PERSPECTIVE

Frontiers of Biogeography
the scientific journal of
the International Biogeography Society

Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society

Robert P. Anderson[1,2,3,*], Miguel B. Araújo[4,5], Antoine Guisan[6], Jorge M. Lobo[4], Enrique Martínez-Meyer[7,8], A. Townsend Peterson[9] and Jorge M. Soberón[9,10]

[1] Department of Biology, City College of New York, City University of New York, New York, NY, 10031, USA; [2] Ph.D. Program in Biology, Graduate Center, City University of New York, New York, NY, 10016, USA; [3] Division of Vertebrate Zoology (Mammalogy), American Museum of Natural History, New York, NY, 10024, USA; [4] Departamento de Biogeografía y Cambio Global, Museo Nacional de Ciencias Naturales (MNCN-CSIC), 28006, Madrid, Spain; [5] "Rui Nabeiro" Biodiversity Research Chair, MED Institute, University of Évora, 7000, Évora, Portugal; [6] Department of Ecology and Evolution & Institute of Earth Surface Dynamics, University of Lausanne, 1015, Lausanne, Switzerland; [7] Centro del Cambio Global y la Sustentabilidad, AC, Villahermosa, Mexico; [8] Departamento de Zoología, Instituto de Biología, Universidad Nacional Autónoma de México, Ciudad de México, Mexico; [9] Biodiversity Institute, University of Kansas, Lawrence, KS, 66045, USA; [10] Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, 66045, USA; *Corresponding author: randerson@ccny.cuny.edu, https://www.andersonlab.ccny.cuny.edu/

Abstract

Vast amounts of Primary Biodiversity Data exist online (~10^9 records, each documenting an individual species at a point in space and time). These data hold immense but unrealized promise for science and society, including use in biogeographic research addressing issues such as zoonotic diseases, invasive species, threatened species and habitats, and climate change. Ongoing and envisioned changes in biodiversity informatics involving data providers, aggregators, and users should catalyze improvements to allow efficient use of such data for diverse analyses. We discuss relevant issues from the perspective of modeling species distributions, currently the most common use of Primary Biodiversity Data. Key cross-cutting principles for progress include harnessing feedback from users and increasing incentives for improving data quality. Critical challenges include: (1) establishing individual and collective stable unique identifiers across all of biodiversity science, (2) highlighting issues regarding data quality and representativeness, and (3) improving feedback mechanisms. Such changes should lead to ever-better data and increased utility and impact, including greater data integration with various research areas within and beyond biogeography (e.g., population demography, biotic interactions, physiology, and genetics). Building on existing pilot functionalities, biodiversity informatics could see transformative changes over the coming decade via a combination of community consensus building, coordinated efforts to justify and secure funding, and technical innovations.

Highlights:

- Online biodiversity data hold great yet untapped potential for biogeographic studies linking to diverse areas of environmental research.
- Human health, agriculture, and the conservation and management of natural systems depend on efficient use of biodiversity data.
- Ongoing progress should be expanded to promote transformative changes in the quality and utility of biodiversity data.
- Data usage in publications and reports can serve as a currency of the utility of biodiversity data and the institutions that provide it.
- Necessary changes related to online portals require consensus-building by various stakeholders, catalysis by funding agencies, innovative pilot solutions, and widespread implementation.

Keywords:
bias, biodiversity, citizen science, environment, herbarium, informatics, natural history museum, occurrence, range, uncertainty

e-ISSN: 1948-6596 https://escholarship.org/uc/fb doi:10.21425/F5FBG47839

© the authors, CC-BY 4.0 license

Introduction

A staggering amount of digital information regarding biodiversity now exists on the Internet, with many ongoing changes aimed at meeting the needs of science and society. Primary Biodiversity Data represent the principal information available for most species on Earth, consisting of individual records with place, time, and taxonomic identification (Soberón and Peterson 2004). The biodiversity informatics community includes three overlapping groups interested in such data: (1) *data providers*, such as natural history museums, herbaria, and networks of citizen scientists; (2) *data aggregators*, initiatives that serve data combined from multiple providers; and (3) *data users*, including scientists, decision-makers, and the general public (Figure 1; Graham et al. 2004). Integrated by standards such as the DarwinCore (Wieczorek et al. 2012), enormous stores of Primary Biodiversity Data now exist online, with the Global Biodiversity Information Facility (GBIF[1]) constituting the largest and most comprehensive aggregator (>1.4 x 10^9 digital records from >1500 providers corresponding to >2.3 x 10^6 species; Robertson et al. 2014).

Ideally, Primary Biodiversity Data lead to synthetic knowledge and real-world applications, especially via association with information regarding diverse organismal attributes (e.g., measurements, images, recordings, and DNA sequences; and physiological, behavioral, ecological, or ethnobiological data; Ratnasingham and Hebert 2007, Cook et al. 2016, Troudet et al. 2018; Box 1). Through diverse biogeographic and environmental research, Primary Biodiversity Data hold tremendous potential for applications to pressing environmental issues—such as understanding zoonotic diseases and invasive species, characterizing threatened species and habitats, planning conservation priorities, and anticipating effects of ongoing climate change (Figure 2; Peterson et al. 2010, Guisan et al. 2013, Hallgren et al. 2016, Johnson et al. 2019). Indeed, many major biodiversity assessments rely heavily on Primary Biodiversity Data and linked information or the results of studies that use them (Pereira et al. 2010, Sarukhán et al. 2015, IPBES 2019). For all of these uses, relevant high-quality data must be readily available for efficient assembly, especially for time-sensitive issues such as an emerging zoonotic disease or recently detected invasive species (Anderson 2012, Johnson et al. 2019).

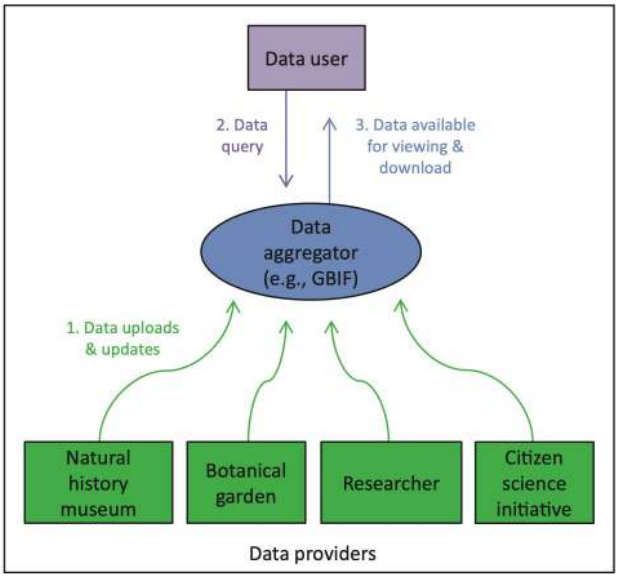

Figure 1. Simplified overview of the interactions and flow of data among providers, aggregators, and users in biodiversity informatics. Numbers indicate the typical order of actions: 1. Aggregator receives data uploads (and periodic updates) from providers; 2. User makes a data query to aggregator's online portal; 3. Aggregator responds to query by making data available on portal (for viewing and/or download). Note that by querying a single aggregator, a user can receive data from multiple providers. Additionally, multiple intermediate aggregators typically exist, feeding into the largest ones most commonly consulted by users (e.g., GBIF).

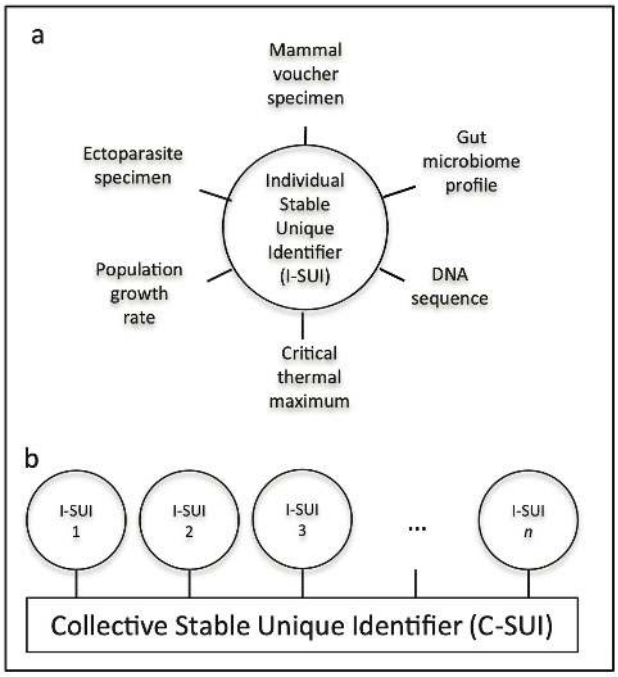

Figure 2. Use of individual and collective Stable Unique Identifiers (e.g., DOIs) in biodiversity informatics. (a) Individual Stable Unique Identifier (I-SUI) allows linking diverse data domains for a given organism. In this example, an I-SUI links the voucher specimen and associated Primary Biodiversity Data (e.g., date and locality) of an individual mammal to information regarding various aspects of molecular- to population-level biology. (b) Collective Stable Unique Identifier (C-SUI) denotes a set (i.e., a list) of individual identifiers. For example, a C-SUI could indicate the *n* individual records used in a given analysis.

1 https://www.gbif.org/, last accessed on 11 March 2020

 © the authors, CC-BY 4.0 license

Box 1. Data realms and research areas within and beyond biogeography that will be promoted by changes to biodiversity informatics focusing on Primary Biodiversity Data.

Important data realms beyond the current DarwinCore fields include those regarding **absences** (Lobo et al. 2010, Howard et al. 2014, Guillera-Arroita et al. 2015), **population demography** (Fordham et al. 2013, Merow et al. 2014, Ehrlén and Morris 2015), **movement** (Brook et al. 2009, Smouse et al. 2010, Franklin et al. 2014), **biotic interactions** (Kissling et al. 2012, Wisz et al. 2013, Morales-Castilla et al. 2015, D'Amen et al. 2018), **physiology** (Clusella-Trullas et al. 2011, Barve et al. 2014, Kearney et al. 2014), and **genetics** (Harris et al. 2013, Valladares et al. 2014, Fitzpatrick and Keller 2015, Exposito-Alonso et al. 2018). Such information can be integrated with Primary Biodiversity Data records: (1) using the flexible "dynamicProperties" field of DarwinCore, (2) directly with an expansion of the DarwinCore, or (3) via links from Primary Biodiversity Data aggregators to external databases. For the latter, stable unique identifiers allow linkages to individual records, but sometimes links only will be possible for taxonomic names and geographic locations.

Data realm	Examples	Research topics
Absences	Field survey effort underlying sets of Primary Biodiversity Data records (allowing discrimination of well vs. poorly sampled spatial units; Soberón et al. 2007, Lobo et al. 2018)	• Building distribution models using sites of relatively reliable absence • Identifying regions with greater uncertainties in model prediction • Prioritizing future survey efforts
Population demography	Population size (abundance and density) and growth rates over space and time (Salguero-Gómez et al. 2015, Salguero-Gómez et al. 2016, Santini et al. 2018)	• Associations between environmental suitability and population biology • Population-level research questions of a temporally dynamic nature (e.g., species range shifts)
Movement	Position of individuals through time, individual movement tracks, and capture–recapture information (Nathan and Muller-Landau 2000, Ovaskainen et al. 2008)	• Consideration of the ability of individuals to move across landscapes • Migratory phenomena and ongoing range shifts (e.g., invasive species)
Biotic interactions	Interactions between individuals of different species (e.g., insect X collected on plant Y); or co-occurrence matrices linked with databases regarding species traits, biotic interactions, and phylogenetic relationships (Jones et al. 2009, Kattge et al. 2011, Poelen et al. 2014, Wilman et al. 2014)	• Effects of biotic interactions on species distributions and community composition • Applied topics that depend on the effects of biotic interactions (e.g., zoonotic diseases)
Physiology	Physiological measurements (*in situ* or *ex situ*; Sunday et al. 2011, Bennett et al. 2018)	• Physiological variation among individuals and across populations • Comparisons between (and integration of) correlative and mechanistic models
Genetics	Gene sequences, expression profiles (Ratnasingham & Hebert 2007, Pelini et al. 2009, O'Neil et al. 2014)	• Geographic and environmental distributions of alleles • Tests for natural selection across populations

However, several limitations currently constrain the utility of Primary Biodiversity Data, and we explain and advocate for ongoing and envisioned changes that could improve data quality dramatically and allow widespread realistic uses for basic and applied science. We provide examples through the lens of adequacy for modeling species distributions—for which they are most commonly employed—but the same issues and solutions hold for myriad other uses (Graham et al. 2004). Although we take advantage of an *ad hoc* online consultation of the community conducted by the GBIF Secretariat (GBIF 2016; Table 1), we cover issues germane to all aggregators. We provide several specific illustrations based on current functionalities of GBIF (Robertson et al. 2014) but also point out innovations by some other aggregators. Below, we summarize principal current limitations in the field, outline ongoing and envisioned solutions, and sketch a

 © the authors, CC-BY 4.0 license

roadmap for implementation. We begin by highlighting two critical cross-cutting principles for improving biodiversity informatics: harnessing feedback from users and promoting improvements in data quality.

Cross-cutting principles for progress

Current and future users can provide the best information regarding Primary Biodiversity Data and its quality, and ongoing changes that link users, providers, and aggregators can help harness their feedback (Suhrbier et al. 2017). Most aggregators integrate periodic updates from providers, adding new records and correcting previous errors. Additionally, users often invest substantial time and resources correcting taxonomic identifications and determining georeferences. Nevertheless, in both the GBIF community consultation and our discussions with colleagues, users indicated that: (1) most aggregated databases lack functionalities allowing users to flag problematic records or suggest improved information within the online interface; and (2) providers do not consistently update records based on user feedback (GBIF 2016, Suhrbier et al. 2017). Fortunately, the situation regarding the former is changing rapidly via pilot implementations, but changes are needed to increase the incentives and resources for the latter.

The biodiversity informatics community can take various actions to promote improvements in data quality. Often with fixed or declining budgets, data providers (especially natural history museums and herbaria) juggle many priorities, including maintaining physical specimens and their associated data. To help increase the resources available for improving data quality, the field needs explicit information flows that document both data quality and use (van Hintum et al. 2011). Importantly, indices of data quality can be tracked over time to assess progress and outstanding needs. Moreover, data usage represents a critical potential currency, with higher-quality information being used more frequently. The usage of individual Primary Biodiversity Data can be quantified via linkage with documentation of their use—for example downloading events and, most importantly, publications or reports based on them (Costello et al. 2013). Standardized quantifications of data quality and use should both help justify improvements to data quality and increase incentives for both providers and funding sources to improve data quality.

Current limitations

Consideration of key data-related issues for models of species ecological niches and geographic distributions (hereafter distribution models) exemplifies current limitations of Primary Biodiversity Data for many kinds of biodiversity analyses (Araújo et al. 2019). Distribution models integrate such data with environmental information to estimate the conditions and places suitable for a species (Franklin 2010, Peterson et al. 2011, Guisan et al. 2017). Nevertheless, data from aggregated databases cannot be used in distribution modeling without substantial data-cleaning and filtering (to fix errors and remove records of insufficient quality), as well as consideration of inherent biases (Beck et al. 2014, Gueta and Carmel 2016). Indeed, the DarwinCore standard was developed to include fields that characterize data limitations and promote appropriate usage (Wieczorek et al. 2012, Otegui et al. 2013). However, current portals do not provide the functionalities necessary for researchers to assemble data suitable for such analyses efficiently, especially because various uses require different data quality needs (GBIF 2016, Veiga et al. 2017). As we outline briefly below, limitations that hinder the use of such data at present correspond to those that are: 1) inherent to the data, 2) affect access to the data, or 3) relate to how the data are used.

Limitations associated with Primary Biodiversity Data themselves include the lack of information, as well as inaccuracies and biases. As frequently mentioned, a few key information fields remain empty for a high proportion of digital records. Although copious records lack digitization or species-level identification, the greatest immediate obstacle concerns the lack of georeferences (Hill et al. 2009, Beaman and Cellinese 2012, Peterson et al. 2015). Furthermore, records include inaccuracies and biases, which are well known but not yet rectified (Meyer et al. 2015, Amano et al. 2016, Troia and McManamay 2016). Taxonomic misidentifications and inaccurate georeferences are highly problematic, compounded by the fact that fields regarding their uncertainty are almost always empty (Wieczorek et al. 2004, Guralnick et al. 2007). In addition, geographic and temporal biases in biological sampling effort pervade Primary Biodiversity Data (with some areas more heavily sampled than others, and effort varying greatly among years and to a lesser degree across annual seasons); such biases negatively affect distribution models unless taken into account (Hortal et al. 2008, Phillips et al. 2009).

Regarding data access, key information is seldom provided in transparent and easily accessible ways, leading to unrealistically high impressions of data quality as well as incorrect inferences regarding species ranges and their shifts over time (Araújo et al. 2009). Some data shielding rightly aims to protect sensitive species from exploitation, and temporary data "embargoes" sometimes protect research interests of those who collected the data (Brooke 2000, Graves 2000). However, existing information regarding the uncertainty of taxonomic identifications and georeferences—as well as characterizations of spatial and temporal biases—are not made immediately obvious to the user in current portals. This situation leads many non-specialists to misconceptions: that identifications and georeferences have little or no error; and that the lack of occurrence records for a species in a region or time period indicates its absence (Ruete 2015).

Limitations regarding data use correspond to both use *per se* as well as documentation. Commonly, researchers use data without adequate cleaning and filtering, often not realizing the high levels of error, bias, and uncertainty or the degree to which such problems adversely affect modeling analyses. Whereas

 © the authors, CC-BY 4.0 license

substantial research has addressed issues related to error and bias in distribution modeling, the field needs substantial advancements regarding how to integrate and characterize information on uncertainty (Rocchini et al. 2011, Lash et al. 2012). With respect to documentation, distribution modeling is part of an ongoing transition in scientific research regarding data access and reproducibility. Increasingly, journals and funding sources require that data used in publications be made openly available (Molloy 2011, Reichmen et al. 2011; e.g., *Nature Scientific Data*, *Biodiversity Data Journal*). Whereas digital deposition is customary for some kinds of data (e.g., GenBank for gene sequences; DRYAD for more diverse data types; Greenberg et al. 2009), no equivalent expectation or standard mechanism yet exists for Primary Biodiversity Data (Table 1; Chavan and Ingwersen 2009, Costello et al. 2013, Guralnick et al. 2015). Similarly, recent years have seen dramatic increases in online supplemental information and external repositories to document methods and provide code (Campbell et al. 2019). Unfortunately, distribution modeling studies still infrequently explain adequately the steps taken to obtain, clean, and filter Primary Biodiversity Data and to conduct analyses (or provide underlying code/workflows), but recent advances in automated documentation and metadata standardization greatly facilitate such goals (Kass et al. 2018, Feng et al. 2019, Merow et al. 2019).

Ongoing and envisioned solutions

Enable universal communication

Several initiatives by providers and aggregators are currently progressing towards the establishment and implementation of stable unique identifiers that allow clear links among data, both for individual records and collective sets of records (Figure 2; Page 2008). Stable unique identifiers (e.g., Digital Object Identifiers) provide unambiguous, long-lasting reference to a particular entity—for Primary Biodiversity Data typically a voucher specimen or observation event. At the individual level, such identifiers help data providers receive and act on feedback from users or aggregators (Table 1) and also allow individual-level linkages both between records (e.g., parasite and host) and between data realms (e.g., Primary Biodiversity Data and gene sequences; Peterson et al. 2010, Cook et al. 2016; Box 1). Fortunately, many aspects of such identifiers have been implemented for some individual aggregators, such as a universally unique identifier (UUID) automatically generated upon upload. Nevertheless, data providers and aggregators need to ensure that a given Primary Biodiversity Data record does not exist more than once under different identifiers (as currently happens in GBIF), for example via checks against other identifier fields in the DarwinCore. Furthermore, a broad consensus must be reached regarding mechanism to achieve a standardized identifier system that can be used across aggregators and throughout biodiversity science (Guralnick et al. 2015, Suhrbier et al. 2017, BCN 2018). We advocate for a single registry service to guarantee that a given identifier indeed is universally unique for all biodiversity uses (Costello et al. 2013).

The field also needs collective stable unique identifiers that each specify a list of individual-level identifiers. For example, a collective identifier can be used to denote all of the records in a particular download from an aggregator, or to all records used in an analysis (Figure 2; Table 1). Many aggregators (including GBIF) support the first functionality, but they and other aggregators currently lack the second (e.g., to receive and integrate information regarding a bundle of records). Collective stable unique identifiers for the records used in a particular analysis (e.g., *after* data cleaning/filtering; Costello et al. 2013) or for a coherent dataset (e.g., sampled in a specific field survey effort) will provide a short way of denoting long lists of records; such identifiers will prove critical by facilitating

Table 1. Summary of responses from GBIF community consultation of users regarding data adequacy for modeling species distributions (n = 137; GBIF 2016). Respondents provided overwhelmingly consistent answers to issues of data access via the online portal and feedback from users, as well as strong majority opinions regarding repositories of occurrence data used in peer-reviewed publications.

Enhancement of online biodiversity informatics portals	Favorable response
Quantification/mapping of sampling effort/data completeness would be useful.	89%
Users should be allowed to annotate data.	99%
Annotations should be transmitted automatically to data providers.	97%
Allowed annotations should include the quality of the taxonomic identification.	100%
Allowed annotations should include the quality of the georeference.	100%
Users should be allowed to provide a quality or "fit for use" tag for individual records.	93%
Providers should spend the time and money required to correct/update data (taxonomically/geographically).	99%
The field would be well served by a single online repository/archive for point occurrence data published in peer-reviewed journals.	77%
GBIF should be one such repository/archive for point occurrence data published in peer-reviewed journals.	90%

 © the authors, CC-BY 4.0 license

documentation, reproducibility, and calculation of statistics regarding the use of Primary Biodiversity Data (Guralnick et al. 2015, Nelson et al. 2018). Importantly, a system for individual and collective identifiers that are unique across all of biodiversity science could catalyze agreement regarding a community standard for expected digital deposition of Primary Biodiversity Data used in publications and reports (analogous to submission to Genbank for DNA sequences).

Highlight data uncertainties and biases

Wise and efficient use of Primary Biodiversity Data also depends on aggregators highlighting issues regarding data quality and representativeness. Users need easy and obvious access to fields documenting the reliability of identifications and georeferences (Figure 3). Both are defined in DarwinCore (Wieczorek et al. 2012). Although most records currently lack any information for these fields, information regarding the latter has been populated densely in a few initiatives (e.g., VertNet[2] and progenitors; Costello and Wieczorek 2014). Similarly, some citizen-science initiatives aim at providing flags based on plausibility upon upload (e.g., INFOFLORA[3]) or have vetting processes built into their posting systems (e.g., eBird[4]). Developing tools that allow easy query and visualization of fields related to uncertainty will help users assess the appropriateness of records for the study at hand (Figure 3; Chapman et al. 2020).

To help users address issues related to sampling biases, aggregators also can facilitate construction and visualization of proxies for sampling effort across space and time (Figure 3; Table 1; Guralnick et al. 2007, Hortal et al. 2008, Otegui et al. 2013, Sousa-Baena et al. 2014). Records for a broad suite of taxa detected with similar techniques ("target groups" for field sampling) can provide a quantitative estimate of the efforts that yielded the records for the particular species of interest. Data for such target groups (e.g., small, non-volant mammals) document the places and times where relevant efforts occurred and can be used to quantify indices that serve as proxies of sampling and its spatial and temporal gaps (Anderson 2003). Some useful implementations exist for visual display of records from a given search. For example, the "Spatial Module" of Symbiota[5] (Gries et al. 2014) provides a heat density visualization of records as well as a "Date Slider" that allows the user to control the display of records by date range. Aggregators should expand such functionalities to make querying, mapping, summarizing, and downloading such records an integral part of their online interfaces, allowing the user to customize the relevant target group by taking into account knowledge of relevant biological sampling protocols (Figure 3). Such quantifications of sampling enable corrections for biases (Phillips et al. 2009, Fithian et al. 2015), and indices of sampling

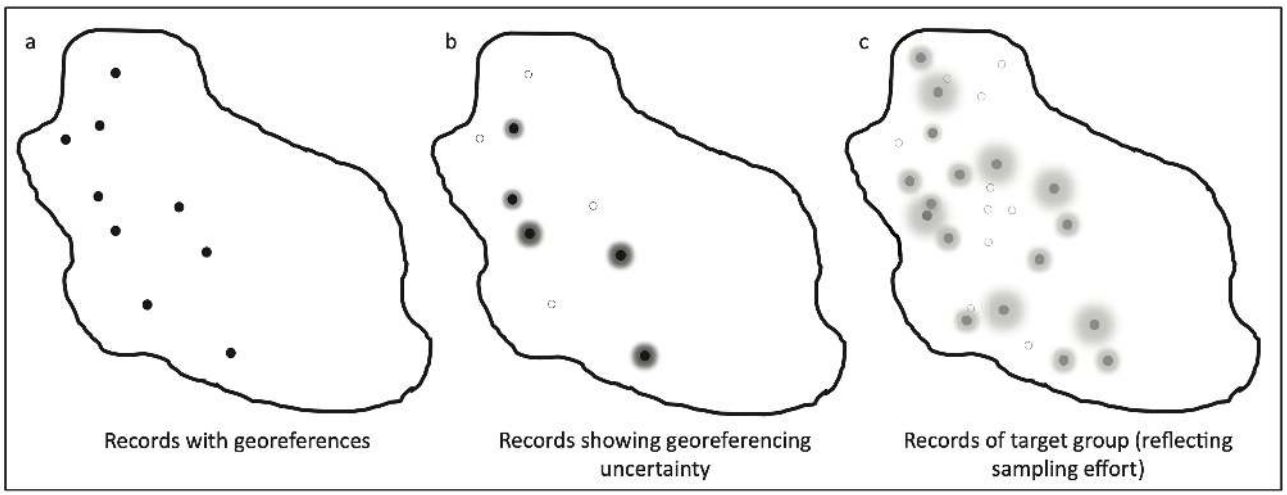

Figure 3. Examples of ways in which aggregators can make uncertainties and biases visually available to users of Primary Biodiversity Data. Such information can be employed to filter data and to quantify and correct for biases in sampling effort, respectively. (a) Georeferenced localities of a given species are simply plotted in geographic space (black dots; current practice). (b) Those same localities appear using symbologies that provide additional information; a hazy cloud indicates the radius of error for localities holding information regarding uncertainty of the georeference, and localities lacking such data appear only as hollow black circles. (c) Information appears that reflects the results of sampling effort, by showing in gray the georeferenced localities for all species belonging to a more inclusive target group (i.e., all species detected with the same techniques as the species of interest; conventions the same as in b). Note that the right-hand side of the study region lacks records for any species of the target group, suggestive of very low sampling effort there.

2 http://vertnet.org/, last accessed on 28 May 2020

3 https://www.infoflora.ch/en/, last accessed on 28 May 2020

4 https://ebird.org/home, last accessed on 28 May 2020

5 http://symbiota.org/docs, last accessed on 28 May 2020

 © the authors, CC-BY 4.0 license

completeness eventually could be populated for the same purpose. Highlighting gaps in sampling also can facilitate priority-setting for digitization, georeferencing, and further sampling efforts.

Improve feedback mechanisms

Finally, aggregators can catalyze improvements in data much more effectively by implementing quality flags and annotations, as well as better quantifications of uncertainty. Automated data-cleaning efforts can discover, document, and flag some problems (e.g., geographic inconsistencies, spatial or environmental outliers, or disagreements with expert maps; García-Roselló et al. 2014, Robertson et al. 2016). For example, GBIF includes a series of known issues and flags discovered by checking procedures during integration (or populated by data providers). However, as mentioned earlier, the best information regarding data quality depends on the expertise of users (Peterson et al. 2004)—both individual researchers (e.g., experts on a given taxon) and groups of users (e.g., national biodiversity agencies; Table 1; Guralnick et al. 2007, Ratnasingham and Hebert 2007). Specifically, aggregators can enlist users to detect and flag problems, suggest improvements, quantify quality, and provide annotations that document the information and methods employed.

To facilitate such data improvements, aggregators have begun introducing functionalities that connect users and providers (Suhrbier et al. 2017). The original architects of biodiversity informatics envisioned that users would notify providers of any issues with the data; providers would then evaluate that input and make changes to data records as they saw fit; and finally the modified records would be passed back to aggregators (and hence become available to users; Figure 4; Soberón et al. 1996). In addition to that original feedback 'loop' of user → provider → aggregator → user, we characterize recent modifications as a 'pendular' feedback pathway of user → aggregator → provider → aggregator → user (Figure 4). Just as users interact with aggregators to receive data from multiple providers, they can send information directly to the aggregator (regarding records corresponding to many providers). Simple implementations of such user feedback mechanisms already exist via open text boxes for commenting (including in GBIF) and should become much more structured (i.e., tied to particular fields). After inspection to remove spam, user feedback can lead to flagging and posting of suggested information and annotations in the database of the aggregator (visible to all users) and transmission of that information to the respective providers for consideration. As a complement to the original feedback loop, this pendular pathway maintains the primacy of decision-making by providers, enlists aggregators in facilitating information flow and availability, and allows users increased access to information regarding data quality and possible improvements.

Implementation and outlook

If implemented widely, these ongoing and envisioned changes could prove transformational, catalyzing increased utility of biodiversity data for myriad scientific uses and applications (Box 1). Importantly, they should promote positive feedback patterns, leading to ever better data and concomitant increases in utility and impact. Implementing these changes can happen via a combination of community consensus-building, coordinated efforts to justify and secure funding, and technical innovations. Because biodiversity informatics depends on diverse data providers, aggregators, and users, the solutions must be feasible for all of these groups. Some advances likely will be achieved by large aggregators and others by smaller ones, yielding pilot implementations subsequently taken up across the field (Canhos et al. 2015). We envision a set of initiatives: (1) to consolidate information regarding existing implementations (to determine what pilot examples exist for each challenge); and (2) to tackle necessary outstanding advances. In designing particular solutions, we suggest consultation with users regarding desired functionalities at the outset, and then again later to test and comment on prototypes. Below, we sketch a roadmap for implementation, organizing items by how quickly they might feasibly be implemented (6–24 months, vs. 2–4 years).

Short-term deliverables

Likely one of the first achievable advances, web interface development for well-funded aggregator portals can highlight the uncertainties and biases of existing data. This includes making the uncertainty of identification and georeferencing for each record obvious to users—including the lack of any such information in a data record. It also entails functionalities to specify relevant target groups (to characterize the results of past sampling), as well as extending functions for mapping and downloading such information. As an example, some innovative implementations for visualizing spatial and temporal biases exist, ripe for expansion (e.g., in Symbiota's "Spatial Module" described above).

Simultaneously, aggregators can summarize simple statistics of data quality—benchmarks to guide, justify, and assess future improvements. For example, GBIF calculates several summary metrics for given providers and higher-level taxonomic groups. In future developments, portals can implement temporal benchmarks for various taxa, geographic or political entities, and data providers. In addition to doing so for higher-level combinations of these categories, we suggest flexibility so that users can tailor reports to their needs. The existence of information regarding species-level taxonomic identification, georeferencing, and their uncertainties constitute the most important fields to be assessed. We anticipate that such information will prove highly useful in advocating for increased investment in data completeness and quality.

Additionally, implementation of stable unique identifiers that can be used universally across all of biodiversity science constitutes a short-term deliverable that will enable many other advancements (Box 1). This requires final consensus among multiple data providers, aggregators, and external databases (likely including a registry to guarantee uniqueness), followed by widespread execution (Guralnick et al. 2015). Once

 © the authors, CC-BY 4.0 license

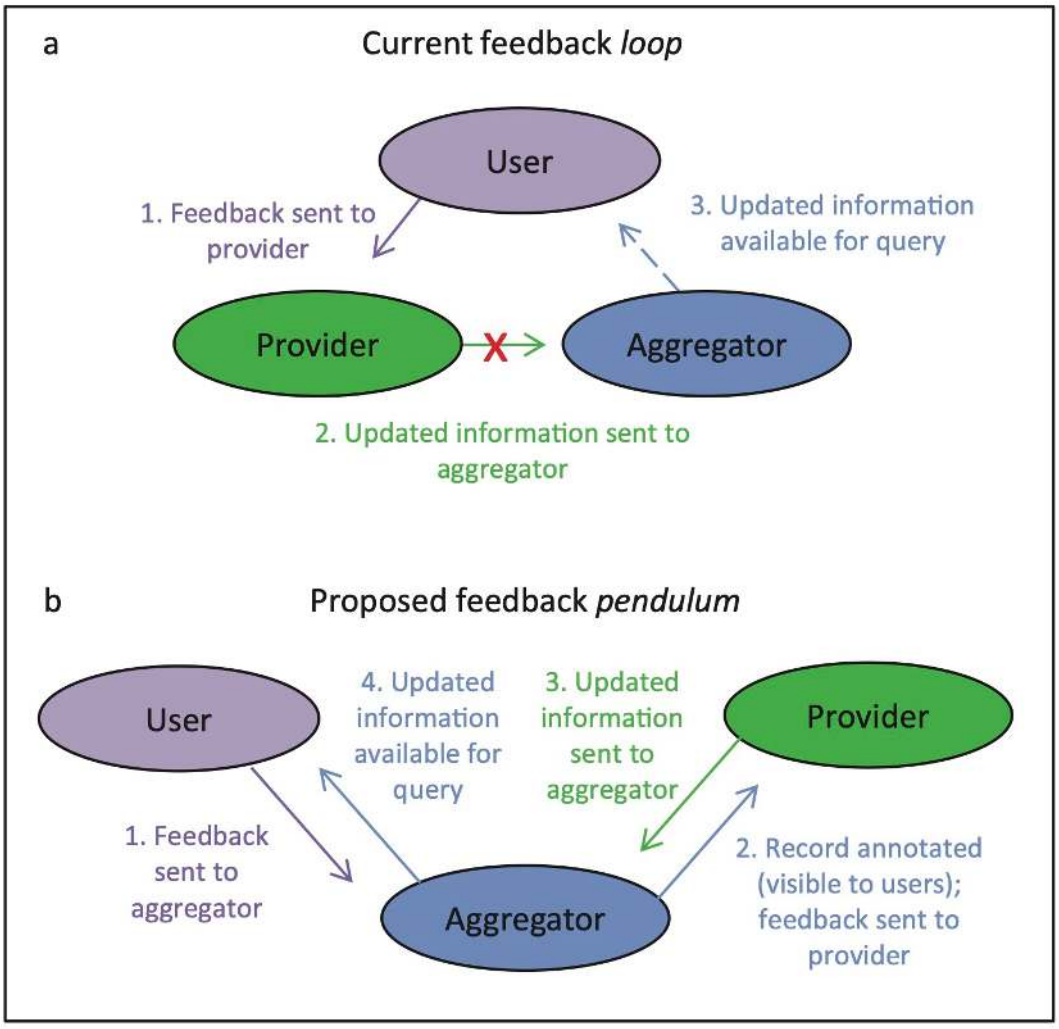

Figure 4. Graphical representation of original and recently modified pathways for feedback regarding Primary Biodiversity Data, showing information transfer among users, providers, and aggregators. Such feedback consists of suggested improvements or additions to data fields, for example a change in species identification or a newly determined georeference. The diagrams contrast two complementary mechanisms: (a) the original feedback *loop* (currently dominant); and (b) the emerging feedback *pendulum* (proposed for expansion). In a: (1) the user sends feedback to the provider (e.g., a given natural history museum); (2) if the provider makes a corresponding change to its database, the updated information is sent to the aggregator; and (3) that information becomes available for query by all users. In practice, because many providers do not consistently make such changes (denoted by an X), users do have access to updated information (dashed line). In b: (1) the user sends feedback to the aggregator; (2) the aggregator simultaneously both annotates the record (visible to all users) and sends the suggested information to the provider; (3) if the provider makes a corresponding change to its database, the updated information is sent to the aggregator; and (4) the aggregator makes the updated information available for query by all users. Note that even if a provider takes no action regarding the suggested information, the annotations placed by the aggregator are nevertheless available to users. Additionally, because the quantifications of data quality and use described in the text allow for benchmarks that can be tracked over time, we anticipate that the feedback pendulum will help providers become more successful in justifying and securing funding to make data improvements based on feedback from users.

achieved for individual-level identifiers, the same protocols can be modified to implement collective ones. Such advances should facilitate development of a community standard for expected digital deposition of Primary Biodiversity Data used in publications and reports (Table 1). In addition to allowing efficient documentation and quantification of data use, these functionalities will also prove essential for medium-term deliverables regarding feedback mechanisms and links to diverse external databases.

Medium-term deliverables

Although launching comprehensive mechanisms for user feedback may take a few years, efforts to determine desired functionalities and identify technical needs and solutions should begin now. First,

 © the authors, CC-BY 4.0 license

interested data providers, aggregators, and user groups can reach consensus regarding the data fields to be included, mechanisms for users to provide feedback to aggregators, and technical vision for how information will be transmitted to aggregators and then to and from providers. We anticipate that feedback from users will include at least: flags for likely taxonomic misidentifications, suggested corrected identifications, level of taxonomic expertise of the person identifying the species (for uncertainty of the identification field), flags for questionable georeferences, suggested new or improved georeferences, and estimated uncertainty of georeferences—as well as annotations regarding the data and resources used and an overall level of confidence regarding the quality of the record (Figure 4; van Hintum et al. 2011). Technical issues to be resolved by aggregators will include how to provide the flags and alternate information, automate the sending of such feedback to providers, and remove flags and alternate information if a provider makes the change. Likely, the solutions for many of these issues will leverage functionalities already implemented in GBIF for some simple standardized flags regarding data quality. Critically, the feedback system also will need to track the history of feedback for each record.

Such feedback machinery could include properties of existing open community platforms that have reputation rewards systems (for example, stackoverflow[6] for the coding community) or more generally online forums such as reddit[7]). For biodiversity informatics, individual 'actors' associated with data providers (e.g., museum curator/collection manager, field collector/observer) could have a login ID and receive a notification when a user provides feedback. Similarly, each user wishing to provide feedback could have a login ID; it also would be possible to implement a system in which such users develop 'reputations' based on community responses to their posts. Many complicated issues come with online forums, including the need to filter spam, and data providers and aggregators undoubtedly will consider user reliability carefully (see user registry protocols in Symbiota[5]).

Outlook

Given concerted engagement by the biodiversity informatics community, we think that many funding agencies, philanthropies, and other organizations supporting biodiversity research and conservation will embrace investments that lead to improved data quality and quantification. Specifically, we foresee successful proposals by groups of stakeholders to develop innovative plans regarding vision and mechanics (where necessary), as well as follow-up ones for implementation. In the immediate term, many entities regularly fund working groups and workshops (relevant for developing detailed plans for needed solutions), either via open calls for proposals or as supplements to current grants. For implementing solutions, various existing funding calls support biodiversity databasing and cyberinfrastructure; critically, we predict that funding agencies will also participate in this rethinking of biodiversity informatics by modifying and expanding their calls to reflect and promote the changing landscape of the field.

Once enabled by stable unique identifiers and valuable information regarding data quality and use, aggregators will be able to catalyze critical data improvements to a degree long envisioned but not yet possible (van Hintum et al. 2011). Aggregated databases will be highly useful for identifying bundles of Primary Biodiversity Data records particularly worthy of improvement, as well as for identifying gaps in data availability to be filled via targeted initiatives (Meyer et al. 2015, Lobo et al. 2018). Often taxonomic and/or geographic in nature, such characterizations can focus and justify efforts to improve the availability and quality of Primary Biodiversity Data (Stein and Wieczorek 2004, Sousa-Baena et al. 2014). Indeed, institutions and consortia of users with common interests and expertise will be particularly well poised to secure funds for collective data improvement initiatives (Anderson 2012, Tobón et al. 2017). For example, institutions and researchers interested in a particular applied topic (e.g., arthropod-borne zoonotic diseases in a given region) should be able to make strong justifications for the benefits of a cooperative project (Peterson 2015). We envision similar situations regarding conservation biology and many other practical applications of Primary Biodiversity Data. In closing, we hope that data providers, aggregators, users, and funding organizations will collaborate to build upon recent advances, leading to high-quality biodiversity data widely available for addressing issues of importance to science and society.

Author contributions

All authors debated the ideas included in the article, participated in organizing them, wrote sections of text, and revised various drafts. JMS headed production of the initial version, and RPA led subsequent reorganization and revision. JML and RPA produced the figures

Acknowledgments

This contribution derives from the *Task Group on GBIF Data Fitness for Use in Distribution Modelling* and the associated symposium and panel discussion *Frontiers of Biodiversity Informatics and Modeling Species Distributions*, which were possible through essential collaboration with Dmitry Schigel and Mary E. Blair. RPA acknowledges funding from the U.S. National Science Foundation (NSF; DBI-1661510) and National Aeronautics and Space Administration (80NSSC18K0406 to Mary E. Blair), and JMS recognizes support from NSF (DBI-1458640). Dmitry Schigel and Andrea Hahn (GBIF) and Benjamin Brandt (Symbiota) provided clear information regarding many current

6 https://stackoverflow.com/, last accessed on 28 May 2020

7 https://www.reddit.com/, last accessed on 28 May 2020

 © the authors, CC-BY 4.0 license

functionalities of those portals. Cecina Babich Morrow, Andrew Bentley, Mathieu Chevalier, Peter J. Galante, Valentina Grisales-Betancur, Lázaro Guevara, Bethany A. Johnson, Erica E. Johnson, Gonzalo E. Pinilla-Buitrago, Robert J. Whittaker, and four anonymous reviewers provided helpful feedback on these ideas and/or previous versions of the manuscript.

References

Amano, T., Lamming, J.D.L. & Sutherland, W.J. (2016) Spatial gaps in global biodiversity information and the role of citizen science. BioScience, 66, 393–400.

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.

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.

Araújo, M.B., Anderson, R.P., Barbosa, A.M., et al. (2019) Standards for models in biodiversity assessments. Science Advances, 5, eaat4858.

Araújo, M.B., Thuiller, W. & Yoccoz, N.G. (2009) Reopening the climate envelope reveals macroscale associations with climate in European birds. Proceedings of the National Academy of Sciences USA, 106, E45-46.

Barve, N., Martin, C., Brunsell, N.A. & Peterson, A.T. (2014) The role of physiological optima in shaping the geographic distribution of Spanish moss. Global Ecology and Biogeography, 23, 633–645.

[BCN] Biodiversity Collections Network (2018) Integration, attribution, and value in the web of natural history museum data: a needs assessment workshop. February 13–14, 2018. Lawrence, KS. Digital resource available at https://bcon.aibs.org/.

Beaman, R.S. & Cellinese, N. (2012) Mass digitization of scientific collections: new opportunities to transform the use of biological specimens and underwrite biodiversity science. ZooKeys, 209, 7–17.

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.

Bennett, J.M., Calosi, P., Clusella-Trullas, S., et al. (2018) GlobTherm, a global database on thermal tolerances for aquatic and terrestrial organisms. Scientific Data, 5, 180022.

Brook, B.W., Akçakaya, H.R., Keith, D.A., Mace, G.M., Pearson, R.G. & Araújo, M.B. (2009) Integrating bioclimate with population models to improve forecasts of species extinctions under climate change. Biology Letters, 5, 723–725.

Brooke, M.L. 2000. Why museums matter. Trends in Ecology and Evolution, 15, 136–137.

Campbell, H.A., Micheli-Campbell, M.A. & Udyawer, V. (2019) Early career researchers embrace data sharing. Trends in Ecology and Evolution, 34, 95–98.

Canhos, D.A.L., Sousa-Baena, M.S., de Souza, S., Maia, L.C., Stehmann, J.R., Canhos, V.P., De Giovanni, R., Bonacelli, M.B.M., Los, W. & Peterson, A.T. (2015) The importance of biodiversity E-infrastructures for megadiverse countries. PLoS Biology, 13, e1002204.

Chapman, A.D., Belbin, L., Zermoglio, P.F., et al. (2020) Developing standards for improved data quality and for selecting fit for use biodiversity data. Biodiversity Information Science and Standards, 4, e50889.

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

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. American Naturalist, 177, 738–751.

Cook, J.A., Greiman, S.E., Agosta, S.J., et al. (2016) Transformational principles for NEON sampling of mammalian parasites and pathogens: a response to Springer and colleagues. BioScience, 66, 917–919.

Costello, M.J., Michener, W.K., Gahegan, M., Zhang, Z.-Q. & Bourne P.E. (2013) Biodiversity data should be published, cited, and peer reviewed. Trends in Ecology and Evolution, 28, 454–461.

Costello, M.J. & Wieczorek, J. (2014) Best practice for biodiversity data management and publication. Biological Conservation, 173, 68–73.

D'Amen, M., Mod, H., Gotelli, N. & Guisan, A. (2018) Disentangling biotic interactions, environmental filters, and dispersal limitation as drivers of species co-occurrence. Ecography, 41, 1233–1244.

 © the authors, CC-BY 4.0 license

Ehrlén, J. & Morris, W.F. (2015) Predicting changes in the distribution and abundance of species under environmental change. Ecology Letters, 18, 303–314.

Exposito-Alonso, M., Vasseur, F., Ding, W., Wang, G., Burbano, H.A. & Weigel, D. (2018) Genomic basis and evolutionary potential for extreme drought adaptation in *Arabidopsis thaliana*. Nature Ecology & Evolution, 2, 352–358.

Feng, X., Park, D.S., Walker, S., Peterson, A.T., Merow, C. & Papeş, M. (2019) A checklist for maximizing reproducibility of ecological niche models. Nature Ecology & Evolution, 3, 1382–1395.

Fithian, W., Elith, J., Hastie, T. & Keith, D.A. (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution, 6, 424–438.

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.

Fordham, D.A., Akçakaya, H.R., Araújo, M.B., Keith, D.A. & Brook, B.W. (2013) Tools for integrating range change, extinction risk and climate change information into conservation management. Ecography, 36, 956–964.

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.

Franklin, J. (2010) Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge. 320 pp.

Franklin, J., Regan, H.M. & Syphard, A.D. (2014) Linking spatially explicit species distribution and population models to plan for the persistence of plant species under global change. Environmental Conservation, 41, 97–109.

García-Roselló, E., Guisande, C., Heine, J., Pelayo-Villamil, P., Manjarrés-Hernández, A., González Vilas, L., González-Dacosta, J., Vaamonde, A. & Granado-Lorencio, C. (2014) Using ModestR to download, import and clean species distribution records. Methods in Ecology and Evolution, 5, 708–713.

[GBIF] Global Biodiversity Information Facility (2016) Report of the Task Group on GBIF data fitness for use in distribution modelling [Anderson, R.P., Araújo, M., Guisan, A., Lobo, J.M., Martínez-Meyer, E., Peterson, A.T. & Soberón, J.]. Digital resource available at https://www.gbif.org/document/82612/report-of-the-task-group-on-gbif-data-fitness-for-use-in-distribution-modelling.

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 and Evolution, 19, 497–503.

Graves, G.R. (2000) Cost and benefits of web access to museum data. Trends in Ecology and Evolution, 15, 374.

Greenberg, J., White, H.C., Carrier, S. & Scherle, R. (2009) A metadata best practice for a scientific data repository. Journal of Library Metadata, 9, 194–212.

Gries, C., Gilbert, E.E. & Franz, N.M. (2014) Symbiota – a virtual platform for creating voucher-based biodiversity information communities. Biodiversity Data Journal, 2, e1114.

Gueta, T. & Carmel, Y. (2016) Quantifying the value of user-level data cleaning for big data: a case study using mammal distribution models. Ecological Informatics, 34, 139–145.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R. & Wintle, B.A. (2015) Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography, 24, 276–292.

Guisan, A., Thuiller, W. & Zimmermann, N.E. (2017) Habitat suitability and distribution models, with applications in R. Cambridge University Press, Cambridge UK. 462 pp.

Guisan, A., Tingley, R., Baumgartner, J.B., et al. (2013) Predicting species distributions for conservation decisions. Ecology Letters, 16, 1424–1435.

Guralnick, R.P., Cellinese, N., Deck, J., et al. (2015) Community next steps for making globally unique identifiers work for biocollections data. ZooKeys, 494, 133–154.

Guralnick, R.P., Hill, A.W. & Lane, M. (2007) Towards a collaborative, global infrastructure for biodiversity assessment. Ecology Letters, 10, 663–672.

Hallgren, W., Beaumont, L., Bowness, A., et al. (2016) The biodiversity and climate change virtual laboratory; where ecology meets big data. Environmental Modelling and Software, 76, 182–186.

 © the authors, CC-BY 4.0 license

Harris, S.E., Munshi-South, J., Obergfell, C. & O'Neill, R. (2013) Signatures of rapid evolution in urban and rural transcriptomes of white-footed mice (*Peromyscus leucopus*) in the New York Metropolitan area. PLoS ONE, 8, e74938.

Hill, A.W., Guralnick, R., Flemons, P., Beaman, R., Wieczorek, J., Ranipeta, A., Chavan, V. & Remsen., D. (2009) Location, location, location: utilizing pipelines and services to more effectively georeference the world's biodiversity data. BMC Bioinformatics, 10, S3.

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.

Howard, C., Stephens, P.A., Pearce-Higgins, J.W., Gregory, R.D. & Willis, S.G. (2014) Improving species distribution models: the value of data on abundance. Methods in Ecology and Evolution, 5, 506–513.

[IPBES] Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (2019) Global Assessment Report on Biodiversity and Ecosystem Services. IPBES Secretariat. Digital resource available at https://ipbes.net/global-assessment.

Johnson, E.E., Escobar, L.E. & Zambrana-Torrelio, C. (2019) An ecological framework for modeling the geography of disease transmission. Trends in Ecology and Evolution, 34, 655–668.

Jones, K.E., Bielby, J., Cardillo, M., et al. (2009) PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology, 90, 2648–2648.

Kass, J.M., Vilela, B., Aiello-Lammens, M.E., Muscarella, R., Merow, C. & Anderson, R.P. (2018) Wallace: a flexible platform for reproducible modeling of species niches and distributions built for community expansion. Methods in Ecology and Evolution, 9, 1151–1156.

Kattge, J., Díaz, S., Lavorel, S. et al. (2011) TRY – A global database of plant traits. Global Change Biology, 17, 2905–2935.

Kearney, M.R., Shamakhy, A., Tingley, R., Karoly, D.J., Hoffmann, A.A., Briggs, P.R. & Porter, W.P. (2014) Microclimate modelling at macro scales: a test of a general microclimate model integrated with gridded continental-scale soil and weather data. Methods in Ecology and Evolution, 5, 273–286.

Kissling, W.D., Dormann, C.F., Groeneveld, J. et al. (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. Journal of Biogeography, 39, 2163–2178.

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.

Lobo, J.M., Hortal, J., Yela, J.L., Millán, A., Sánchez-Fernández, D., García-Roselló, E., González-Dacosta, J., Heine, J., González-Vilas, L. & Guisande, C. (2018) KnowBR: An application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. Ecological Indicators, 91, 241–248.

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.

Merow, C., Latimer, A.M., Wilson, A.M., McMahon, S.M., Rebelo, A.G. & Silander, J.A. (2014) On using integral projection models to generate demographically driven predictions of species' distributions: development and validation using sparse data. Ecography, 37, 1167–1183.

Merow, C., Maitner, B.S., Owens, H.L., Kass, J.M., Enquist, B.J., Jetz, W. & Guralnick, R. (2019) Species' range model metadata standards: RMMS. Global Ecology and Biogeography, 28, 1912–1924.

Meyer, C., Kreft, H., Guralnick, R. & Jetz, W. (2015) Global priorities for an effective information basis of biodiversity distributions. Nature Communications, 6, 8221.

Molloy, J.C. (2011) The open knowledge foundation: open data means better science. PLoS Biology, 9, e1001195.

Morales-Castilla, I., Matias, M.G., Gravel, D. & Araújo, M.B. (2015) Inferring biotic interactions from proxies. Trends in Ecology and Evolution, 30, 347–356.

Nathan, R. & Muller-Landau, H. (2000) Spatial patterns of seed dispersal, their determinants and consequences for recruitment. Trends in Ecology and Evolution, 15, 278–285.

Nelson, G., Sweeney, P. & Gilbert, E. (2018) Use of globally unique identifiers (GUIDs) to link herbarium specimen records to physical

 © the authors, CC-BY 4.0 license

specimens. Applications in Plant Sciences, 6, e1027.

O'Neil, S.T., Dzurisin, J.D.K., Williams, C.M., et al. (2014) Gene expression in closely related species mirrors local adaptation: consequences for responses to a warming world. Molecular Ecology, 23, 2686–2698.

Otegui, J., Ariño, A.H., Chavan, V. & Gaiji, S. (2013) On the dates of GBIF mobilised primary biodiversity records. Biodiversity Informatics, 8, 173–184.

Ovaskainen, O., Smith, A.D., Osborne, J.L., Reynolds, D.R., Carreck, N.L., Martin, A.P., Niitepöld, K. & Hanski, I. (2008) Tracking butterfly movements with harmonic radar reveals an effect of population age on movement distance. Proceedings of the National Academy of Sciences USA, 105, 19090–19095.

Page, R.D.M. (2008) Biodiversity informatics: the challenge of linking data and the role of shared identifiers. Briefings in Bioinformatics, 9, 345–354.

Pelini, S.L., Dzurisin, J.D.K., Prior, K.M., Williams, C.M., Marsico, T.D., Sinclair, B.J. & Hellmann, J.J. (2009) Translocation experiments with butterflies reveal limits to enhancement of poleward populations under climate change. Proceedings of the National Academy of Sciences USA, 106, 11160–11165.

Pereira, H.M., Belnap, J., Brummitt, N., et al. (2010) Global biodiversity monitoring. Frontiers in Ecology and the Environment, 8, 459–460.

Peterson, A.T. (2015) Mapping disease transmission risk: enriching models using biogeography and ecology. Johns Hopkins University Press, Baltimore USA. 210 pp.

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

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.

Peterson, A.T., Soberón, J. & Krishtalka, L. (2015) A global perspective on decadal challenges and priorities in biodiversity informatics. BMC Ecology, 15, 15.

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. Monographs in Population Biology, 49. Princeton University Press, Princeton USA. 314 pp.

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.

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.

Ratnasingham, S. & Hebert, P.D. (2007) BOLD: The Barcode of Life Data System. Molecular Ecology Notes, 7, 355–364.

Reichmen, O.J., Jones, M.B. & Schildhauer, M.P. (2011) Challenges and opportunities of open data in ecology. Science, 331, 703–705.

Robertson, M.P., Visser, V. & Hui, C. (2016) Biogeo: an R package for assessing and improving data quality of occurrence record datasets. Ecography, 39, 394–401.

Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., Otegui, J., Russell, L. & Desmet, P. (2014) The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. PLoS ONE, 9, e102623.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G. & Chiarucci, A. (2011) Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. Progress in Physical Geography, 35, 211–226.

Ruete, A. (2015) Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. Biodiversity Data Journal, 3, e5361.

Salguero-Gómez, R., Jones, O.R., Archer, C.R., et al. (2015) The COMPADRE plant matrix database: an open online repository for plant demography. Journal of Ecology, 103, 202–218.

Salguero-Gómez, R., Jones, O.R., Archer, C.R., et al. (2016) COMADRE: a global data base of animal demography. Journal of Animal Ecology, 85, 371–384.

Santini, L., Isaac, N.J.B. & Ficetola, G.F. (2018) TetraDENSITY: a database of population density estimates in terrestrial vertebrates. Global Ecology and Biogeography, 27, 787–791.

Sarukhán, J., Urquiza-Haas, T., Koleff, P., Carabias, J., Dirzo, R., Ezcurra, E., Cerdeira-Estrada, S. &

 © the authors, CC-BY 4.0 license

Soberón, J. (2015) Strategic actions to value, conserve, and restore the natural capital of megadiversity countries: the case of Mexico. BioScience, 65, 164–173.

Smouse, P.E., Focardi, S., Moorcroft, P.R., Kie, J.G., Forester, J.D. & Morales, J.M. (2010) Stochastic modelling of animal movement. Philosophical Transactions of the Royal Society B, 365, 2201–2211.

Soberón, J., Jimenez, R., Golubov, J. & Koleff, P. (2007) Assessing completeness of biodiversity databases at different spatial scales. Ecography, 30, 152–160.

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

Soberón J. & Peterson A.T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society B, 359, 689–698.

Sousa-Baena, M.S., Garcia, L.C. & Peterson, A.T. (2014) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. Diversity and Distributions, 20, 369–381.

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.

Suhrbier, L., Kusber, W.-H., Tschöpe, O., Güntsch, A. & Berendsohn, W.G. (2017) AnnoSys—implementation of a generic annotation system for schema-based data using the example of biodiversity collection data. Database, 2017, bax018.

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, 278, 1823–1830.

Tobón, W., Urquiza-Haas, T., Koleff, P., Schröter, M., Ortega-Álvarez, R., Campo, J., Lindig-Cisneros, R., Sarukhán, J. & Bonn, A. (2017) Restoration planning to guide Aichi targets in a megadiverse country. Conservation Biology, 31, 1086–1097.

Troia, M.T. & McManamay, R.A. (2016) Filling in the GAPS: evaluating completeness and coverage of open-access biodiversity databases in the United States. Ecology and Evolution, 6, 4654–4669.

Troudet, J., Vignes-Lebbe, R., Grandcolas, P. & Legendre, F. (2018) The increasing disconnection of primary biodiversity data from specimens: how does it happen and how to handle it? Systematic Biology, 67, 1110–1119.

Valladares, F., Matesanz, S., Guilhaumon, F., et al. (2014) The effects of phenotypic plasticity and local adaptation on forecasts of species range shifts under climate change. Ecology Letters, 17, 1351–1364.

van Hintum, T., Menting, F. & van Strien, E. (2011) Quality indicators for passport data in *ex situ* genebanks. Plant Genetic Resources, 9, 478–485.

Veiga, A.K., Saraiva, A.M., Chapman, A.D., Morris, P.J., Gendreau, C., Schigel, D. & Robertson, T.J. (2017) A conceptual framework for quality assessment and management of biodiversity data. PLoS ONE, 12, e0178731.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T. & Vieglais, D. (2012) Darwin Core: an evolving community-developed biodiversity data standard. PLoS ONE, 7, e29715.

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.

Wilman, H., Belmaker, J., Simpson, J., de la Rosa, C., Rivadeneira, M.M. & Jetz, W. (2014) EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. Ecology, 95, 2027–2027.

Wisz, M.S., Pottier, J. & Kissling, W.D., et al. (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. Biological Reviews, 88, 15–30.

Submitted: 23 April 2020
First decision: 29 April 2020
Accepted: 24 May 2020

Edited by Robert J. Whittaker

 © the authors, CC-BY 4.0 license