Review

Title: A review of the heterogeneous landscape of biodiversity databases: opportunities and challenges for a synthesized biodiversity knowledge base

Authors: Xiao Feng[1*], Brian J. Enquist[2,3], Daniel S. Park[4], Brad Boyle[2], David D. Breshears[5], Rachael V. Gallagher[6], Aaron Lien[5], Erica Newman[2,7], Joseph R. Burger[8], Brian S. Maitner[9], Cory Merow[9], Yaoqi Li[10], Kimberly M. Huynh[7], Kacey Ernst[11], Elizabeth Baldwin[12], Wendy Foden[13,14.15], Lee Hannah[16], Peter M. Jørgensen[17], Nathan J. B. Kraft[18], Jon C. Lovett[19,20], Pablo A. Marquet[3,21,22], Brian J. McGill[23], Naia Morueta-Holme[24], Danilo M. Neves[25], Mauricio M. Núñez-Regueiro[26], Ary T. Oliveira-Filho[27], Robert K. Peet[28], Michiel Pillet[2], Patrick R. Roehrdanz[16], Brody Sandel[29], Josep M. Serra-Diaz[30,31], Irena Šímová[32,33], Jens-Christian Svenning[31,34], Cyrille Violle[35], Trang D. Weitemier[7], Susan Wiser[36], Laura López-Hoffman[5]

Affiliations:
[1]Department of Geography, Florida State University, Tallahassee, FL 32306, USA
[2]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA
[3]The Santa Fe Institute, USA, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA
[4]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA
[5]School of Natural Resources and the Environment, University of Arizona, Tucson, AZ 85721, USA
[6]Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109 Australia
[7]Arizona Institutes for Resilience, University of Arizona, Tucson, AZ 85721, USA
[8]Department of Biology, University of Kentucky, Lexington, KY, 40506 USA
[9]Eversource Energy Center and Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA
[10]Institute of Ecology and Key Laboratory for Earth Surface Processes of the Ministry of Education, College of Urban and Environmental Sciences, Peking University, Beijing 100871, China
[11]Department of Epidemiology and Biostatistics, University of Arizona, Tucson, AZ 85724, USA
[12]School of Government and Public Policy, University of Arizona, Tucson, AZ 85721, USA
[13]Cape Research Centre, South African National Parks, Tokai, 7947, Cape Town, South Africa
[14]Global Change Biology Group, Department of Botany & Zoology, University of Stellenbosch, Matieland, 7600, Stellenbosch, South Africa
[15]Climate Change Specialist Group, IUCN Species Survival Commission, Gland, 1196, Switzerland
[16]The Moore Center for Science, Conservation International, Arlington, VA 22202, USA
[17]Missouri Botanical Garden, St. Louis, MO 63110, USA
[18]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA
[19]School of Geography, University of Leeds, Leeds LS2 9JT, UK
[20]Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AB, UK
[21]Departamento de Ecología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, CP 8331150, Santiago, Chile

[22]Instituto de Ecología y Biodiversidad (IEB), Laboratorio Internacional en Cambio Global (LINCGlobal) & Centro de Cambio Global UC
[23]School of Biology and Ecology & Mitchell Center for Sustainability Solutions, University of Maine, Orono, ME 04473 USA
[24]Center for Macroecology, Evolution and Climate, GLOBE Institute, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark
[25]Department of Botany, Federal University of Minas Gerais, Belo Horizonte, MG 31270-901, Brazil
[26]Instituto de Bio y Geociencias del NOA (IBIGEO), Universidad Católica de Salta (UNSa), Consejo Nacional de; Investigaciones Cientificas y Tecnicas (CONICET), Avenida Bolivia 5150, Salta, Argentina
[27]Department of Botany, Federal University of Minas Gerais, Belo Horizonte, MG 31270-901, Brazil
[28]Department of Biology, CB#3280, University of North Carolina, Chapel Hill, NC 27599-3280, USA
[29]Department of Biology, Santa Clara University, 500 El Camino Real, Santa Clara CA 95053, USA
[30]Université de Lorraine, AgroParisTech, INRAE, Silva, 54000 Nancy, France
[31]Center for Biodiversity Dynamics in a Changing World (BIOCHANGE), Department of Biology, Aarhus University, Ny Munkegade 114, DK-8000 Aarhus C, Denmark
[32]Center for Theoretical Study, Charles University and The Czech Academy of Sciences, 110 00 Praha, Czech Republic
[33]Department of Ecology, Faculty of Science, Charles University, 128 44 Praha, Czech Republic
[34]Section for Ecoinformatics and Biodiversity, Department of Biology, Aarhus University, Ny Munkegade 114, DK-8000 Aarhus C, Denmark
[35]CEFE, Univ Montpellier, CNRS, Univ Paul Valéry Montpellier 3, EPHE, IRD, Montpellier, France
[36]Manaaki Whenua -- Landcare Research, Lincoln, New Zealand

[*]Corresponding author: Xiao Feng, fengxiao.sci@gmail.com

Abstract

Aim: Addressing global environmental challenges requires access to biodiversity data across wide spatial, temporal and biological scales. Recent decades have witnessed an exponential increase of biodiversity information aggregated by biodiversity databases (hereafter 'databases'). However, heterogeneous coverage, protocols, and standards of databases hampered the data integration among databases. To stimulate the next stage of data integration, here we present a synthesis of major databases, and investigate i) how the coverages of databases vary across taxonomy, space, and record type; ii) the degree of integration among databases; iii) how integration of databases can increase biodiversity knowledge; iv) the barriers to databases integration.

Location: Global

Time period: Contemporary

Major taxa studied: Plants and Vertebrates

Methods: We reviewed the scope of twelve well-established databases and assessed the status of their integration. We synthesized information from these databases to assess major knowledge gaps and barriers to fully integration. We estimated how improved integration can increase the coverage and depth of biodiversity knowledge.

Results: Each reviewed database had unique focus of data coverages. Data flows were common among databases, though not always clearly documented. Functional trait databases were more isolated than those pertaining to species distributions. Poor compatibility between taxonomic systems used by different databases posed a major challenge to integration. We demonstrated that integration of distribution databases can lead to greater taxonomic coverage that corresponds to 23 years' advancement in knowledge accumulation, and improvement in taxonomic coverage could be as high as 22.4% for trait databases.

Main conclusions: Rapid increase of biodiversity knowledge can be achieved through the integration of databases, providing the data necessary to address critical environmental challenges. Our synthesis provides an overview of the integration status of databases. Full integration across databases will require tackling the major impediments to data integration – taxonomic incompatibility, lags in data exchange, barriers to effective data synchronization, and isolation of individual initiatives.

Keywords: Big Data, Biodiversity Informatics, Biogeography, Database integration, Functional trait, Taxonomic System

1. Introduction

In the face of rapid global changes, a grand challenge is how to efficiently catalogue, assess, anticipate, and respond to changes in biodiversity and associated ecosystem services (Chapin *et al.*, 2000; Ceballos *et al.*, 2015; Díaz *et al.*, 2019). Addressing this challenge requires unprecedented access to biodiversity data across fine to broad spatial, temporal and biological scales (Beck *et al.*, 2012). The past few decades have witnessed fast growth of biodiversity information (Bisby, 2000; Hardisty *et al.*, 2013; Hobern *et al.*, 2019). Rapid digitization of existing biodiversity collections and ongoing collection of new information are expanding data availability worldwide (Sullivan *et al.*, 2014; Page *et al.*, 2015; Chandler *et al.*, 2017b). Indeed, the Global Biodiversity Information Facility (GBIF) – the world's leading repository of biodiversity observations – recently reached 1.6 billion records (accessed March 2021). However, we are still a long way from fully characterizing the taxonomy, geographic ranges and functions of all species on Earth (Lomolino, 2004; Hortal *et al.*, 2015; Stork, 2018). Addressing these shortfalls requires novel efforts in data synthesis to integrate the information held in the world's biodiversity projects, some 600+ of which had been created as of 2014 (Belbin, 2014) and nearly half of which are essentially invisible or inaccessible to the research community due to lack of cataloguing and integration (Blair *et al.*, 2020).

Data aggregation has been an ongoing goal of the biodiversity community (Nelson & Ellis, 2019), and a tremendous amount of work has been done by existing biodiversity data aggregators, such as GBIF, iDigBio, and VertNet. However, the challenges are many: existing biodiversity data aggregators often have singular objectives and consequently adhere to different protocols and standards (Mesibov, 2018) (termed "data domains" in (König *et al.*, 2019)), and datasets are highly heterogeneous spatially, temporally, and taxonomically (Reichman *et al.*, 2011; Cornwell *et al.*, 2019). The differences among biodiversity data aggregators can accumulate over time; thus, biodiversity data aggregators run the risk of "speciating," or becoming isolated, which can impede data sharing and integration. In response, the community has been calling for greater alignment between efforts and actively working on coordination mechanisms for developing shared roadmaps for biodiversity informatics (Hobern *et al.*, 2019). We therefore assert that a new synthesis is needed for the next stage of biodiversity data integration, i.e., information from existing biodiversity data aggregators should be further integrated to reduce shortfalls in biodiversity knowledge and achieve a more complete picture of Earth's biodiversity (Hobern *et al.*, 2019; König *et al.*, 2019; Kattge *et al.*, 2020).

To facilitate better integration among biodiversity data domains, we first need to assess the current state of connectivity and integration among databases. Though biodiversity data generally are well organized in individual databases, overlaps in their data coverage and the extent of communication between databases remains unclear. Indeed, attention has rarely been paid to the post-aggregation processes and interactions among commonly used databases (such as nontransparent data flows between two databases) and synthesis studies of biodiversity data from multiple databases are still scarce in the literature (Cornwell *et al.*, 2019; König *et al.*, 2019). To address this gap, we conducted a synthesis of existing biodiversity databases, and aimed to answer four questions: **(i)** How does the coverage of a suite of major biodiversity databases differ across taxon, space, and record type? **(ii)** How are existing biodiversity databases integrated? **(iii)** How would the integration of databases increase biodiversity knowledge? and **(iv)** What are the barriers that prevent data integration? To answer these questions, we first reviewed the scope of existing major biodiversity databases and assessed the status of their integration. We also demonstrated that the integration of biodiversity databases could rapidly narrow major knowledge gaps. Finally, we discussed barriers that need to be overcome to obtain a more complete picture of the biodiversity on Earth.

2. Review of biodiversity databases

Many biodiversity databases have been built over the past two decades, with varying emphases on taxonomy, spatial location, and record type. To synthesize the major attributes of existing biodiversity databases, we selected twelve well-established biodiversity databases: Atlas of Living Australia (ALA; Belbin & Williams, 2016), Botanical Information and Ecology Network (BIEN; Enquist *et al.*, 2016), Biodiversity Information Serving Our Nation (BISON; U.S. Geological Survey, 2018), eBird (Sullivan *et al.*, 2014), Encyclopedia of Life (EOL; Parr *et al.*, 2014), Global Biodiversity Information Facility (GBIF), Global Inventory of Floras and Traits (GIFT; Weigelt *et al.*, 2017), Integrated Digitized Biocollections (iDigBio, 2018a), iNaturalist (iNaturalist), Map of Life (MOL; Jetz *et al.*, 2012), a global database of plant traits (TRY; Kattge *et al.*, 2011), and VertNet (Constable *et al.*, 2010). Our selection can not cover every notable database because of limited effort and the accessibility of database content or documentations, though they were chosen to represent the breadth of the most commonly used, well-established large-scale biodiversity databases (MacFadden & Guralnick, 2016; Chandler *et al.*, 2017a; James *et al.*, 2018; Singer *et al.*, 2018; Cornwell *et al.*, 2019; König *et al.*, 2019) to maximize the generalizability of our results and conclusions. We acknowledge that these databases are typically under active development; thus our synthesis is based on a snapshot of their status on the access date (March 2021; see Appendix 1).

2.1 Varied focuses among biodiversity databases

We reviewed associated metadata for biodiversity databases from project websites or publications. We recorded database name, taxonomic scope, taxonomic system, record type, number of records, and spatial coverage. We classified the record types into three categories: geographic distribution, media type, and biological information (standardized trait databases or generalized text descriptions). Within geographic distribution, we further classified the information as specimen records, observations, checklists of geographic regions, or distribution maps. Specimen records and observations both have information on specific occurrences of a species at a georeferenced point location, but only specimen records are associated with physical specimens. Checklists usually contain lists of species known to be present in defined geographic regions (e.g., political divisions or protected areas). Distribution maps are those that were drawn by experts or generated through models with various degrees of complexity. Media data type were classified as image, audio, and video. Biological information included standardized trait and generalized text descriptions.

Our review showed that each of these biodiversity databases holds unique scientific value because they cover different spatial extents, taxonomic groups, and record types (Fig. 1a). The databases could be grouped into different clusters based on similarities of focus and data coverage. For example, EOL, iNaturalist, and eBird form a cluster of databases that indexes media data and biological descriptions, while also sharing public education objective (Fig. 1b). TRY and GIFT form another cluster that mainly focuses on indexing functional traits of plants. GBIF, BISON, iDigBio, and VertNet form yet another cluster that emphasizes indexing species occurrences. The cluster of ALA, MOL, and BIEN share the property of indexing both species occurrences and geographic range maps. Here our grouping of databases considered the different attributes equally, though assigning different weights on the attributes can lead to different grouping outcomes. For example, many of the databases seek to document all taxa across the globe (e.g., GBIF, EOL, eBird) or to index many types of data (e.g., EOL, ALA, iNaturalist).

2.2 Data integration status among biodiversity databases

To understand how existing biodiversity databases are integrated, we reviewed the data flow among the databases. Biodiversity databases (e.g., GBIF) are typically data aggregators of digitalized information from data providers, such as museums, herbariums, and research data repositories, and detailed information about data providers are usually acknowledged on a databases' website (e.g., BIEN data contributors-https://web.archive.org/web/20210511034441/https://bien.nceas.ucsb.edu/bien/data-contributors/). However, it is usually not straight forward to understand whether one database is aggregated by another database, probably because of the concern of losing uniqueness of data coverage, i.e. acknowledging to be aggregated by another aggregator can be interpreted as one database becoming a subset of the other database. Regardless, understanding such relationships among databases is important for users, as this immediately affects the determination of most comprehensive data coverage (e.g., whether or not GBIF has the most complete occurrence set of a species) or evaluation of data quality (e.g., whether or not to consider duplicated records when using multiple databases). Therefore, we assessed data integration among biodiversity databases based on their documentation and publications.

Overall, the data flows between biodiversity databases are not always clearly documented and at times the relationships need to be inferred. Key technical details of data flow, such as time and frequency of data exchange/flow, and the version or date of the imported data, are usually lacking. The lack of 'snapshot' data archives hinders the reproduction of data content, as well as the reproducibility of associated scientific research (Feng *et al.*, 2019). Unclear documentation of data exchange may also lead to compliance issues with data licensing, and can prevent assignment of proper credit to data collectors.

We found that data flow, unidirectional or bidirectional, is common among biodiversity databases (Fig. 2 & Table S1). Among the network of databases, GBIF serves as a central aggregator at a global scale that ingests species occurrence data from many databases, such as BISON, iDigBio, and eBird. ALA and BISON have bidirectional data flows with GBIF – they both i) aggregate biodiversity data collected from their focal regions (i.e., Australia and North America respectively) and pass the data to GBIF, and ii) import other data collected from Australia or North America from GBIF to their respective databases (Table S1). There are also cases of unidirectional data flow from GBIF to specialized databases. For example, MOL aggregates multiple types of information of species geographic distributions, including occurrences from GBIF; as does BIEN.

We summarized the status of data integration across databases into four categories: synced, lagged, impeded, and isolated (Fig. 3). Ideally, information in databases could be fully integrated in either one or multiple directions in real (or near-real) time (i.e., *synced*). For example, data published to iDigBio is automatically published to GBIF (iDigBio, 2018b; Singer *et al.*, 2018), thus the content of iDigBio is considered synced with GBIF (Fig. 3). However, differences may arise between otherwise fully integrated databases in the time between synchronization events (*lagged*). For example, BIEN imports and integrates data from GBIF and other sources at annual or longer intervals, which provides more stable and easily archived datasets, but the imported GBIF content can be different from the most up-to-date GBIF data until the next synchronization. This lag can be addressed by increasing the frequency of data exchange, shared data import protocols, or developing novel database architecture designed for data integration (LeBauer *et al.*, 2013). Differences between databases may also arise from obstacles that prevent subsets of data from being shared (*impeded*). For example, iNaturalist only publishes data to GBIF that are properly licensed (iNaturalist, 2018)). Differences in data licensing is one of the major impediments to integration and is a problem that was rarely emphasized in biodiversity data aggregation prior to the last decade. For example, GBIF initialized a license requirement in 2014 (GBIF, 2014) and excluded approximately 49 million existing records without appropriate licenses. Clearly defined data licenses will make future data use and integration legally straightforward, and will also provide a cornerstone for the Open Science movement (Escribano *et al.*, 2018). Creative commons licenses are the most widely used mechanism to ensure proper attribution while allowing others to copy and distribute data (Fitzgerald *et al.*, 2007).

Unlike the distribution databases discussed above, trait databases are characterized by isolation status. These databases typically capture data within particular taxa or focus on a single trait, such as GlobTherm for thermal tolerance (Bennett *et al.*, 2018) and AmphiBIO for amphibian ecological traits (Oliveira *et al.*, 2017) (Fig. 3). A degree of isolation is unavoidable due to the complex nature of trait data, which varies greatly in terms of data types, units, and measurement methods (Deans *et al.*, 2015) and the taxon-specific nature of many traits (e.g., seed traits apply only to seed plants). Such complexity is not resolved by following existing standard commonly used by occurrence data such as Darwin Core (Wieczorek *et al.*, 2012). Effective synthesis and integration of trait information will require trait-specific specifications such as trait ontologies (Walls *et al.*, 2012), trait data standards (Schneider *et al.*, 2019) and embracing of Open Science principles via initiatives like the Open Traits Network (Gallagher *et al.*, 2020).

Poor compatibility between taxonomic systems adopted by different databases has posed a major impediment for database integration (Fig. 2 & Table S2). As biodiversity information is generally indexed by species' scientific names, a crucial step is to index information based on one unified or multiple compatible taxonomic systems. Taxonomic systems reflect decisions of database developers; some databases maintain flexibility in nomenclature, especially when the taxa are in flux (e.g., vertebrate species stored in VertNet), whereas some databases impose stronger rules. For example, EOL maintains multiple independent taxonomic systems to avoid potential conflicts between non-compatible nomenclature; GBIF and COL have both employed a comprehensive but single-backbone system designed to be compatible with different taxonomic systems; MOL developed a backbone that includes Catalogue of Life (a global effort to compile existing catalogued species) and manually curated taxonomic datasets for synonym issues; BIEN standardizes taxon names according to external, expert-curated taxonomic reference databases (Boyle *et al.*, 2013). The different approaches and strategies to accommodating taxonomic systems among biodiversity databases may solve taxonomic issues locally for that specific database (Jorge & Peterson, 2004), but deepen differences that prevent future data integration, thus facilitating the "speciation" of databases. Still, resolving differences between existing taxonomic systems is just an initial step. Creation of a single authoritative list of names will take time; full reconciliation of synonyms and distinct taxon concepts may take decades (Berendsohn, 1997; Franz & Peet, 2009; Boyle *et al.*, 2013; Wiser, 2016; Garnett *et al.*, 2020). This will require a global effort, as envisioned by the Global Taxonomy Initiative (Samper, 2004).

3. Enhanced data coverage via database integration

To quantify the improvement of combining multiple databases, we compared leading databases that focus on similar taxonomic groups and similar record types. We used terrestrial plants (Embryophyta; hereafter "plants") and vertebrates (Vertebrata) as test cases, because these taxonomic groups are comparatively well collected and documented in biodiversity databases compared to others (Clark & May, 2002; Fazey *et al.*, 2005; Hecnar, 2009; Titley *et al.*, 2017; Cornwell *et al.*, 2019; König *et al.*, 2019; Kattge *et al.*, 2020). We did not use taxon, such as microbes or invertebrates, that account for large portions of biodiversity on Earth but face huge data gaps (Locey & Lennon, 2016). Specifically, we combined (i) the distribution of terrestrial plants from GBIF and non-GBIF sources, and (ii) one crucial and commonly measured trait for plants and vertebrates, respectively: maximum height (Moles *et al.*, 2009; Guralnick *et al.*, 2016) using the Botanical Information and Ecology Network (BIEN (Enquist *et al.*, 2016)), TRY initiative (Kattge *et al.*, 2011), and EOL (Parr *et al.*, 2014), and body length using VertNet (Constable *et al.*, 2010) and EOL (see Appendix 1). Our study goes beyond recent gap analyses of biodiversity data (Meyer *et al.*, 2016; Cornwell *et al.*, 2019; König *et al.*, 2019), by expanding the scope to multiple data aggregators with similar missions, in two major clades (i.e., plants and vertebrates), and using an ecological trait characterized by continuous values.

3.1 Better coverage through data integration

3.1.1 Overall trend in data collection

We found that the total number of distribution records (spatial coordinates) for plants has increased exponentially since the 1750s (Lomolino *et al.*, 2010) (Fig. 4a) as documented in GBIF and the combined dataset. A similar exponential increase was found when only spatially unique records were examined (Fig. 4b). This pattern is also supported by a model selection analysis among linear, exponential, and logistic functions (Table S3). This trend in the growth of biodiversity data is analogous to many accelerating processes in the Anthropocene (Steffen *et al.*, 2015), such as urbanization, globalization, transportation, and telecommunications. One prominent example in Information Technology (IT) is the exponential growth in the number of transistors in a dense integrated circuit, which doubles roughly every two years (Moore, 1965). This pattern, termed "Moore's Law", is also evident in the accelerating development of cyber infrastructures for many disciplines in science. Based on the similar exponential curve for biodiversity data, we estimated that the total number of plant distribution records doubles every 17 years and the number of spatially unique records doubles every 21 years. The high speed of biodiversity data accumulation represents the great power of data collection, digitization, processing, and publishing, which lays the basis for and presents the opportunities for biodiversity database integration.

In contrast to the number of distribution records, the number of species identified is gradually reaching saturation (Fig. 4c). Based on a fitted logistic curve (Table S3), we predicted that the number of catalogued plant species in distribution databases would be saturated at 365,519 ±

2,233 (mean ± SD of the coefficient from the fitted logistic model), i.e. the saturation point of predicted number of terrestrial plant species in the integrated biodiversity distribution databases, with species names resolved using the Taxonomic Name Resolution Service (TNRS; version 5.0) (Boyle *et al.*, 2013). This estimate is higher than the current catalogued number of terrestrial plants in Catalogue of Life (COL; 354,327), though within the previously estimated range for the total number of plant species on Earth (334,000 - 403,911) (Lughadha *et al.*, 2016). The slowing trend in plant species discovery started in ~1949 (the inflection point of the logistic curve of the cumulative number of species in GBIF; Table S1), and is in line with previous estimations (Christenhusz & Byng, 2016). Such trends may suggest that we are gradually reaching saturation and closing the *Linnean shortfall*, the lack of knowledge in describing and cataloging species (Hortal et al., 2015), for plants. The slowing trend could also be caused by species extinctions, reduced funding for natural history studies, and increasing difficulties in detecting the remaining rare species (Joppa *et al.*, 2011).

3.1.2 Improvement in distribution data

Integration of biodiversity databases would powerfully increase our knowledge of biodiversity. For instance, GBIF is the world’s largest biodiversity repository, but adding ~15 million records from additional sources (compiled by BIEN) would improve its coverage by ~3.7 million spatially unique records and ~20 thousand species (Fig. 4d-f). The number of distribution records per taxon in GBIF could be increased by 4.4% – an average of 19 additional records per species. The improvement of taxonomic coverage in GBIF would be equivalent to 23 years of new data accumulation, based on extrapolation of the fitted logistic curve (Fig. 4c, Table S3). GBIF and non-GBIF datasets together provide distribution data for ~ 307,985 species (76-92% of the estimated richness of all plants (Lughadha *et al.*, 2016)), suggesting we are gradually decreasing the *Wallacean shortfall*, the lack of knowledge in species distribution, for plant species, in accordance with findings in Cornwell *et al.* (2019).

3.1.3 Improvement in trait data

Database integration also substantially improves the taxonomic coverage of trait information (i.e., maximum height in plants; body length in vertebrates; see Methods). Under standardized taxonomy, we found that individual plant and vertebrate trait databases always include unique species-trait combinations and cover different portions of taxonomic diversity (Fig. 5). For instance, trait knowledge increased in 69-82 plant orders and 86-124 vertebrate orders through database integration, while the range of increase varied by database. The average improvement of species-trait combination across these databases ranged from 2.0 to 8.7% for plant orders and 21.5-22.4% for vertebrate orders. The number of plant orders that were sparsely-sampled in BIEN (i.e., <10% of species with trait observations), for example, decreased from 99 to 65 through data integration; a similar decrease was seen for sparsely-sampled vertebrate orders in EOL from 53 down to nine (Fig. 5).

3.1.4 Limitations of our assessment

Data integration can effectively decrease the gaps in our knowledge, and the resulting more comprehensive data can facilitate global scale studies of biodiversity and help identify and reduce potential data biases (Reddy & Dávalos, 2003). We note that our assessment of the possibilities for data integration does not address how different data sources (or “data resolutions,” as defined in (König *et al.*, 2019)) should be best integrated for different study objectives. These mismatches are apparent in cases, such as distribution data represented by presences vs. abundances, or a trait value measured at individual level vs. species level. However, indexing the availability of trait data for a focal species is a major step toward more rigorous data integration and scientific research. With the integrated data, one could cross-validate the values from different sources to ask questions such as: "Do trait values vary by methods of measurements?" or "Can species-level trait data well represent the range of values measured at the individual level?" Cross-validations will be especially useful if the user of one database is mainly the general public while the user of the other is the science community, so that more rigorous information is delivered from the science community to the general public. With the integrated data, one could also conduct scientific research at broader scales and study, for example, trait variation across time or across spatial or environmental gradients (Siefert *et al.*, 2015), or species-trait combinations within communities.

3.2 A clearer picture of what we do not know

Importantly, database integration can provide an improved assessment of gaps in biodiversity knowledge (Meyer *et al.*, 2015; Cornwell *et al.*, 2019; König *et al.*, 2019). Following our integration of various databases (Appendix 1), approximately 58,000 plant species still lacked publicly available distribution records. This gap corresponds to approximately 15.8% of the species in Catalogue of Life – a global effort to compile existing catalogued species. The coverage of distribution records in plant orders varied from 47% (in order Hypnales) to fully covered in some orders with small number of extant species (Cornwell *et al.*, 2019) (e.g. Ceratophyllales). Further, 30.8 million km^2 of ice-free land surface, as assessed using Eckert IV equal area projection, currently has no valid plant geolocations (Fig. 4g). These areas are mainly in Russia (despite the considerable recent progress of data sharing by the Russian GBIF community (Shashkov & Ivanova, 2019)), central Asia, and northern Africa, and are approximately 13% of the Earth's land area.

Trait data have considerably larger gaps: height information is absent for 333,597 plant species from 102 orders from BIEN, TRY and EOL, and body length information is absent for 38,992 vertebrate species from 127 orders from VertNet and EOL. In total, height data is unavailable for approximately 92.6% of plant species and body length for 56.8% of vertebrate species in Catalogue of Life. The data coverages were mostly below 60% for plant orders and percentages were relatively higher for vertebrate orders. Plant height and vertebrate body length are commonly used traits in ecological research that are frequently recorded in databases (Moles *et al.*, 2009; Guralnick *et al.*, 2016), suggesting other biological traits (e.g., life span, metabolic rate) or essential biodiversity variables (e.g., population abundances) (Pereira *et al.*, 2013) will likely have much larger *shortfalls* (but see analyses of plant growth form in (König *et al.*, 2019)). In the face of accelerating increases in biodiversity data availability, recognizing the remaining knowledge gaps could help guide future data compilation efforts (e.g. the gap filling activity in eBird (eBird, 2014)) and potentially turn our enhanced power of compiling information into efforts that generate critically needed knowledge (Cornwell *et al.*, 2019).

4. Challenges and Opportunities

4.1 A catalogue and synthesis of biodiversity databases

To achieve global integration of biodiversity knowledge, we would first need to know what databases are available. To facilitate this process, we need a catalogue of biodiversity databases with their metadata recorded, such as spatial, temporal, taxonomic scope, as well as the types of data aggregated, so that existing or new databases can be easily known, compared, and effectively used. Lee Belbin has maintained the Biodiversity Information Projects of the World (Belbin, 2014) – essentially containing metadata of 685 biodiversity projects. The recorded metadata includes project summary, geographic, temporal, and taxonomic scope, and key technique attributes (though this list is no longer accessible after 2019; but see (Blair *et al.*, 2020)). Similarly, GBIF has a registry system that indexes the metadata of GBIF participants, institutions, and datasets; however, data associated with this registry mainly focuses on a few record types, including occurrences, checklists, and sampling events (https://web.archive.org/web/20210514141441/https://www.gbif.org/article/5FlXBKbirSiq0ascKYiA8q/gbif-infrastructure-registry). Another example is Global Index of Vegetation Plot Databases that indexes the metadata of vegetation-plot data that are publicly available (Dengler *et al.*, 2011). In contrast, DataONE has a broader scope that indexes the metadata of large variety of biological and environmental data (Michener *et al.*, 2012). Those existing efforts form a good basis for a catalogue of biodiversity databases that can continuously keep track of existing data aggregators and index new aggregation efforts. Still, the relationships among the biodiversity databases are not always obvious. Therefore, a synthesis, ideally updated regularly, would be helpful to clarify the relationships among the biodiversity databases, in particular what is the unique data coverage of one database and what are the data flows among biodiversity databases.

4.2 Overcoming the barriers to database integration

After cataloguing the metadata and synthesizing the relationships among biodiversity databases, many technical barriers remain. As a prerequisite to integration, the data in a database should be openly available with proper data licenses to minimize impediments to data sharing (see section 2.2); another major barrier is the incompatible taxonomic systems. A promising effort is Catalogue of Life Plus (Banki *et al.*, 2019) that builds upon existing but disconnected efforts (such as the COL and GBIF backbone taxonomy) to create an open, shared and sustainable consensus taxonomy, which can serve as the infrastructure for individual biodiversity databases or database integration. Thirdly, existing databases adopt different mechanisms of data standards and database architecture (Hardisty *et al.*, 2019), thus leading to incompatibilities for database integration. For example, during the data cleaning stage, one collection of a specimen without coordinates could be georeferenced differently based on different georeferencing algorithms, thus likely leading to two different coordinates, and therefore appear to be two different records after data integration. One solution could be creating a community-wide standard and tools for data evaluation and cleaning (e.g. Belbin *et al.*, 2018; Serra-Diaz *et al.*, 2018). Community-driven standards for biodiversity data, such as Darwin Core (Wieczorek *et al.*, 2012), Humboldt Core (Guralnick *et al.*, 2018), and trait-data standard (Schneider *et al.*, 2019) have emerged; expanding the use of those community-developed data standards by individual databases would enable more effective database integration. Overall, the essential goal is to maximize compatibility, and thus minimize barriers to data flow and synthesis. After solving the technical barriers, the integrated content from multiple databases could be organized in multiple non-exclusive ways: i) a single centralized database, ii) some decentralized but connected databases (Gallagher *et al.*, 2020), or iii) multiple synced databases (LeBauer *et al.*, 2013).

4.3 Challenges for individual aggregators after database integration

It is also worth thinking the uniqueness and destiny of individual databases after integration. Seemingly, integration may render individual databases irrelevant, e.g., an individual database may be considered a subset of an integrated database. However, this should not the case. While data integration occurs at shared data element (e.g., taxon, place, time) and data standard, each individual database could still have unique domain information. For example, while GBIF aggregates species occurrence data from iNaturalist, the latter still uniquely host the media data. Also, an individual database can make a unique contribution by aiming to fill data gaps (e.g., spatial or taxon gaps revealed by the integrated knowledge base).

On the other side, there has been a process of specialization of databases along the whole workflow of data aggregation. Specifically, the developers of some databases have expanded their scope to development of infrastructure, such as tools for data integration, data cleaning, and hosting data portals. There are prominent examples among the databases that have close relationships with GBIF. For example, ALA develops open-access modules for the platform that can be implemented by other biodiversity initiatives (Belbin *et al.*, 2021). VertNet has been actively providing data maintenance services, including data cleaning and indexing, among the network of collaborative biodiversity databases (Constable *et al.*, 2010).

Besides specialized roles in data aggregation or tool development, individual databases can also play unique roles for users, even when based on the same shared knowledge base. For example, ALA is prominent in the education of Australian biodiversity to its Australian users, as well as in facilitating scientific research by putting this biodiversity in the context of its environment.

5. Concluding remarks

The accelerating increase of biodiversity data offers numerous exciting prospects and challenges for documenting and forecasting the location, status, function and potential fate of species on the planet. However, increases in biodiversity data do not directly translate to similar increases in the knowledge needed to address many fundamental and applied questions. In the face of urgent environmental challenges, new approaches are urgently needed to increase biodiversity knowledge and accessibility of the knowledge. We demonstrate that rapid progress can be made toward better biodiversity knowledge through the integration of database infrastructures. Integration can lead to large and rapid increases in knowledge of species distributions and traits (see (Conde *et al.*, 2019; König *et al.*, 2019)), but the benefit goes beyond just more complete knowledge: it can reduce biases and doubled efforts in biodiversity research, allow cross-validations to compare conclusions drawn from different sources, and provide a clearer picture of where gaps remain, thereby helping to focus future sampling and research (König *et al.*, 2019). To address the shortfalls in biodiversity knowledge and achieve full integration across databases, we need to fund and maintain the foundations of biodiversity information science including biological surveys, taxonomic assessment (Australian Academy of Science, 2018), and digitization of legacy data (Ariño, 2010), as well as tackle the major impediments to data integration – taxonomic incompatibility, lags in data exchange, barriers to effective synthesis, and isolation of individual initiatives.

References

Ari ño, A.H. (2010) Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics*, **7**, 81-92.

Australian Academy of Science (2018) Discovering Biodiversity: a decadal plan for taxonomy and biosystematics in Australia and New Zealand 2018–2027. In:

Banki, O., Hobern, D., D öring, M. & Remsen, D. (2019) Catalogue of Life Plus: A collaborative project to complete the checklist of the world's species. *Biodiversity Information Science and Standards*, **3**, e37652.

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.

Belbin, L. (2014) *Biodiversity Information Projects of the World*. Retrieved from: https://web.archive.org/web/20180609082447/http://www.tdwg.org/biodiv-projects/ (accessed 1 May 2018).

Belbin, L. & Williams, K.J. (2016) Towards a national bio-environmental data facility: experiences from the Atlas of Living Australia. *International Journal of Geographical Information Science*, **30**, 108-125.

Belbin, L., Wallis, E., Hobern, D. & Zerger, A. (2021) The Atlas of Living Australia: History, current state and future directions. *Biodiversity data journal*, **9**, e65023-e65023.

Belbin, L., Chapman, A., Wieczorek, J., Zermoglio, P., Thompson, A. & Morris, P. (2018) Data Quality Task Group 2: Tests and Assertions. *Biodiversity Information Science and Standards*, **2**, e25608.

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**, 180022-180022.

Berendsohn, W.G. (1997) A Taxonomic Information Model for Botanical Databases: The IOPI Model. *Taxon*, **46**, 283-309.

Bisby, F.A. (2000) The Quiet Revolution: Biodiversity Informatics and the Internet. *Science*, **289**, 2309.

Blair, J., Gwiazdowski, R., Borrelli, A., Hotchkiss, M., Park, C., Perrett, G. & Hanner, R. (2020) Towards a catalogue of biodiversity databases: An ontological case study. *Biodiversity Data Journal*, **8**, e32765.

Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J.A., Mozzherin, D., Rees, T., Matasci, N., Narro, M.L., Piel, W.H., McKay, S.J., Lowry, S., Freeland, C., Peet, R.K. & Enquist, B.J. (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics*, **14**, 16.

Catalogue of Life (2021) *Species 2000 & ITIS Catalogue of Life, 2021-04-05.Digital resource at www.catalogueoflife.org. Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-8858.*

Ceballos, G., Ehrlich, P.R., Barnosky, A.D., Garcia, A., Pringle, R.M. & Palmer, T.M. (2015) Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances*, **1**, e1400253.

Chamberlain, S.A. & Szocs, E. (2013) taxize: taxonomic search and retrieval in R. *F1000Res*, **2**, 191.

Chandler, M., See, L., Copas, K., Bonde, A.M.Z., L ópez, B.C., Danielsen, F., Legind, J.K., Masinde, S., Miller-Rushing, A.J., Newman, G., Rosemartin, A. & Turak, E. (2017a)

Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, **213**, 280-294.

Chandler, M., See, L., Copas, K., Bonde, A.M.Z., Lopez, B.C., Danielsen, F., Legind, J.K., Masinde, S., Miller-Rushing, A.J., Newman, G., Rosemartin, A. & Turak, E. (2017b) Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, **213**, 280-294.

Chapin, F.S., 3rd, Zavaleta, E.S., Eviner, V.T., Naylor, R.L., Vitousek, P.M., Reynolds, H.L., Hooper, D.U., Lavorel, S., Sala, O.E., Hobbie, S.E., Mack, M.C. & Díaz, S. (2000) Consequences of changing biodiversity. *Nature*, **405**, 234-242.

Christenhusz, M.J.M. & Byng, J.W. (2016) The number of known plants species in the world and its annual increase. *Phytotaxa*, **261**, 201-217.

Clark, J.A. & May, R.M. (2002) Taxonomic Bias in Conservation Research. *Science*, **297**, 191.

Conde, D.A., Staerk, J., Colchero, F., et al. (2019) Data gaps and opportunities for comparative and conservation biology. *Proceedings of the National Academy of Sciences*, **116**, 9658.

Constable, H., Guralnick, R., Wieczorek, J., Spencer, C., Peterson, A.T. & VertNet Steering, C. (2010) VertNet: a new model for biodiversity data sharing. *PLoS Biology*, **8**, e1000309.

Cornwell, W.K., Pearse, W.D., Dalrymple, R.L. & Zanne, A.E. (2019) What we (don't) know about global plant diversity. *Ecography*, **0**

Deans, A.R., Lewis, S.E., Huala, E., et al. (2015) Finding Our Way through Phenotypes. *PLoS Biology*, **13**, e1002033.

Dengler, J., Jansen, F., Glöckler, F., Peet, R.K., De Cáceres, M., Chytrý, M., Ewald, J., Oldeland, J., Lopez-Gonzalez, G., Finckh, M., Mucina, L., Rodwell, J.S., Schaminée, J.H.J. & Spencer, N. (2011) The Global Index of Vegetation-Plot Databases (GIVD): a new resource for vegetation science. *Journal of Vegetation Science*, **22**, 582-597.

Díaz, S., Settele, J., Brondízio, E., et al. (2019) Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. In:

eBird (2014) *eBird's missing species*. Retrieved from: https://ebird.org/news/ebirds-missing-species/ (accessed 1 January 2020).

Enquist, B.J., Condit, R., Peet, R.K., Schildhauer, M. & Thiers, B.M. (2016) Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. *PeerJ Preprints*, **4**, e2615v2.

Enquist, B.J., Feng, X., Donoghue, J.C.I., et al. The commonness of rarity: global distribution across the land plants. In prep.

Escribano, N., Galicia, D. & Ariño, A.H. (2018) The tragedy of the biodiversity data commons: a data impediment creeping nigher? *Database*, **2018**

Fazey, I., Fischer, J. & Lindenmayer, D.B. (2005) What do conservation biologists publish? *Biological Conservation*, **124**, 63-73.

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

Fitzgerald, B.F., Coates, J.M. & Lewis, S.M. (2007) *Open Content Licensing: Cultivating the Creative Commons*. Sydney University Press, Sydney, Australia.

Franz, N.M. & Peet, R.K. (2009) Towards a language for mapping relationships among taxonomic concepts. *Systematics and Biodiversity*, **7**, 5-20.

Gallagher, R.V., Falster, D.S., Maitner, B.S., et al. (2020) Open Science principles for accelerating trait-based science across the Tree of Life. *Nature Ecology & Evolution*, **4**, 294-303.

Garnett, S.T., Christidis, L., Conix, S., et al. (2020) Principles for creating a single authoritative list of the world's species. *PLOS Biology*, **18**, e3000736.

GBIF (2014) *New approaches to data licensing and endorsement*. Retrieved from: https://www.gbif.org/news/82363/new-approaches-to-data-licensing-and-endorsement (accessed 1 May 2018).

Guralnick, R., Walls, R. & Jetz, W. (2018) Humboldt Core – toward a standardized capture of biological inventories for biodiversity monitoring, modeling and assessment. *Ecography*, **41**, 713-725.

Guralnick, R.P., Zermoglio, P.F., Wieczorek, J., LaFrance, R., Bloom, D. & Russell, L. (2016) The importance of digitized biocollections as a source of trait data and a new VertNet resource. *Database*, **2016**, baw158-baw158.

Hardisty, A., Roberts, D. & The Biodiversity Informatics, C. (2013) A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecology*, **13**, 16.

Hardisty, A.R., Belbin, L., Hobern, D., McGeoch, M.A., Pirzl, R., Williams, K.J. & Kissling, W.D. (2019) Research infrastructure challenges in preparing essential biodiversity variables data products for alien invasive species. *Environmental Research Letters*, **14**, 025005.

Hecnar, S.J. (2009) Human bias and the biodiversity knowledge base: An examination of the published literature on vertebrates. *Biodiversity*, **10**, 18-24.

Hobern, D., Baptiste, B., Copas, K., et al. (2019) Connecting data and expertise: a new alliance for biodiversity knowledge. *Biodiversity data journal*, **7**, e33679-e33679.

Hortal, J., de Bello, F., 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.

iDigBio (2018a) *Integrated Digitized Biocollections (iDigBio)*. Retrieved from: https://www.idigbio.org (accessed 1 May 2018).

iDigBio (2018b) *Data Ingestion Guidance*. Retrieved from: https://www.idigbio.org/wiki/index.php/Data_Ingestion_Guidance (accessed).

iNaturalist Retrieved from: https://www.inaturalist.org/ (accessed 1 May 2018).

iNaturalist (2018) *Research Grade Obserations*. Retrieved from: https://www.inaturalist.org/posts/16429-research-grade-obserations (accessed 20 October 2018).

James, S.A., Soltis, P.S., Belbin, L., Chapman, A.D., Nelson, G., Paul, D.L. & Collins, M. (2018) Herbarium data: Global biodiversity and societal botanical needs for novel research. *Applications in Plant Sciences*, **6**, e1024.

Jetz, W., McPherson, J.M. & Guralnick, R.P. (2012) Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology and Evolution*, **27**, 151-159.

Joppa, L.N., Roberts, D.L. & Pimm, S.L. (2011) How many species of flowering plants are there? *Proceedings of the Royal Society B: Biological Sciences*, **278**, 554-559.

Jorge, S. & Peterson, A.T. (2004) Biodiversity Informatics: Managing and Applying Primary Biodiversity Data. *Philosophical Transactions: Biological Sciences*, **359**, 689-698.

Kattge, J., D ́az, S., Lavorel, S., et al. (2011) TRY - a global database of plant traits. *Global Change Biology*, **17**, 2905-2935.

Kattge, J., Bönisch, G., Díaz, S., et al. (2020) TRY plant trait database – enhanced coverage and open access. *Global Change Biology*, **26**, 119-188.

König, C., Weigelt, P., Schrader, J., Taylor, A., Kattge, J. & Kreft, H. (2019) Biodiversity data integration—the significance of data resolution and domain. *PLoS Biology*, **17**, e3000183.

LeBauer, D.S., Wang, D., Richter, K.T., Davidson, C.C. & Dietze, M.C. (2013) Facilitating feedbacks between field measurements and ecosystem models. *Ecological Monographs*, **83**, 133-154.

Locey, K.J. & Lennon, J.T. (2016) Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, **113**, 5970.

Lomolino, M.V. (2004) Conservation biogeography. *Frontiers of biogeography: new directions in the geography of nature*, 293-296.

Lomolino, M.V., Riddle, B.R., Whittaker, R.J. & Brown, J.H. (2010) *Biogeography*, 4th edn. Sinauer Associates, Sunderland, Massachusetts.

Lughadha, E.N., Govaerts, R., Belyaeva, I., Black, N., Lindon, H., Allkin, R., Magill, R.E. & Nicolson, N. (2016) Counting counts: revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates. *Phytotaxa*, **272**, 82-88.

MacFadden, B.J. & Guralnick, R.P. (2016) Horses in the Cloud: big data exploration and mining of fossil and extant Equus (Mammalia: Equidae). *Paleobiology*, **43**, 1-14.

Mesibov, R. (2018) An audit of some processing effects in aggregated occurrence records. *ZooKeys*, **751**, 129-146.

Meyer, C., Weigelt, P. & Kreft, H. (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, **19**, 992-1006.

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

Michener, W.K., Allard, S., Budden, A., Cook, R.B., Douglass, K., Frame, M., Kelling, S., Koskela, R., Tenopir, C. & Vieglais, D.A. (2012) Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics*, **11**, 5-15.

Moles, A.T., Warton, D.I., Warman, L., Swenson, N.G., Laffan, S.W., Zanne, A.E., Pitman, A., Hemmings, F.A. & Leishman, M.R. (2009) Global patterns in plant height. *Journal of Ecology*, **97**, 923-932.

Moore, G.E. (1965) Cramming more components onto integrated circuits. *Electronics*, **38**, 114-117.

Nelson, G. & Ellis, S. (2019) The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **374**, 20170391.

Oliveira-Filho, A.T. (2017) *NeoTropTree, Flora arbórea da Região Neotropical: Um banco de dados envolvendo biogeografia, diversidade e conservação*. Retrieved from: http://www.neotroptree,info (accessed 7 May 2019).

Oliveira, B.F., São-Pedro, V.A., Santos-Barrera, G., Penone, C. & Costa, G.C. (2017) AmphiBIO, a global database for amphibian ecological traits. *Sci Data*, **4**, 170123.

Page, L.M., MacFadden, B.J., Fortes, J.A., Soltis, P.S. & Riccardi, G. (2015) Digitization of biodiversity collections reveals biggest data on biodiversity. *BioScience*, **65**, 841-842.

Parr, C.S., Wilson, N., Leary, P., Schulz, K.S., Lans, K., Walley, L., Hammock, J.A., Goddard, A., Rice, J., Studer, M., Holmes, J.T.G. & Corrigan, R.J., Jr. (2014) The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. *Biodivers Data J*, e1079.

Pereira, H.M., Ferrier, S., Walters, M., et al. (2013) Essential Biodiversity Variables. *Science*, **339**, 277.

Reddy, S. & Dávalos, L.M. (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, **30**, 1719-1727.

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

Samper, C. (2004) Taxonomy and environmental policy. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **359**, 721-728.

Schneider, F.D., Fichtmueller, D., Gossner, M.M., Güntsch, A., Jochum, M., König-Ries, B., Le Provost, G., Manning, P., Ostrowski, A., Penone, C. & Simons, N.K. (2019) Towards an ecological trait-data standard. *Methods in Ecology and Evolution*, **10**, 2006-2019.

Serra-Diaz, J.M., Enquist, B.J., Maitner, B., Merow, C. & Svenning, J.-C. (2018) Big data of tree species distributions: how big and how good? *Forest Ecosystems*, **4**, 30.

Shashkov, M. & Ivanova, N. (2019) Considerable Progress in Russian GBIF Community. *Biodiversity Information Science and Standards*, **3**, e37015.

Siefert, A., Violle, C., Chalmandrier, L., et al. (2015) A global meta-analysis of the relative extent of intraspecific trait variation in plant communities. *Ecology Letters*, **18**, 1406-1419.

Singer, R.A., Love, K.J. & Page, L.M. (2018) A survey of digitized data from U.S. fish collections in the iDigBio data aggregator. *PLOS ONE*, **13**, e0207636.

Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O. & Ludwig, C. (2015) The trajectory of the Anthropocene: The Great Acceleration. *The Anthropocene Review*, **2**, 81-98.

Stork, N.E. (2018) How many species of insects and other terrestrial arthropods are there on Earth? *Annual Review of Entomology*, **63**, 31-45.

Sullivan, B.L., Aycrigg, J.L., Barry, J.H., et al. (2014) The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, **169**, 31-40.

Titley, M.A., Snaddon, J.L. & Turner, E.C. (2017) Scientific research on animal biodiversity is systematically biased towards vertebrates and temperate regions. *PLOS ONE*, **12**, e0189577.

U.S. Department of Agriculture Forest Service *Forest Inventory and Analysis Database*. Retrieved from: https://www.fia.fs.fed.us/ (accessed 1 May 2018).

U.S. Geological Survey (2018) *Biodiversity Information Serving Our Nation (BISON)*. Retrieved from: https://bison.usgs.gov (accessed 1 May 2018).

Walls, R.L., Athreya, B., Cooper, L., Elser, J., Gandolfo, M.A., Jaiswal, P., Mungall, C.J., Preece, J., Rensing, S., Smith, B. & Stevenson, D.W. (2012) Ontologies as integrative tools for plant science. *American journal of botany*, **99**, 1263-1275.

Weigelt, P., König, C. & Kreft, H. (2017) GIFT - a global inventory of Floras and traits for macroecology and biogeography. *bioRxiv*, 535005.

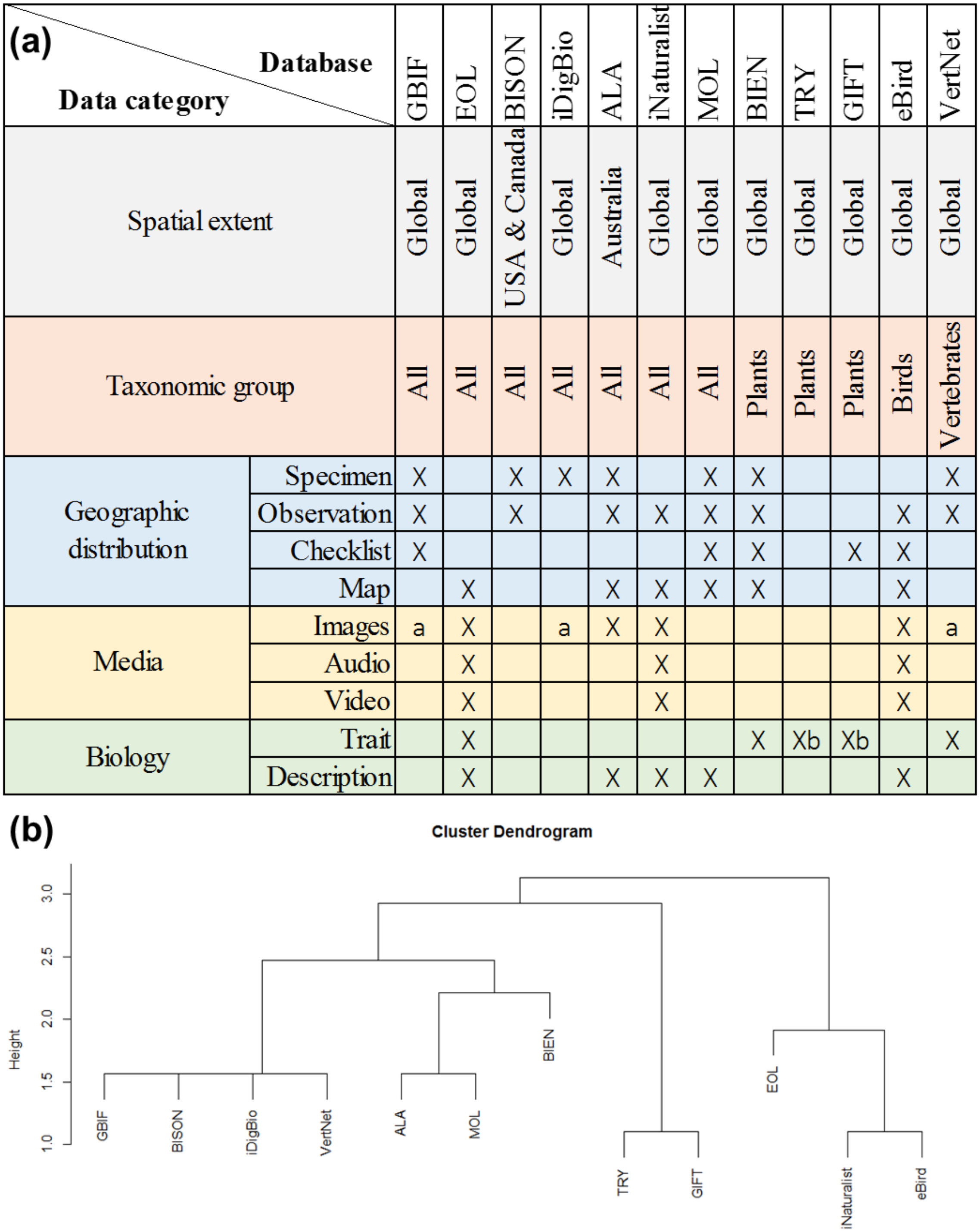
Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T. & Vieglais, D. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*, **7**, e29715.

Wiser, S.K. (2016) Achievements and challenges in the integration, reuse and synthesis of vegetation plot data. *Journal of Vegetation Science*, **27**, 868-879.

Zermoglio, P.F., Guralnick, R.P. & Wieczorek, J.R. (2016) A Standardized Reference Data Set for Vertebrate Taxon Name Resolution. *PLOS ONE*, **11**, e0146894.

Acknowledgment: n/a. **Funding:** X.F., B.J.E., D.S.P., D.D.B., A.L., E.N., J.R.B., K.M.H., K.E., E.B., M.M.N., T.D.W., and L.L.H. acknowledge the support from The University of Arizona Office of Research, Discovery, and Innovation, Institute of the Environment, the Udall Center for Studies in Public Policy, and the College of Science on the postdoctoral cluster initiative — Bridging Biodiversity and Conservation Science. R.V.G. acknowledges support from ARC grant DE170100208. L.H. and P.R. acknowledge support from Global Environment Facility (GEF). N.J.B.K. was partially supported by NSF DEB #1644641. B.J.M. acknowledges support from USDA Hatch grant MAFES #1011538 and NSF ABI grant #1660000. C.M. acknowledges support from NSF Grant DBI-1913673. N.MH. acknowledges support from the Carlsberg Foundation, Marie Skłodowska-Curie Individual Fellowship (grant 746334), and support from the Danish, and National Research Foundation to the Center for Macroecology, Evolution and Climate (grant DNRF96). I. Š. acknowledges support from GAČR EXPRO 20-29554X. J.C.S. considers this work a contribution to his VILLUM Investigator project "Biodiversity Dynamics in a Changing World" funded by VILLUM FONDEN (grant 16549) and the TREECHANGE project funded by the Danish Council for Independent Research | Natural Sciences (grant 6108-00078B). C.V. was supported by the European Research Council (ERC) Starting Grant Project "Ecophysiological and biophysical constraints on domestication in crop plants" (Grant ERC-StG-2014-639706-CONSTRAINTS). S.W. acknowledges support from Strategic Science Investment Fund from the NZ Ministry of Business, Innovation and Employment. **Author contribution:** X.F., D.P., and B.J.E. formulated the study. B.B., B.S.M., B.J.E., X.F., Y.L., K.M.H., L.H., and P.R. assembled data. X.F. developed and implemented the analyses and wrote the first draft. X.F., B.J.E., D.S.P., D.D.B., and L.L.H. designed the data visualizations with constructive discussion from M.M.N., T.D.W., K.E. and E.B.. X.F., B.J.E., D.S.P., D.D.B., R.V.G., A.L., J.R.B., E.A.N., B.B., B.S.M., K.E., E.B., and L.L.H. contributed to interpretation and writings. L.L.H. and B.J.E. provided support of supervision. L.L.H., B.J.E., K.E., and D.D.B. led to the acquisition of the financial support for the project. W.F., L.H., P.M.J., N.J.B.K., J.C.L., P.A.M., B.J.M., N.M.H., D.M.N., A.T.O., R.K.P., M.P., P.R., B.S., J.M.S., I.Š., J.C.S., C.V., S.W. contributed to data assembling and development of BIEN database. All authors contributed to interpreting the results and the editing of manuscript drafts. **Competing interests:** authors declare no conflicts of interest.

Data and materials availability: The plant distribution data from Global Biodiversity Information Facility are accessible from https://doi.org/10.15468/dl.87zyez. Trait data from Encyclopedia of Life are accessible from https://eol.org/docs/what-is-eol/traitbank. Trait data from VertNet are accessible from http://portal.vertnet.org/search. Plant distribution and trait data from Botanical Information and Ecology Network are accessible from RBIEN package. Trait data from TRY are accessible from https://try-db.org/TryWeb/dp.php. The data from Catalogue of Life are accessible from https://download.catalogueoflife.org/col/monthly/2021-04-05_dwca.zip. The administrative boundary dataset is accessible from https://biogeo.ucdavis.edu/data/gadm3.6/gadm36_shp.zip.



(a) Database / Data category		GBIF	EOL	BISON	iDigBio	ALA	iNaturalist	MOL	BIEN	TRY	GIFT	eBird	VertNet
Spatial extent		Global	Global	USA & Canada	Global	Australia	Global	Global	Global	Global	Global	Global	Global
Taxonomic group		All	All	All	All	All	All	All	Plants	Plants	Plants	Birds	Vertebrates
Geographic distribution	Specimen	X		X	X	X		X	X				X
	Observation	X		X		X	X	X	X			X	X
	Checklist	X						X	X		X	X	
	Map		X			X	X	X	X			X	
Media	Images	a	X		a	X	X					X	a
	Audio		X				X					X	
	Video		X				X					X	
Biology	Trait		X						X	Xb	Xb		X
	Description		X			X	X	X				X	

Figure 1. Overview of biodiversity databases reviewed in this paper. The coverages of their data are shown in panel (a) indicated by “X”. Based on the data coverages, the biodiversity databases are grouped into several clusters (b), where the height of the dendrogram is the relative distance between clusters. Notes: a) GBIF, iDigBio, and VertNet indexes and displays images on its website, while the images are mainly hosted by external institutions or facilities. b) TRY and GIFT also stores geographic information about where the trait was measured.

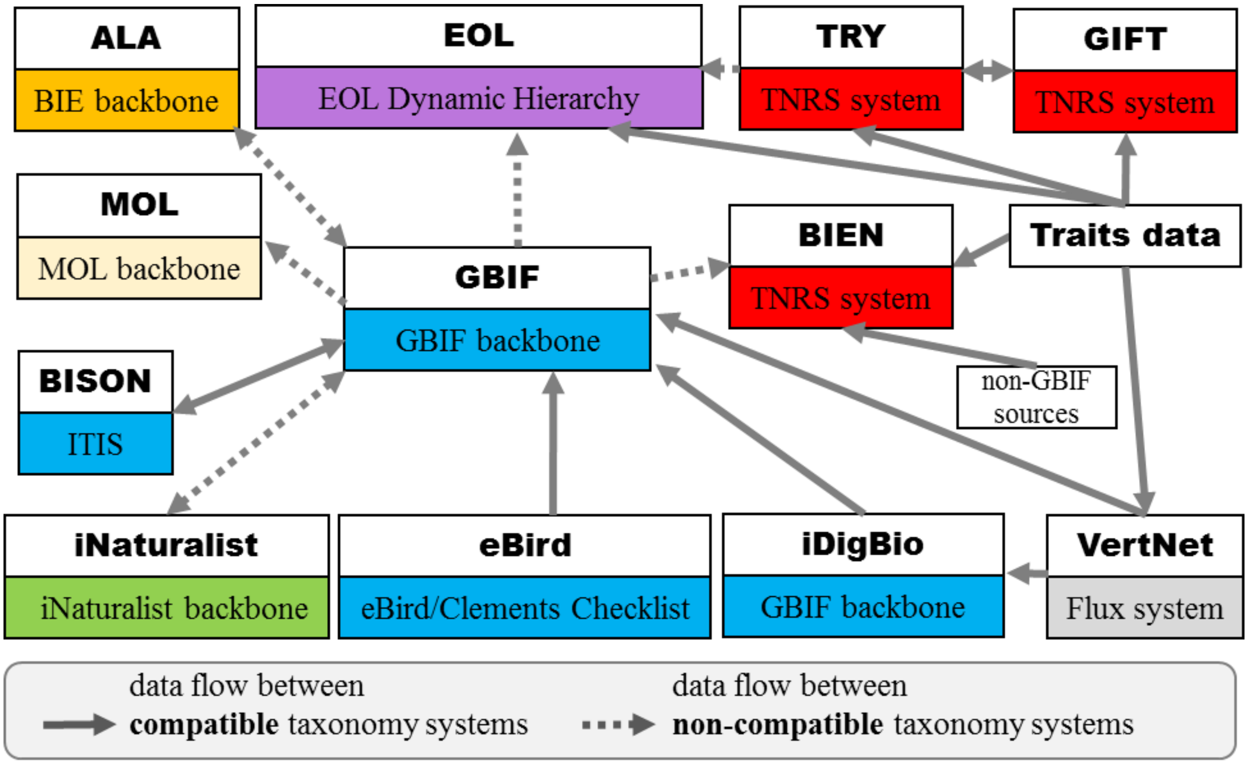


Figure 2. Data exchange between biodiversity databases with different taxonomic systems. Each box represents one database and its adopted taxonomic system (lower half). The taxonomic systems are shown in different colors, while the same color represents compatible systems. A variety of taxonomic systems exist: some databases develop backbone systems (e.g. BIE backbone, GBIF backbone, MOL backbone), some databases adopt a name scrubbing tool that standardizes names towards pre-selected taxonomic systems (e.g. BIEN, GIFT, TRY), some rely on multiple taxonomic systems (e.g. iNaturalist, EOL), and some do not implement a strong regulation on taxonomic names (e.g. VertNet). The one-way or two-way arrow represents unidirectional or bidirectional data flow between databases. ALA: Atlas of Living Australia; BIE: Biodiversity Information Explorer; BIEN: Botanical Information and Ecology Network; BISON: Biodiversity Information Serving Our Nation; EOL: Encyclopedia of Life; GBIF: Global Biodiversity Information Facility; GIFT: Global Inventory of Floras and Traits; iDigBio: Integrated Digitized Biocollections; ITIS: Integrated Taxonomic Information System; IUCN: International Union for Conservation of Nature; MOL: Map of Life; TNRS: Taxonomic Name Resolution Service; TRY: TRY, a global database of plant traits; uBio: Universal Biological Indexer and Organizer. As the databases continue to grow and develop, this figure represents the best of our knowledge as of March 2021.

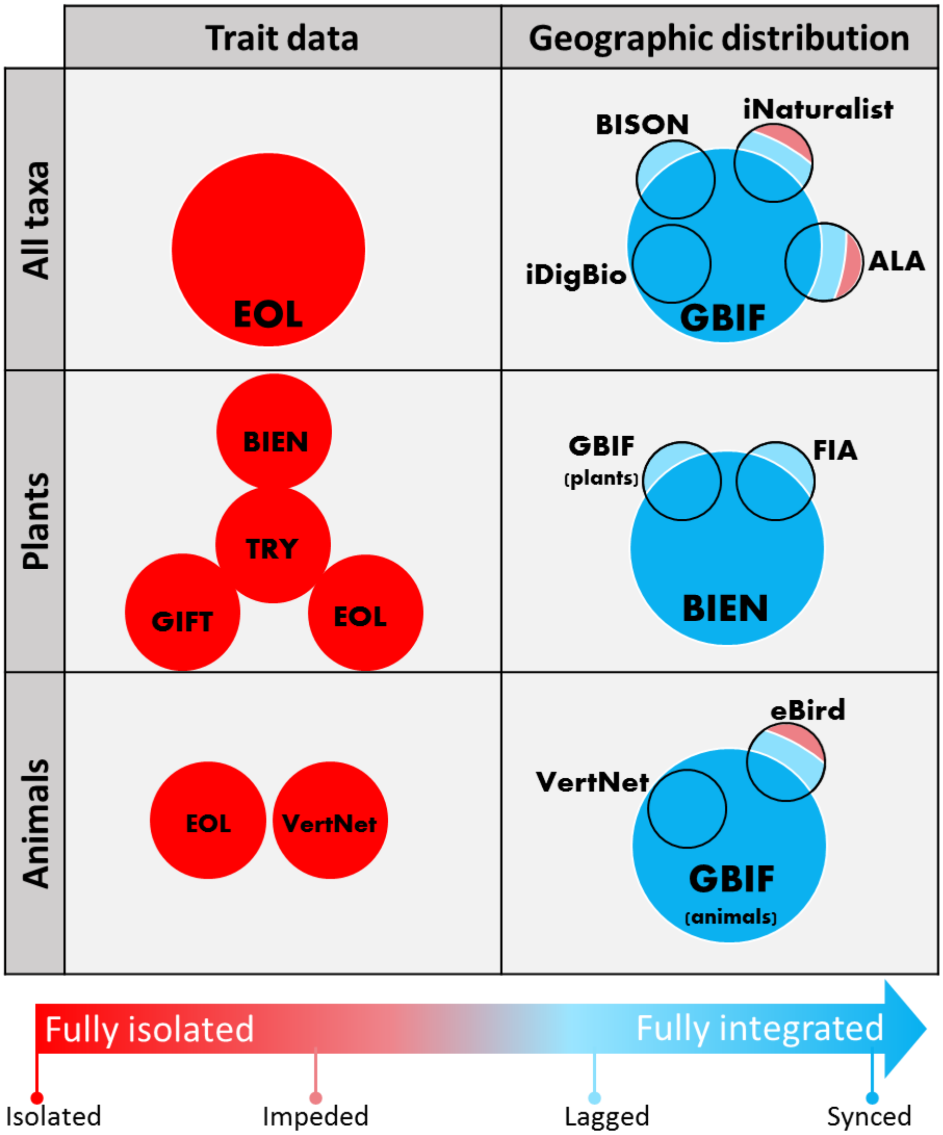


Figure 3. Data integration among biodiversity databases. The status of data integration is classified as four categories: synced, lagged, impeded, and isolated . *Synced* refers to the status of full integration, in either one or multiple directions, between different databases in or near real-time. For example, data published to iDigBio is automatically published to GBIF. *Lagged* refers to the difference between otherwise fully integrated databases between two sync events. For example, BIEN imports and integrates data from GBIF and other sources (e.g., The Forest Inventory and Analysis or FIA) annually or at longer intervals and publishes the results as versioned database releases. The most recent data in those sources will not be available via BIEN until the next import and versioned release. Impeded refers to differences between databases caused by barriers that prevent subsets of the data from being shared. For example, iNaturalist only publishes data to GBIF that are properly licensed for open sharing (iNaturalist, 2018). Contrary to distribution databases, trait databases are generally isolated from one another in different databases, though there are flows/exchanges of plant trait data between TRY and GIFT, and TRY and EOL (Table S1). We caution that the data flow between or among databases is not well documented, and this figure represents the best of our knowledge as of March 2021.

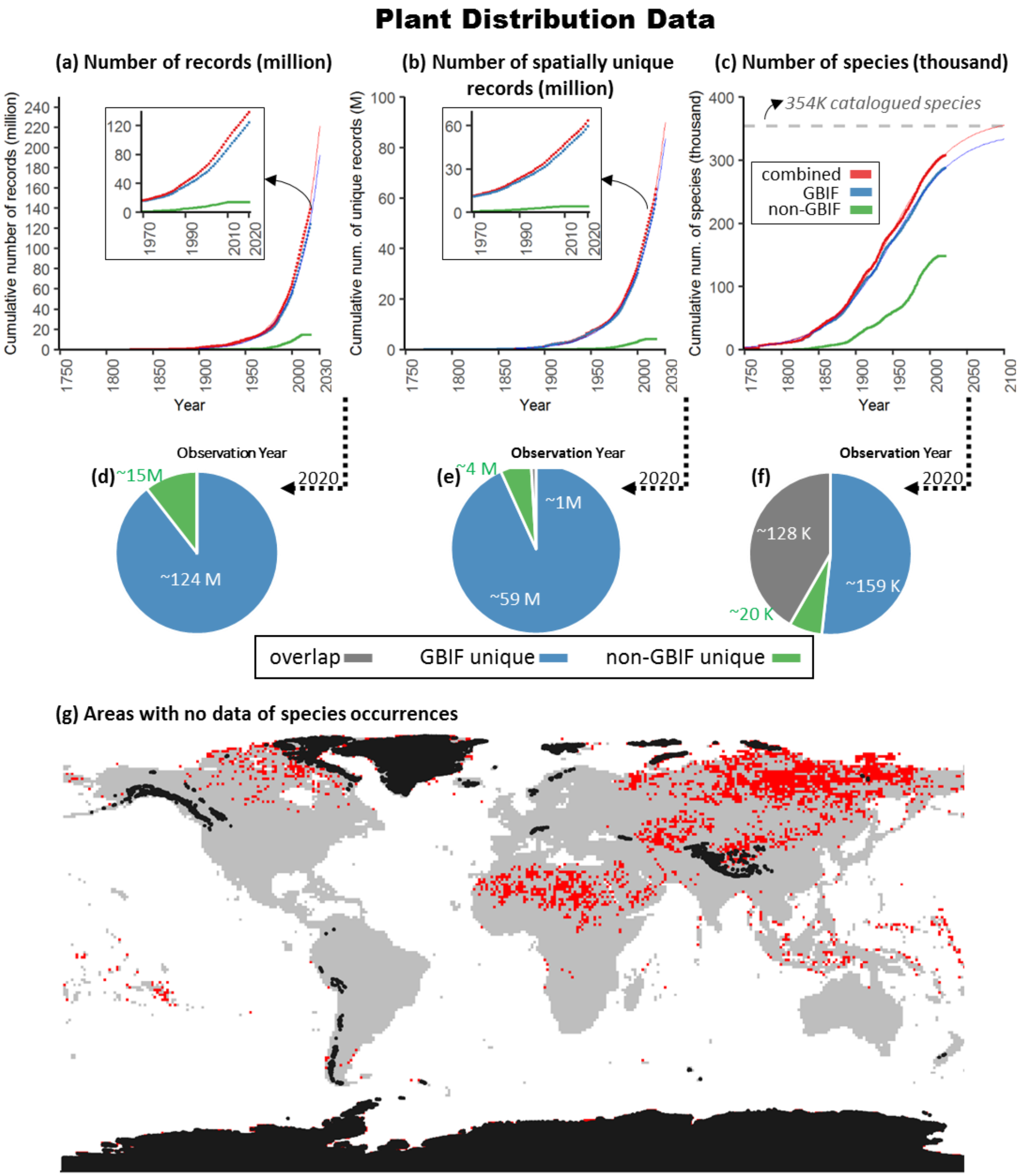


Figure 4. Spatial and taxonomic coverage of terrestrial plant occurrence data. Georeferenced plant observations, as illustrated by observation dates in GBIF, the largest biodiversity informatics infrastructure, have increased exponentially over the past 200 years (panel a,b), though the number of species recorded in these databases is reaching saturation (panel c). By integrating additional data sources compiled by BIEN (i.e. non-GBIF sources; ~15 million records; panel d), the georeferenced plant observations in GBIF can be expanded by an additional ~4 million spatially unique records (panel e) and ~20 thousand species (panel f). Still, the gaps in plant distributions warrant our attention: large areas in Russia, central Asia, and northern Africa (red area in panel g) are missing publicly available occurrences. The black color in panel g represents ice covered areas.

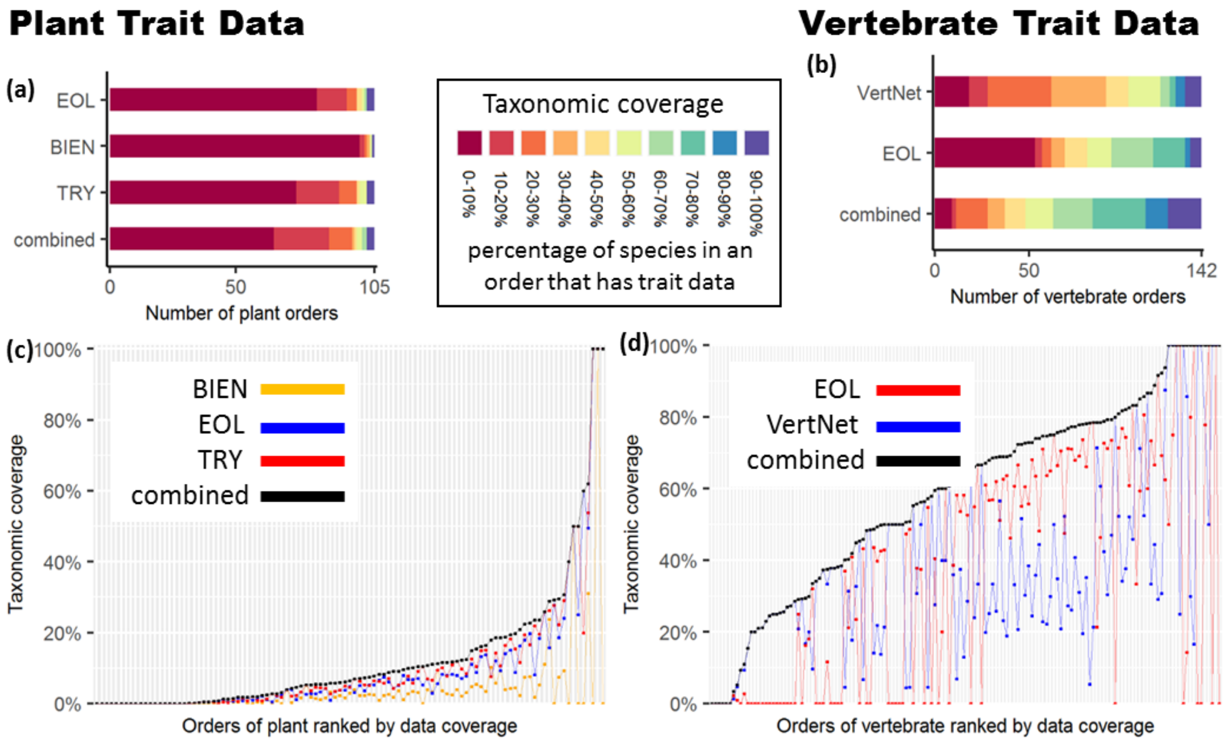


Figure 5. Increased taxonomic coverage of plant and vertebrate trait data through data integration. By combining trait databases, coverage could be expanded in 69-82 plant orders (panel a) and 86-124 vertebrate orders (panel b) compared to individual data sources (panel c & d). The taxonomic coverage of a database is measured as the percentage of the species in that plant or vertebrate order that are represented. Panels c & d show the taxonomic coverages of individual databases and the combined dataset; the positions of the points on the x-axis are re-ordered from low to high based on the combined taxonomic coverage (orders with low coverage on the left and orders with high coverage on the right).

Table S1. Summary of data flow among biodiversity databases.

From	**To**	**Details**	**References/Links**
ALA	GBIF	ALA is a GBIF publisher, though data hosted by ALA may not be fully available on GBIF because of, for example, data licenses.	*https://web.archive.org/web/20210506151646/https://www.gbif.org/publisher/3c5e4331-7f2f-4a8d-aa56-81ece7014fc8*
GBIF	ALA	ALA includes exported data from GBIF that occur in Australia.	*https://web.archive.org/web/20210407034945/https://collections.ala.org.au/public/showDataResource/dr695*
GBIF	MOL	MOL includes exported data from GBIF.	*https://web.archive.org/web/20210506152723/https://mol.org/datasets/9905692e-6a28-4310-b01e-476a471e5bf8*
BISON	GBIF	BISON is a product of the United States Geological Survey (USGS) (Administrator of the U.S. Node of GBIF), and thus works closely and shares data with GBIF.	*https://bison.usgs.gov/#help*
GBIF	BISON	The Canadian and U.S. data added directly to GBIF would become available through BISON.	*https://bison.usgs.gov/#help*
iNaturalist	GBIF	iNaturalist is a GBIF publisher.	*https://web.archive.org/web/20210506161424/https://www.gbif.org/publisher/28eb1a3f-1c15-4a95-931a-4af90ecb574d*
GBIF	iNaturalist	iNaturalist displays data from GBIF on the interactive map.	*https://www.inaturalist.org/taxa/71130-Polyphaga*
GBIF	EOL	EOL incorporates data from GBIF.	*https://web.archive.org/web/20210506162446/https://opendata.eol.org/dataset/gbif-data-summaries*
eBird	GBIF	eBird Observational Dataset is published on GBIF.	*https://web.archive.org/web/20210329225357/https://ebird.org/news/gbif/*
TRY	EOL	TRY summarized records are available from EOL.	*https://web.archive.org/web/20210326174302/https://eol.org/resources/504*
TRY	GIFT	Co-develop and exchange trait data on plant growth form.	*(Kattge et al., 2020)*
GIFT	TRY	Co-develop and exchange trait data on plant growth form.	*(Kattge et al., 2020)*
GBIF	BIEN	BIEN includes data exported from GBIF.	*https://web.archive.org/web/20210506163327/https://bien.nceas.ucsb.edu/bien/biendata/bien-2/sources/*

iDigBio	GBIF	iDigBio is a GBIF publisher.	*https://web.archive.org/web/20210506164312/https://www.gbif.org/publisher/2053a639-84c3-4be5-b8bc-96b6d88a976c*
VertNet	GBIF	VertNet is a GBIF publisher.	*https://web.archive.org/web/20210329192932/http://vertnet.org/join/ipt.html*
VertNet	iDigBio	The majority of the data in the datasets published by VertNet are available in other portals such as GBIF, Canadensys, and iDigBio.	*https://web.archive.org/web/20201012204516/vertnet.org/resources/datalicensingguide.html*

Table S2. Summary of taxonomic system of biodiversity databases.

Name	**Taxonomic system**	**References**
GBIF	GBIF backbone	*https://doi.org/10.15468/39omei*
ALA	Biodiversity Information Explorer (BIE) backbone	*https://web.archive.org/web/20210407032823/https://www.ala.org.au/blogs-news/updates-to-alas-name-and-taxonomy-index/*
MOL	MOL developed a backbone that includes Catalogue of Life and manually curated taxonomic datasets for synonym issues.	*Anonymous reviewer*
BISON	Integrated Taxonomic Information System (ITIS)	*https://web.archive.org/web/20210505185337/https://bison.usgs.gov/*
iNaturalist	iNaturalist backbone is composed of global taxonomic authorities. regional taxonomic authorities, primary literature, and other name providers including Catalogue of Life and uBio.	*https://web.archive.org/web/20210505185713/https://www.inaturalist.org/pages/curator+guide*
EOL	The EOL Dynamic Hierarchy is curated by EOL staff based on a suite of classification providers (including Catalog of Life, the International Union for Conservation of Nature (IUCN), the National Center for Biotechnology Information (NCBI) and the World Register of Marine Species (WoRMS)) for different branches and layers of the tree of life, and can be manually patched and curated.	*https://web.archive.org/web/20210505190456/https://eol.org/docs/what-is-eol/whats-new*
TRY	Plant taxonomy of the TRY database is consolidated using the Taxonomic Names Resolution Service (TNRS) with a taxonomic backbone based on the Plant List, Tropicos, the Global Compositae Checklist, the International Legume Database and Information Service, and USDA's Plants Database.	*(Kattge et al., 2020)*
GIFT	The GIFT database standardized non-hybrid species names in The Plant List 1.1 and additional resources available via iPlant's Taxonomic Name Resolution Service (TNRS).	*(Weigelt et al., 2017)*
BIEN	Taxon names were corrected and standardized using the Taxonomic Name Resolution Service v5.0 (TNRS) with Tropicos, The Plant List and USDA Plants as taxonomic references, and all other options at their default settings.	*(Enquist et al.)*
eBird	eBird/Clements Checklist The eBird species and subspecies taxonomy follows the Clements Checklist. In addition to the formal taxonomic concepts that are included in the Clements Checklist, the eBird taxonomy includes an expanded list of other bird taxa that birders may report.	*https://web.archive.org/web/20210505232653/https://ebird.org/science/use-ebird-data/the-ebird-taxonomy*
iDigBio	The scientific names are matched to the GBIF backbone to correct typos and older names.	*https://web.archive.org/web/20210505233105/https://www.idigbio.org/wiki/index.php/Data_Ingestion_Guidance*
Vertnet	Flux system VertNet does not have a simple taxon resolution mechanism, and vertebrate species names are particularly in flux.	*(Zermoglio et al., 2016)*

Table S3. Summaries of model fitting for the temporal trend in plant distribution data.

Data source	Data	Model	AIC	Inflection point
combined	number of records	exponential	-1686	n/a
		linear	-239	n/a
		logistic	NA	NA
	number of spatially unique records	exponential	-1916	n/a
		linear	-258	n/a
		logistic	NA	NA
	number of species	exponential	-739	n/a
		linear	-510	n/a
		logistic	-1682	1947
GBIF	number of records	exponential	-1816	n/a
		linear	-315	n/a
		logistic	NA	2059
	number of spatially unique records	exponential	-1957	n/a
		linear	-301	n/a
		logistic	NA	NA
	number of species	exponential	-804	n/a
		linear	-552	n/a
		logistic	-1762	1949

Appendix 1. Materials and Methods

Metadata review

Many biodiversity databases have been built over the past decade, with varying emphases on taxonomy, spatial location, and record type. Associated metadata for biodiversity databases is typically found in publications or project websites. To synthesize the major attributes of existing biodiversity databases, we selected 12 well-established biodiversity databases: Atlas of Living Australia (ALA (Belbin & Williams, 2016)), Botanical Information and Ecology Network (BIEN version 4.1 (Enquist *et al.*, 2016)), Biodiversity Information Serving Our Nation (BISON (U.S. Geological Survey, 2018)), eBird (Sullivan *et al.*, 2014), Encyclopedia of Life (EOL (Parr *et al.*, 2014)), Global Biodiversity Information Facility (GBIF), Global Inventory of Floras and Traits (GIFT (Weigelt *et al.*, 2017)), Integrated Digitized Biocollections (iDigBio (iDigBio, 2018a)), iNaturalist (iNaturalist), Map of Life (MOL (Jetz *et al.*, 2012)), a global database of plant traits (TRY version 1.0 (Kattge *et al.*, 2011)), and VertNet (Constable *et al.*, 2010). The twelve databases we examined were chosen among the most commonly used, well-established, large-scale biodiversity databases (MacFadden & Guralnick, 2016; Chandler *et al.*, 2017a; James *et al.*, 2018; Singer *et al.*, 2018; Cornwell *et al.*, 2019; König *et al.*, 2019) to maximize the generalizability of our results and conclusions. Selections were also limited to databases from which we could either access the entirety of the data or the ones with clear documentations. We compiled information from online documentation and relevant publications, though the design and architecture of a database can be in continuous development. Specifically, we recorded database name, taxonomic scope, taxonomic system, record type, number of records, and spatial coverage. We classified the record types into three categories: geographic distribution, media (image, audio, or video), and biological information (standardized trait databases or generalized text descriptions). Within geographic distribution, we further classified the information as specimen records, observations, checklists of geographic regions, and distribution maps. Specimen records and observations both have information on species' geolocations, but only specimen records are associated with physical specimens. Checklists usually contain lists of species known to be present in certain geographic regions (e.g., political divisions or protected areas). Distribution maps are either drawn by experts or generated through models. There are frequent data exchanges among biodiversity databases, but many are not transparent to database users. Consequently, we compiled data exchange information and assessed the status of data integration between databases. We used geographic distribution and trait data as examples, which are the most prominent record type among the reviewed databases. We assessed the integration status by taxonomy groups, which are all organisms, plants, or vertebrates

Improvement of data coverage by database integration

To quantify the improvement gained by combining multiple databases, we compared leading databases that focus on similar taxonomic groups and record type. We used terrestrial plants (Embryophyta) and vertebrates as test cases, because these are the taxonomic groups that are comparatively better collected and documented in biodiversity databases compared to other taxonomic groups (Clark & May, 2002; Fazey *et al.*, 2005; Hecnar, 2009; Titley *et al.*, 2017; Cornwell *et al.*, 2019; König *et al.*, 2019; Kattge *et al.*, 2020). We did not use taxoa, such as microbes, that account for large portions of biodiversity on Earth but face huge data gaps (Locey & Lennon, 2016). More specifically, we compared (1) plant distribution data from GBIF and non-GBIF sources compiled by BIEN (Enquist *et al.*, 2016), (2) plant trait data (i.e. plant height) from BIEN, TRY, GIFT, and EOL, and (3) animal trait data (i.e. vertebrate body length) from VertNet and EOL.

We obtained plant distribution data from BIEN (version 4.2; accessed March 2021) that compiled plant distribution data from GBIF (*https://doi.org/10.15468/dl.87zyez*) and non-GBIF sources, such as the *Forest Inventory and Analysis* (U.S. Department of Agriculture Forest Service) (FIA) and *NeoTropTree* (Oliveira-Filho, 2017). The GBIF and non-GBIF sources have been fused through a series of data scrubbing and standardization workflows (e.g. TNRS (Boyle *et al.*, 2013)) and here we only included data with valid collection year and spatial coordinates. We classified the data into three groups: data from GBIF, data from non-GBIF sources, and the combined full dataset. We quantified the numbers of distribution records, numbers of spatially unique records, and numbers of species with distribution records in all three data sources. A spatially unique record is defined as a record of the distribution of a species (a pixel at 30 arc-seconds resolution in WGS84 coordinate reference system that its coordinate corresponds to) that is unique to a dataset. We standardized all species names against multiple reference taxonomies, including *Tropicos* and *The Plant List*, through the *TNRS* (Boyle *et al.*, 2013). The standardization process parses and corrects misspelled names and authorities, standardizes variant spellings, and converts nomenclatural synonyms to currently accepted names. To reveal the temporal trend of data accumulation, we quantified the cumulative numbers of observations made over time, from 1750 to present (2020).

To describe and quantify those temporal trends, we fitted the cumulative numbers (dependent variable) and years (independent variable) with simple linear (eqn 1), exponential (eqn 2), and logistic regression (eqn 3) using ordinary least squares (“nls” function in stats package version 3.4.2 in R version 3.4.2):

$$y = a + b * x \ (eqn\ 1)$$

$$y = e^{a+b*x} \ (eqn\ 2)$$

$$y = \frac{a}{1 + e^{-b-c*x}} \ (eqn\ 3)$$

where *x* represents time and *y* represents either number of records, number of spatially unique records, or the number of species. We determined the best model fit from the lowest Akaike Information Criterion value (AIC). To reveal the contribution of GBIF or non-GBIF sources to the combined dataset, we quantified the commonalities and uniqueness of GBIF and non-GBIF subsets in terms of number of records, number of spatially unique records, and number of species with distribution data. For our quantification of the temporal trend in the number of species observed, we also retained only currently accepted names to reduce uncertainty (Berendsohn, 1997; Franz & Peet, 2009; Boyle *et al.*, 2013), which yield comparable temporal pattern.
We identified knowledge gaps in two ways. We showed the pixels (at 30 arc-seconds resolution in WGS84 coordinate reference system) for which there were no valid plant geolocation data, and quantified the geographic area of those pixels (in Eckert IV equal area projection). We caution that the gap here may be an overestimation because the plant distribution data compiled by BIEN (including the data exported from GBIF) do not include all possible data sources, but rather shareable data that are mainly publicly available. We then calculated the taxonomic completeness of the distribution data at the level of plant orders. We obtained a list of accepted names of extant terrestrial plant species from the *Catalogue of Life* (Catalogue of Life, 2021) and considered that as the master list of known species. All taxonomic names were standardized

through TNRS (Boyle *et al.*, 2013). We obtained the order level completeness by calculating the percentage of species in a plant order that have distribution information in the combined dataset.

In addition to distribution data, we also investigated the improvement in taxonomic coverage of trait data through database integration, specifically terrestrial plant height and vertebrate body length. We downloaded plant height data from BIEN, EOL, and TRY (accessed March 2021). We also obtained a list of accepted names of extant terrestrial plant species from *Catalogue of Life* (accessed March 2021) and considered that as the master list of known species. All taxonomic names were standardized through TNRS (Boyle *et al.*, 2013). We calculated the taxonomic completeness of species trait information at the species and order levels. We obtained the species level completeness by checking species whose heights were recorded in BIEN, EOL, TRY, or the combined dataset, against the names recorded in COL. We obtained the order level completeness by calculating the percentage of species in a plant order that have height information in either dataset. We calculated the improvement in percentages by comparing individual datasets to the combined dataset. The improvement in taxonomic coverage represents the benefit of using multiple databases.

Following the same workflow, we quantified the taxonomic coverage of animal trait and percentage improvement between individual dataset and the combined dataset. Body length of vertebrates were downloaded from VertNet and EOL (accessed March 2021). Accepted names of extant vertebrates were obtained from *Catalogue of Life*. The taxonomic names were standardized through Global Names Resolver using the *Taxize* package (Chamberlain & Szocs, 2013) (version 0.9.4.9100) in R (version 3.4.2). The Global Names Resolver resolves names against specific name databases, which is *Catalogue of Life* in this study. The resolution process includes a series of exact and fuzzy matches based on the full or part of the name input (see more details in *https://resolver.globalnames.org/about*). The matching process also considers the context of taxonomy and reduces the likelihood of matches to taxonomic homonyms. The matching process yields a series of confidence scores for all possible matches; here we only kept the best matching records. However, the creation of a single authoritative list of names will take time; full reconciliation of synonyms and distinct taxon concepts may take decades (Berendsohn, 1997; Franz & Peet, 2009; Boyle *et al.*, 2013). The standardization of taxonomic names based on either TNRS or Global Names Resolver will not solve all issues of taxonomic name integration, but this step represents the state-of-the-art in standardizing taxonomy names in biodiversity databases and provides a baseline for the comparisons of different biodiversity databases.