Check for updates

Received: 29 June 2021 | Revised: 1 March 2022 | Accepted: 14 March 2022

DOI: 10.1111/geb.13497

Global Ecology and Biogeography | A Journal of Macroecology | WILEY

REVIEW ARTICLE

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

Xiao Feng[1] | Brian J. Enquist[2,3] | Daniel S. Park[4,5] | Brad Boyle[2] | David D. Breshears[6] | Rachael V. Gallagher[7] | Aaron Lien[6] | Erica A. Newman[2,8] | Joseph R. Burger[9] | Brian S. Maitner[10] | Cory Merow[10] | Yaoqi Li[11] | Kimberly M. Huynh[8] | Kacey Ernst[12] | Elizabeth Baldwin[13] | Wendy Foden[14,15,16] | Lee Hannah[17] | Peter M. Jørgensen[18] | Nathan J. B. Kraft[19] | Jon C. Lovett[20,21] | Pablo A. Marquet[3,22,23] | Brian J. McGill[24] | Naia Morueta-Holme[25] | Danilo M. Neves[26] | Mauricio M. Núñez-Regueiro[27] | Ary T. Oliveira-Filho[26] | Robert K. Peet[28] | Michiel Pillet[2,29] | Patrick R. Roehrdanz[17] | Brody Sandel[30] | Josep M. Serra-Diaz[31,32] | Irena Šímová[33,34] | Jens-Christian Svenning[32,35] | Cyrille Violle[36] | Trang D. Weitemier[8] | Susan Wiser[37] | Laura López-Hoffman[6]

[1]Department of Geography, Florida State University, Tallahassee, Florida, USA

[2]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona, USA

[3]The Santa Fe Institute, Santa Fe, New Mexico, USA

[4]Department of Biological Sciences, Purdue University, West Lafayette, Indiana, USA

[5]Purdue Center for Plant Biology, Purdue University, West Lafayette, Indiana, USA

[6]School of Natural Resources and the Environment, University of Arizona, Tucson, Arizona, USA

[7]Hawkesbury Institute for the Environment, Western Sydney University, Richmond, New South Wales, Australia

[8]Arizona Institutes for Resilience, University of Arizona, Tucson, Arizona, USA

[9]Department of Biology, University of Kentucky, Lexington, Kentucky, USA

[10]Eversource Energy Center and Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut, USA

[11]Department of Health and Environmental Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China

[12]Department of Epidemiology and Biostatistics, University of Arizona, Tucson, Arizona, USA

[13]School of Government and Public Policy, University of Arizona, Tucson, Arizona, USA

[14]Cape Research Centre, South African National Parks, Cape Town, South Africa

[15]Global Change Biology Group, Department of Botany & Zoology, University of Stellenbosch, Stellenbosch, South Africa

[16]Climate Change Specialist Group, IUCN Species Survival Commission, Gland, Switzerland

[17]The Moore Center for Science, Conservation International, Arlington, Virginia, USA

[18]Missouri Botanical Garden, St. Louis, Missouri, USA

[19]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California, USA

[20]School of Geography, University of Leeds, Leeds, UK

[21]Royal Botanic Gardens, Kew, Richmond, Surrey, UK

[22]Departamento de Ecología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago, Chile

[23]Instituto de Ecología y Biodiversidad (IEB), Centro de Modelamiento Matemático (CMM), Universidad de Chile - IRL 2807 CNRS Beauchef 851 & Centro de Cambio Global UC, Santiago, Chile

[24]School of Biology and Ecology & Mitchell Center for Sustainability Solutions, University of Maine, Orono, Maine, USA

[25]Center for Macroecology, Evolution and Climate, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

[26]Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

 © 2022 John Wiley & Sons Ltd

[27]Instituto de Bio y Geociencias del NOA (IBIGEO) Universidad Nacional de Salta (UNSa), Universidad Católica de Salta, Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Salta, Argentina

[28]Department of Biology, CB#3280, University of North Carolina, Chapel Hill, North Carolina, USA

[29]Species Survival Commission, Cactus and Succulent Plants Specialist Group, International Union for Conservation of Nature, Cambridge, UK

[30]Department of Biology, Santa Clara University, Santa Clara, California, USA

[31]Université de Lorraine, AgroParisTech, INRAE, Silva, Nancy, France

[32]Center for Biodiversity Dynamics in a Changing World (BIOCHANGE), Department of Biology, Aarhus University, Aarhus C, Denmark

[33]Center for Theoretical Study, Charles University and The Czech Academy of Sciences, Prague, Czech Republic

[34]Department of Ecology, Faculty of Science, Charles University, Prague, Czech Republic

[35]Section for Ecoinformatics and Biodiversity, Department of Biology, Aarhus University, Aarhus C, Denmark

[36]CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France

[37]Manaaki Whenua – Landcare Research, Lincoln, New Zealand

Correspondence
Xiao Feng, Department of Geography, Florida State University, Tallahassee, FL, USA.
Email: fengxiao.sci@gmail.com

Funding information
X.F., B.J.E., D.S.P., D.D.B., A.L., E.A.N., J.R.B., K.M.H., K.E., E.B., M.M.N.-R., 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 Australian Research Council (ARC) Grant DE170100208. L.H. and P.R.R. acknowledge support from Global Environment Facility (GEF). N.J.B.K. was partially supported by National Science Foundation (NSF) Division of Environmental Biology (DEB) #1644641. B.J.M. acknowledges support from United States Department of Agriculture (USDA) Hatch grant MAFES #1011538 and NSF EPSCOR 2019470. C.M. acknowledges support from NSF Grant DBI-1913673. N.M.-H. acknowledges support from the Carlsberg Foundation and support from the Danish, and National Research Foundation to the Center for Macroecology, Evolution and Climate (grant DNRF96). B.J.E. and D.M.N. acknowledge support from NSF DEB (grant 1556651). D.M.N. acknowledges support from Instituto Serrapilheira/Brazil (grant Serra-1912-32082). 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

Abstract

Aim: Addressing global environmental challenges requires access to biodiversity data across wide spatial, temporal and taxonomic scales. Availability of such data has increased exponentially recently with the proliferation of biodiversity databases. However, heterogeneous coverage, protocols, and standards have hampered integration among these databases. To stimulate the next stage of data integration, here we present a synthesis of major databases, and investigate (a) how the coverage of databases varies across taxonomy, space, and record type; (b) what degree of integration is present among databases; (c) how integration of databases can increase biodiversity knowledge; and (d) the barriers to database integration.

Location: Global.

Time period: Contemporary.

Major taxa studied: Plants and vertebrates.

Methods: We reviewed 12 established biodiversity databases that mainly focus on geographic distributions and functional traits at global scale. We synthesized information from these databases to assess the status of their integration and major knowledge gaps and barriers to full integration. We estimated how improved integration can increase the data coverage for terrestrial plants and vertebrates.

Results: Every database reviewed had a unique focus of data coverage. Exchanges of biodiversity information were common among databases, although not always clearly documented. Functional trait databases were more isolated than those pertaining to species distributions. Variation and potential incompatibility of taxonomic systems used by different databases posed a major barrier to data integration. We found that integration of distribution databases could lead to increased taxonomic coverage that corresponds to 23 years' advancement in data accumulation, and improvement in taxonomic coverage could be as high as 22.4% for trait databases.

Main conclusions: Rapid increases in biodiversity knowledge can be achieved through the integration of databases, providing the data necessary to address critical environmental challenges. 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

plants' (Grant ERC-StG-2014-639706-CONSTRAINTS). S.W. acknowledges support from the Strategic Science Investment Fund from the NZ Ministry of Business, Innovation and Employment.

Handling Editor: Allen Hurlbert

1 | INTRODUCTION

In the face of rapid global changes, a grand challenge is efficiently cataloguing, assessing, and responding to changes in biodiversity and associated ecosystem services (Ceballos et al., 2015; Chapin et al., 2000; Díaz et al., 2019). Addressing this challenge requires unprecedented access to biodiversity data across spatial, temporal, and taxonomic 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 (Ball-Damerow et al., 2019; Chandler et al., 2017; Page et al., 2015; Sullivan et al., 2014). 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 (Hortal et al., 2015; Lomolino, 2004; 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), nearly half of which are essentially invisible or inaccessible to the research community due to lack of cataloguing and integration (Blair et al., 2020). There are also large volumes of 'dark data' that need to be catalogued and integrated (Heidorn, 2008).

Data aggregation has been an ongoing goal of the biodiversity community (Ball-Damerow et al., 2019; Nelson & Ellis, 2018). A tremendous amount of work has been done by existing biodiversity data aggregators. For example, building upon a community designed standard (Darwin Core; Wieczorek et al., 2012) and specialized tools (e.g., Integrated Publishing Toolkit; Robertson et al., 2014), GBIF, iDigBio, and VertNet have made over a billion records of species occurrences available online over the past two decades. However, the challenges to integration are many: existing biodiversity data aggregators often have singular objectives and consequently adhere to different protocols and standards (Mesibov, 2018), such as Darwin Core used by GBIF (Wieczorek et al., 2012), Veg-X used by Botanical Information and Ecology Network (BIEN; Wiser et al., 2011), eBird Checklist format used by eBird (eBird, 2021), and structured property graph data used by Encyclopedia of Life (EOL, 2014); and datasets are highly heterogeneous spatially, temporally, and taxonomically (Cornwell et al., 2019; Reichman et al., 2011). As new data are continuously aggregated, and data processing protocols and standards (e.g., georeferencing of missing coordinates and autocorrection of taxonomic names) are further developed along different trajectories, 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; 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; Kattge et al., 2020; König et al., 2019).

To facilitate better integration among biodiversity data domains (König et al., 2019), 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 across 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 that mainly focus on geographic distributions and functional traits at global scale, and aimed to answer four questions: (a) How does the coverage of a suite of major biodiversity databases differ across taxa, space, and record type? (b) How are existing biodiversity databases integrated? (c) How would the integration of databases increase biodiversity knowledge? and (d) 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 examined 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 12 well-established

biodiversity databases: the 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., 2020), Integrated Digitized Biocollections (iDigBio, 2018), iNaturalist (https://www.iNaturalist.org), 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 cannot cover every notable database because of limited resources and accessibility of database content and documentation. Our selections were chosen to represent the breadth of the most commonly used, well-established large-scale biodiversity databases (Chandler et al., 2017; Cornwell et al., 2019; James et al., 2018; König et al., 2019; MacFadden & Guralnick, 2017; Singer et al., 2018) to maximize the generalizability of our results and conclusions. Note that some of the databases share considerable amounts of data, but they can also be different in certain aspects (e.g., distribution data from VertNet are available in GBIF, but the trait data from VertNet are not). More details of their similarities and differences were investigated by compiling information from online documentation and relevant publications (see Sections 2.1 and 2.2). 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).

2.1 | Variation in coverage and data types within biodiversity databases

We reviewed metadata for biodiversity databases from project websites or publications, and recorded the 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 data or generalized text descriptions). Within the category of 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 were classified by type as either image, audio, or video. Biological information included standardized trait data and generalized text descriptions.

Our review showed that each biodiversity database holds unique scientific value because they cover different spatial extents, taxonomic groups, and record types (Figure 1a). The databases could be grouped into different clusters based on similarities of focus and data coverage. For example, iNaturalist and eBird are two citizen science projects where anyone can submit their original observations. EOL, iNaturalist, and eBird form a cluster of databases that indexes media data and biological descriptions, while also sharing public education objectives (Figure 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 shares the property of indexing both species occurrences and geographic range maps. Here, we considered the different attributes equally, though assigning different weights to the attributes can lead to different database groupings. For example, many of the databases seek to document all taxa across the globe (e.g., GBIF, EOL) 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-flows among the databases, that is unidirectional flow of biodiversity data from one database to another, or bidirectional flow between two databases. Biodiversity databases (e.g., GBIF) typically aggregate digitized information from data providers, such as museums, herbaria, and research data repositories, and the detailed information about data providers is usually acknowledged on a database's website (e.g., BIEN data contributors – https://web.archive.org/web/20210228121556/https://bien.nceas.ucsb.edu/bien/data-contributors/all/). However, it is usually not straightforward to understand whether one database is aggregated by another. This may be partially due to concerns of appearing redundant and losing uniqueness, as acknowledging a database to be aggregated by another could be interpreted as one database becoming a subset of the other (larger) 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 seemingly duplicate records when using data from multiple databases). Therefore, we found it important to assess the degree of data sharing and integration among biodiversity databases, which we accomplished by reviewing their documentation and associated publications.

Overall, the data-flows (i.e., the exchange of primary data) 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 the 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 (Figure 2 and Supporting Information 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 (a) aggregate biodiversity data collected from their focal regions (i.e., Australia and North America, respectively) and pass the data to GBIF, and (b) import other data collected from Australia or North America from GBIF to their respective databases (Supporting Information Table S1). There are also cases of unidirectional data-flow from GBIF to specialized databases. For example, MOL aggregates multiple types of information on species geographic distributions, including occurrence records from GBIF; as does BIEN.

We summarized the status of data integration across databases into four categories: *synced*, *lagged*, *impeded*, and *isolated* (Figure 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 are automatically published in GBIF (iDigBio, 2021; Singer et al., 2018), thus the content of iDigBio is considered synced with GBIF (Figure 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 that are properly licensed on GBIF (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).

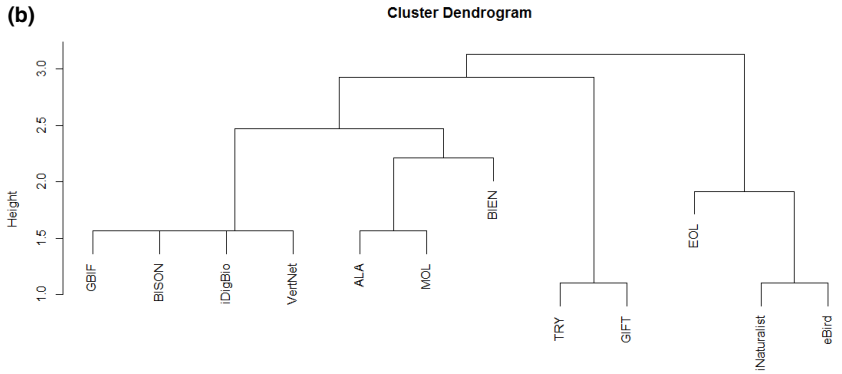
Unlike the distribution databases discussed above, trait databases are characterized by isolation. These databases typically capture data within particular taxa or focus on a single trait, such as GlobTherm for thermal tolerance (Bennett et al., 2018) and AmphiBIO for amphibian ecological traits (Oliveira et al., 2017; Figure 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 standards 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).

Information in a biodiversity database is generally indexed by species' scientific names. However, the dynamic nature of taxonomic research presents challenges to such indexing. With the exception of a few well-established groups such as birds (whose taxonomy is relatively stable and subject to regular external review and standardization; Chesser et al., 2021), conflicting taxonomic concepts, homonyms, outdated synonyms, and ambiguous alternative spellings are prevalent across the tree of life (Boyle et al., 2013; Franz et al., 2008). Furthermore, in addition to the roughly 18,000 new species discovered each year, the taxonomy of the > 2 million species currently described (Mora et al., 2011) is in a state of constant flux, as old species definitions are re-examined and relationships are updated to reflect new insights. Finally, no single, universally agreed-upon taxonomy exists for the tree of life. Therefore, biodiversity databases have implemented different strategies (here termed taxonomic systems) to handle this taxonomic churn (Figure 2 and Supporting Information Table S2). Some databases maintain flexibility in nomenclature, especially when the taxonomy is in flux (e.g., vertebrate species stored in VertNet), whereas some databases impose stronger rules. For example, EOL allows multiple independent taxonomic sources to coexist to avoid potential conflicts between non-compatible nomenclature; GBIF and Catalogue of Life (COL) have both employed a comprehensive but single-backbone system designed to be compatible with different taxonomic sources; MOL developed a backbone that includes Catalogue of Life (a global effort to compile existing catalogued species) and manually curated taxonomic datasets addressing synonyms. For land plants, The Plant List (http://www.theplantlist.org/) adopts periodically revised, comprehensive checklists of accepted species and synonyms for major taxonomic groups. While such static checklists provide relatively stable taxonomies that are easily portable among databases, they have been criticized as imposing arbitrary taxonomic opinions, implying certainty where none exists and perpetuating the use of invalid or illegitimate names in the guise of poorly vetted 'unresolved names' (Kalwij, 2012). An alternative approach is to align species names using one or more dynamic, actively curated taxonomic sources such as the Missouri Botanical Garden's Tropicos database (https://www.tropicos.org/). This is the method used by databases such as BIEN and TRY that resolve their taxonomy using the Taxonomic Name Resolution Service (TNRS; Boyle et al., 2013). However, while actively curated taxonomic references generally provide the most up-to-date taxonomic opinions, their dynamism can also result in taxonomic instability and decreased compatibility among databases (e.g., taxonomic databases such as Tropicos can change daily as curators update content).

Still, the different approaches and strategies may solve taxonomic issues locally within a database (Soberón & Peterson, 2004),

(a) Data category	Database	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	*	X		*	X	X					X	*
	Audio		X				X					X	
	Video		X				X					X	
Biology	Trait		X						X	X†	X†		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. (*) Global Biodiversity Information Facility (GBIF), Integrated Digitized Biocollections (iDigBio), and VertNet index and display images on their websites, while the images are mainly hosted by external institutions or facilities. (†) TRY and Global Inventory of Floras and Traits (GIFT) also store geographic information about where the trait was measured. EOL = Encyclopedia of Life; BISON = Biodiversity Information Serving Our Nation; MOL = Map of Life; ALA = Atlas of Living Australia; BIEN = Botanical Information and Ecology Network



but could deepen differences among different databases that prevent future data integration, thus facilitating the 'speciation' of databases (Figure 2 and Supporting Information Table S2). Ultimately, the solution to this impasse will likely be the use of static, versioned 'snapshots' of actively curated taxonomic databases maintained through a collaboration of global taxonomic experts and biodiversity institutions. An example is the recently initiated World Flora Online (http://www.worldfloraonline.org/). In the meantime, the current incompatibilities of existing databases will need to be resolved either by adherence to a standard set of static checklists (however imperfect) or the development and deployment of tools that allow users to align taxonomies to different taxonomic sources on the fly.

3 | ENHANCED DATA COVERAGE VIA DATABASE INTEGRATION

To quantify the improvement in data coverage provided by 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 (Ball-Damerow et al., 2019; Clark & May, 2002; Cornwell et al., 2019; Fazey et al., 2005; Hecnar, 2009; Kattge et al., 2020; König et al., 2019; Titley et al., 2017). We did not use taxa that account for large portions of biodiversity on Earth but face huge data gaps, such as microbes or invertebrates (Locey & Lennon, 2016). Specifically, we combined (a) the distribution of terrestrial plants from GBIF and non-GBIF sources, and (b) one crucial and commonly measured trait for plants and vertebrates, respectively: maximum height (Guralnick et al., 2016; Moles et al., 2009) 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). Our study goes beyond recent gap analyses of biodiversity data (Cornwell et al., 2019; König et al., 2019; Meyer et al., 2016), by expanding the scope to multiple data aggregators with similar missions, in two major clades (i.e., terrestrial plants and vertebrates), and using an ecological trait characterized by continuous values.

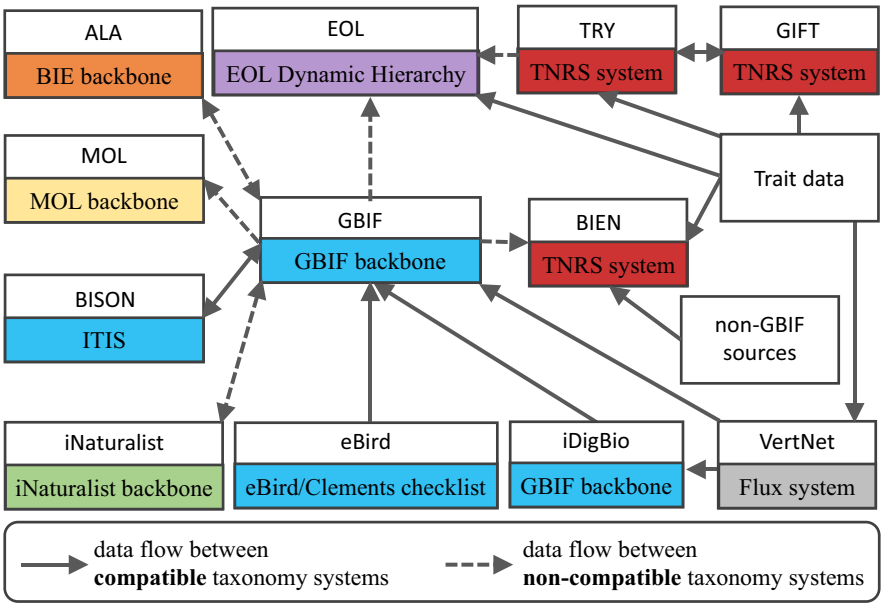


FIGURE 2 Data exchanges between biodiversity databases with different taxonomic systems. Each box represents one database and its adopted taxonomic system (lower half). The taxonomic systems are shown in different colours, with the same colour representing compatible systems. A variety of taxonomic systems 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 arrows represent unidirectional or bidirectional exchanges of primary biodiversity data between databases, respectively. 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; MOL = Map of Life; TNRS = Taxonomic Name Resolution Service; TRY = TRY, a global database of plant traits. As the databases continue to grow and develop, this figure represents the best of our knowledge as of March 2021

3.1 | Better coverage through data integration

3.1.1 | Overall trends 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; Figure 4a) as documented in GBIF and the combined dataset. A similar exponential increase was found when only spatially unique records were examined (Figure 4b). This pattern is also supported by a model selection analysis among linear, exponential, and logistic functions (Supporting Information 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 2 years (Moore, 2006). 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 (Figure 4c). Based on a fitted logistic curve (Supporting Information 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), that is, 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; Supporting Information 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 cataloguing 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).

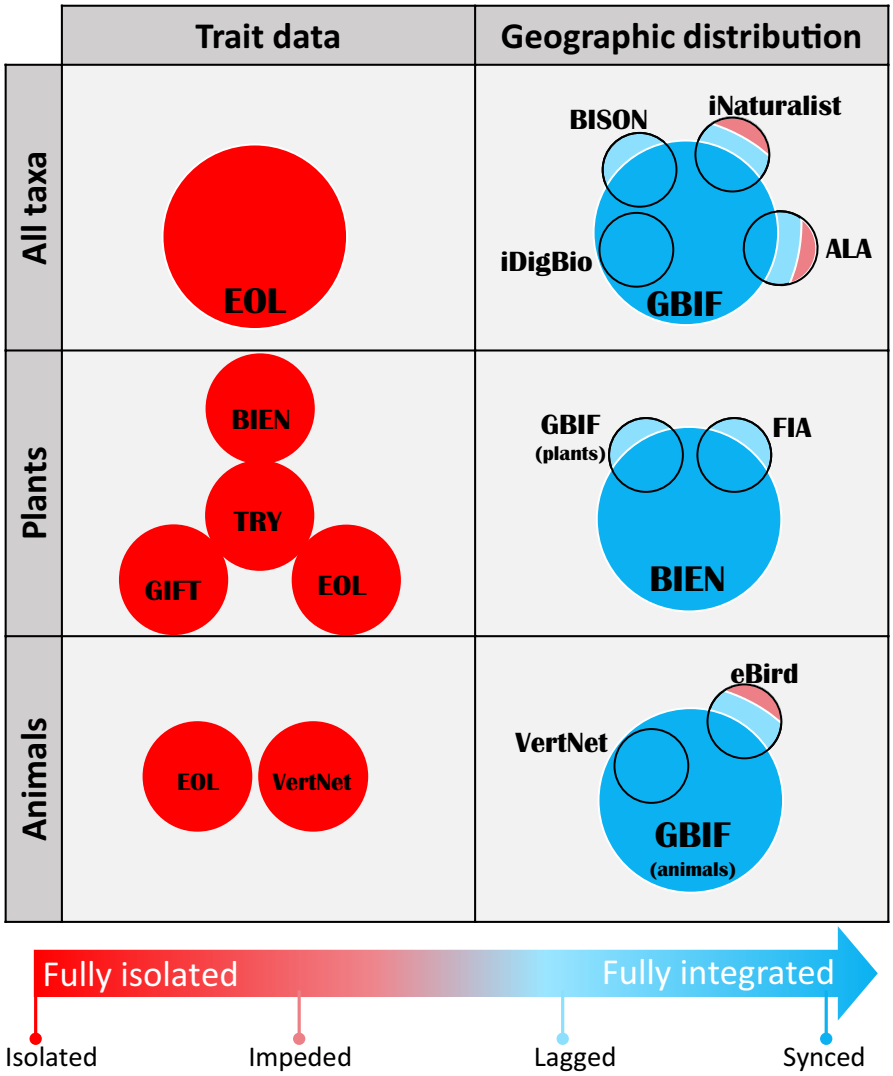


FIGURE 3 Data integration among biodiversity databases. The status of data integration is classified as four categories: synced, lagged, impeded, and isolated. *Synced* refers to the status of full integration, in either one or multiple directions, between different databases in or near real-time. For example, data published to Integrated Digitized Biocollections (iDigBio) are automatically published to Global Biodiversity Information Facility (GBIF). *Lagged* refers to the difference between otherwise fully integrated databases between two sync events. For example, Botanical Information and Ecology Network (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, although there are flows/exchanges of plant trait data between TRY and Global Inventory of Floras and Traits (GIFT), and TRY and Encyclopedia of Life (EOL) (Supporting Information 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. ALA = Atlas of Living Australia; BISON = Biodiversity Information Serving Our Nation

3.1.2 | Improvement in distribution data

Integration of biodiversity databases would increase our knowledge of biodiversity greatly. For instance, adding ~15 million records from additional sources (compiled by BIEN) to GBIF, the world's largest biodiversity repository, would improve its coverage by ~3.7 million spatially unique records and ~20,000 species (Figure 4d–f). The number of distribution records per taxon in GBIF could be increased by 4.4%, which is 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 (or aggregation), based on extrapolation of the fitted logistic curve (Figure 4c, Supporting Information 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). Nonetheless, the complete geographic distributions of many species could remain poorly known. For example, it has been estimated that 36.5% of land plant species are represented by five or

fewer observations (Enquist et al., 2019), and five is much below the number of occurrences used in the inference of species geographic ranges. Even for well-known groups like trees, it has been estimated that only 26% of the species had more than 20 occurrences with high quality information (Serra-Diaz et al., 2018). Therefore, substantial efforts are still needed to increase the quality and quantity of records for many species to fully address the *Wallacean shortfall*.

3.1.3 | Improvement in trait data

Database integration also substantially improves the taxonomic coverage of trait information (maximum height in plants; body length in vertebrates; see Appendix: Materials and 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 (Figure 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 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 (Figure 5).

3.1.4 | Limitations of our assessment

Data integration can effectively decrease gaps in our knowledge, resulting in more comprehensive data that 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 versus abundances, or a trait value measured at individual level versus the species level. However, indexing trait data availability 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 accurately represent the range of values measured at the individual level?' Cross-validations will be especially useful if one database is mainly used by the general public, while other databases are more heavily used by the scientific community, such that more rigorous information is delivered by the scientific 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 (Park et al., 2021; 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 (Cornwell et al., 2019; König et al., 2019; Meyer et al., 2015). Following our integration of various databases (Appendix), approximately 58,000 plant species still lacked publicly available distribution records (i.e., presence records with coordinates). 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 a small number of extant species (Cornwell et al., 2019; e.g., Ceratophyllales). In addition to true knowledge gaps where information on the distributions of species does not exist in any form, there are instances where: (a) locality information exists but is not digitized, (b) locality information is only available at large spatial scales (e.g., country level), (c) locality descriptions are digitized but not georeferenced, (d) coordinates are available but not publicly accessible. Therefore, gaps estimated from publicly available distribution records may represent an overestimation of the true knowledge gap. From a spatial perspective, distribution records are known to have strong biases across regions, usually driven by human factors instead of species richness (Daru et al., 2017; Enquist et al., 2019; Hughes et al., 2021; Park et al., 2021). While North America and Europe are more intensely sampled, there are still ~30.8 million km^2 of ice-free land surface, as assessed using Eckert IV equal area projection, that currently have no valid plant geolocations (Figure 4g). These areas are mainly located in Russia (despite the considerable recent progress of data sharing by the Russian GBIF community; Shashkov & Ivanova, 2019), central and Southeast Asia, and northern Africa, and collectively cover approximately 13% of the Earth's land area. With the exception of the tropical forests of Southeast Asia, these poorly-sampled areas are generally known to have to low plant richness, such as Sahara desert, Taklimakan desert, Siberian tundra, Siberian taiga, and Arctic tundra (Barthlott et al., 2007), suggesting the spatial gap of plant distribution knowledge could be much smaller.

Trait data have considerably larger gaps compared to species distribution data. Plant height and vertebrate body length are commonly used traits in ecological research that are frequently recorded in databases (Guralnick et al., 2016; Moles et al., 2009). However, 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 within 127 orders of vertebrate species from VertNet and EOL (Figure 5). In total, height data are unavailable for approximately 92.6% of plant species, and body length is unavailable 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. Assuming that a trait that is of obvious biological importance and is easily measurable shall have more information in the literature or the biodiversity databases, other biological traits (e.g., life span, metabolic rate, 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

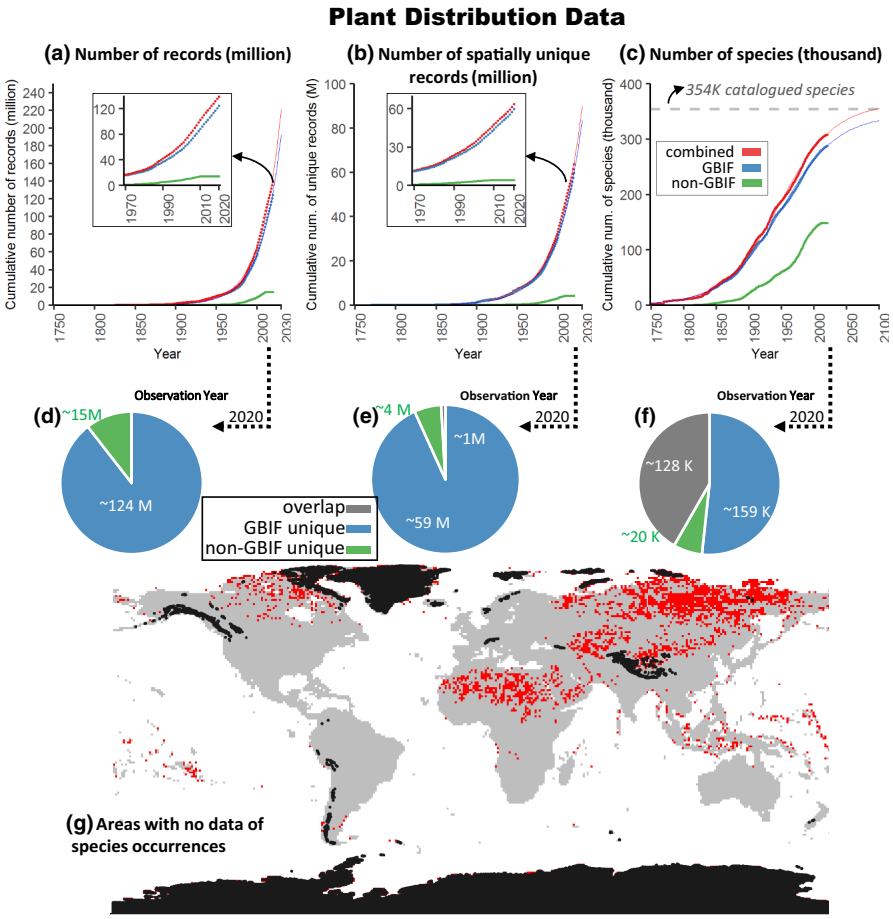


FIGURE 4 Spatial and taxonomic coverage of terrestrial plant occurrence data. Georeferenced plant observations, as illustrated by observation dates in Global Biodiversity Information Facility (GBIF; the largest biodiversity informatics infrastructure), have increased exponentially over the past 200 years (panels a, b), though the number of species recorded in these databases is reaching saturation (panel c). By integrating additional data sources compiled by Botanical Information and Ecology Network (BIEN; i.e., non-GBIF sources comprising ~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 ~20,000 species (panel f). Still, the gaps in plant distributions warrant our attention: areas in Russia, central Asia, and northern Africa (red colour in panel g) are missing publicly available occurrences. The grey colour in panel (g) represents the presence of plant data, and the black colour represents ice-covered areas

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 located, 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 include project summary, geographic, temporal, and taxonomic scope, and key technique attributes (though this has not been accessible since 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 are mainly focused on a few record types, including occurrences, checklists, and sampling events (https://web.archive.org/web/20210514141441/https://www.gbif.org/article/5FIXBKbirSiq0ascKYiA8q/gbif-infrastructure-registry). Another example is the 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 a large variety of biological and environmental data (Michener et al., 2012). These existing efforts form a good basis for a catalogue of biodiversity databases that can continuously keep track of existing data aggregators and index new

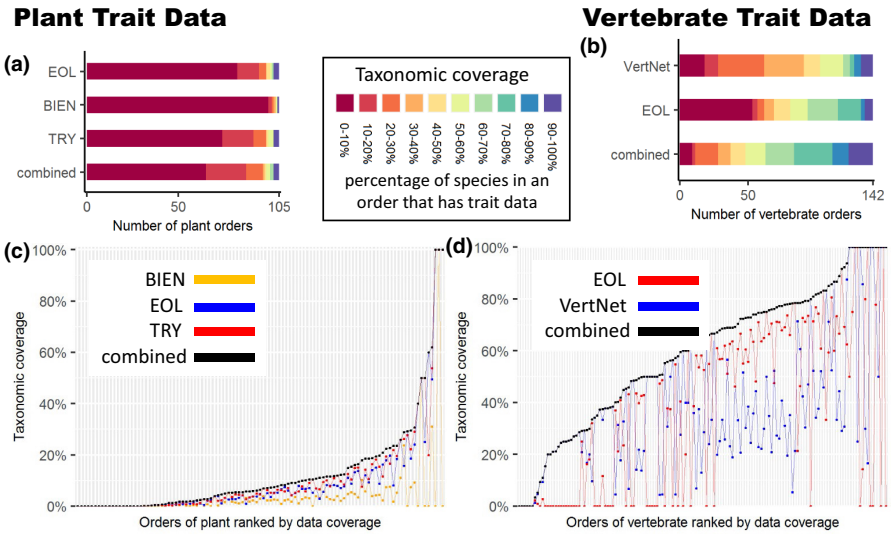


FIGURE 5 Potential for increased taxonomic coverage of plant and vertebrate trait data through data integration. The taxonomic coverage of a database is measured as the percentage of the species in a plant or vertebrate order that has trait data. By combining trait databases, coverage could be expanded in 69–82 plant orders (panel a) and 86–124 vertebrate orders (panel b) compared to individual data sources (panels c & d). Panels (c) and (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). EOL = Encyclopedia of Life; BIEN = Botanical Information and Ecology Network

aggregation efforts. Still, the relationships among biodiversity databases are not always obvious. Therefore, a synthesis, ideally updated regularly, would be helpful to clarify the relationships among biodiversity databases, in particular what is the unique data coverage of each 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 data aggregators, 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 incompatible taxonomic systems. A promising effort is Catalogue of Life Plus (Bánki et al., 2018), which 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 data processing methods and storing standards (Mesibov, 2018), 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 by two databases based on different, commonly automated, georeferencing algorithms, thus likely leading to two different sets of coordinates for the same observation, therefore appearing to represent 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 these community-developed data standards by individual databases would enable more effective database integration. The exemplified issue of one specimen being assigned different coordinates, or more broadly speaking, data duplication due to different processing standards, could be better resolved by implementing permanent, consistent, globally unique identifiers for all records, in particular for new records to be collected, digitized, and integrated (Clark et al., 2004; Guralnick et al., 2015; Page, 2008). The implementation of globally unique identifiers could also provide enhanced support for database integration by facilitating the detection of field discrepancies, such as variation in scientific names and coordinates that derived from one single record, thus avoiding data duplication, as well as facilitating the documentation of data provenance, the exchange of relevant (meta)data, and the linkage between biological and environmental data.

The essential goal of the approaches discussed here is to maximize compatibility, and thus minimize barriers to data-flow and synthesis. Essentially, solving technical barriers will lead to an enhanced ability to find, access, integrate, and reuse biodiversity data. Once such technical barriers are addressed, the integrated content from multiple databases, either for similar types of data or across different research domains, could be organized in multiple non-exclusive ways, including (a) a single centralized database, (b) some decentralized but connected databases (Gallagher et al., 2020), or (c) multiple synced databases (LeBauer et al., 2013). The implementation

of database integration will not be a trivial effort; it needs special skills and considerable computation capacity and needs to be well planned and coordinated by the biodiversity informatics community (Hobern et al., 2019).

4.3 | Outlook for individual aggregators after database integration

Finally, it is worth thinking about the uniqueness and destiny of individual data aggregators post-integration. An individual aggregator may simply be considered a subset of a larger integrated database, though the relevance of individual aggregators could be maintained in several aspects. First, while data integration can occur for particular data elements (e.g., taxon, place, time) facilitated through common use of (meta)data standards, each individual aggregator could still retain unique domain information. For example, while GBIF aggregates species occurrence data from iNaturalist, the latter still uniquely hosts the media data, which are not aggregated by GBIF. Similarly, while eBird frequently publishes bird observations in GBIF, the media data (images, sounds, videos), species count data, as well as range maps are uniquely hosted by eBird. Second, individual aggregators will still perform an irreplaceable role in the initial step of data aggregations, especially if that involves data curation and standardization, which is usually automated but also commonly requires intensive computational effort and relies on expert domain knowledge or interactive supervision. Such effort is not trivial and it forms the basis for downstream data/database integration. Actually, GBIF, a downstream data aggregator, does not allow individual users to directly provide data to the database (with the exception of data papers), but only takes standardized data from organizations, including upstream data aggregators (GBIF, 2021). Third, individual aggregators can also play unique roles for users, even when based on the same shared knowledge base. For example, while ALA and GBIF share data that were collected from Australia (Supporting Information Table S1), ALA comprises a prominent education component with respect to Australian biodiversity for its Australian users, as well as in facilitating scientific research by putting the biodiversity data in the context of their environment. Therefore, the individual aggregators can retain their relevance even post-integration with other databases. Nonetheless, the possibility that smaller data generators and aggregators will be treated as irrelevant in the face of large data aggregators is a legitimate concern. More-intentional approaches are needed to ensure that the relevance of smaller data providers is appropriately maintained and their contributions are adequately acknowledged. One way the relevance and contribution of small data providers could be emphasized more is through the development of data citation mechanisms and tools (e.g., Owens et al., 2021).

On the other hand, there has been a process of specialization of data aggregators 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 data aggregators 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).

5 | CONCLUDING REMARKS

The accelerating influx 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 pressing environmental challenges, new approaches are urgently needed to facilitate the ability to find, access, integrate, and reuse biodiversity data. We demonstrate that rapid progress can be made toward better biodiversity knowledge through overcoming the barriers for data integration from diverse biodiversity 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 duplicate 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 (Taxonomy Decadal Plan Working Group, 2018), and digitization of legacy data (Ariño, 2010), as well as tackle the major impediments to data integration, including taxonomic incompatibility, lags in data exchange, barriers to effective synthesis, and isolation of individual initiatives.

ACKNOWLEDGMENTS

B.J.E., B.J.M., B.B. and C.M. were supported by NSF ABI-1565118. B.J.E., C.M., B.S.M., B.B. were supported by NSF HDR-1934790. B.J.E., B.M., N.J.B.K., C.V., and B.J.M. acknowledge the FREE group funded by the synthesis center CESAB of the French Foundation for Research on Biodiversity (FRB) and EDF. J.-C.S. and B.J.E. acknowledge support from the Center for Informatics Research on Complexity in Ecology (CIRCE), funded by the Aarhus University Research Foundation under the AU Ideas program. This work was conducted as a part of the BIEN Working Group, 2008–2012. We thank all the data contributors and numerous herbaria who have contributed their data to various data compiling organizations (see the Supplementary Materials) for the invaluable data and support provided to BIEN. We thank the New York

Botanical Garden; Missouri Botanical Garden; Utrecht Herbarium; the UNC Herbarium; and GBIF, REMIB, and SpeciesLink. The staff at CyVerse provided critical computational assistance. We thank the more than 50 scientists who participated in our various BIEN working group and subgroup meetings since 2008 including B. Blonder, K. Engemann, E. Fegraus, J. Cavender-Bares, B. Dobrin, K. Gendler, R. Jorgensen, G. Lopez-Gonzalez, L. Zhenyuan, S. McKay, O. Phillips, J. Pickering, N. Swenson, C. Vriesendorp, and K. Woods, who participated in a working group meeting, and D. Ackerly, E. Garnier, R. Guralnick, W. Jetz, J. Macklin, N. Matasci, S. Ramteke, and A. Zanne who participated in subgroup meetings. We also acknowledge the critical support of the University of Arizona high-performance computing resources via the Research Data Center as well as iPlant and CyVerse support from R. Jorgensen, S. Goff, N. Matasci, N. Merchant, M. Narrow, and R. Walls. Furthermore, the long-term vision, encouragement, and computational support of F. Davis, S. Hampton, M. Jones, N. Outin, and the ever-helpful staff at NCEAS were critical for the completion of this first stage of the BIEN working group. Special thanks to K. Koenig for cartographic support. We acknowledge the herbaria that contributed data to this work: A, AAH, AAS, AAU, ABH, ACAD, ACOR, AD, AFS, AK, AKPM, ALCB, ALTA, ALU, AMD, AMES, AMNH, AMO, ANGU, ANSM, ANSP, AQP, ARAN, ARIZ, AS, ASDM, ASU, AUT, AV, AWH, B, BA, BAA, BAB, BABY, BACP, BAF, BAFC, BAI, BAJ, BAL, BARC, BAS, BBB, BBS, BC, BCMEX, BCN, BCRU, BEREA, BESA, BG, BH, BHCB, BIO, BISH, BLA, BM, BOCH, BOL, BOLV, BONN, BOON, BOTU, BOUM, BPI, BR, BREM, BRI, BRIT, BRLU, BRM, BSB, BUT, C, CALI, CAN, CANB, CANU, CAS, CATA, CATIE, CAY, CBM, CDA, CDBI, CEN, CEPEC, CESJ, CGE, CGMS, CHAM, CHAPA, CHAS, CHR, CHSC, CIB, CICY, CIIDIR, CIMI, CINC, CLEMS, CLF, CMM, CMMEX, CNPO, CNS, COA, COAH, COCA, CODAGEM, COFC, COL, COLO, CONC, CORD, CP, CPAP, CPUN, CR, CRAI, CRP, CS, CSU, CSUSB, CTES, CTESN, CU, CUVC, CUZ, CVRD, DAO, DAV, DBG, DBN, DES, DLF, DNA, DPU, DR, DS, DSM, DUKE, DUSS, E, EA, EAC, EAN, EBUM, ECON, EIF, EIU, EMMA, ENCB, ER, ERA, ESA, ETH, F, FAA, FAU, FAUC, FB, FCME, FCO, FCQ, FEN, FHO, FI, FLAS, FLOR, FM, FR, FRU, FSU, FTG, FUEL, FULD, FURB, G, GAT, GB, GDA, GENT, GES, GH, GI, GLM, GMDRC, GMNHJ, GOET, GRA, GUA, GZU, H, HA, HAC, HAL, HAM, HAMAB, HAO, HAS, HASU, HB, HBG, HBR, HCIB, HEID, HGM, HIB, HIP, HNT, HO, HPL, HRCB, HRP, HSC, HSS, HU, HUA, HUAA, HUAL, HUAZ, HUCP, HUEFS, HUEM, HUFU, HUJ, HUSA, HUT, HXBH, HYO, IAA, IAC, IAN, IB, IBGE, IBK, IBSC, IBUG, ICEL, ICESI, ICN, IEA, IEB, ILL, ILLS, IMSSM, INB, INEGI, INIF, INM, INPA, IPA, IPRN, IRVC, ISC, ISKW, ISL, ISTC, ISU, IZAC, IZTA, JACA, JBAG, JBGP, JCT, JE, JEPS, JOTR, JROH, JUA, JYV, K, KIEL, KMN, KMNH, KOELN, KOR, KPM, KSC, KSTC, KSU, KTU, KU, KUN, KYO, L, LA, LAGU, LBG, LD, LE, LEB, LIL, LINC, LINN, LISE, LISI, LISU, LL, LMS, LOJA, LOMA, LP, LPAG, LPB, LPD, LPS, LSU, LSUM, LTB, LTR, LW, LYJB, LZ, M, MA, MACF, MAF, MAK, MARS, MARY, MASS, MB, MBK, MBM, MBML, MCNS, MEL, MELU, MEN, MERL, MEXU, MFA, MFU, MG, MGC, MICH, MIL, MIN, MISSA, MJG, MMMN, MNHM, MNHN, MO, MOL, MOR, MPN, MPU, MPUC, MSB, MSC, MSUN, MT, MTMG, MU, MUB, MUR, MVFA, MVFQ, MVJB, MVM, MW, MY, N, NA, NAC, NAS, NCU, NE, NH, NHM, NHMC, NHT, NLH, NM, NMB, NMNL, NMR, NMSU, NSPM, NSW, NT, NU, NUM, NY, NZFRI, O, OBI, ODU, OS, OSA, OSC, OSH, OULU, OWU, OXF, P, PACA, PAMP, PAR, PASA, PDD, PE, PEL, PERTH, PEUFR, PFC, PGM, PH, PKDC, PLAT, PMA, POM, PORT, PR, PRC, PRE, PSU, PY, QCA, QCNE, QFA, QM, QRS, QUE, R, RAS, RB, RBR, REG, RELC, RFA, RIOC, RM, RNG, RSA, RYU, S, SACT, SALA, SAM, SAN, SANT, SAPS, SASK, SAV, SBBG, SBT, SCFS, SD, SDSU, SEL, SEV, SF, SFV, SGO, SI, SIU, SJRP, SJSU, SLPM, SMDB, SMF, SNM, SOM, SP, SPF, SPSF, SQF, SRFA, STL, STU, SUU, SVG, TAES, TAI, TAIF, TALL, TAM, TAMU, TAN, TASH, TEF, TENN, TEPB, TEX, TFC, TI, TKPM, TNS, TO, TOYA, TRA, TRH, TROM, TRT, TRTE, TU, TUB, U, UADY, UAM, UAMIZ, UB, UBC, UC, UCMM, UCR, UCS, UCSB, UCSC, UEC, UESC, UFG, UFMA, UFMT, UFP, UFRJ, UFRN, UFS, UGDA, UH, UI, UJAT, ULM, ULS, UME, UMO, UNA, UNB, UNCC, UNEX, UNITEC, UNL, UNM, UNR, UNSL, UPCB, UPEI, UPNA, UPS, US, USAS, USF, USJ, USM, USNC, USP, USZ, UT, UTC, UTEP, UU, UVIC, UWO, V, VAL, VALD, VDB, VEN, VIT, VMSL, VT, W, WAG, WAT, WELT, WFU, WII, WIN, WIS, WMNH, WOLL, WS, WTU, WU, XAL, YAMA, Z, ZMT, ZSS, and ZT. The BIEN working group was supported by the National Center for Ecological Analysis and Synthesis, a center funded by NSF EF-0553768 at the University of California, Santa Barbara and the State of California. Additional support for the BIEN working group was provided by iPlant/CyVerse via NSF DBI-0735191.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

X.F., D.S.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.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., E.A.N. and L.L.-H. designed the data visualizations with constructive discussion from M.M.N.-R., 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., B.B., B.S.M., K.E., E.B., and L.L.-H. contributed to interpretation and writing. L.L.-H. and B.J.E. provided supervision. L.L.-H., B.J.E., K.E., and D.D.B. led 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.-F., R.K.P., M.P., P.R.R., B.S., J.M.S.-D., I.Š., J.-C.S., C.V., S.W. contributed to data assembly and development of the BIEN database. All authors contributed to interpreting the results and the editing of drafts.

DATA AVAILABILITY STATEMENT

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 the

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

ORCID

Xiao Feng https://orcid.org/0000-0003-4638-3927
Brian J. Enquist https://orcid.org/0000-0002-6124-7096
Daniel S. Park https://orcid.org/0000-0003-2783-530X
Rachael V. Gallagher https://orcid.org/0000-0002-4680-8115
Erica A. Newman https://orcid.org/0000-0001-6433-8594
Cory Merow https://orcid.org/0000-0003-0561-053X
Yaoqi Li https://orcid.org/0000-0001-6540-395X
Pablo A. Marquet https://orcid.org/0000-0001-6369-9339
Naia Morueta-Holme https://orcid.org/0000-0002-0776-4092
Robert K. Peet https://orcid.org/0000-0003-2823-6587
Patrick R. Roehrdanz https://orcid.org/0000-0003-4047-5011
Brody Sandel https://orcid.org/0000-0003-2162-6902
Jens-Christian Svenning https://orcid.org/0000-0002-3415-0862
Cyrille Violle https://orcid.org/0000-0002-2471-9226

REFERENCES

Ariño, A. H. (2010). Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics*, *7*, 81–92. https://doi.org/10.17161/bi.v7i2.3991

Ball-Damerow, J. E., Brenskelle, L., Barve, N., Soltis, P. S., Sierwald, P., Bieler, R., LaFrance, R., Ariño, A. H., & Guralnick, R. P. (2019). Research applications of primary biodiversity databases in the digital age. *PLoS One*, *14*, e0215794. https://doi.org/10.1371/journal.pone.0215794

Bánki, O., Döring, M., Holleman, A., & Addink, W. (2018). Catalogue of life plus: Innovating the CoL systems as a foundation for a clearinghouse for names and taxonomy. *Biodiversity Information Science and Standards*.

Barthlott, W., Hostert, A., Kier, G., Küper, W., Kreft, H., Mutke, J., Rafiqpoor, M. D., & Sommer, J. H. (2007). Geographic patterns of vascular plant diversity at continental to global scales (Geographische Muster der Gefäßpflanzenvielfalt im kontinentalen und globalen Maßstab). *Erdkunde*, *61*, 305–315. https://doi.org/10.3112/erdkunde.2007.04.01

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. https://doi.org/10.1111/j.1600-0587.2012.07364.x

Belbin, L. (2014). Biodiversity Information Projects of the World. http://old.tdwg.org/biodiv-projects/

Belbin, L., Chapman, A., Wieczorek, J., Zermoglio, P. F., Thompson, A. B., & Morris, P. J. (2018). Data quality task group 2: Tests and assertions. *Biodiversity Information Science and Standards*, *2*, e25608. https://doi.org/10.3897/biss.2.25608

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. https://doi.org/10.3897/BDJ.9.e65023

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: IJGIS*, *30*, 108–125. https://doi.org/10.1080/13658816.2015.1077962

Bennett, J. M., Calosi, P., Clusella-Trullas, S., Martínez, B., Sunday, J., Algar, A. C., Araújo, M. B., Hawkins, B. A., Keith, S., Kühn, I., Rahbek, C., Rodríguez, L., Singer, A., Villalobos, F., Ángel Olalla-Tárraga, M., & Morales-Castilla, I. (2018). GlobTherm, a global database on thermal tolerances for aquatic and terrestrial organisms. *Scientific Data*, *5*, 180022. https://doi.org/10.1038/sdata.2018.22

Berendsohn, W. G. (1997). A taxonomic information model for botanical databases: The IOPI Model. *Taxon*, *46*, 283–309. https://doi.org/10.2307/1224098

Bisby, F. A. (2000). The quiet revolution: Biodiversity informatics and the internet. *Science*, *289*, 2309–2312. https://doi.org/10.1126/science.289.5488.2309

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. https://doi.org/10.3897/BDJ.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. https://doi.org/10.1186/1471-2105-14-16

Catalogue of Life. (2021). *Species 2000 & ITIS catalogue of life*. Catalogue of Life.

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances*, *1*, e1400253. https://doi.org/10.1126/sciadv.1400253

Chamberlain, S. A., & Szöcs, E. (2013). taxize: Taxonomic search and retrieval in R. *F1000Research*, *2*, 191. https://doi.org/10.12688/f1000research.2-191.v1

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., Rosemartin, A., & Turak, E. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *213*, 280–294. https://doi.org/10.1016/j.biocon.2016.09.004

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. https://doi.org/10.1038/35012241

Chesser, R. T., Billerman, S. M., Burns, K. J., Cicero, C., Dunn, J. L., Hernández-Baños, B. E., Kratter, A. W., Lovette, I. J., Mason, N. A., Rasmussen, P. C., Remsen, J. V. Jr, Stotz, D. F., & Winker, K. (2021). Sixty-second supplement to the American ornithological society's check-list of north American birds. *The Auk*, *138*, ukab037.

Christenhusz, M. J. M., & Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa*, *261*, 201. https://doi.org/10.11646/phytotaxa.261.3.1

Clark, J. A., & May, R. M. (2002). Taxonomic bias in conservation research. *Science*, *297*, 191–192. https://doi.org/10.1126/science.297.5579.191b

Clark, T., Martin, S., & Liefeld, T. (2004). Globally distributed object identification for biological knowledgebases. *Briefings in Bioinformatics*, *5*, 59–70. https://doi.org/10.1093/bib/5.1.59

Conde, D. A., Staerk, J., Colchero, F., da Silva, R., Schöley, J., Baden, H. M., Jouvet, L., Fa, J. E., Syed, H., Jongejans, E., Meiri, S., Gaillard, J.-M., Chamberlain, S., Wilcken, J., Jones, O. R., Dahlgren, J. P., Steiner, U. K., Bland, L. M., Gomez-Mestre, I., … Vaupel, J. W. (2019). Data gaps and opportunities for comparative and conservation biology. *Proceedings of the National Academy of Sciences*, *116*, 9658–9664. https://doi.org/10.1073/pnas.1816367116

Constable, H., Guralnick, R., Wieczorek, J., Spencer, C., Peterson, A. T., & Committee, V. N. S. (2010). VertNet: A new model for biodiversity

data sharing. *PLoS Biology*, *8*, e1000309. https://doi.org/10.1371/journal.pbio.1000309

Cornwell, W. K., Pearse, W. D., Dalrymple, R. L., & Zanne, A. E. (2019). What we (don't) know about global plant diversity. *Ecography*, *42*, 1819–1831. https://doi.org/10.1111/ecog.04481

Daru, B. H., Park, D. S., Primack, R. B., Willis, C. G., Barrington, D. S., Whitfeld, T. J. S., Seidler, T. G., Sweeney, P. W., Foster, D. R., Ellison, A. M., & Davis, C. C. (2017). Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist*, *217*, 939–955. https://doi.org/10.1111/nph.14855

Deans, A. R., Lewis, S. E., Huala, E., Anzaldo, S. S., Ashburner, M., Balhoff, J. P., Blackburn, D. C., Blake, J. A., Burleigh, J. G., Chanet, B., Cooper, L. D., Courtot, M., Csösz, S., Cui, H., Dahdul, W., Das, S., Dececchi, T. A., Dettai, A., Diogo, R., … Mabee, P. (2015). Finding our way through phenotypes. *PLoS Biology*, *13*, e1002033. https://doi.org/10.1371/journal.pbio.1002033

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. https://doi.org/10.1111/j.1654-1103.2011.01265.x

Díaz, S., Settele, J., Brondízio, E. S., Ngo, H. T., Guèze, M., Agard, J., Arneth, A., Balvanera, P., Brauman, K., Butchart, S. H., Chan, K. M., Garibaldi, L. A., Ichii, K., Liu, J., Mazhenchery Subramanian, S., Midgley, G., Miloslavich, P., Molnár, Z., Obura, D., … Zayas, C. (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*. Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services.

eBird. (2014). eBird's missing species. *eBird*.

eBird. (2021). eBird record format. *eBird*.

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.

Enquist, B. J., Feng, X., Boyle, B., Maitner, B., Newman, E. A., Jørgensen, P. M., Roehrdanz, P. R., Thiers, B. M., Burger, J. R., Corlett, R. T., Couvreur, T. L. P., Dauby, G., Donoghue, J. C., Foden, W., Lovett, J. C., Marquet, P. A., Merow, C., Midgley, G., Morueta-Holme, N., … McGill, B. J. (2019). The commonness of rarity: Global and future distribution of rarity across land plants. *Science Advances*, *5*, eaaz0414. https://doi.org/10.1126/sciadv.aaz0414

EOL. (2014). TraitBank. *Encyclopedia of Life*.

Escribano, N., Galicia, D., & Ariño, A. H. (2018). The tragedy of the biodiversity data commons: A data impediment creeping nigher? *Database: the Journal of Biological Databases and Curation*, *2018*, bay033. https://doi.org/10.1093/database/bay033

Fazey, I., Fischer, J., & Lindenmayer, D. B. (2005). What do conservation biologists publish? *Biological Conservation*, *124*, 63–73. https://doi.org/10.1016/j.biocon.2005.01.013

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. https://doi.org/10.1038/s41559-019-0972-5

Fitzgerald, B. F., Coates, J. M., & Lewis, S. M. (2007). *Open content licensing: Cultivating the creative commons*. Sydney University Press.

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

Franz, N. M., Peet, R. K., & Weakley, A. S. (2008). On the use of taxonomic concepts in support of biodiversity research and taxonomy. *Systematics Association Special Volume*, *76*, 63.

Gallagher, R. V., Falster, D. S., Maitner, B. S., Salguero-Gómez, R., Vandvik, V., Pearse, W. D., Schneider, F. D., Kattge, J., Poelen, J. H., Madin, J. S., Ankenbrand, M. J., Penone, C., Feng, X., Adams, V. M., Alroy, J., Andrew, S. C., Balk, M. A., Bland, L. M., Boyle, B. L., … Enquist, B. J. (2020). Open Science principles for accelerating trait-based science across the Tree of Life. *Nature Ecology & Evolution*, *4*, 294–303. https://doi.org/10.1038/s41559-020-1109-6

GBIF. (2014). New approaches to data licensing and endorsement. *GBIF*.

GBIF. (2021). Quick guide to publishing data through GBIF.org. *GBIF*.

Guralnick, R. P., Cellinese, N., Deck, J., Pyle, R. L., Kunze, J., Penev, L., Walls, R., Hagedorn, G., Agosti, D., Wieczorek, J., Catapano, T., & Page, R. D. M. (2015). Community next steps for making globally unique identifiers work for biocollections data. *ZooKeys*, *494*, 133–154. https://doi.org/10.3897/zookeys.494.9352

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. https://doi.org/10.1111/ecog.02942

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: The Journal of Biological Databases and Curation*, *2016*, baw158. https://doi.org/10.1093/database/baw158

Hardisty, A., Roberts, D., & Community, B. I. (2013). A decadal view of biodiversity informatics: Challenges and priorities. *BMC Ecology*, *13*, 16. https://doi.org/10.1186/1472-6785-13-16

Hecnar, S. J. (2009). Human bias and the biodiversity knowledge base: An examination of the published literature on vertebrates. *Biodiversity*, *10*, 18–24. https://doi.org/10.1080/14888386.2009.9712633

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*, 280–299. https://doi.org/10.1353/lib.0.0036

Hobern, D., Baptiste, B., Copas, K., Guralnick, R., Hahn, A., van Huis, E., Kim, E.-S., McGeoch, M., Naicker, I., Navarro, L., Noesgaard, D., Price, M., Rodrigues, A., Schigel, D., Sheffield, C. A., & Wieczorek, J. (2019). Connecting data and expertise: A new alliance for biodiversity knowledge. *Biodiversity Data Journal*, *7*, e33679. https://doi.org/10.3897/BDJ.7.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. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., Yang, Q., Zhu, C., & Qiao, H. (2021). Sampling biases shape our view of the natural world. *Ecography*, *44*, 1259–1269. https://doi.org/10.1111/ecog.05926

iDigBio. (2018). Integrated digitized biocollections. *Integrated Digitized Biocollections*.

iDigBio. (2021). Data ingestion guidance. *iDigBio*.

iNaturalist. (2018). Research grade observations. *iNaturalist*.

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. https://doi.org/10.1002/aps3.1024

Jetz, W., McPherson, J. M., & Guralnick, R. P. (2012). Integrating biodiversity distribution knowledge: Toward a global map of life. *Trends in Ecology & Evolution*, *27*, 151–159. https://doi.org/10.1016/j.tree.2011.09.007

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. https://doi.org/10.1098/rspb.2010.1004

Kalwij, J. M. (2012). Review of "The Plant List, a working list of all plant species". *Journal of Vegetation Science*, *23*, 998–1002. https://doi.org/10.1111/j.1654-1103.2012.01407.x

Kattge, J., Bönisch, G., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Tautenhahn, S., Werner, G. D. A., Aakala, T., Abedi, M., Acosta, A. T. R., Adamidis, G. C., Adamson, K., Aiba, M., Albert, C. H., Alcántara, J. M., Alcázar C, C., Aleixo, I., Ali, H., … Wirth, C. (2020). TRY plant

trait database – Enhanced coverage and open access. *Global Change Biology*, *26*, 119–188. https://doi.org/10.1111/gcb.14904

Kattge, J., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P. B., Wright, I. J., Cornelissen, J. H. C., Violle, C., Harrison, S. P., Van Bodegom, P. M., Reichstein, M., Enquist, B. J., Soudzilovskaia, N. A., Ackerly, D. D., Anand, M., … Wirth, C. (2011). TRY – A global database of plant traits. *Global Change Biology*, *17*, 2905–2935. https://doi.org/10.1111/j.1365-2486.2011.02451.x

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. https://doi.org/10.1890/12-0137.1

Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, *113*, 5970–5975. https://doi.org/10.1073/pnas.1521291113

Lomolino, M. (2004). Conservation biogeography. *Frontiers of Biogeography*, *293*, 1–3.

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

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. https://doi.org/10.11646/phytotaxa.272.1.5

MacFadden, B. J., & Guralnick, R. P. (2017). Horses in the Cloud: Big data exploration and mining of fossil and extant Equus (Mammalia: Equidae). *Paleobiology*, *43*, 1–14.

Mackenzie, D. I. (2005). Was it there? Dealing with imperfect detection for species presence/absence data+. *Australian & New Zealand Journal of Statistics*, *47*, 65–74. https://doi.org/10.1111/j.1467-842X.2005.00372.x

Mesibov, R. (2018). An audit of some processing effects in aggregated occurrence records. *ZooKeys*, *751*, 129–146. https://doi.org/10.3897/zookeys.751.24791

Meyer, C., Kreft, H., Guralnick, R. P., & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*, *6*, 82218. https://doi.org/10.1038/ncomms9221

Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, *19*, 992–1006. https://doi.org/10.1111/ele.12624

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. https://doi.org/10.1016/j.ecoinf.2011.08.007

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. *The Journal of Ecology*, *97*, 923–932. https://doi.org/10.1111/j.1365-2745.2009.01526.x

Moore, G. E. (2006). Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff. *IEEE Solid-State Circuits Society Newsletter*, *11*, 33–35.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., & Worm, B. (2011). How many species are there on earth and in the ocean? *PLoS Biology*, *9*, e1001127. https://doi.org/10.1371/journal.pbio.1001127

Nelson, G., & Ellis, S. (2018). The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *374*, 20170391. https://doi.org/10.1098/rstb.2017.0391

Oliveira, B. F., São-Pedro, V. A., Santos-Barrera, G., Penone, C., & Costa, G. C. (2017). AmphiBIO, a global database for amphibian ecological traits. *Scientific Data*, *4*, 170123. https://doi.org/10.1038/sdata.2017.123

Owens, H. L., Merow, C., Maitner, B. S., Kass, J. M., Barve, V., & Guralnick, R. P. (2021). occCite: Tools for querying and managing large biodiversity occurrence datasets. *Ecography*, *44*, 1228–1235. https://doi.org/10.1111/ecog.05618

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. https://doi.org/10.1093/biosci/biv104

Page, R. D. M. (2008). Biodiversity informatics: The challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics*, *9*, 345–354. https://doi.org/10.1093/bib/bbn022

Park, D. S., Newman, E. A., & Breckheimer, I. K. (2021). Scale gaps in landscape phenology: Challenges and opportunities. *Trends in Ecology & Evolution*, *36*, 709–721. https://doi.org/10.1016/j.tree.2021.04.008

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. *Biodiversity Data Journal*, *2*, e1079. https://doi.org/10.3897/BDJ.2.e1079

Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., Bruford, M. W., Brummitt, N., Butchart, S. H. M., Cardoso, A. C., Coops, N. C., Dulloo, E., Faith, D. P., Freyhof, J., Gregory, R. D., Heip, C., Höft, R., Hurtt, G., Jetz, W., … Wegmann, M. (2013). Essential biodiversity variables. *Science*, *339*, 277–278. https://doi.org/10.1126/science.1229931

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reddy, S., & Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, *30*, 1719–1727. https://doi.org/10.1046/j.1365-2699.2003.00946.x

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, *331*, 703–705. https://doi.org/10.1126/science.1197962

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. https://doi.org/10.1371/journal.pone.0102623

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. https://doi.org/10.1111/2041-210X.13288

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. https://doi.org/10.1186/s40663-017-0120-0

Shashkov, M., & Ivanova, N. (2019). Considerable progress in Russian GBIF community. *Biodiversity Information Science and Standards*, *3*, e37015. https://doi.org/10.3897/biss.3.37015

Siefert, A., Violle, C., Chalmandrier, L., Albert, C. H., Taudiere, A., Fajardo, A., Aarssen, L. W., Baraloto, C., Carlucci, M. B., Cianciaruso, M. V., L. Dantas, V., Bello, F., Duarte, L. D. S., Fonseca, C. R., Freschet, G. T., Gaucherand, S., Gross, N., Hikosaka, K., Jackson, B., … Wardle, D. A. (2015). A global meta-analysis of the relative extent of intraspecific trait variation in plant communities. *Ecology Letters*, *18*, 1406–1419. https://doi.org/10.1111/ele.12508

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.

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

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. https://doi.org/10.1177/2053019614564785

Stork, N. E. (2018). How many species of insects and other terrestrial arthropods are there on earth? *Annual Review of Entomology*, *63*, 31–45. https://doi.org/10.1146/annurev-ento-020117-043348

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W. M., Iliff, M. J., Lagoze, C., La Sorte, F. A., … Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

Taxonomy Decadal Plan Working Group. (2018). *Discovering biodiversity: A decadal plan for taxonomy and biosystematics in Australia and New Zealand 2018–2027*. Australian Academy of Science, Royal Society of New Zealand.

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. https://doi.org/10.1371/journal.pone.0189577

U.S. Geological Survey. (2018). Biodiversity Information Serving Our Nation (BISON). https://bison.usgs.gov

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. https://doi.org/10.3732/ajb.1200222

Weigelt, P., König, C., & Kreft, H. (2020). GIFT – A Global Inventory of Floras and Traits for macroecology and biogeography. *Journal of Biogeography*, *47*, 16–43. https://doi.org/10.1111/jbi.13623

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. https://doi.org/10.1371/journal.pone.0029715

Wiser, S. K., Spencer, N., De Cáceres, M., Kleikamp, M., Boyle, B., & Peet, R. K. (2011). Veg-X – An exchange standard for plot-based vegetation data. *Journal of Vegetation Science*, *22*, 598–609. https://doi.org/10.1111/j.1654-1103.2010.01245.x

Zermoglio, P. F., Guralnick, R. P., & Wieczorek, J. R. (2016). A standardized reference data set for vertebrate taxon name resolution. *PLoS One*, *11*, e0146894. https://doi.org/10.1371/journal.pone.0146894

BIOSKETCH

The authors share common research interest in biodiversity informatics, with special interest in developing informatics infrastructures to best integrate data and to serve the community in order to advance the discovery, study, and preservation of biodiversity.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Feng, X., Enquist B. J., Park D. S., Boyle B., Breshears D. D., Gallagher R. V., Lien A., Newman E. A., Burger J. R., Maitner B. S., Merow C., Li Y., Huynh K. M., Ernst K., Baldwin E., Foden W., Hannah L., Jørgensen P. M., Kraft N. J. B., … López-Hoffman L. (2022). A review of the heterogeneous landscape of biodiversity databases: Opportunities and challenges for a synthesized biodiversity knowledge base. *Global Ecology and Biogeography*, *31*, 1242–1260. https://doi.org/10.1111/geb.13497

APPENDIX A 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 are 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; https://www.gbif.org/), Global Inventory of Floras and Traits (GIFT; Weigelt et al., 2020), Integrated Digitized Biocollections (iDigBio, 2018), iNaturalist (https://www.inaturalist.org/), 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 12 databases we examined were chosen to maximize the generalizability of our results and conclusions. They were also among the most commonly used, well-established, and large-scale biodiversity databases (Chandler et al., 2017; Cornwell et al., 2019; James et al., 2018; König et al., 2019; MacFadden & Guralnick, 2017; Singer et al., 2018). Selections were 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, although we note that 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 type (image, audio, or video), and biological information (standardized trait databases or generalized text descriptions). Within the category of 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 types among the reviewed databases. We assessed the integration status by taxonomic groups, which are all organisms, plants, or vertebrates.

IMPROVEMENT OF DATA COVERAGE THROUGH DATABASE INTEGRATION

To quantify the improvement of data coverage that could be gained by combining multiple databases, we compared leading databases that focus on similar taxonomic groups and record types. 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 (Ball-Damerow et al., 2019; Clark & May, 2002; Cornwell et al., 2019; Fazey et al., 2005; Hecnar, 2009; Kattge et al., 2020; König et al., 2019; Titley et al., 2017). We did not use taxa that account for large portions of biodiversity on Earth but face huge data gaps, such as microbes and invertebrates (Locey & Lennon, 2016). We compared (a) plant distribution data from GBIF and non-GBIF sources compiled by BIEN (Enquist et al., 2016), (b) plant trait data (i.e., plant height) from BIEN, TRY, GIFT, and EOL, and (c) animal trait data (i.e., vertebrate body length) from VertNet and EOL.

We obtained 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* (https://www.fia.fs.fed.us/) and *NeoTropTree* (http://www.neotroptree.info/). The GBIF and non-GBIF sources have been fused through a series of data scrubbing and standardization workflows [e.g., Taxonomic Name Resolution Service (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 counted 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 the World Geodetic System 1984 (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*, using *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 temporal trends of data accumulation, we quantified the cumulative numbers of observations made over time, from 1750 to present (2020). Note that we only considered presence data and not absence data. This is because true absences are relatively rare (Mackenzie, 2005), and gathering absences is not usually the goal of field surveys. Also, the observed absences may not accurately represent a species' absence, but rather an artefact of species' detection or limited dispersal ability rather than environmental unsuitability.

To describe and quantify those temporal trends, we fitted the cumulative numbers (dependent variable) and years (independent variable) with simple linear (Equation A1), exponential (Equation A2), and logistic regression (Equation A3) using ordinary least squares ['nls' function in stats package version 3.4.2 in R version 3.4.2 (R Core Team 2021)]:

$$y = a + b * x \tag{A1}$$

$$y = e^{a+b*x} \tag{A2}$$

$$y = \frac{a}{1 + e^{-b-c*x}} \tag{A3}$$

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 (AIC) value. 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 retained only currently accepted names to reduce uncertainty (Berendsohn, 1997; Franz & Peet, 2009; e.g., TNRS; Boyle et al., 2013), which yield comparable temporal patterns.

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 the 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* (*COL*; 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 has 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 has 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 data and percentage improvement by using individual datasets versus the combined dataset. Body length data 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 & Szöcs, 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; Boyle et al., 2013; Franz & Peet, 2009). 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.