

ESSAY

Biodiversity data integration—the significance of data resolution and domain

Christian König^{1*}, Patrick Weigelt¹, Julian Schrader¹, Amanda Taylor¹, Jens Kattge^{2,3}, Holger Kreft^{1,4}

1 Biodiversity, Macroecology & Biogeography, University of Goettingen, Goettingen, Germany, **2** Research Group Functional Biogeography, Max Planck Institute for Biogeochemistry, Jena, Germany, **3** German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany, **4** Centre of Biodiversity and Sustainable Land Use (CBL), University of Goettingen, Goettingen, Germany

* chr.koenig@outlook.com



Abstract

Recent years have seen an explosion in the availability of biodiversity data describing the distribution, function, and evolutionary history of life on earth. Integrating these heterogeneous data remains a challenge due to large variations in observational scales, collection purposes, and terminologies. Here, we conceptualize widely used biodiversity data types according to their domain (what aspect of biodiversity is described?) and informational resolution (how specific is the description?). Applying this framework to major data providers in biodiversity research reveals a strong focus on the disaggregated end of the data spectrum, whereas aggregated data types remain largely underutilized. We discuss the implications of this imbalance for the scope and representativeness of current macroecological research and highlight the synergies arising from a tighter integration of biodiversity data across domains and resolutions. We lay out effective strategies for data collection, mobilization, imputation, and sharing and summarize existing frameworks for scalable and integrative biodiversity research. Finally, we use two case studies to demonstrate how the explicit consideration of data domain and resolution helps to identify biases and gaps in global data sets and achieve unprecedented taxonomic and geographical data coverage in macroecological analyses.

OPEN ACCESS

Citation: König C, Weigelt P, Schrader J, Taylor A, Kattge J, Kreft H (2019) Biodiversity data integration—the significance of data resolution and domain. *PLoS Biol* 17(3): e3000183. <https://doi.org/10.1371/journal.pbio.3000183>

Academic Editor: Georgina M. Mace, University College London, UNITED KINGDOM

Received: June 25, 2018

Published: March 18, 2019

Copyright: © 2019 König et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: API, Application Programming Interface; BIEN, Botanical Information and Ecology Network; CHELSA, Climatologies at High resolution for the Earth's land Surface Areas; DataONE, Data Observation Network for Earth; DOI, Digital Object Identifier; GBIF, Global Biodiversity Information Facility; GIFT, Global Inventory of Floras and Traits; ILTER, International Long Term Ecological Research network; LSID, Life Science Identifier;

The biosphere is facing unprecedented pressure from habitat loss, climate change, and the introduction of nonnative species [1–3]. To better understand how biodiversity will be affected under changing environmental conditions, data from multiple ecological disciplines have to be integrated across a wide range of spatiotemporal scales [4,5]. Significant progress towards this objective has been made in recent years. Initiatives such as the Global Biodiversity Information Facility (GBIF) [6], TRY [7], sPlot [8], and GenBank [9] provide access to massive collections of biological data that drive increasingly comprehensive macroecological analyses [10–12]. At the same time, it is becoming clearer that the naïve accumulation of evermore data will not resolve the widespread gaps and biases in ecological data sets [13–15]. A more systematic understanding

ORCID, Open Researcher and Contributor ID; USGS, United States Geological Survey; WCSP, World Checklist of Selected Plant families.

of biodiversity data types, their applications, and their synergetic potentials is needed. Focusing on vascular plants, our aim here is to help build such an understanding in order to improve the coverage, representativeness, and usefulness of global biodiversity data.

Data domains, types, and resolutions

Biodiversity research is organized into domains that cover distinct spheres of knowledge, e.g., of the taxonomy, geographical distribution, or functional traits of organisms [16]. For an effective integration and utilization of biodiversity data, two domains are of particular importance. Biogeography, on the one hand, studies the distribution of life across space and time [17], providing a key link between organisms and their environment. Biogeographical data can therefore be linked to a wide range of organismic (e.g., taxonomic, functional, phylogenetic) and environmental (e.g., climate, soil, topography) information. Functional ecology, on the other hand, aims to approximate the ecological strategy of organisms by means of measurable traits that vary across species [18] and offers a mechanistic approach to understanding ecological patterns and processes. Functional biogeography—the combination of these two domains—allows for the systematic study of trait variation along biotic and abiotic gradients at different scales and represents a promising approach to build a mechanistic understanding of plant diversity [19]. For these reasons, we focus our discussion on the key domains of biogeography and functional ecology.

A domain is typically associated with a set of domain-specific data types (Fig 1). Species distributions, e.g., can be represented by point occurrences, vegetation plots, checklists, or range maps. Functional trait data may come as field measurements for individual organisms or as aggregated values for populations, species, higher taxa (e.g., genera, families) or functional groups (e.g., plant functional types). Additionally, some biodiversity data types combine information from multiple domains, e.g., regional Floras representing sources of both distributional and functional information.

Biodiversity data types provide information at varying resolutions. Although the concept of resolution has substantially improved our understanding of spatial biodiversity patterns [20,21], it is less commonly used in other contexts. However, resolution is a general property of biodiversity data that can be understood as the degree of ecological generalization represented by a given data type. Highly disaggregated data, e.g., point occurrences or trait measurements, represent a single sampling event for a particular individual at a certain location and time. In contrast, highly aggregated data, e.g., Floras or taxonomic monographs, provide a more general account of biodiversity across large spatial, temporal, and taxonomic scales. There is a fundamental trade-off between fine-scale precision and large-scale representativeness across the data resolution spectrum. Although disaggregated data provide the necessary detail to address questions at the level of populations or communities [8,22], they tend to be less complete and representative at macroecological scales [13,15]. Aggregated data, on the other hand, are limited in their capacity to resolve fine-grained ecological patterns but usually provide higher completeness and representativeness at large scales. This trade-off, which, too, has been mostly described in geographical contexts [23,24], is highly relevant for the precision and accuracy of macroecological inferences [16,25].

Most projects for the integration of biodiversity data focus on the disaggregated end of the data spectrum (e.g., GBIF [6], Botanical Information Network and Ecology Network (BIEN) [26], sPlot [29], or TRY [7]). Given the above-mentioned trade-offs and their implications for large-scale coverage and representativeness of biodiversity data, a stronger consideration of aggregated data (e.g., World Checklist of Selected Plant families [WCSP] [27] or Global Inventory of Floras and Traits [GIFT] [28]) seems instrumental for establishing robust global

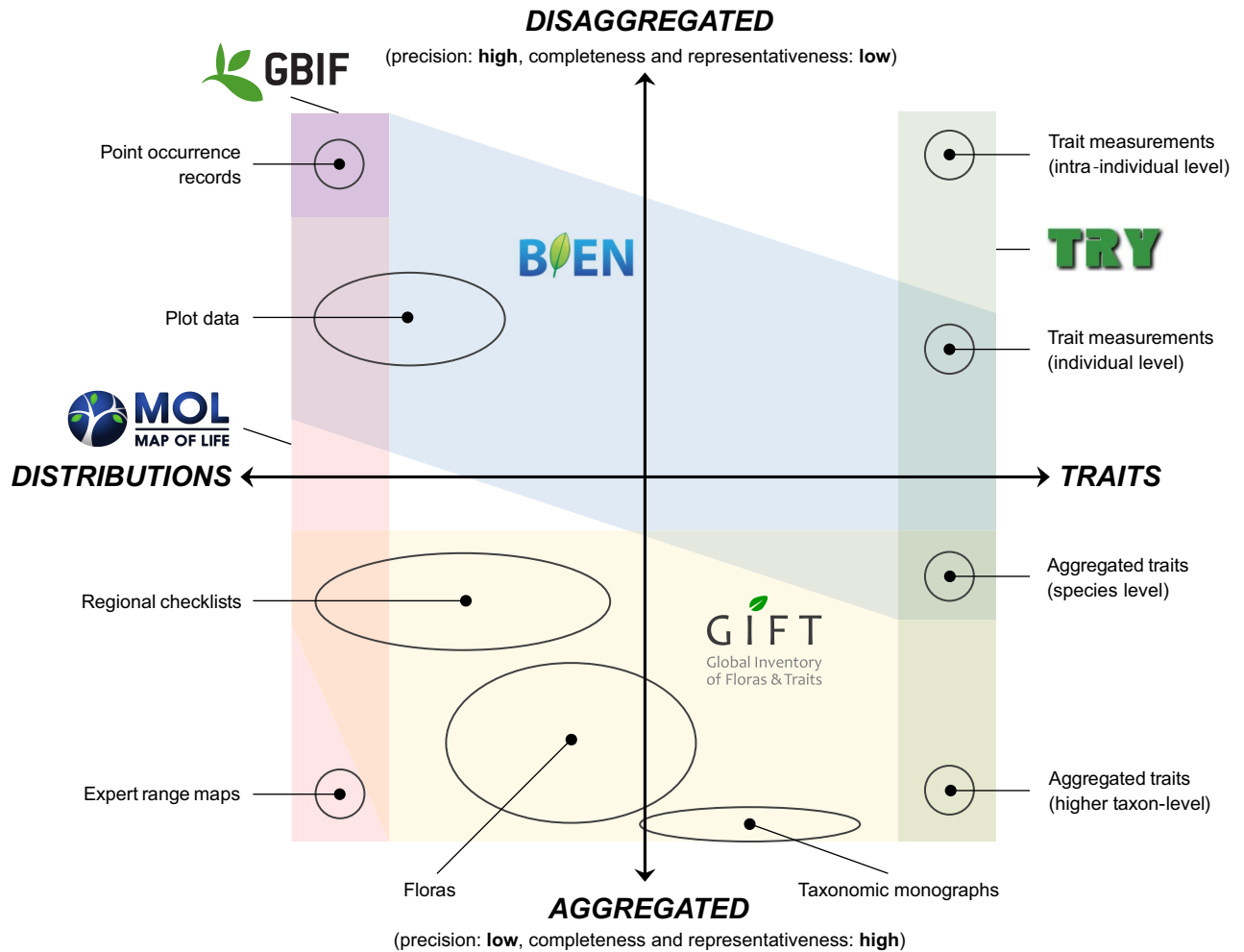


Fig 1. Selected biodiversity data types, arranged according to their primary domain (here, species distributions versus functional traits) and informational resolution (disaggregated versus aggregated). Projects that integrate global plant diversity data are often domain-specific (e.g., Map of Life [24]; TRY [7]) or focus on the disaggregated end of the data spectrum (e.g., GBIF [6], BIEN [26]). Complementing the ecological data landscape with aggregated data (e.g., GIFT [28]) creates strong synergies and facilitates biodiversity data integration across domains and resolutions. BIEN, Botanical Information Network and Ecology Network; GBIF, Global Biodiversity Information Facility; GIFT, Global Inventory of Floras and Traits.

<https://doi.org/10.1371/journal.pbio.3000183.g001>

baselines in plant diversity research. This not only opens up new opportunities but also poses new challenges with respect to data collection, mobilization, and sharing, as well as the utilization of synergies across data types.

Data collection and processing

The integration of biodiversity data starts in the field, with the primary data collected in surveys, experiments, and other research efforts. Such data are usually tailored to answer a particular research question. Therefore, robust ecological generalizations require large quantities of disaggregated or aggregated data that are organized and integrated in biodiversity databases. The quality and coverage of such databases can be greatly improved when primary research projects put strong emphasis on the utility and reusability of collected data for secondary scientific purposes [30].

The utility of primary data for data integration efforts can be increased in several ways. First, focusing on regions, ecosystems, plant groups, or functional traits that are currently

underrepresented in global biodiversity databases increases the general interest in the collected data and the study itself. Coverage analyses based on integrated biodiversity resources can provide guidance by identifying knowledge gaps and setting research priorities [15]. Second, cross-institutional coordination of research projects creates synergies through standardized methods and complementary research foci. Research networks such as International Long Term Ecological Research network (ILTER) [31] provide an ideal framework to utilize these synergies [32]. Third, an efficient study design helps to maximize the data output given the available resources. This can be aided, e.g., by optimizing study logistics and surveying effort [33], applying power analyses to estimate required sample sizes [34], and cooperating closely with local field guides and botanists [35]. Throughout the process of data collection, digital solutions such as Open Data Kit [36] can help to conveniently enter, cross-check, annotate, and aggregate field data. This increases data integrity and provides crucial meta-information for subsequent quality assessments and integration efforts.

The reusability of primary data can be ensured by adopting existing data standards and protocols. Species names—the most critical common identifier for data integration—can be standardized using pertinent software packages [37,38], which spell-check input names and match them against authoritative taxonomic resources. Moreover, trait measurement protocols [39] and terminologies [40] facilitate interoperability across research projects. The exchange of diversity data is supported by open data standards like Darwin Core [41] or Humboldt Core [42]. Finally, innovative publishing frameworks, such as the Biodiversity Data Journal [43] or the GBIF Integrated Publishing Toolkit [6], allow for a quick publication of standardized, annotated, and easily accessible data sets.

Data mobilization

A prime example of successful data mobilization is the massive extraction of distributional information from preserved specimens within the last 15 years. However, specimen records hold other types of information as well [44]. In particular, the (semi-)automated extraction of traits from herbarium specimens represents an area of largely unused potential. Standardized measurements on collected plant material may be incorporated into digitization workflows, potentially yielding thousands of geographically defined records of, e.g., specific leaf area [45] or phenological plant information [46]. Also, images of already digitized specimens can be used to retrieve functional traits [47]. Nonetheless, the set of traits that can be (nondestructively) obtained from herbarium specimens excludes many important characteristics, e.g., plant growth form, vegetative height, or stem specific density.

Another way to mobilize substantial amounts of ecological data—mainly from the aggregated end of the data spectrum—lies in botanical literature, e.g., Floras, checklists, and taxonomic monographs. Such resources are broadly available [48] and provide expert-validated distributional information, often including the biogeographical status of the species listed (e.g., endemic, native, introduced). Moreover, descriptions of morphology, life history, flowers, fruits, seeds, phenology, and other functionally relevant features are often available. Considering the wealth of information contained in published floristic literature, the development of general, scalable methods for data extraction seems to be central for improving the coverage of biodiversity databases. First projects that implement such workflows show promising results [49,50] and could contribute significantly to gap-filling of primary biodiversity data.

Data imputation

Data imputation is a technique in which missing or inconsistent data items are replaced with estimated values [51] and represents an inexpensive yet powerful way to improve data

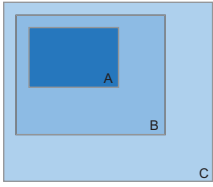
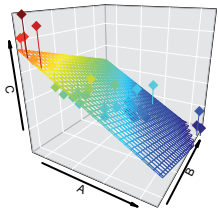
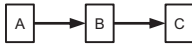
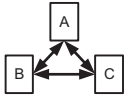
	Logical imputation	Statistical imputation
Data relationship	Hierarchical (one-to-many) or bijective (one-to-one) 	Correlative (many-to-many) 
Imputation method	Logical deduction 	Statistical prediction 
Gap-filling potential	Limited	Very high
Certainty of results	Very high (depending on correctness of input data and specified relationships)	Variable (depending on correlative structure of input data and model performance)

Fig 2. Comparison of logical and statistical data imputation. Logical imputation infers a limited quantity of highly certain data (e.g., deducing woodiness status from growth form), whereas statistical imputation yields large quantities of less certain data (e.g., predicting a suite of functional traits or species occurrences from sparse records).

<https://doi.org/10.1371/journal.pbio.3000183.g002>

coverage in ecological data sets. A conceptual distinction can be made between logical and statistical imputation methods (Fig 2).

Logical imputation uses unequivocal relationships among data to infer new values. This is possible either when data are categorically nested, e.g., trees always being woody [52], or linked by mathematical relationships, e.g., leaf mass per unit area being the inverse of specific leaf area. Although the considerations underlying logical imputation seem rather trivial, many applications in biodiversity data science remain underexploited, e.g., the propagation of information from complex functional traits to more simple ones, the “inheritance” of uniform traits from higher to lower taxonomic groups, or the improvement of regional species checklists based on geographically nested occurrence records or plot data. A major advantage of logical imputation is that the results can be treated with the same certainty as the input data. This makes it a particularly suitable approach for building and extending repositories of primary data. At the same time, logical imputation helps to harmonize data that uses differing terminologies by embedding it in a logical hierarchy (e.g., bee pollination, insect pollination, and animal pollination form nested subsets of pollination syndromes). However, considering that such clear hierarchical relationships are scarce among biodiversity data, the gap-filling potential of logical imputation is limited.

Statistical imputation, on the other hand, utilizes correlative relationships among data to predict new values. Because statistical imputation is based on statistical models, it can incorporate a variety of additional data to refine prediction accuracy. Gap filling techniques for functional traits, e.g., take into account trait–trait, trait–environment, and trait–phylogeny relationships to predict full trait matrices from sparse data [13,53]. Analogously, species distribution models make use of environmental information, species-specific characteristics, or biotic interactions to predict continuous species distributions from point occurrence records [54,55]. Statistical imputation methods allow for the prediction of any number of missing values, but the accuracy of these predictions is always dependent on the quality (i.e., correctness, representativeness, and completeness) of observations and predictor variables as well as the performance of the underlying statistical model. Therefore, statistical imputation is a valuable

tool for improving data coverage in specific use cases [10,56,57] but cannot be considered an expansion of primary data.

Strong synergies arise from combining logical imputation, which maximizes the amount of quasiprimary data, with statistical imputation, which may utilize these additional data to improve prediction accuracy. The potential of logical imputation for deducing simple functional traits such as woodiness or growth form is substantial (see case study 1). Although improved knowledge on these traits is of broad ecological interest in itself [58,59], it might be particularly useful to enhance the performance of statistical imputation techniques [13,60]. Similarly, logically imputed distributional information can help to improve species distribution models, e.g., by flagging and removing inconsistent occurrence records [24] or deriving pseudoabsences for species distribution models from regional checklists [61,62].

Data sharing

Data sharing is a basic condition for data integration [5,63], and although open science initiatives have started to gain traction in ecology, considerable institutional and sociocultural challenges remain [64,65]. Publishers, universities, and funding agencies have a central responsibility for creating an environment in which data sharing is a scientific asset not a disadvantage. Corresponding measures comprise a range of obligations and incentives for data sharing [66,67]. One example for an effective obligation is that many journals now require all data that were used to conduct a study to be stored in open repositories [68]. Likewise, funding agencies strive to improve data quality and long-term accessibility by requiring data management plans [69]. The most important measure, however, is the establishment of adequate incentives for data sharing, primarily by increasing the academic credit gained from doing so. Data citations have been pointed out as a fair and effective way of incentivizing and acknowledging data contributions [67,70] but also alternative measures of research impact and a generally stronger appreciation of data as scientific output will help to open up the ecological research culture [65,67].

Data integration

Biodiversity data are typically collated and integrated in domain-specific databases that allow fast extraction, exploration, and visualization of normalized data. This approach has transformed the ecological research landscape in the past decades and catalyzed ecological synthesis [4]. However, the scope of any single project is bound to limited technical, financial, and human resources. Building a scalable, dynamic infrastructure to integrate the wealth of existing environmental and ecological data thus requires bundling existing efforts within a unifying framework [32,71].

Distributed networks facilitate the organization of data, resources, and expertise from diverse data holders in a single, collaborative infrastructure that allows for the discovery, acquisition, citation, and (re)use of data [32,72]. A shared data portal acts as a central access point, whereas more specialized databases remain in charge of data aggregation and warehousing [30]. This organizational model has the potential to integrate the heterogeneous ecological data landscape but is also strongly dependent on the broad adoption of data standards. These include, e.g., universal identifiers ranging from standardized species names to digital identifiers for documents, data, and persons (e.g., digital object identifiers [DOIs], Life Science Identifiers [LSIDs], Open Researcher and contributor IDs [ORCID]) [73], compatible database structures as well as the implementation of standardized application program interfaces [APIs] and exchange formats [74], rich and well-structured metadata [63,75], and the formalization of existing ecological concepts in ontologies and thesauri [40,76].

The Data Observation Network for Earth (DataONE; <https://www.dataone.org>, [72]) already provides the basic infrastructure for building an open and distributed network of biodiversity data holders. However, currently the majority of member nodes consists of generic data repositories (e.g., DRYAD) and regional projects (e.g., United States Geological Survey [USGS]), whereas the participation of major aggregators of global plant diversity data (e.g., GBIF) has yet to be realized. Consequently, DataONE currently does not leverage the full potential of its powerful organizational model [63,72]. Some of the future challenges for distributed infrastructures such as DataONE are, e.g., the continuing promotion and development of data standards, the improvement of web-based visualization and analysis capabilities, the incorporation of machine learning for improved data discovery and utilization [77], and the robust implementation of dynamic cross-checking and data imputation workflows for parallel data streams.

Case studies

We present two case studies based on the GIFT database [28] to demonstrate the importance of data resolution and cross-domain data integration for addressing key questions in macroecology. Considering that GIFT focuses on aggregated data on plant distributions and functional traits only, these case studies provide an outlook on the full potential of an integrated biodiversity data landscape.

Case study 1: Global patterns in plant growth form

Plant functional types such as growth form capture fundamental axes of ecological variation in a simple way [10,78]. Consequently, knowledge of plant growth form is an important aspect of many ecological applications, ranging from local studies of plant diversity [79] to dynamic global vegetation models [80]. However, despite being a relatively simple and easily determinable trait, data on growth form is still surprisingly scarce and scattered both taxonomically and geographically. Here, we showcase opportunities arising from a systematic integration of aggregated functional and distributional data by predicting growth form spectra across the globe.

We combined angiosperm checklists and growth form data (distinguishing between herb, shrub, and tree) available in GIFT [28]. Oceanic islands as well as geographical units with more than 33% of species lacking growth form information were excluded. From the remaining 818 regional checklists, we included only those species with known growth form status, resulting in a data set containing 1,472,024 species-by-region combinations and 162,300 validated species. We calculated average climatic conditions for all 818 geographical units based on Climatologies at High resolution for the Earth's Land Surface areas (CHELSA) climate layers [81]. To assess the effects of climate on the relative proportion of growth forms, we used multinomial logistic regression as implemented in the R-package *nnet* [82]. Because our objective was predictive accuracy, not statistical inference, we used all 19 CHELSA bioclimatic variables as predictors without accounting for collinearity [83]. In fitting the model, each observation was weighted by the inverse log-area of the corresponding geographical unit to account for the decreasing representativeness of averaged climatic conditions for larger, climatically more variable regions. We then used the fitted model to predict growth form spectra across a global equal-area hexagon grid (R-package *dggridR* [84]; cell size = 23,300 km²).

Globally, herbs represented the most frequent growth form (Fig 3A and 3C), accounting for 56% of the species and 68% of the species-by-region combinations. Shrubs and trees were less frequent with 23% and 21% of species and 17% and 18% of species-by-region combinations, respectively. Regionally, however, shrubs and trees reached relatively high proportions of

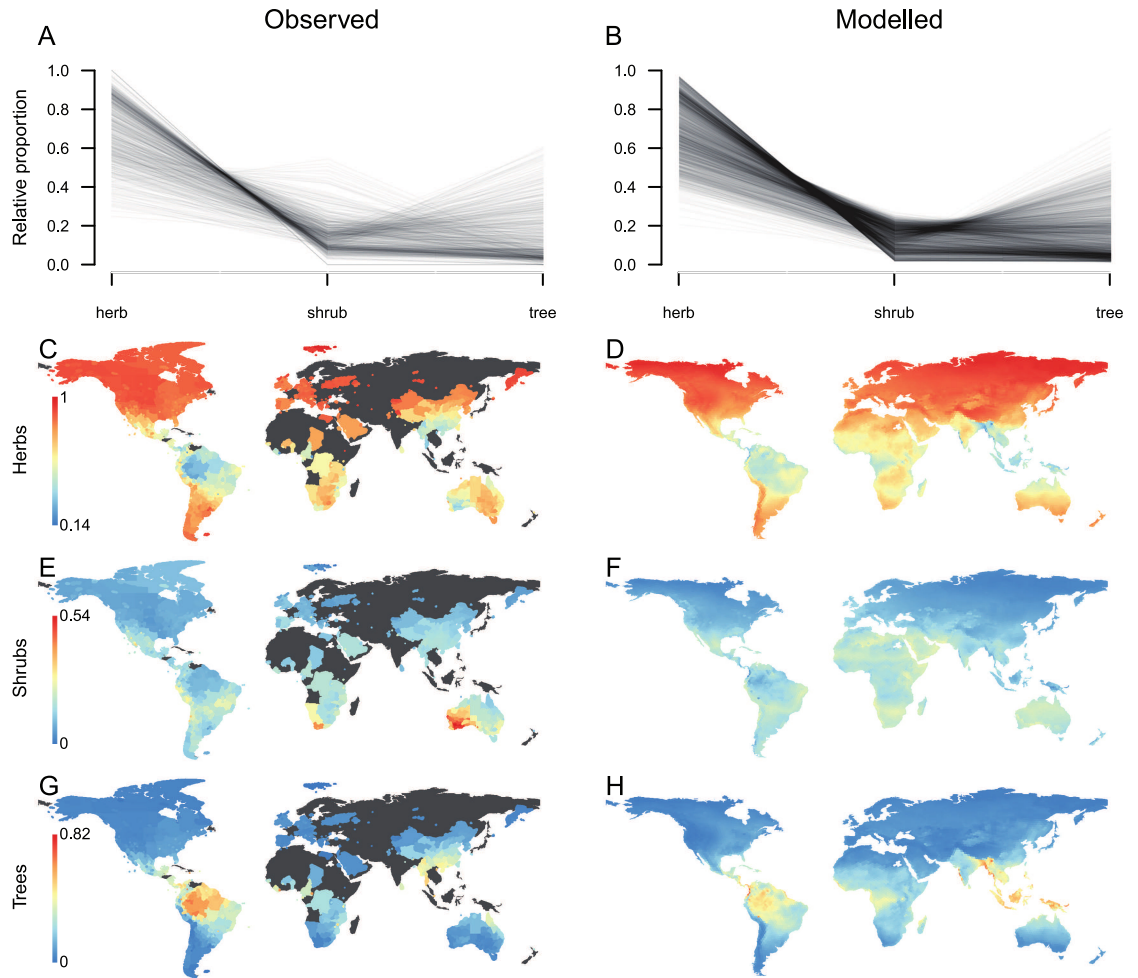


Fig 3. The global composition in plant growth form as observed for 818 angiosperm floras (left) and modeled for 6,495 equal-area grid cells (right). Upper plots summarize the growth form spectra across all observed (A) and modeled (B) geographical units, with each line representing a single flora. Lower plots (C–H) show the observed and modeled geographic variation in the proportion of herbs, shrubs, and trees individually. Note that the range of values varies across growth forms. The underlying data and data references for this figure can be found in [S1 Data](#).

<https://doi.org/10.1371/journal.pbio.3000183.g003>

species, particularly in Australian scrublands (Fig 3E) and the Amazonian rainforest (Fig 3G). Except for a few local deviations, our predictions of growth form composition were in strong agreement with the observed data (McFadden’s Pseudo- $R^2 = 0.91$). Moreover, our results are strongly supported by plot-based analyses of the African and American floras [85,86], which reveal similar geographical trends in growth form composition (S1 Fig and S2 Fig).

In the context of biodiversity data integration, this case study has two implications. First, a near-complete characterization of plant species with respect to fundamental categorical plant traits such as growth form is within reach when exploiting the full potential of data mobilization and imputation. This marks a critical step towards utilizing functional approaches at macroecological scales. Second, aggregated data types reveal remarkably similar but generally smoother biogeographical patterns when compared to comprehensive disaggregated datasets (S1 Fig and S2 Fig). This demonstrates that, due to their high global completeness and representativeness, aggregated data types capture fundamental ecological relationships and produce realistic predictions of plant diversity from regional to global scales.

Case study 2: The latitudinal gradient in seed mass revisited

Latitude is strongly correlated with numerous environmental characteristics such as temperature, precipitation, seasonality, and long-term climatic stability. Consequently, many aspects of biodiversity show systematic variation along latitude as well [87–89]. Moles and colleagues [90] provide an analysis of the latitudinal variation in seed mass based on a dataset of 11,481 species-by-sites combinations. The authors found a 320-fold decrease in seed mass between the equator and 60 degrees latitude as well as a sudden, 7-fold drop at 23 degrees latitude. These results were linked to changes in vegetation type and growth form composition, leading the authors to posit an abrupt change in plant strategy at the edge of the tropics. Here, our aim is to replicate these findings.

We extracted species lists from GIFT [28] for all mainland units with a complete survey of seed plants. In cases in which geographical units overlapped by more than 5%, we removed the larger unit if floristic data was available at a higher spatial resolution (e.g., preferring federal state- over country-level data); otherwise, we removed the smaller units (e.g., preferring continuous country-level data over patchy national park inventories). Furthermore, we only kept species with information on both seed mass and growth form, yielding a final data set of 519,812 species-by-region combinations and 563 distinct geographical units. In reassessing the relationship between seed mass and latitude, we followed the methodology of Moles and colleagues [90] and used linear regression and piecewise regression.

We found that the overall decrease in mean seed mass between the equator and 60 degrees latitude was only 11-fold according to linear regression (Fig 4, solid black line) and 8.8-fold according to piecewise regression, the latter indicating a 1.5-fold drop at 27 degrees latitude (Fig 4, dashed black line). Both models had low explanatory power ($R^2_{linear} = 0.045$, $R^2_{piecewise} = 0.048$), reflecting the substantial variation in seed mass at any given latitude. When examining latitudinal variation in seed mass for individual growth forms (Fig 4, colored lines), only trees showed a pronounced decrease towards the poles (12.5-fold), whereas shrubs and herbs exhibited little latitudinal variation (2.1- and 1.3-fold decrease, respectively). In agreement with Moles and colleagues [90], we found a strong latitudinal pattern in the relative proportion of growth forms, with herbs being increasingly dominant at higher latitudes (Fig 4, upper plot). Considering that the logarithmic mean seed mass differs significantly among growth forms (herbs: 0.99 mg, shrubs: 4.59 mg, trees: 48.95 mg; Fig 4, right-hand plot), the overall poleward decrease in seed mass seems to be mostly driven by the replacement of large-seeded trees by small-seeded herbs. According to our data, however, there is no evidence for an abrupt change in plant strategy. In conclusion, we find that the latitudinal gradient in seed mass is considerably less steep than previously reported and lacks a pronounced drop at the edge of the tropics.

This case study illustrates that the quantification of large-scale diversity patterns is highly dependent on the representativeness of the underlying data. In this respect, functional representativeness has been a largely neglected dimension of sample quality. Indeed, the data underlying the original results of Moles and colleagues show remarkably high proportions of tree-dominated biomes and tree species at tropical latitudes (Fig 2B and 2C in [90]). Values of 100% tropical rainforest and 90% tree species at the equator are neither consistent with existing literature [91,92] nor with our data set, which comprises about 50 times more data points (S3 Fig). The representation of trees in Moles and colleagues [90] decreases at the exact point—between 20 and 25 degrees latitude—at which the authors find a sudden drop in seed mass. Thereafter, the latitudinal gradients in growth form composition and seed mass are highly consistent with our data (S3 Fig), suggesting that an uneven latitudinal representation of biomes and growth forms amplified the magnitude and distorted the shape of the latitudinal gradient in seed mass in the study of Moles and colleagues [90]. Integrated biodiversity

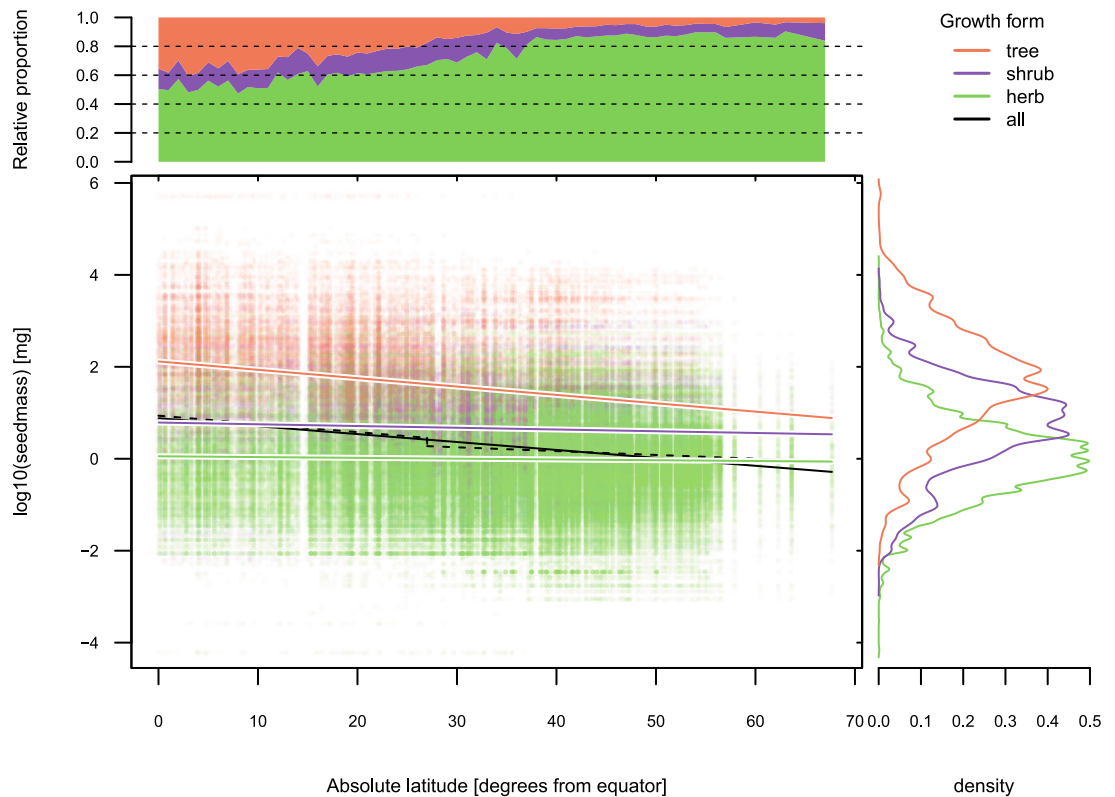


Fig 4. Latitudinal gradient in seed mass for 519,812 species-region combinations. Piecewise regression (dashed black line) was compared against linear models for the entire data set (solid black line) and individual growth forms (colored lines). Upper plot shows the relative proportion of growth forms in each 1-degree latitudinal band. Right-hand plot depicts the frequency distribution of seed mass for individual growth forms. The underlying data and data references for this figure can be found in [S2 Data](#).

<https://doi.org/10.1371/journal.pbio.3000183.g004>

resources and a targeted utilization of aggregated data types can help to detect and resolve such latent biases to facilitate more robust descriptions of macroecological patterns.

Conclusion and future directions

The availability, quality, and interoperability of data is paramount to the progress of biogeography and ecology as increasingly data-driven disciplines [5,30,93]. We demonstrate how the explicit consideration of data resolution offers new perspectives on the compilation and integration of plant diversity data. Our results show that a coarse-grained but near-complete knowledge of global plant distributions and basic functional traits is within reach when exploiting the full potential of data mobilization and imputation. This offers exciting opportunities for plant diversity research.

Currently, studies and projects integrating global plant diversity data are mostly based on disaggregated data types. Although this approach has been a successful line of research [10,94,95], the pervasiveness of biases and gaps in disaggregated biodiversity data is of increasing concern to ecologists [14,15,96,97]. We have shown that the systematic utilization of aggregated data can help address this problem (see case studies 1 and 2). First, aggregated data provide a coarse but more complete and less biased picture of geographical variation in taxonomic, functional, and phylogenetic diversity. This offers much-needed baselines against which the completeness of disaggregated data can be evaluated in order to quantify and map gaps in global biodiversity knowledge [16,93]. Second, aggregated data provide prior information about the geographical and statistical

distribution of more highly resolved but potentially incomplete or biased ecological variables. This knowledge can be used, e.g., to inform analyses in functional biogeography, to improve species distribution and niche models [98], or to parametrize ancestral state reconstructions [99] and dynamic global vegetation models [100]. Third, aggregated data capitalizes on expert knowledge to compensate for the varying availability and quality of primary (disaggregated) data. Consequently, aggregated data types are not mere compilations of disaggregated data but provide valuable additional information, e.g., reliable species absences or uniform functional traits for higher taxa. These potentials extend to other clades, e.g., mammals, birds, or certain arthropod groups, for which a wealth of literature exists.

Data integration has to bridge not only multiple resolutions but also domains. Satellite-borne, multispectral imagery has become a crucial component of biodiversity research and monitoring, providing global high-resolution data of, e.g., net primary productivity, vegetation cover, or canopy height [101]. Advanced instruments will soon enable the derivation of similar data products for selected functional traits that require integration with in situ observations [102]. Vegetation plot databases are another key source of plant diversity data, holding crucial information on species abundances and interactions. Initiatives like BIEN and sPlot demonstrate how the integration of specimen- and plot data with taxonomic, functional, phylogenetic, and environmental information helps bridge the gap between local-, regional-, and continental-scale ecological processes [8,85,103]. Finally, a better integration of paleontological and socioeconomic data sources with existing biodiversity data resources bears great potential to improve our understanding of biogeography and inform questions concerning conservation planning and alien species management.

The unparalleled pressure on the biosphere renders a full utilization of all available biodiversity data imperative. Rapid advancements in information technology have brought down the technological barriers to this objective. It is now up to ecologists to keep pace with this development and to work collaboratively on creating an integrated biodiversity data landscape that bridges the gap between fine-scale precision and global representativeness.

Supporting information

S1 Fig. Relative frequency of plant growth forms (herb, shrub, or tree) across the New World derived from disaggregated (left panel, BIEN) versus aggregated (right panel, GIFT) plant diversity data. (Left) Data from BIEN were obtained through the BIEN r-package by downloading species lists and trait information for 399 geographical units from the New World available in GIFT. The BIEN data set comprised 131,041 species, 969,625 species-by-region combinations, and 69,070 species-by-trait combinations. (Right) The GIFT data set was assembled according to the methodology described in case study 1 and comprised 117,163 species, 940,541 species-by-region combinations, and 89,515 species-by-trait combinations. BIEN, Botanical Information Network and Ecology Network; GIFT, Global Inventory of Floras and Traits. (DOCX)

S2 Fig. Relative frequency of plant growth forms (herb, shrub, or tree) across central Africa derived from disaggregated plant diversity data (left panel, RAINBIO) versus model predictions derived from aggregated plant diversity data (right panel, GIFT). (Left) High-resolution plot data from RAINBIO were aggregated to varying spatial resolutions following Watson and colleagues and matched with growth form data available in RAINBIO. (Right) Predictions of growth form composition are based on multinomial logistic regression of a global data set of species checklists and growth form information extracted from the GIFT database (see case study 1 for methodology). GIFT, Global Inventory of Floras and Traits. (DOCX)

S3 Fig. Data set comparison for case study 2 between Moles and colleagues (2007) (11,481 species-by-sites combinations, upper plot) and GIFT (519,812 species-by-region combinations, lower plot). GIFT, Global Inventory of Floras and Traits.
(DOCX)

S1 Data. Supporting Data for case study 1.
(ZIP)

S2 Data. Supporting Data for case study 2.
(ZIP)

Acknowledgments

Acknowledgments to Brian Enquist, Walter Jetz, and Daniel Noesgaard for permitting the use organizational logos of BIEN, Map of Life, and GBIF, respectively. We also want to thank the organizers and participants of the Macro2018 conference for the opportunity to present and discuss our results with the scientific community.

References

1. Watson JEM, Jones KR, Fuller RA, Di Marco M, Segan DB, Butchart SHM, et al. Persistent Disparities between Recent Rates of Habitat Conversion and Protection and Implications for Future Global Conservation Targets. *Conservation Letters*. 2016; 9: 413–421. <https://doi.org/10.1111/cons.12295>
2. Pachauri RK, Allen MR, Barros VR, Broome J, Cramer W, Christ R, et al. Climate change 2014. Synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change: IPCC; 2014.
3. Seebens H, Blackburn TM, Dyer EE, Genovesi P, Hulme PE, Jeschke JM, et al. No saturation in the accumulation of alien species worldwide. *Nature Communications*. 2017; 8: 14435. <https://doi.org/10.1038/ncomms14435> PMID: 28198420
4. Kelling S, Hochachka WM, Fink D, Riedewald M, Caruana R, Ballard G, et al. Data-intensive Science. A New Paradigm for Biodiversity Studies. *BioScience*. 2009; 59: 613–620. <https://doi.org/10.1525/bio.2009.59.7.12>
5. Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, et al. Big data and the future of ecology. *Frontiers in Ecology and the Environment*. 2013; 11: 156–162. <https://doi.org/10.1890/120103>
6. GBIF. The Global Biodiversity Information Faculty; 2018. Available from: <http://www.gbif.org>. [cited 2017 July 27].
7. Kattge J, Díaz S, Lavorel S, Prentice IC, Leadley P, Bönisch G, et al. TRY—a global database of plant traits. *Global Change Biology*. 2011; 17: 2905–2935.
8. Bruelheide H, Dengler J, Purschke O, Lenoir J, Jiménez-Alfaro B, Hennekens SM, et al. Global trait–environment relationships of plant communities. *Nature Ecology & Evolution*. 2018. <https://doi.org/10.1038/s41559-018-0699-8> PMID: 30455437
9. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res*. 2005; 33: D34–8. <https://doi.org/10.1093/nar/gki063> PMID: 15608212
10. Díaz S, Kattge J, Cornelissen JHC, Wright IJ, Lavorel S, Dray S, et al. The global spectrum of plant form and function. *Nature*. 2016; 167–171. <https://doi.org/10.1038/nature16489> PMID: 26700811
11. Smith SA, Brown JW. Constructing a broadly inclusive seed plant phylogeny. *Am J Bot*. 2018; 105: 1–13. <https://doi.org/10.1002/ajb2.1015>
12. Zanne AE, Pearse WD, Cornwell WK, McGlenn DJ, Wright IJ, Uyeda JC. Functional biogeography of angiosperms. *Life at the extremes*. *New Phytol*. 2018; 218: 1697–1709. <https://doi.org/10.1111/nph.15114> PMID: 29603243
13. Schrod F, Kattge J, Shan H, Fazayeli F, Joswig J, Banerjee A, et al. BHPMF—a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. *Glob Ecol Biogeogr*. 2015; 24: 1510–1521. <https://doi.org/10.1111/geb.12335>

14. Engemann K, Enquist BJ, Sandel B, Boyle B, Jørgensen PM, Morueta-Holme N, et al. Limited sampling hampers “big data” estimation of species richness in a tropical biodiversity hotspot. *Ecol Evol*. 2015; 5: 807–820. <https://doi.org/10.1002/ece3.1405> PMID: 25692000
15. Meyer C, Weigelt P, Kreft H. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol Lett*. 2016; 19: 992–1006. <https://doi.org/10.1111/ele.12624> PMID: 27250865
16. Hortal J, Bello F de, Diniz-Filho JAF, Lewinsohn TM, Lobo JM, Ladle RJ. Seven shortfalls that beset large-scale knowledge of biodiversity. *Annu. Rev. Ecol. Evol. Syst*. 2015; 46: 523–549. <https://doi.org/10.1146/annurev-ecolsys-112414-054400>
17. Brown JH, Lomolino MV. *Biogeography*. 2nd ed.: Sunderland MA: Sinauer Associates; 1998.
18. McGill BJ, Enquist BJ, Weiher E, Westoby M. Rebuilding community ecology from functional traits. *Trends Ecol Evol*. 2006; 21: 178–185. <https://doi.org/10.1016/j.tree.2006.02.002> PMID: 16701083
19. Violle C, Reich PB, Pacala SW, Enquist BJ, Kattge J. The emergence and promise of functional biogeography. *Proc Natl Acad Sci U S A*. 2014; 111: 13690–13696. <https://doi.org/10.1073/pnas.1415442111> PMID: 25225414
20. Wiens JA. Spatial scaling in ecology. *Funct Ecol*. 1989; 3: 385–397.
21. Rahbek C. The role of spatial scale and the perception of large-scale species-richness patterns. *Ecol Lett*. 2005; 8: 224–239. <https://doi.org/10.1111/j.1461-0248.2004.00701.x>
22. Bolnick DI, Amarasekare P, Araújo MS, Bürger R, Levine JM, Novak M, et al. Why intraspecific trait variation matters in community ecology. *Trends Ecol Evol*. 2011; 26: 183–192. <https://doi.org/10.1016/j.tree.2011.01.009> PMID: 21367482
23. Rondinini C, Wilson KA, Boitani L, Grantham H, Possingham HP. Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecol Lett*. 2006; 9: 1136–1145. <https://doi.org/10.1111/j.1461-0248.2006.00970.x> PMID: 16972877
24. Jetz W, McPherson JM, Guralnick RP. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol Evol*. 2012; 27: 151–159. <https://doi.org/10.1016/j.tree.2011.09.007> PMID: 22019413
25. Walther BA, Moore JL. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*. 2005; 28: 815–829. <https://doi.org/10.1111/j.2005.0906-7590.04112.x>
26. Enquist BJ, Condit R, Peet RK, Schildhauer M, Thiers B. Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. *PeerJ Preprints*. 2016; 4: e2615v2. <https://doi.org/10.7287/peerj.preprints.2615v2>
27. WCSP. World Checklist of Selected Plant Families; 2014. Available from: <http://apps.kew.org/wcsp/home.do>. [cited 2014 Dec 1].
28. Weigelt P, König C, Kreft H. GIFT—A Global Inventory of Floras and Traits for macroecology and biogeography. *bioRxiv*. 2019. <https://doi.org/10.1101/535005>
29. sPlot Core Team. sPlot—The global vegetation Database; 2017. Available from: https://www.idiv.de/en/sdiv/working_groups/wg_pool/splot.html. [cited 2017 Oct 2].
30. Michener WK, Jones MB. Ecoinformatics. Supporting ecology as a data-intensive science. *Trends Ecol Evol*. 2012; 27: 85–93. <https://doi.org/10.1016/j.tree.2011.11.016> PMID: 22240191
31. Vanderbilt K, Gaiser E. The International Long Term Ecological Research Network. A platform for collaboration. *Ecosphere*. 2017; 8. <https://doi.org/10.1002/ecs2.1697>
32. Peters DPC, Loescher HW, SanClements MD, Havstad KM. Taking the pulse of a continent. Expanding site-based research infrastructure for regional- to continental-scale ecology. *Ecosphere*. 2014; 5: art29. <https://doi.org/10.1890/ES13-00295.1>
33. Moore AL, McCarthy MA. Optimizing ecological survey effort over space and time. *Methods Ecol Evol*. 2016; 7: 891–899. <https://doi.org/10.1111/2041-210X.12564>
34. Johnson PCD, Barry SJE, Ferguson HM, Müller P. Power analysis for generalized linear mixed models in ecology and evolution. *Methods Ecol Evol*. 2015; 6: 133–142. <https://doi.org/10.1111/2041-210X.12306> PMID: 25893088
35. Elbroch M, Mwampamba TH, Santos MJ, Zylberberg M, Liebenberg L, Minye J, et al. The value, limitations, and challenges of employing local experts in conservation research. *Conserv Biol*. 2011; 25: 1195–1202. <https://doi.org/10.1111/j.1523-1739.2011.01740.x> PMID: 21966985
36. Brunette W, Sundt M, Dell N, Chaudhri R, Breit N, Borriello G. Open data kit 2.0. Expanding and refining information services for developing regions. In: Agarwal S, Varshavsky A, editors. *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*: ACM; 2013.
37. Boyle B, Hopkins N, Lu Z, Raygoza Garay JA, Mozzherin D, Rees T, et al. The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics*. 2013; 14: 16. <https://doi.org/10.1186/1471-2105-14-16> PMID: 23324024

38. Chamberlain SA, Szöcs E. taxize. Taxonomic search and retrieval in R. *F1000Res*. 2013. <https://doi.org/10.12688/f1000research.2-191.v1> PMID: 24555091
39. Pérez-Harguindeguy N, Díaz S, Garnier E, Lavorel S, Poorter H, Jaureguiberry P, et al. New handbook for standardised measurement of plant functional traits worldwide. *Aust. J. Bot.* 2013; 61: 167–234. <https://doi.org/10.1071/BT12225>
40. Garnier E, Stahl U, Laporte M-A, Kattge J, Mougnot I, Kühn I, et al. Towards a thesaurus of plant characteristics. An ecological contribution. *J Ecol.* 2017; 105: 298–309. <https://doi.org/10.1111/1365-2745.12698>
41. Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. Darwin Core. An evolving community-developed biodiversity data standard. *PLoS ONE*. 2012; 7: e29715. <https://doi.org/10.1371/journal.pone.0029715> PMID: 22238640
42. Guralnick R, Walls R, Jetz W. Humboldt Core—toward a standardized capture of biological inventories for biodiversity monitoring, modeling and assessment. *Ecography*. 2017; 1297: 8. <https://doi.org/10.1111/ecog.02942>
43. Pensoft. Biodiversity Data Journal. a peer review open-access journal; 2017. <https://bdj.pensoft.net/>
44. Beaman RS, Cellinese N. Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. *Zookeys*. 2012: 7–17. <https://doi.org/10.3897/zookeys.209.3313> PMID: 22859875
45. Queenborough SA, Porras C. Expanding the coverage of plant trait databases. A comparison of specific leaf area derived from fresh and dried leaves. *Plant Ecology & Diversity*. 2014; 7: 383–388.
46. Gallinat AS, Russo L, Melaas EK, Willis CG, Primack RB. Herbarium specimens show patterns of fruiting phenology in native and invasive plant species across New England. *Am J Bot.* 2018; 97: 1296. <https://doi.org/10.1002/ajb2.1005>
47. Corney D, Clark JY, Tang HL, Wilkin P. Automatic extraction of leaf characters from herbarium specimens. *Taxon*. 2012; 61: 231–244.
48. Frodin DG. *Guide to Standard Floras of the World*. Cambridge: Cambridge University Press; 2001.
49. Hoehndorf R, Alshahrani M, Gkoutos GV, Gosline G, Groom Q, Hamann T, et al. The flora phenotype ontology (FLOPO): tool for integrating morphological traits and phenotypes of vascular plants. *J Biomed Semantics*. 2016; 7: 65. <https://doi.org/10.1186/s13326-016-0107-8> PMID: 27842607
50. Endara L, Cui H, Burleigh JG. Extraction of phenotypic traits from taxonomic descriptions for the tree of life using natural language processing. *Appl Plant Sci*. 2018; 6: e1035. <https://doi.org/10.1002/aps3.1035> PMID: 29732265
51. OECD. Glossary of statistical terms. Definition of “Data imputation”; 2013. Available from: <https://stats.oecd.org/glossary/detail.asp?ID=3406>. [cited 2017 Oct 10].
52. Beentje H. *The Kew Plant Glossary. An illustrated dictionary of plant terms*. 2nd ed. Royal Botanic Gardens, Kew; 2016.
53. Penone C, Davidson AD, Shoemaker KT, Di Marco M, Rondinini C, Brooks TM, et al. Imputation of missing data in life-history trait datasets: which approach performs the best. *Methods Ecol Evol*. 2014; 5: 961–970. <https://doi.org/10.1111/2041-210X.12232>
54. Elith J, Leathwick JR. *Species Distribution Models. Ecological Explanation and Prediction Across Space and Time*. *Annu Rev Ecol Evol Syst*. 2009; 40: 677–697. <https://doi.org/10.1146/annurev.ecolsys.110308.120159>
55. Peterson AT. *Ecological niches and geographic distributions*: Princeton University Press; 2011.
56. Paine CET, Baraloto C, Chave J, Hérault B. Functional traits of individual trees reveal ecological constraints on community assembly in tropical rain forests. *Oikos*. 2011; 120: 720–727. <https://doi.org/10.1111/j.1600-0706.2010.19110.x>
57. Syfert MM, Joppa L, Smith MJ, Coomes DA, Bachman SP, Brummitt NA. Using species distribution models to inform IUCN Red List assessments. *Biol Conservation*. 2014; 177: 174–184. <https://doi.org/10.1016/j.biocon.2014.06.012>
58. Scheffer M, Vergnon R, Cornelissen JHC, Hantson S, Holmgren M, van Nes EH, et al. Why trees and shrubs but rarely trubs. *Trend Ecol Evol*. 2014; 29: 433–434. <https://doi.org/10.1016/j.tree.2014.06.001>
59. Beech E, Rivers M, Oldfield S, Smith PP. GlobalTreeSearch. The first complete global database of tree species and country distributions. *Journal of Sustainable Forestry*. 2017: 1–36.
60. van Buuren S, Groothuis-Oudshoorn K. mice. Multivariate Imputation by Chained Equations in R. *J Statistical Software*. 2011; 45: 1–67. <https://doi.org/10.18637/jss.v045.i03>
61. VanDerWal J, Shoo LP, Graham C, Williams SE. Selecting pseudo-absence data for presence-only distribution modeling. How far should you stray from what you know. *Ecological Modelling*. 2009; 220: 589–594. <https://doi.org/10.1016/j.ecolmodel.2008.11.010>

62. Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. Selecting pseudo-absences for species distribution models. How, where and how many. *Methods Ecol Evol.* 2012; 3: 327–338. <https://doi.org/10.1111/j.2041-210X.2011.00172.x>
63. Reichman OJ, Jones MB, Schildhauer MP. Challenges and opportunities of open data in ecology. *Science.* 2011; 331: 703–705. <https://doi.org/10.1126/science.1197962> PMID: 21311007
64. Michener WK. Ecological data sharing. *Ecological Informatics.* 2015; 29: 33–44. <https://doi.org/10.1016/j.ecoinf.2015.06.010>
65. Gewin V. Data sharing. An open mind on open data. *Nature.* 2016; 529: 117–119. <https://doi.org/10.1038/nj7584-117a> PMID: 26744755
66. Whitlock MC. Data archiving in ecology and evolution. Best practices. *Trends Ecol Evol.* 2011; 26: 61–65. <https://doi.org/10.1016/j.tree.2010.11.006> PMID: 21159406
67. Kattge J, Díaz S, Wirth C. Of carrots and sticks. *Nature Geoscience.* 2014; 7: 778–779.
68. Mills JA, Teplitsky C, Arroyo B, Charmantier A, Becker PH, Birkhead TR, et al. Archiving Primary Data: Solutions for Long-Term Studies. *Trends Ecol Evol.* 2015; 30: 581–589. <https://doi.org/10.1016/j.tree.2015.07.006> PMID: 26411615
69. Michener WK. Ten Simple Rules for Creating a Good Data Management Plan. *PLoS Comput Biol.* 2015; 11: e1004525. <https://doi.org/10.1371/journal.pcbi.1004525> PMID: 26492633
70. Data Citation Synthesis Group. Joint Declaration of Data Citation Principles. Mertone M. (ed.), San Diego, Force11. 2014; <https://doi.org/10.25490/a97f-egy4>
71. La Salle J, Williams KJ, Moritz C. Biodiversity analysis in the digital era. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 2016; 371. <https://doi.org/10.1098/rstb.2015.0337> PMID: 27481789
72. Michener W, Vieglais D, Vision T, Kunze J, Cruse P, Janée G. DataONE. Data Observation Network for Earth Preserving Data and Enabling Innovation in the Biological and Environmental Sciences. *D-Lib Magazine.* 2011; 17. <https://doi.org/10.1045/january2011-michener>
73. Page RDM. Biodiversity informatics. The challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics.* 2008; 9: 345–354. <https://doi.org/10.1093/bib/bbn022> PMID: 18445641
74. Kattge J, Ogle K, Bönisch G, Díaz S, Lavorel S, Madin J, et al. A generic structure for plant trait databases. *Methods Ecol Evol.* 2011; 2: 202–213. <https://doi.org/10.1111/j.2041-210X.2010.00067.x>
75. Feigraus EH, Andelman S, Jones MB, Schildhauer M. Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. *Bulletin of the Ecological Society of America.* 2005; 86: 158–168. [https://doi.org/10.1890/0012-9623\(2005\)86\[158:MTVOED\]2.0.CO;2](https://doi.org/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2)
76. Mouquet N, Lagadeuc Y, Devictor V, Doyen L, Duputié A, Eveillard D, et al. Predictive ecology in a changing world. *Journal of Applied Ecology.* 2015; 52: 1293–1310. <https://doi.org/10.1111/1365-2664.12482>
77. Peters DPC, Havstad KM, Cushing J, Tweedie C, Fuentes O, Villanueva-Rosales N. Harnessing the power of big data. Infusing the scientific method with machine learning to transform ecology. *Ecosphere.* 2014; 5: art67. <https://doi.org/10.1890/ES13-00359.1>
78. Leishman MR, Westoby M. Classifying plants into groups on the basis of associations of individual traits—evidence from Australian semi-arid woodlands. *J Ecol.* 1992; 80: 417–424.
79. Knapp AK, Briggs JM, Collins SL, Archer SR, Bret-Harte MS, Ewers BE, et al. Shrub encroachment in North American grasslands. Shifts in growth form dominance rapidly alters control of ecosystem carbon inputs. *Global Change Biology.* 2008; 14: 615–623. <https://doi.org/10.1111/j.1365-2486.2007.01512.x>
80. Wullschleger SD, Epstein HE, Box EO, Euskirchen ES, Goswami S, Iversen CM, et al. Plant functional types in Earth system models. Past experiences and future directions for application of dynamic vegetation models in high-latitude ecosystems. *Ann Bot.* 2014; 114: 1–16. <https://doi.org/10.1093/aob/mcu077> PMID: 24793697
81. Karger DN, Conrad O, Böhrner J, Kawohl T, Kreft H, Soria-Auza RW, et al. Climatologies at high resolution for the earth's land surface areas. *Scientific Data.* 2017; 4: 170122 EP. <https://doi.org/10.1038/sdata.2017.122> PMID: 28872642
82. Venables WN, Ripley BD. *Modern Applied Statistics with S.* New York: Springer; 2002.
83. Morrissey MB, Ruxton GD. Multiple regressions: the meaning of multiple regression and the non-problem of collinearity. *Philosophy, Theory and Practice in Biology.* 2018; 10. <https://doi.org/10.3998/ptpbio.16039257.0010.003>
84. Barnes R. dggridR: Discrete Global Grids for R; 2018. <https://github.com/r-barnes/dggridR/>
85. Engemann K, Sandel B, Enquist BJ, Jørgensen PM, Kraft NJB, Marcuse-Kubitza A, et al. Patterns and drivers of plant functional group dominance across the Western Hemisphere. A macroecological

- re-assessment based on a massive botanical dataset. *Bot J Linn Soc.* 2016; 180: 141–160. <https://doi.org/10.1111/boj.12362>
86. Sosef MSM, Dauby G, Blach-Overgaard A, van der Burgt X, Catarino L, Damen T, et al. Exploring the floristic diversity of tropical Africa. *BMC Biology.* 2017; 15: 15. <https://doi.org/10.1186/s12915-017-0356-8> PMID: 28264718
 87. Stevens GC. The latitudinal gradient in geographical range. How so many species coexist in the tropics. *The American Naturalist.* 1989; 133: 240–256.
 88. Cramer W, Kicklighter DW, Bondeau A, Iii BM, Churkina G, Nemry B, et al. Comparing global models of terrestrial net primary productivity (NPP). Overview and key results. *Global Change Biology.* 1999; 5: 1–15. <https://doi.org/10.1046/j.1365-2486.1999.00009.x>
 89. Hillebrand H. On the generality of the latitudinal diversity gradient. *The American Naturalist.* 2004; 163: 192–211. <https://doi.org/10.1086/381004> PMID: 14970922
 90. Moles AT, Ackerly DD, Tweddle JC, Dickie JB, Smith R, Leishman MR, et al. Global patterns in seed size. *Glob Ecol Biogeogr.* 2007; 16: 109–116.
 91. Ewel JJ, Bigelow SW. Plant life-forms and tropical ecosystem functioning. In: Orians GH, Dirzo R, Cushman JH, editors. *Biodiversity and ecosystem processes in tropical forests.* Berlin, New York: Springer; 1996.
 92. Walter H, Breckle S-W. *Walter's Vegetation of the Earth. The Ecological Systems of the Geo-Biosphere.* 4th ed. Berlin/Heidelberg: Springer; 2002.
 93. Franklin J, Serra-Diaz JM, Syphard AD, Regan HM. Big data for forecasting the impacts of global change on plant communities. *Glob Ecol Biogeogr.* 2017; 26: 6–17. <https://doi.org/10.1111/geb.12501>
 94. Swenson NG, Enquist BJ, Pither J, Kerkhoff AJ, Boyle B, Weiser MD, et al. The biogeography and filtering of woody plant functional diversity in North and South America. *Glob Ecol Biogeogr.* 2012; 21: 798–808. <https://doi.org/10.1111/j.1466-8238.2011.00727.x>
 95. Moles AT, Perkins SE, Laffan SW, Flores-Moreno H, Awasthy M, Tindall ML, et al. Which is a better predictor of plant traits: temperature or precipitation. *Journal of Vegetation Science.* 2014; 25: 1167–1180. <https://doi.org/10.1111/jvs.12190>
 96. Boakes EH, McGowan PJK, Fuller RA, Chang-qing D, Clark NE, O'Connor K, et al. Distorted Views of Biodiversity. Spatial and Temporal Bias in Species Occurrence Data. *PLoS Biol.* 2010; 8: e1000385. <https://doi.org/10.1371/journal.pbio.1000385> PMID: 20532234
 97. Sandel B, Gutiérrez AG, Reich PB, Schrödt F, Dickie J, Kattge J, et al. Estimating the missing species bias in plant trait measurements. *Journal of Vegetation Science.* 2015; 26: 828–838. <https://doi.org/10.1111/jvs.12292>
 98. Merow C, Allen JM, Aiello-Lammens M, Silander JA. Improving niche and range estimates with MaxEnt and point process models by integrating spatially explicit information. *Glob Ecol Biogeogr.* 2016; 25: 1022–1036. <https://doi.org/10.1111/geb.12453>
 99. Pagel M, Meade A, Barker D, Thorne J. Bayesian Estimation of Ancestral Character States on Phylogenies. *Syst Biol.* 2004; 53: 673–684. <https://doi.org/10.1080/10635150490522232> PMID: 15545248
 100. Scheiter S, Langan L, Higgins SI. Next-generation dynamic global vegetation models. Learning from community ecology. *New Phytol.* 2013; 198: 957–969. <https://doi.org/10.1111/nph.12210> PMID: 23496172
 101. Kuenzer C, Ottinger M, Wegmann M, Guo H, Wang C, Zhang J, et al. Earth observation satellite sensors for biodiversity monitoring. Potentials and bottlenecks. *International Journal of Remote Sensing.* 2014; 35: 6599–6647. <https://doi.org/10.1080/01431161.2014.964349>
 102. Jetz W, Cavender-Bares J, Pavlick R, Schimel D, Davis FW, Asner GP, et al. Monitoring plant functional diversity from space. *Nat Plants.* 2016; 2: 16024. <https://doi.org/10.1038/nplants.2016.24> PMID: 27249357
 103. Blonder B, Nogués-Bravo D, Borregaard MK, Donoghue JC, Jørgensen PM, Kraft NJB, et al. Linking environmental filtering and disequilibrium to biogeography with a community climate framework. *Ecol.* 2015; 96: 972–985. <https://doi.org/10.1890/14-0589.1> PMID: 26230018