups of species, or entire floras or faunas. These models can be used to design prospective sampling in a framework embedded within biodiversity databases that allows using models to improve sampling, update or generate new data, which are then used to update the models. This sequence can be repeated in an adaptive framework until data and models are improved to reach given standards (Guisan *et al*. 2006). Models to design new or complementary sampling can follow standard sampling theories (Hirzel & Guisan 2002; Edwards *et al*. 2005; Albert *et al*. 2010), such as using model predictions to design a random-stratified sampling (Hirzel & Guisan 2002), where the stratifying variable are the predictions themselves (reclassified in two or more classes, from suitable to non-suitable). Models can also be used to support the improvement of data completeness (Mokany *et al*. 2011), e.g. using macro-ecological model predictions (e.g. species richness) to set the expected number of species to be found in a given cell, and compare these predictions to the actual number recorded for these cells in GBIF or other databases to identify the cells likely to be deficient in data, and ideally identify which species are missing (Pineda & Lobo, 2009).

[14] http://biodiversity-informatics-training.org

[15] https://www.facebook.com/JRSBDF

5.2 New data realms

The basic suite of fields that has been part of the idea of PBD is largely limited to the documentation of a taxon, where it occurred, and when the record was noted. However, we envision improvements and innovations in SDM, and propose a series of links to additional data realms that will almost certainly enrich and broaden the impact of this suite of tools in science. That is, each of the following points constitutes a set of information not presently manifested within the DarwinCore data architecture, but could be added to the data resource, either by means of additional data sets or data fields hosted by GBIF, or via partnership and linkage with entities already working in these realms.

The rationale for addition of these new data realms is rooted in the fact that primary biodiversity data will continue to be biased, even if all existing information in the scientific literature and natural history collections is compiled. This permanent bias exists because the ultimate cause of these biases is the heterogeneous distribution of taxonomic resources and the systematics workforce, a shortcoming that is difficult to solve. In one sense, building PBD resources and SDMs are strategies whose successes or failures are mutually dependent. Precise knowledge of the biodiversity distribution on Earth will require use of models to predict and interpolate distributions of species from the partial information in PBD records, and improvement of SDM predictions should involve the use of a representative subset of data capable of accurately representing the entire unknown "universe."

Documenting sampling biases and absences of species

Identification of gaps in PBD is a basic requirement for designing new explorations, but is also indispensable for providing a necessary correction in distributional hypotheses derived from their use. Unfortunately, most data available through PBD portals lack associated measures of sampling effort, and very few provide explicit references to absence of species at sites. In practice, this problem has been solved for SDMs by resorting to so-called "pseudoabsences," whereby absence information is created artificially to enable use of regression or other discriminant methods, or samples from the background are used thus contrasting the environments in the presences with those in the region in question (Phillips *et al*. 2009). Both methods have their own problems (Pearce & Boyce 2006; Ward *et al*. 2009; Royle *et al*. 2012) and impede the rigorous estimation of occurrence probabilities (Royle *et al*. 2012), and a crucial step in using such pseudoabsence approaches is that of delimiting the sampling background based on rigorous and explicit biogeographic bases (Barve *et al*. 2011). Were GBIF to implement tools capable of detecting and characterizing gaps, well-surveyed sites, and uncertain sites, it would be a great asset for progress towards the development of more efficient distribution models.

Nevertheless, on the user side, several procedures can be used on presence-only data to identify well-surveyed regions or localities that would reinforce the assumption that a species is absent (Lobo & Martín-Piera *et al*. 2002; Anderson *et al*. 2003; Hortal *et al*. 2008). In all such analyses, assumed characteristics of well-surveyed localities (its species composition, the shape of their collector curves, etc.) are compared to those at the remaining localities, and a threshold or criteria used to discriminate between them (Soberón *et al*. 2007). With this simple procedure, modelers can weight more heavily those sites that have been sampled well in development of models (Lobo *et al*. 2010), but also the location of uncertain areas not well-surveyed, which would be accorded less weight. For this purpose, it would be recommended not to reject the frequently redundant information about species occurrences coming from intensively and long-time studied localities; although this information may not add relevant species distribution information, it will be very important to demonstrate the intensive character of the survey carried out in this area. Another key functionality will be to allow the user to specify the relevant sets of taxa that have been sampled similarly to the species of interest (Anderson *et al*. 2003).

Adding data on abundance and numbers of individuals

Abundance data would enable a whole new set of statistical techniques to be applied to PBD records (Franklin 2010), and, if collected frequently, enable questions of a dynamic nature to be addressed (Peterson *et al.* 2008), something that is either impossible or difficult using traditional correlative SDMs, which are much more static in nature (Svenning & Skov 2004, Guisan & Thuiller 2005). At present, digitally available PBD of species abundance is largely restricted to birds in the United States, Canada and western Europe (although this situation is changing) that have large communities of nature observers, such as eBird[16]; EuroBird[17] and other similar networks. Countries like Norway, the United Kingdom, Sweden and Switzerland have very dense and well-populated networks of observations of many taxonomic groups, often documenting abundance, but these cases are exceptional. These datasets are huge in size, and are expanding in geographic coverage, so their potential to answer dynamic questions cannot be underestimated; GBIF should ensure that such databases continue to be served.

There are also data portals that provide access to data on abundances of species, like the NCEAS/Imperial College Global Population Dynamics database[18]. Although these data are sparse geographically, as are most ecological data (Martin *et al*. 2012), data about species abundances published by ecologists represent a valuable resource that can be used in SDMs, as illustrated in several recent studies (Guisan & Harrell 2000, Iverson *et al*. 2008; Peterson *et al*. 2008, Randin *et al*. 2009). For such ancillary data, data aggregators need to ensure that linking fields exist, that they are populated where abundance data exist, and that their existence is well documented.

Genetic and genomic data

A clear awareness has emerged that models developed based on recognized species-level entities may be a gross oversimplification. For instance, significant niche differentiation has been documented within a single recognized species of triatomine bug that vectors Chagas disease (*Triatoma dimidiata*), such that a single, species-level model may well be overly broad and inclusive, and not representative of the niche of any one population (Gómez-Palacio *et al*. 2015). In such situations, linking primary biodiversity data records to data on gene or genome sequences (e.g., GenBank), such that niche models correspond to evolutionary lineages, may be highly informative, allowing identification of monophyletic evolutionary lineages for analysis. A further approach may involve modeling and tracking specific genetic elements, perhaps even from ecological genomic data sets that do not necessarily even link to particular organisms, but rather to presence of genetic elements in a place at a point in time (Fournier-Level *et al*. 2011; Fitzpatrick & Keller 2015).

Movements and dispersal

The recent hybrid models integrating dynamic processes in SDMs are very data hungry, and require information related to physiology, interactions, and movements (Cabral & Schurr 2010; Smolik *et al*. 2010; Barve *et al*. 2011; Dullinger *et al*. 2012, Schurr *et al*. 2012). Movement-related information is of several types (migratory, home-range, tracking, dispersal, etc.; Matthysen 2012). Several more or less publicly available data resources currently provide access to such data[19]. However, to our knowledge no databases exist that document the key numbers required to define dispersal kernels of species (Nathan & Muller-Landau 2000), which would be what is required to parameterize the movement part of process-oriented SDMs. The data exist in large quantities in the literature (e.g. Vittoz &

[16] http://ebird.org/content/ebird

[17] http://www.eurobirdportal.org/ebp/en

[18] http://www3.imperial.ac.uk/cpb/databases/gpdd

[19] https://migbirdapps.fws.gov/mbdc/databases/db_selection.html

Engler 2007), but have never been organized as digitally available knowledge. It is an open question whether GBIF should embark on creating and populating such a database, with all the work and effort that would entail; certainly, though, providing the unique identifiers that will permit individual PBD records to be linked to documentation of particular movements (e.g., a record of an individual that was ringed at one spot and recovered at another as more than two records of occurrence of that particular species) is a crucial enabling step.

Biotic interactions

The presence and absence of a species in any given locality are not independent of the distributions of other species. The mechanisms controlling co-existence of species are still the focus of intensive research. Although increasingly clear that predicting distributions of species in space and time often requires understanding the effects of biotic interactions (Davis *et al*. 1998; Gilman *et al*. 2010; González-Salazar *et al*. 2013; Wisz *et al*. 2013; Araújo & Rozenfeld 2014; Mod *et al*. 2015), documentation of the direct and indirect interactions among species is a daunting task. For example, identifying direct species interactions within a system with only seven species would require documentation of 42 potential links and up to 13,650 links if indirect interactions were considered (Dodds & Nelson 2006). Given that most systems have more than seven species, documentation of all biotic interactions at any site (let alone across the world) is beyond reach (Morales-Castilla *et al*. 2015).

An alternative to documenting interactions among species extensively is to use ecological theory and models to predict the backbone of biotic interactions networks (Kissling *et al*. 2012). A conceptual framework was recently proposed that enables inferences of backbones of interaction networks by sequentially pruning potential networks from forbidden and unlikely species links. Such pruning of networks can be made using prior knowledge on the functional, phylogenetic and geographical relationships among species (Morales-Castilla *et al*. 2015).

Inferences of interactions thus require that relevant datasets exist on species traits (e.g., PanTHERIA, Try, Jones *et al*. 2009; Kattge *et al*. 2011), species distributions (e.g., GBIF, Map of Life), and phylogenetic relationships among species (e.g., GenBank, Tree of Life), as well as that effective linkages are established among them, which will require the individual, record-level identifiers discussed above. Effective structuring that permits specification of one-to-one interactions at the individual level (e.g., insect specimen number XXXX was collected on plant specimen number YYYY) will be an effective path to such detailed documentation of interactions, at least in the long run (see the example analyzed in Estrada-Peña *et al.* 2015). Even if comprehensive documentation of the biotic interactions that are established among organisms on Earth is beyond current capabilities, existing data on biotic interactions should be linked to trait, distribution, and phylogenetic databases to enable training of the models and testing of their inferences. Initiatives such as Globis[20] (Poelen *et al*. 2014) or the Web of Life[21], that record primary observations of interactions between individuals of different species in different parts of the world, should be linked with and integrated within major biodiversity data portals.

Physiological information

Species distributional limits are constrained primarily by aspects of the environment, such as climate, and species' physiological responses to those conditions (Thomas 2010). SDMs have been utilized for making inferences about the environmental factors controlling species distributional limits, but recent research suggests that modelled estimates of species-environmental tolerances are highly conservative, while physiological limits of species being much broader (Araújo *et al*. 2013). Integrating SDMs with specific physiological measurements and models can offer deep insights into factors controlling distributions of

[20] http://www.globalbioticinteractions.org

[21] http://www.web-of-life.es

species (Barve *et al*. 2014, GEB). Improving understanding of mechanisms governing species distributions thus requires that data on species physiological environmental tolerances be linked on an individual basis to PBD records. At the moment, such physiological data are scattered in the literature with a relatively small number of reviews compiling small parts of a larger body of data available (Addo-Bediako *et al*. 2000; Chown *et al*. 2002; Clusella-Trullas *et al*. 2011; Sunday *et al*. 2011; Hoffmann *et al*. 2013). Therefore, individual identifiers permitting the linking of physiological data to PBD databases such as GBIF once again become crucial.

Literature Cited

Addo-Bediako A., Chown S.L. & Gaston K.J. (2000) Thermal tolerance, climatic variability and latitude. *Proceedings of the Royal Society B.*, 267: 739–45. doi: 10.1098/rspb.2000.1065

Adhikari D., Tiwary R. & Barik S.K. (2015) Modelling Hotspots for Invasive Alien Plants in India. *PloS ONE*, 10: e0134665. doi: 10.1371/journal.pone.0134665

Albert C.H., Yoccoz N.G., Edwards T.C., Graham C.H., Zimmermann N.E. & Thuiller W. (2010) Sampling in ecology and evolution - bridging the gap between theory and practice. *Ecography*, 33: 1028–1037. doi: 10.1111/j.1600-0587.2010.06421.x

Allan J.D., McIntyre P.B., Smith S.D.P., Halpern B.S., Boyer G.L., Buchsbaum A., Burton G.A., Campbell L.M., Chadderton W.L., Ciborowski J.J.H., Doran P.J., Eder T., Infante D.M., Johnson L.B., Joseph C.A., Marino A.L., Prusevich A., Read J.G., Rose J.B., Rutherford E.S., Sowa S.P. & Steinman A.D. (2013) Joint analysis of stressors and ecosystem services to enhance restoration effectiveness. *Proceedings of the National Academy of Sciences of the United States of America*, 110: 372–7. doi: 10.1073/pnas.1213841110

Anderson R. P. (2003). Real vs. artefactual absences in species distributions: Tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. *Journal of Biogeography,* 30: 591–605. doi: 10.1046/j.1365-2699.2003.00867.x

Anderson R. P. (2012). Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. Annals of the New York Academy of Sciences, 1260: 66–80. doi: 10.1111/j.1749-6632.2011.06440.x

Anderson R.P. (2013) A framework for using niche models to estimate impacts of climate change on species distributions. *Annals of the New York Academy of Sciences*, 1297: 8–28. doi: 10.1111/nyas.12264

Antonelli A., Humphreys A.M., Lee W.G. & Linder H.P. (2010) Absence of mammals and the evolution of New Zealand grasses. *Proceedings of the Royal Society of London B,* 278: 695-701. doi: 10.1098/rspb.2010.1145

Araújo M.B., Ferri-Yáñez F., Bozinovic F., Marquet P.A., Valladares F. & Chown S.L. (2013) Heat freezes niche evolution. *Ecology Letters*, 16: 1206–1219. doi: 10.1111/ele.12155

Araújo M. & Rozenfeld A. (2014) The geographic scaling of biotic interactions. *Ecography*, 37: 406–415. doi: 10.1111/j.1600-0587.2013.00643.x

Barbosa F. & Chneck F. (2015) Characteristics of top-cited papers in species distribution predictive models. *Ecological Modelling*, 313: 77–83. doi:10.1016/j.ecolmodel.2015.06.014

Barve N., Barve V., Jiménez-Valverde A., Lira-Noriega A., Maher S.P., Peterson A.T., Soberón J. & Villalobos F. (2011) The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, 222: 1810–1819. doi:10.1016/j.ecolmodel.2011.02.011

Beck J., Ballesteros-Mejia L., Buchmann C.M., Dengler J., Fritz S.A., Gruber B., Hof C., Jansen F., Knapp S., Kreft H., Schneider A.-K., Winter M. & Dormann C.F. (2012) What's on the horizon for macroecology? *Ecography*, 35: 673–683. doi: 10.1111/j.1600-0587.2012.07364.x

Beck J., Ballesteros-Mejia L., Nagel P. & Kitching I.J. (2013) Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? *Diversity and Distributions*, 19: 1043–1050. doi: 10.1111/ddi.12083

Beck J., Böller M., Erhardt A. & Schwanghart W. (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, 19: 10–15. doi:10.1016/j.ecoinf.2013.11.002

Bojórquez-Tapia L.A., Azuara I., Ezcurra E. & Flores-Villela O. (1995) Identifying conservation priorities in Mexico through geographic information systems and modeling. *Ecological Applications*, 5: 215–231. Available online at http://www.jstor.org/stable/1942065

Cabral J.S. & Schurr F.M. (2010) Estimating demographic models for the range dynamics of plant species. *Global Ecology and Biogeography*, 19: 85–97. doi: 10.1111/j.1466-8238.2009.00492.x

Chapman A. D. (2005) Principles and Methods of Data Cleaning - Primary Species and Species-Occurrence Data. version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. Available online at http://www.gbif.org/orc/?doc_id=1262

Chapman A.D. & Wieczorek J. (eds). (2006) *Guide to best practices for georeferencing*. Copenhagen: Global Biodiversity Information Facility. Available online at http://www.gbif.org/orc/?doc_id=1288

Chown S.L., Addo-Bediako A. & Gaston K.J. (2002) Physiological variation in insects: large-scale patterns and their implications. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 131: 587–602. doi:10.1016/S1096-4959(02)00017-9

Clusella-Trullas S., Blackburn T.M., Chown S.L. (2011) Climatic Predictors of Temperature Performance Curve Parameters in Ectotherms Imply Complex Responses to Climate Change. *The American Naturalist*, 177: 738–751. doi: 10.1086/660021

CONABIO (2012) CONABIO: Two Decades of History, 1992–2012. In: (ed. .), pp. 1–36. Comision Nacional para el Conocimiento y Uso de la Biodiversidad, Mexico D. F., Mexico.

Constable H., Guralnick R., Wieczorek J., Spencer C., Peterson A.T. & The VertNet Steering Committee, 2010. VertNet: a new model for biodiversity data sharing. PLoS Biol 8: e1000309. doi:10.1371/journal.pbio.1000309

Costello M.J., Coll M., Danovaro R., Halpin P., Ojaveer H. & Miloslavich P. (2010) A Census of Marine Biodiversity Knowledge, Resources, and Future Challenges. *PLoS ONE*, 5: e12110. doi:10.1371/journal.pone.0012110

CRIA (2012) Data Cleaning. URL http://splink.cria.org.br/dc/http://splink.cria.org.br/dc/

Davis A.J., Jenkinson L.S., Lawton J.H., Shorrocks B. & Wood S. (1998) Making mistakes when predicting shifts in species range in response to global warming. *Nature*, 391: 783–786. doi:10.1038/35842

Dawson M.N., Algar A.C., Antonelli A., Dávalos L.M., Davis E., Early R., Guisan A., Jansson R., Lessard J.-P., Marske K.A., McGuire J. L., Stigall A. L., Swenson N. G., Zimmermann N. E. & Gavin D. G. (2013). An horizon scan of biogeography. Frontiers of Biogeography 5. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3972886/

Dullinger S., Gattringer A., Thuiller W., Moser D., Zimmermann N.E., Guisan A., Willner W., Plutzar C., Leitner M., Mang T., Caccianiga M., Dirnbock T., Ertl S., Fischer A., Lenoir J., Svenning J.C., Psomas A., Schmatz D.R., Silc U., Vittoz P. & Hulber K. (2012) Extinction

debt of high-mountain plants under twenty-first-century climate change. Nature Climate Change, 2: 619–622. doi: 10.1038/nclimate1514

Edwards T.C., Cutler D.R., Zimmermann N.E., Geiser L. & Alegría J. (2005) Model-based stratifications for enhancing the detection of rare ecological events. *Ecology*, 86: 1081–1090. doi: 10.1890/04-0608

Elith J. & Leathwick J.R. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology Evolution and Systematics*, 40: 677–697. doi: 10.1146/annurev.ecolsys.110308.120159

Engler R., Guisan A. & Rechsteiner L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology 41: 263–274. doi: 10.1111/j.0021-8901.2004.00881.x

Estrada-Peña A., de La Fuente J., Ostfeld R.S. & Cabezas-Cruz A., 2015. Interactions between tick and transmitted pathogens evolved to minimise competition through nested and coherent networks. Scientific reports 5. doi: 10.1038/srep10361

Fitzpatrick M.C. & Keller S.R. (2015) Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecology letters*, 18: 1–16. doi: 10.1111/ele.12376

Fournier-Level A., Korte A., Cooper M.D., Nordborg M., Schmitt J. & Wilczek A.M. (2011) A Map of Local Adaptation in Arabidopsis thaliana. *Science*, 334: 86–89. doi: 10.1126/science.1209271

Franklin J. (2010) *Mapping species distributions: spatial inference and prediction*. Cambridge University Press.

Richards K., White R., Nicolson N. & Pyle R. GBIF (2011) A Beginner's Guide to Persistent Identifiers, version 1.0. Available at: http://links.gbif.org/persistent_identifiers_guide_en_v1.pdf

GBIF (2015), GBIF Annual Report 2014, Copenhagen: Global Biodiversity Information Facility, 34 pp. Available online at http://www.gbif.org/resource/annual_report_2014.

Gilman S.E., Urban M.C., Tewksbury J., Gilchrist G.W. & Holt R.D. (2010) A framework for community interactions under climate change. *Trends in Ecology & Evolution*, 25: 325–331. doi:10.1016/j.tree.2010.03.002

Gómez-Palacio A., Arboleda S., Dumonteil E. & Townsend Peterson A. (2015) Ecological niche and geographic distribution of the Chagas disease vector, Triatoma dimidiata (Reduviidae: Triatominae): Evidence for niche differentiation among cryptic species. *Infection, Genetics and Evolution*, 36: 15–22. doi:10.1016/j.meegid.2015.08.035

Graham C. H., Elith J., Hijmans R. J., Guisan A., Peterson A. T. & Loiselle B. A.. 2008. The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology, 45: 239–247. doi: 10.1111/j.1365-2664.2007.01408.x

Graham C.H., Ferrier S., Huettman F., Moritz C. & Peterson A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in ecology & evolution*, 19: 497–503. doi:10.1016/j.tree.2004.07.006

González-Salazar, C., Stephens, C.R., Marquet. P.A. 2013. Comparing the relative contributions of biotic and abiotic factors as mediators of species' distributions, Ecological Modelling 248: 57–70. doi:10.1016/j.ecolmodel.2012.10.007

Guisan A., Broennimann O., Engler R., Vust M., Yoccoz N.G., Lehmann A. & Zimmermann N.E. (2006) Using niche-based models to improve the sampling of rare species. *Conservation biology: the journal of the Society for Conservation Biology*, 20: 501–511. doi: 10.1111/j.1523-1739.2006.00354.x

Guisan A. & Harrell F.E. (2000) Ordinal response regression models in ecology. Journal of Vegetation Science, 11: 617–626. doi: 10.2307/3236568

Guisan A., Tingley R., Baumgartner J.B., Naujokaitis-Lewis I., Sutcliffe P.R., Tulloch A.I., Regan T.J., Brotons L., McDonald-Madden E., Mantyka-Pringle C., Martin T.G., Rhodes, J. R., Maggini R., Setterfield S. A., Elith J., Schhwartz M. W., Wintle B. A., Broenningmann O., Austin M., Ferrier S., Kearney M. R., Possingham H. P. & Buckley Y.M. (2013) Predicting species distributions for conservation decisions. *Ecology Letters* 16: 1424–1435. doi: 10.1111/ele.12189

Guisan A. & Zimmermann N. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, 135: 147–186. doi:10.1016/S0304-3800(00)00354-9

Guralnick R.P., Wieczorek J., Beaman R., Hijmans R.J. & BioGeomancer Working G. (2006) BioGeomancer: automated georeferencing to map the world's biodiversity data. *PLoS Biol*, 4: e381. doi:10.1371/journal.pbio.0040381

Hefley T.J., Tyre A.J., Baasch D.M. & Blankenship E.E. (2013) Nondetection sampling bias in marked presence-only data. *Ecology and evolution*, 3: 5225–5236. doi: 10.1002/ece3.887

Hijmans R.J. (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, 93: 679–688. doi: 10.1890/11-0826.1

Hirzel A. & Guisan A. (2002) Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*, 157: 331–341. doi:10.1016/S0304-3800(02)00203-X

Hoffmann A.A., Chown S.L. & Clusella-Trullas S. (2013) Upper thermal limits in terrestrial ectotherms: how constrained are they? *Functional Ecology*, 27: 934–949. doi: 10.1111/j.1365-2435.2012.02036.x

Hortal J., Bello F.d., Diniz-Filho J.A.F., Lewinsohn T.M., Lobo J.M. & Ladle R.J. (2015) Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, 46: 523–549. doi: 10.1146/annurev-ecolsys-112414-054400

Hortal J., Jiménez-Valverde A., Gómez J.F., Lobo J.M. & Baselga A. (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, 117: 847–858. doi: 10.1111/j.0030-1299.2008.16434.x

Iverson L.R., Prasad A.M., Matthews S.N. & Peters M. (2008) Estimating potential habitat for 134 eastern US tree species under six climate scenarios. *Forest Ecology and Management*, 254: 390–406. doi:10.1016/j.foreco.2007.07.023

Jamevich C., Stohlgren T.J., Kumar S., Morisette J. T. & Holcombe T.R. (2015) Caveats for correlative species distribution modeling. *Ecological Informatics*, 29: 6–15. doi:10.1016/j.ecoinf.2015.06.007

Jiménez-Valverde, A., Peterson, A.T., Soberón, J., Overton, J.M., Aragón, P. & Lobo J.M. (2011) Use of niche models in invasive species risk assessments. Biological Invasions, 13: 2785–2797. doi: 10.1007/s10530-011-9963-4

Jones K.E., Bielby J., Cardillo M., Fritz S.A., O'Dell J., Orme C.D.L., Safi K., Sechrest W., Boakes E.H., Carbone C., Connolly C., Cutts M.J., Foster J.K., Grenyer R., Habib M., Plaster C.A., Price S.A., Rigby E.A., Rist J., Teacher A., Bininda-Emonds O.R.P., Gittleman J.L., Mace G.M. & Purvis A. (2009) PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, 90: 2648–2648. doi: 10.1890/08-1494.1

Joppa, L.N., McInerny G., Harper R., Salido L., Takeda K., O'Hara K., Gavaghan D. & Emmott S. (2013) Troubling trends in scientific software use. *Science (New York, N.Y.)*, 340: 814–815. doi: 10.1126/science.1231535

Dodds W.K. & Nelson J. A. (2006) Redefining the community: a species-based approach. *Oikos*, 112: 464–472. doi: 10.1111/j.0030-1299.2006.13558.x

Kadmon R., Farber O. & Danin A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, 14: 401–413. doi: 10.1890/02-5364

Kattge J., Díaz S., Lavorel S., Prentice I.C., Leadley P., Bönisch G., Garnier E., Westoby M., Reich P.B., Wright I.J., Cornelissen J.H.C., Violle C., Harrison S.P., Van Bodegom P.M., Reichstein M., Enquist B.J., Soudzilovskaia N.A., Ackerly D.D., Anand M., Atkin O., Bahn M., Baker T.R., Baldocchi D., Bekker R., Blanco C.C., Blonder B., Bond W.J., Bradstock R., Bunker D.E., Casanoves F., Cavender-Bares J., Chambers J.Q., Chapin Iii F.S., Chave J., Coomes D., Cornwell W.K., Craine J.M., Dobrin B.H., Duarte L., Durka W., Elser J., Esser G., Estiarte M., Fagan W.F., Fang J., Fernández-Méndez F., Fidelis A., Finegan B., Flores O., Ford H., Frank D., Freschet G.T., Fyllas N.M., Gallagher R.V., Green W.A., Gutierrez A.G., Hickler T., Higgins S.I., Hodgson J.G., Jalili A., Jansen S., Joly C.A., Kerkhoff A.J., Kirkup D., Kitajima K., Kleyer M., Klotz S., Knops J.M.H., Kramer K., Kühn I., Kurokawa H., Laughlin D., Lee T.D., Leishman M., Lens F., Lenz T., Lewis S.L., Lloyd J., Llusià J., Louault F., Ma S., Mahecha M.D., Manning P., Massad T., Medlyn B.E., Messier J., Moles A.T., Müller S.C., Nadrowski K., Naeem S., Niinemets Ü., Nöllert S., Nüske A., Ogaya R., Oleksyn J., Onipchenko V.G., Onoda Y., Ordoñez J., Overbeck G., Ozinga W.A., Patiño S., Paula S., Pausas J.G., Peñuelas J., Phillips O.L., Pillar V., Poorter H., Poorter L., Poschlod P., Prinzing A., Proulx R., Rammig A., Reinsch S., Reu B., Sack L., Salgado-Negret B., Sardans J., Shiodera S., Shipley B., Siefert A., Sosinski E., Soussana J.F., Swaine E., Swenson N., Thompson K., Thornton P., Waldram M., Weiher E., White M., White S., Wright S.J., Yguel B., Zaehle S., Zanne A.E. & Wirth C. (2011) TRY – a global database of plant traits. Global Change Biology, 17: 2905–2935. doi: 10.1111/j.1365-2486.2011.02451.x

Kissling W.D., Dormann C.F., Groeneveld J., Hickler T., Kuhn I., McInerny G.J., Montoya J.M., Romermann C., Schiffers K., Schurr F.M., Singer A., Svenning J.C., Zimmermann N.E. & O'Hara, R.B. (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. Journal of Biogeography, 39: 2163–2178. doi: 10.1111/j.1365-2699.2011.02663.x

Koffi K.J., Kouassi A.F., Adou Yao C.Y., Bakayoko A., Ipou I.J. & Bogaert J. (2015) The present state of botanical investigations in Côte d'Ivoire. *Biodiversity Informatics*, 10: 56–64. doi: 10.17161/bi.v10i2.5007

Lash R.R., Carroll D.S., Hughes C.M., Nakazawa Y., Karem K., Damon I.K. & Peterson A.T. (2012) Effects of georeferencing effort on mapping monkeypox case distributions and transmission risk. *International journal of health geographics*, 11: 23. doi: 10.1186/1476-072X-11-23

Le Lay G., Engler R., Franc E. & Guisan A. (2010) Prospective sampling based on model ensembles improves the detection of rare species. *Ecography*, 33: 1015–1027. doi: 10.1111/j.1600-0587.2010.06338.x

Lobo J.M. & Martín-Piera, F. (2002) Searching for a Predictive Model for Species Richness of Iberian Dung Beetle Based on Spatial and Environmental Variable. *Conservation Biology* 16: 158–173. **doi: 10.1046/j.1523-1739.2002.00211.x**

Lobo J.M., Jiménez-Valverde A. & Hortal J. (2010) The uncertain nature of absences and their importance in species distribution Modelling. *Ecography* 33: 103–114. doi: 10.1111/j.1600-0587.2009.06039.x

Lyal C., Kirk P., Smith D. & Smith R. (2008) The value of taxonomy to biodiversity and agriculture. *(2008) Biodiversity* 9: 8-14. Available at http://r4d.dfid.gov.uk/PDF/Outputs/CABI/BiodiversityandAgriculture.pdf

Martin L., Blossey B. & Ellis E. (2012) Mapping where the ecologists work: biases in the global distribution of terrestrial ecological observations. *Frontiers in Ecology and the Environment*, 10: 195–201. doi: 10.1890/110154

Matthysen E. (2012) Multicausality of dispersal: a review. In: *Dispersal Ecology and Evolution* (eds. Clobert J, Baguette M, Benton TG & Bullock JM), pp. 3–18. Oxford University Press, Oxford, UK. doi: 10.1093/acprof:oso/9780199608898.003.0001

Meyer C., Jetz W., Guralnick R.P., Fritz S.A. & Kreft H. (2015a) Global drivers of species variation in mobilized point-occurrence information. *PeerJPrePrints*, 3: e1493. doi: 10.7287/peerj.preprints.1218v2

Meyer C., Kreft H., Guralnick R.P. & Jetz W. (2015b) Global priorities for an effective information basis of biodiversity distributions. *PeerJ PrePrints* 3: e1057. doi: 10.7287/peerj.preprints.856v1

Mod H.K., le Roux P.C., Guisan A. & Luoto M. (2015) Biotic interactions boost spatial models of species richness. *Ecography*, 38: 913–921. doi: 10.1111/ecog.01129

Mokany K., Harwood T.D., Overton J.M., Barker G.M. & Ferrier S. (2011) Combining α - and β -diversity models to fill gaps in our knowledge of biodiversity. *Ecology Letters*, 14, 1043–1051. doi: 10.1111/j.1461-0248.2011.01675.x

Morales-Castilla I., Matias M.G., Gravel D. & Araújo M.B. (2015) Inferring biotic interactions from proxies. *Trends in Ecology & Evolution*, 30: 347–356. doi:10.1016/j.tree.2015.03.014

Nathan R. & Muller-Landau H. (2000) Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends in Ecology & Evolution*, 15: 278-285. doi:10.1016/S0169-5347(00)01874-7

Navarro A.G., Peterson A.T. & Gordillo-Martinez, A. (2003) Museums working together: the atlas of the birds of Mexico. *Bulletin of the British Ornithologists' Club*, 123A: 207–225. Available at http://www.biodiversitylibrary.org/item/131291#page/209/mode/1up

Newbold T. (2010) Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, 34: 3–22. doi: 10.1177/0309133309355630

Otegui J., Ariño A.H., Chavan V. & Gaiji S. (2013a) On the dates of GBIF mobilised primary biodiversity records. Biodiversity Informatics 8: 173-184. doi: 10.17161/bi.v8i2.4125

Otegui J., Ariño A.H., Encinas M.A. & Pando F. (2013b). Assessing the primary data hosted by the Spanish node of the Global Biodiversity Information Facility (GBIF). PloS ONE 8: e55144. doi: 10.1371/journal.pone.0055144

Pearce J. L. & Boyce M.S. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, 43: 405–412. doi: 10.1111/j.1365-2664.2005.01112.x

Pelayo-Villamil P., Guisande C., Vari R.P., Manjarrés-Hernández A., García-Roselló E., González-Acosta J., Heine J., González Vilas L., Patti B., Quince E.M., Jiménez L.F., Granado-Lorencio C., Tedesco P.A. & Lobo J.M. (2015) Global diversity patterns of freshwater fishes – potential victims of their own success. *Diversity and Distributions* 21: 345–356. doi: 10.1111/ddi.12271

Peterson A.T., Moses L.M. & Bausch D.G. (2014) Mapping Transmission Risk of Lassa Fever in West Africa: The Importance of Quality Control, Sampling Bias, and Error Weighting. *PLoS ONE*, 9: e100711. doi: 10.1371/journal.pone.0100711

Peterson A.T. (2015) Mapping Disease Transmission Risk: Enriching Models Using Biogeography and Ecology. *Emerging Infectious Diseases*, 21: 1498–. doi: 10.3201/eid2108.150665

Peterson A.T., Barve N., Bini L.M., Diniz-Filho J.A., Jimenez-Valverde A., Lira-Noriegaa A., Lobo J., Maher S., de Marco P., Martinez-Meyer E., Nakazawa Y. & Soberon J. (2009) The climate envelope may not be empty. *Proceedings of the National Academy of Sciences of the United States of America*, 106: E47–E47. doi: 10.1073/pnas.0809722106

Peterson A.T., Knapp S., Guralnick R., Soberón J. & Holder M. (2010) The big questions for biodiversity informatics. *Systematics and Biodiversity*, 8: 159–168. doi: 0.1080/14772001003739369

Peterson A.T., Navarro-Sigüenza A. & Pereira R.S. (2004) Detecting errors in biodiversity data based on collectors itineraries. *Bulletin of the British Ornithologists Club*, 124: 143–151. Available at http://www.biodiversitylibrary.org/page/40056120#page/155/mode/1up

Peterson A.T., Navarro-Sigüenza A.G., Martínez-Meyer E., Cuervo-Robayo A.P., Berlanga H. & Soberón J. (2015) Twentieth century turnover of Mexican endemic avifaunas: Landscape change versus climate drivers. *Science advances*, 1: e1400071. doi: 10.1126/sciadv.1400071

Peterson A.T., Soberón J., Pearson R.G., Anderson R. P., Martínez–Meyer E., Nakamura M. & Araújo M.B. (2011) *Ecological Niches and Geographic Distributions*. Princeton University Press, Princeton. Available at http://press.princeton.edu/titles/9641.html

Phillips S.J., Dudík M., Elith J., Graham C.H., Lehman A., Leathwick J. & Ferrier S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19: 181–197. doi: 10.1890/07-2153.1

Pineda E. & Lobo J.M. (2009) Assessing the accuracy of species distribution models to predict amphibian species richness patterns. *Journal of Animal Ecology* 78: 182–190. doi: 10.1111/j.1365-2656.2008.01471.x

Poelen J.H., Simons J.D. & Mungall C.J. (2014) Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics*, 24: 148–159. doi:10.1016/j.ecoinf.2014.08.005

Randin C.F., Jaccard H., Vittoz P., Yoccoz N.G. & Guisan A. (2009) Land use improves spatial predictions of mountain plant abundance but not presence-absence. *Journal of Vegetation Science*, 20: 996–1008. doi: 10.1111/j.1654-1103.2009.01098.x

Rios N. & Bart L. (2008) Community Building and Collaborative Georeferencing using GeoLocate. In: *The Proceedings of TDWG: Biodiversity Information Standards* (eds. Weitzmann A & Belbin L). TDWG, Freemantle, Australia. Available at http://www.tdwg.org/fileadmin/2008conference/documents/Proceedings2008.pdf

Royle J.A., Chandler R.B., Yackulic C. & Nichols J.D. (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, 3: 545–554. doi: 10.1111/j.2041-210X.2011.00182.x

Ruete A. (2015) Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. *Biodiversity data journal* 3: e5361. doi: 10.3897/BDJ.3.e5361

Scholes R.J., Mace G.M., Turner W., Geller G.N., Jurgens N., Larigauderie A., Muchoney D., Walther B.A. & Mooney H.A. (2008) Ecology. Toward a global biodiversity observing system. *Science*, 321: 1044–1045. doi: 10.1126/science.1162055

Schurr F.M., Pagel J., Cabral J.S., Groeneveld J., Bykova O., O'Hara R.B., Hartig F., Kissling W.D., Linder H.P., Midgley G.F., Schröder B., Singer A. & Zimmermann N.E. (2012) How to understand species' niches and range dynamics: a demographic research agenda for biogeography. *Journal of Biogeography*, 39: 2146–2162. doi: 10.1111/j.1365-2699.2012.02737.x

Smolik M.G., Dullinger S., Essl F., Kleinbauer I., Leitner M., Peterseil J., Stadler L.-M. & Vogl G. (2010) Integrating species distribution models and interacting particle systems to predict the spread of an invasive alien plant. *Journal of Biogeography*, 37: 411–422. doi: 10.1111/j.1365-2699.2009.02227.x

Soberón J. (2014) The Global Biodiversity Information Facility: a case study of biodiversity data sharing. In: *DNA Banking for the 21st Century* (eds. Applequist W & Campbell L). The William L. Brown Center at the Missouri Botanical Garden,153–164.

Soberón J., Arriaga L. & Lara L. (2002a) Issues of quality control in large, mixed-origin entomological databases. In: *Towards a Global Biological Information Infrastructure,* 15–22. European Environment Agency, Copenhagen. Available at http://www.eea.europa.eu/publications/technical_report_2001_70

Soberón J., Arriaga L. & Lara L. (2002b) Issues of quality control in large, mixed-origin entomological databases. In: *Towards a Global Biological Information infrastructure*, 15–22. European Environment Agency, Copenhagen. Available at http://www.eea.europa.eu/publications/technical_report_2001_70

Soberón J., Davila P. & Golubov J. (2004) Targeting sites for biological collections. In: *Seed Storage: Turning Science into Practice.* (eds. Smith RD, Dickie JB, Linington SH, Pritchard HW & Probert RJ). Kew Royal Botanical Gardens, London.

Soberón J., Jimenez R., Golubov J. & Koleff P. (2007) Assessing completeness of biodiversity databases at different spatial scales. *Ecography*, 30: 152–160. doi: 10.1111/j.0906-7590.2007.04627.x

Soberón J., Llorente J. & Benítez H. (1996) An international view of national biological surveys. *Annals of the Missouri Botanical Garden*, 83: 562–573. doi: 10.2307/2399997

Soberón J. & Peterson T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359: 689–698. doi: 10.1098/rstb.2003.1439

Sousa-Baena M.S., Garcia L.C. & Peterson A.T. (2013) Completeness of Digital Accessible Knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions*, 20: 369–381. doi: 10.1111/ddi.12136

Stein B.R. & Wieczorek J. (2004) Mammals of the world: MaNIS as an example of data integration in a distributed network environment. *Biodiversity Informatics*, 1: 14–22. doi: 10.17161/bi.v1i0.7

Sunday J.M., Bates A.E. & Dulvy N.K. (2011) Global analysis of thermal tolerance and latitude in ectotherms. *Proceedings of the Royal Society of London B: Biological Sciences*, 278: 1823–1830. doi: 10.1098/rspb.2010.1295

Svenning J.-C. & Skov F. (2004) Limited filling of the potential range in European tree species. *Ecology Letters*, 7: 565–573. doi: 10.1111/j.1461-0248.2004.00614.x

Thomas C.D. (2010) Climate, climate change and range boundaries. *Diversity and Distributions*, 16: 488–495. doi: 10.1111/j.1472-4642.2010.00642.x

Vittoz P. & Engler R. (2007) Seed dispersal distances: a typology based on dispersal modes and plant traits. Botanica Helvetica, 117: 109–124. doi: 10.1007/s00035-007-0797-8

Ward G., Hastie T., Barry S., Elith J. & Leathwick J.R. (2009) Presence-only data and the EM algorithm. *Biometrics*, 65: 554–563. doi: 10.1111/j.1541-0420.2008.01116.x

Warren R., VanDerWal J., Price J., Welbergen J.A., Atkinson I., Ramirez-Villegas J., Osborn T.J., Jarvis A., Shoo L.P., Williams S.E. & Lowe J. (2013) Quantifying the benefit of early climate change mitigation in avoiding biodiversity loss. *Nature Climate Change*, 3: 678–682. doi: 10.1038/nclimate1887

Wheeler Q.D., Knapp S., Stevenson D.W., Stevenson J., Blum S.D., Boom B.M., Borisy G.G., Buizer J.L., De Carvalho M.R., Cibrian A., Donoghue M.J., Doyle V., Gerson E.M., Graham C.H., Graves P., Graves S.J., Guralnick R.P., Hamilton A.L., Hanken J., Law W., Lipscomb D.L., Lovejoy T.E., Miller H., Miller J.S., Naeem S., Novacek M.J., Page L.M., Platnick N.I., Porter-Morgan H., Raven P.H., Solis M.A., Valdecasas A.G., Van Der Leeuw S., Vasco A., Vermeulen N., Vogel J., Walls R.L., Wilson E.O. & Woolley J.B. (2012) Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. *Systematics and Biodiversity*, 10: 1–20. doi: 10.1080/14772000.2012.665095

Whittaker R.J., Araújo M.B., Jepson P., Ladle R.J., Watson J.E.M. & Willis K.J. (2005) Conservation Biogeography: assessment and prospect. *Diversity and Distributions*, 11: 3–23. doi: 10.1111/j.1366-9516.2005.00143.x

Wieczorek J., Guo Q. & Hijmans R.J. (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, 18: 745–767. doi: 10.1080/13658810412331280211

Wisz M.S., Pottier J., Kissling W.D., Pellissier L., Lenoir J., Damgaard C.F., Dormann C.F., Forchhammer M.C., Grytnes J.A., Guisan A., Heikkinen R.K., Hoye T.T., Kuhn I., Luoto M., Maiorano L., Nilsson M.C., Normand S., Ockinger E., Schmidt N.M., Termansen M., Timmermann A., Wardle D.A., Aastrup P. & Svenning J.C. (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. Biological Reviews, 88: 15–30. doi: 10.1111/j.1469-185X.2012.00235.x

Yesson C., Brewer P.W., Sutton T., Caithness N., Pahwa J.S., Burgess M., Gray W.A., White R.J., Jones A.C., Bisby F.A. & Culham A. (2007) How Global Is the Global Biodiversity Information Facility? *PLoS ONE*, 2: e1124. doi:10.1371/journal.pone.0001124

Appendix. A survey of fitness for use of GBIF data for SDM.

The members of the Task Group created a survey for the biodiversity informatics community, aimed specifically on those likely to use GBIF data for modeling species distributions. They compiled a database of names of researchers considered particularly relevant, and invited those persons to take the survey. Additionally, the survey was opened to the biodiversity informatics community worldwide, and publicized through email, social media (Facebook, Twitter, etc.), and word of mouth. This effort yielded 137 responses from scientists in31 countries[22]. Responders provided overwhelmingly consistent answers to issues of data shortcomings and presentation, annotations from users, feedback to providers, and repositories of occurrence data used in peer-reviewed publications. A variety of practical problems surfaced regarding the interface itself . Regarding principal problems of the data, 78% noted issues with the georeferences. Importantly, respondees mentioned how the GBIF portal could be improved in two regards. First, 89% would find quantification/mapping of sampling effort/data completeness useful. Second, annotations of data quality (and communication *from* users and *to* data providers) were seen as critical, in the following particular ways. Almost all suggested that users be allowed to annotate data (56% “very important” + additional 43% “important”), and that those annotations be transmitted automatically to data providers (56% “very important” + additional 41 “important”). To rounding error, 100% of respondees saw great utility in GBIF transmitting information to data providers, including annotations of ID quality (80% “very important” + additional 20%

[22] *Argentina, Australia, Belgium, Benin, Brazil, Cameroon, Canada, Chile, Colombia, Denmark, Ecuador, Finland, France, Germany, India, Ireland, Italy, Kenya, Mexico, Netherlands, New Zealand, Peru, Portugal, South Africa, Spain, Sweden, Switzerland, UK, Uruguay, USA, Venezuela*

“important”) and georeferences (85% “very important + additional 15% important”). Nevertheless, most (39% “very important” + additional 37% “important”) suggested that users be allowed to annotate only certain parts of the data (taxonomy/geography), and almost all (57% “very important” + 36% “important”) advocated allowing users to provide a quality or “fit for use” tag for individual records. Closing the loop regarding feedback and data quality, they considered it very important (55%) or important (44%) that data providers spend the time and money required to correct/update data (taxonomically/geographically) as per observations provided by users. Finally, a large majority (77%) thought that the field would be well served by a single online repository/archive for point occurrence data published in peer-reviewed journals. An even larger number were highly supportive (59%) or supportive (31%) of GBIF being *a* repository for such data.