11-2021

# Open Data Practices among Users of Primary Biodiversity Data

Caitlin P. Mandeville

# Open Data Practices among Users of Primary Biodiversity Data

CAITLIN P. MANDEVILLE, WOUTER KOCH, ERLEND B. NILSEN, AND ANDERS G. FINSTAD

*Presence-only biodiversity data are increasingly relied on in biodiversity, ecology, and conservation research, driven by growing digital infrastructures that support open data sharing and reuse. Recent reviews of open biodiversity data have clearly documented the value of data sharing, but the extent to which the biodiversity research community has adopted open data practices remains unclear. We address this question by reviewing applications of presence-only primary biodiversity data, drawn from a variety of sources beyond open databases, in the indexed literature. We characterize how frequently researchers access open data relative to data from other sources, how often they share newly generated or collated data, and trends in metadata documentation and data citation. Our results indicate that biodiversity research commonly relies on presence-only data that are not openly available and neglects to make such data available. Improved data sharing and documentation will increase the value, reusability, and reproducibility of biodiversity research.*

*Keywords: applied ecology, biodiversity, informatics, monitoring and mapping, publication practices*

**Biodiversity data are increasingly made openly** available, facilitated by extensive digital infrastructures that support data standardization and publication (Farley et al. 2018, Anderson et al. 2020, Kays et al. 2020). There is growing recognition that this open sharing of biodiversity data is critical for advancing biodiversity research (Farley et al. 2018). Some of the primary benefits of open biodiversity data include enhanced reproducibility of research (Alston and Rick 2021); making data available for reuse in new research applications (Chawinga and Zinn 2019); enabling researchers to receive credit, in the form of citations, for their efforts producing and sharing data sets (Costello et al. 2013, Brown 2021); and minimizing the duplication of research effort, enabling researchers to prioritize new data collection that fills research gaps (Troudet et al. 2017). As data sharing continues to become normalized, best practices have developed for the sharing of biodiversity data (Kühl et al. 2020). The FAIR data principles, for instance, outline four key attributes of effectively shared data: findable, accessible, interoperable, and reusable (Wilkinson et al. 2016). Specific practices have been developed to implement biodiversity data sharing in accordance with FAIR data principles. For example, global data aggregators such as the Global Biodiversity Information Facility (GBIF) provide a central location for aggregated data sets, ensuring that they will be findable and accessible (Robertson et al. 2014), and standardization schemes such as Darwin Core provide a mechanism for researchers to improve interoperability (Wieczorek et al. 2012). Such innovations support extensive data reuse; for example, the GBIF currently enables integrated data searches of nearly 1.7 billion species records from diverse sources around the world and has facilitated data reuse in thousands of publications (Heberling et al. 2021).

Although any type of data can be openly shared, the biodiversity data type most readily associated with open data sharing is presence-only occurrence data (König et al. 2019, Anderson et al. 2020, Wüest et al. 2020, Gadelha et al. 2021). Presence-only data consist of the taxonomic identification and location of an organism, often with the time of observation but without further information about species abundance, sampling design, or sites at which the species was not observed. The quantity of presence-only data aggregated in open biodiversity data repositories is immense and continuing to grow rapidly (Peterson et al. 2018, Ball-Damerow et al. 2019). This growth has been driven in large part by two simultaneous trends: the increasing popularity of citizen science platforms through which the public submit opportunistic observations to centralized databases (Theobald et al. 2015, Amano et al. 2016, Sullivan et al. 2017) and the digitization and aggregation of historical records and museum specimens (Speed et al. 2018, Nelson and Ellis 2019, Hedrick et al. 2020, Miller et al. 2020). The growing volume of openly shared presence-only data is also driven by characteristics of the data type itself: It is relatively simple and is easily standardized within existing best practices for data sharing (Anderson et al. 2020). Presence-only occurrence data now

*BioScience* 71: 1128–1147. 
https://doi.org/10.1093/biosci/biab072 Advance Access publication 18 August 2021

offer greater spatial, temporal, and taxonomic coverage on a global scale than other biodiversity data types and are often less costly and time intensive to collect (Tulloch et al. 2013, Bayraktarov et al. 2019).

As presence-only biodiversity data have grown in volume and accessibility, they have become increasingly common in biodiversity research (Peterson et al. 2018, Heberling et al. 2021). The open availability of massive modern and historical biodiversity data sets has contributed to a wide range of research areas, including ecology, biogeography, global change, and conservation (James et al. 2018, Ball-Damerow et al. 2019, Heberling et al. 2021). But the analysis of presence-only data is not without challenges; both historical and modern presence-only data are associated with limitations and biases that are distinct from other data types, both because of the lack of absence data and also because of the opportunistic collection process frequently associated with presence-only data (James et al. 2018, Støa et al. 2018, Gelfand and Shirota 2019, Grimmett et al. 2020, Sicacha-Parada et al. 2020, Johnston et al. 2021, Petersen et al. 2021). Further biases, errors, and limitations can be introduced in the processes of data preparation, publishing, and long-term maintenance (Tessarolo et al. 2017, Mesibov 2018), including the issues of data leakage (Peterson et al. 2018) and data obsolescence (Escribano et al. 2016). In response to these challenges, the growing application of presence-only data has been paralleled by an explosion of innovation in approaches to assess and improve both data accessibility and quality (Ball-Damerow et al. 2019) and also analysis methods that account for the specific limitations associated with this data type (Araújo et al. 2019, Kelling et al. 2019). As the development of analysis approaches for presence-only data continues, there is broad consensus that the documentation of metadata that details the study protocol, including information about sampling design or effort, allows for greater inference and also greater data reuse and reproducibility of analyses (Huettmann 2009, Kelling et al. 2019, Dobson et al. 2020, Foster et al. 2021). Open biodiversity data repositories commonly encourage the publishing of metadata (Poisot et al. 2019), but in practice the quality and amount of documented metadata varies widely (Peterson et al. 2018, Bishop et al. 2019, Anderson et al. 2020).

Although presence-only biodiversity data are reported and analyzed extensively in the traditional peer-reviewed literature, they are not restricted to it. In particular, authors who publish or access openly accessible biodiversity data may be more likely to seek out alternative outlets for research publication, such as preprint servers and journals with novel publishing models, because of their emphasis on free sharing of scientific information. Furthermore, biodiversity data are likely reported and analyzed often in gray literature and conference proceedings. Still, because a great deal of biodiversity data are reported and analyzed in the traditional peer-reviewed literature, it is important to understand the role that this literature plays in either facilitating or hindering the open sharing of biodiversity data. In this review we consider the extent of and barriers to the adoption of open data sharing practices within the traditional peer-reviewed literature, represented by the set of journals indexed by the Web of Science Core Collection.

Many aspects of the sharing and reuse of openly accessible biodiversity data in the peer-reviewed literature have been characterized, including common research applications of open data, taxonomic and spatial trends in open data, persistence of data stored in open databases, and current citation practices for open data (Troudet et al. 2017, Escribano et al. 2018, Ball-Damerow et al. 2019, Heberling et al. 2021, Luo et al. 2021). These studies make it clear that openly shared presence-only biodiversity data are foundational to a large body of biodiversity research. Still, many data go unshared. Earlier in the open data movement, it was widely recognized that open data formed just a small portion of the total biodiversity data known to exist (Ariño 2010, Amano et al. 2016, Peterson et al. 2018). But the current volume of presence-only data that are not openly shared, despite being presented and analyzed in the literature, is unknown. The concept of data sources and sinks can be helpful to conceptualize this issue; publication approaches that generate or perpetuate openly shared data can act as sources for continued data reuse, whereas publication approaches that entail a single use of data with no means for open access or reuse can be thought of as data sinks.

In the present article, we examine a broad cross section of the traditional peer-reviewed literature to assess the degree to which it serves as a source or sink for open presence-only biodiversity data. Our goal is to provide insight into the current adoption of open data practices among users of presence-only biodiversity data in journals indexed by the Web of Science Core Collection. To our knowledge, this is the first review of open data practices to be broadly defined by the presence-only data type, rather than by a particular type of data source, such as open databases. We focus on the following questions: How commonly does research published in articles indexed by the Web of Science Core Collection rely on presence-only data from open sources, and how commonly does it rely on data that are newly generated or compiled from other sources? To what extent do articles indexed by the Web of Science Core Collection serve as a data source for open presence-only biodiversity data; that is, are newly generated or compiled data made openly available, and are open data analyzed, documented, and cited in a way that supports continued reuse?

We identify both successes and challenges in the open sharing of presence-only biodiversity data, finding that the sharing of presence-only biodiversity data is overall increasing but that there is ample room for improvement in adherence to many data sharing best practices. We compare these findings with those of other recent reviews of the biodiversity literature, discussing trends that may be distinct to the presence-only data type, as well as new patterns that may be emerging within open data sharing practices. Because presence-only data are the biodiversity data type most

**Box 1. The search string used to query the Web of Science Core Collection to obtain literature.**

(((TS = ("presence-only" OR "presence only" OR "opportunistic observation*" OR "opportunistic species observation*" OR "opportunistic occurrence*" OR "opportunistic distribution*" OR "opportunistic species occurrence*" OR "opportunistic species distribution*" OR "pseudo-absence*" OR "pseudoabsence*" OR "inferred absence*" OR "presence-background" OR "presence background" OR "citizen science" OR "community science" OR "participatory science" or "ad hoc data" OR "ad hoc collection" OR "ad hoc method*" OR "incidental data" OR "incidental sighting*" OR "incidentally collected" OR "geographic one-class data" OR "incidental detection*" OR "opportunistic detection*" OR "primary biodiversity data*" OR "occurrence record*" OR "atlas data" OR "unstructured occurrence data" OR "unstructured species observation" OR "unstructured biodiversity data"))

AND (TS = ("distribution" OR "species" OR "biodiversity" OR "habitat*" OR "niche*")))

*AND* LANGUAGE: (English) *AND* DOCUMENT TYPES: (Article)

*Indexes = SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI Timespan = All years*

commonly associated with open data sharing, they can serve as an early indicator to illustrate the developing state of data sharing more broadly in the related fields of biodiversity, ecology, and conservation. Therefore, our characterization of current practices in presence-only data sharing can illuminate successes, challenges, and barriers to the adoption of data sharing practices that may be of growing relevance to the greater biodiversity research community.

## Review of the presence-only biodiversity data literature

We searched the Web of Science Core Collection to target all scholarly articles that report on the application of presence-only biodiversity occurrence data. Our search targeted articles whose titles, abstracts, or keywords contained any of 31 terms commonly used in the literature to indicate presence-only data as well as any of 5 terms used to indicate biodiversity (box 1). We screened the abstracts of all returned articles and retained those that demonstrated the analysis or reporting of presence-only occurrence data. After screening, a total of 2151 articles were included in the review (see the extended methods description in supplemental file S1). Data management and bibliometric summary statistics were conducted in part with the bibliometrix package in R (Aria and Cuccurullo 2017).

To identify broad trends in applications of presence-only data, we classified all included articles into three topic clusters using latent dirichlet allocation (LDA) topic modeling. LDA topic modeling uses word associations within a corpus to identify topic clusters and assigns documents to the topic clusters on the basis of word frequency within each document (Westgate 2019). We classified each document on the basis of the words in the abstract and title. LDA topic modeling requires the desired number of clusters to be defined, so to select a number of topic clusters we conducted LDA analysis six times, each time producing a different number of clusters ranging from three to eight. We used two criteria to select the number of clusters in our final topic model: First, we assessed the clusters for lack of redundancy in an ordination of all articles by their highest rated topic classification, and, second, we assessed the redundancy and interpretability of the sets of most highly weighted words in each set of clusters (see supplemental file S2; Asmussen and Møller 2019, Westgate 2019). The modeling iteration that produced three topic clusters was least redundant and most interpretable. The topic clusters were assigned descriptive names on the basis of the words most characteristic of each cluster: *methodological* articles were characterized by terms related to the application and assessment of analysis methods; *applied* articles were characterized by terms related to topics in biodiversity science, conservation, and related fields; and *records* articles were characterized by terms related to the collection and reporting of occurrence data (figure 1). Topic modeling was conducted with the revtools package in R (Westgate 2019).

A subset of 300 articles randomly selected from the included articles was read in full and coded according to a standardized data sheet (see supplemental files S3 and S4). The 300-article subset was representative of the full data set in terms of publication year and topic area (figure 2). For each article read in full, we recorded information on 10 fields: taxa, study system, study and author region, sample size, study scale, sampling design, analysis approach, data source, and data publication (see supplemental file S3). For all data fields except for study region and author region, the classifications were not mutually exclusive; each article was tagged with all applicable responses. Such classification is a common approach in descriptive literature reviews (e.g., Ball-Damerow et al. 2019, Hao et al. 2019). All data management and analyses were conducted with R version 4.0.2 (R Core Team 2020), and data and R scripts are available online (Mandeville 2021).

## Broad trends in the presence-only biodiversity literature

The literature relying on presence-only biodiversity occurrence data has grown steadily since the mid-2000s, maintaining an average annual growth rate that exceeds that of the biodiversity literature as a whole (Stork and Astrin 2014). This literature has seen a shift in recent years from

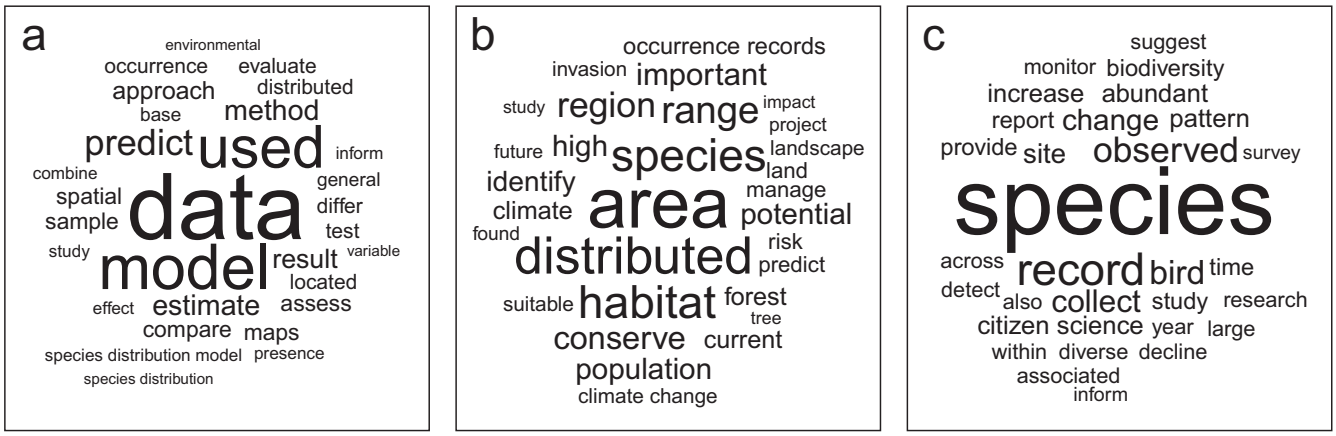



***Figure 1. The articles were classified into three topic areas using latent dirichlet allocation (LDA) topic modeling, which uses word frequencies to cluster articles by topic. The 30 most heavily weighted words in (a) the methodological topic (n = 641), (b) the applied topic (n = 753), and (c) the records topic (n = 757) are shown in the present figure. Word size indicates relative weight within each topic.***

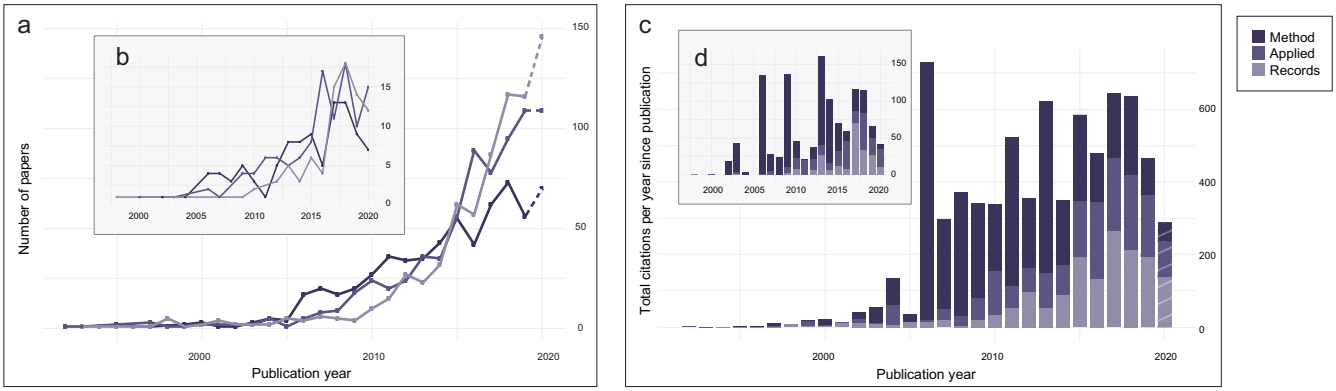



***Figure 2. The number of articles published per year in each topic area within (a) the full set of 2151 articles and (b) the 300-article subset; the total citations per year since publication in each topic area within (c) the full set of 2151 articles and (d) the 300-article subset. 2020 is indicated with dashed lines because the results for 2020 may be less complete than those for other years; although the set of articles was obtained with a search on 4 January 2021, some articles with a 2020 publication date may not yet have been indexed by journals or the Web of Science.***

a focus on methodological research to data sharing and applied analyses, as is evidenced by both the number of articles published and the citations obtained by articles in each topic area (figure 2). The *methodological* topic area was most common from the mid-2000s through 2015. From 2015 to 2020, the frequency of articles within the *methodological* topic area remained relatively constant, whereas the frequency of *applied* and *records* articles increased rapidly. *Methodological* articles are overall the most highly cited, but the relative citation rate has declined since 2015 (figure 2). The shifting distribution of topic areas suggests that there are two distinct eras in the presence-only data literature: an era focused on methodological developments, which lasted from approximately 2005–2015 and an era with a greater focus on applications that began in 2015 and continues today. A similar trend has been reported among articles that rely on GBIF-mediated data (Heberling et al. 2021).

The increase in articles focused on simple reports of occurrence is likely due to an increase in infrastructure and incentivization for data papers in recent years (Chavan and Penev 2011, Ball-Damerow et al. 2019, Li et al. 2020), and the parallel increase in applied research may indicate that presence-only approaches are being used more frequently to address issues of relevance to conservation and management (Guisan et al. 2013, Tulloch et al. 2018, Bayraktarov et al. 2019). The decline of methodological articles in terms of relative frequency and citation rate might suggest that applied researchers are using more established analysis methods more often than they are adopting newer approaches.

As a whole, the literature relying on presence-only biodiversity data is relatively decentralized and young. Its influence, as was measured by citations, is still growing; just a small number of the reviewed articles were highly cited, with a median of six citations per article. Unsurprisingly, *methodological* articles made up the majority of the 89 articles cited more than 100 times (figure 2; see supplemental file S5). The average author contributed to just 1.3 of the reviewed articles, which aligns with trends reported in the biodiversity literature (Stork and Astrin 2014) but is substantially lower than authorship rates in the biological sciences overall (Fanelli and Larivière 2016). Articles were published in a wide range of outlets, with 482 distinct journals represented in our review. The relative lack of common references is a further indicator of the varied scope of the presence-only biodiversity literature (see supplemental file S5). This is likely due to specialization among biodiversity researchers within many distinct research areas, defined for example by taxon of interest, geographic region, or scientific subdiscipline. Nevertheless, it may indicate a challenge to the efficient sharing of information regarding best practices for biodiversity data sharing.

## Using complementary reviews to build a more complete picture of the biodiversity literature

All efforts to systematically review literature contain trade-offs and biases introduced by the strategy used to search the literature, including search terms, search platform, and screening protocol. Therefore, efforts to characterize a body of literature are most informative when complementary reviews are considered alongside one another to form a more complete picture of the literature as a whole. We expect that this is particularly true for rapidly expanding research areas, including the presence-only biodiversity data literature; reviews of presence-only biodiversity data are complicated by the broad and rapidly developing variety of ways that this data type is accessed, analyzed, and referred to in the literature. To this end, we conducted a small test of the similarity of our search results to those of two recently published complementary reviews: Ball-Damerow and colleagues (2019) and the 2019 GBIF Science Review (GBIF Secretariat 2019). Each of these reviews used a search strategy and platform that complements our own, targeting a distinct subset of the literature on applications of presence-only biodiversity data (figure 3).

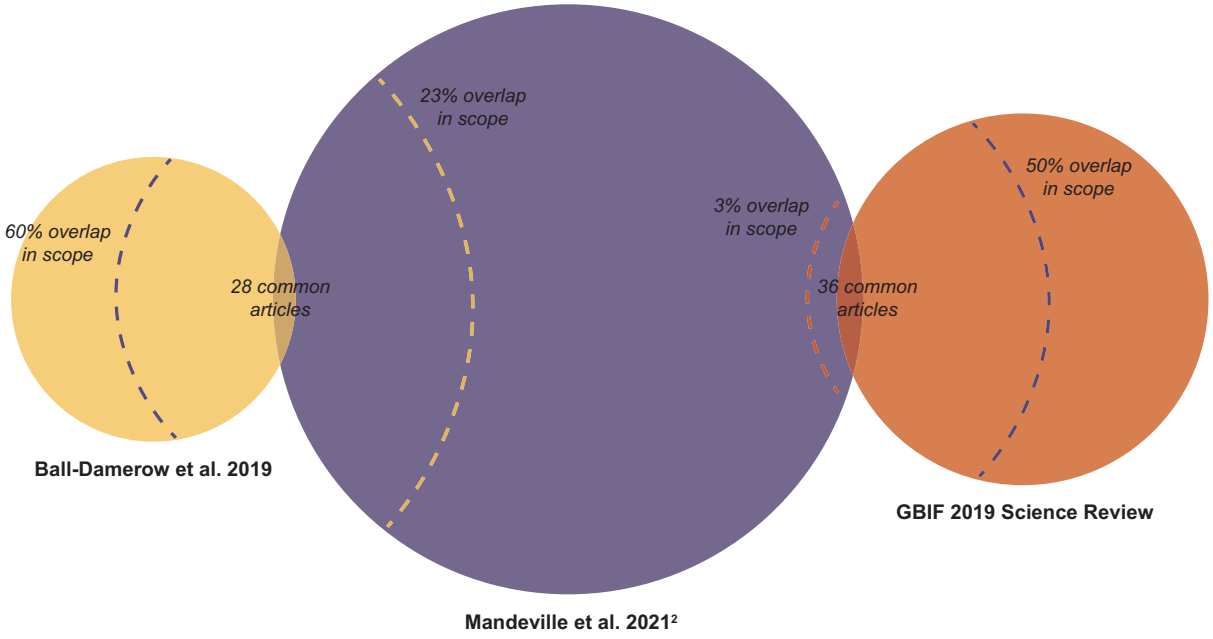
For this test, we identified the articles from our review that met the inclusion criteria defined for each of the other two reviews, screened the abstracts of 50 articles randomly selected from each of the other reviews according to our own inclusion criteria, and identified the percentage of articles that were common to our review and each of the complementary reviews. There was relatively little overlap between the articles in our review and the other two reviews (figure 3). The lack of overlap illustrates the importance of considering complementary reviews alongside one another. Although other recent reviews, including the two considered in the present article, have focused largely on applications of presence-only biodiversity data known to be accessed from open sources, our review fills a key knowledge gap by characterizing a broad set of the traditional literature with an as yet unknown reliance on open databases.

**Comparison of basic study characteristics with trends in biodiversity research.** Our review joins several recent studies in identifying trends in basic characteristics of the biodiversity literature, including taxonomic focus, study domain, and study region (Tydecks et al. 2018, Ball-Damerow et al. 2019, Heberling et al. 2021). We found that the articles in our review align some general trends in the biodiversity literature, including an emphasis on terrestrial settings (figures 4 and 5; Tydecks et al. 2018, Ball-Damerow et al. 2019, Heberling et al. 2021). Still, there are some distinct trends associated with the articles in our review: vertebrates—and, to a lesser extent, invertebrates—are better represented among our reviewed articles than in other reviews of the biodiversity literature, whereas plants and the freshwater domain are underrepresented (figure 4; Tydecks et al. 2018, Ball-Damerow et al. 2019, Heberling et al. 2021). The overrepresentation of vertebrates in our review is primarily due to their prevalence in reviewed articles that did not use data from open databases, suggesting that the range of vertebrate data available from open databases may not be as aligned with research needs as data from other taxonomic groups. On the other hand, the relative underrepresentation of freshwater and marine studies in our review was consistent between articles that did and did not rely on open data. This suggests that the presence-only data type as a whole may be less common in freshwater and marine domains, likely because many freshwater and marine species are not as easily detected via opportunistic observation.

The global distribution of studies in our review aligns closely with trends in the biodiversity literature (Tydecks et al. 2018, Heberling et al. 2021). The largest number of articles were authored by researchers based in Europe, followed by North America (figure 4). Alignment between study region and author region was uneven; articles that addressed Europe and North America were written by first authors based at institutions in the same region in respectively 98% and 95% of cases, whereas articles that addressed study regions in other parts of the world were less likely to have been written by first authors based in the focus region (figure 6). The uneven global distribution of biodiversity research reflects the greater coverage of biodiversity data in North America, Europe, and Australia relative to much of the rest of the world (Serra-Diaz et al. 2017, Pelayo-Villamil et al. 2018, Wüest et al. 2020) and is also partially explained by the less frequent publication of ecological research conducted in the Global South in journals that are indexed by major databases (Nuñez et al. 2019). It is critical that the field of biodiversity advances to better represent and support researchers based in underrepresented global regions in the international academic

| | Ball-Damerow et al. 2019 | Mandeville et al. 2021 | GBIF 2019 Science Review |
|---|---|---|---|
| **Number of reviewed articles** | 501 articles | 2151 articles (300 screened in greater detail) | 854 articles |
| **Search platform** | • Google Scholar<br>• Selected a predetermined number of returned articles | • Web of Science Core Collection | • GBIF literature tracking programme[1] |
| **Article inclusion criteria** | • Primary biodiversity data accessed from openly accessible online database<br>• Published between 2010 and April 2017 | • Presence-only biodiversity occurrence data<br>• Published before January 2021 | • Mention or citation of GBIF or GBIF data<br>• Published in 2018 |

[1] Draws from Google Scholar, Scopus, Wiley Online Library, SpringerLink, NCBI Pubmed, and bioRxiv




[2] Circle size refers to the 2151 articles used in a portion of analyses; 300 of these were screened in greater detail for further analyses.

***Figure 3. The Venn diagram indicates the overlap between articles included in this review and two complementary reviews. The circle size corresponds to review sample size; it should be noted that only a portion of the analyses reported in Mandeville (2021) were conducted on the full article set, whereas the remaining analyses were conducted on a subset of 300 samples chosen randomly from the full set. The overlap between the circles indicates the overlap in articles included in each review, and the dotted lines indicate the estimated overlap in targeted articles according to the reviews' described inclusion criteria. The inset table indicates the inclusion criteria and search strategy of each review.***

literature (Ramirez et al. 2018, Nuñez et al. 2019, Pettorelli et al. 2021). It has been shown that international collaborations are often inequitable, with European and North American researchers gaining more benefits in terms of publications and reputation than collaborators in the Global South (Boshoff 2009, Habel et al. 2014, Di Marco et al. 2017, Tydecks et al. 2018, Heberling et al. 2021). This trend should prompt caution in the growing open data

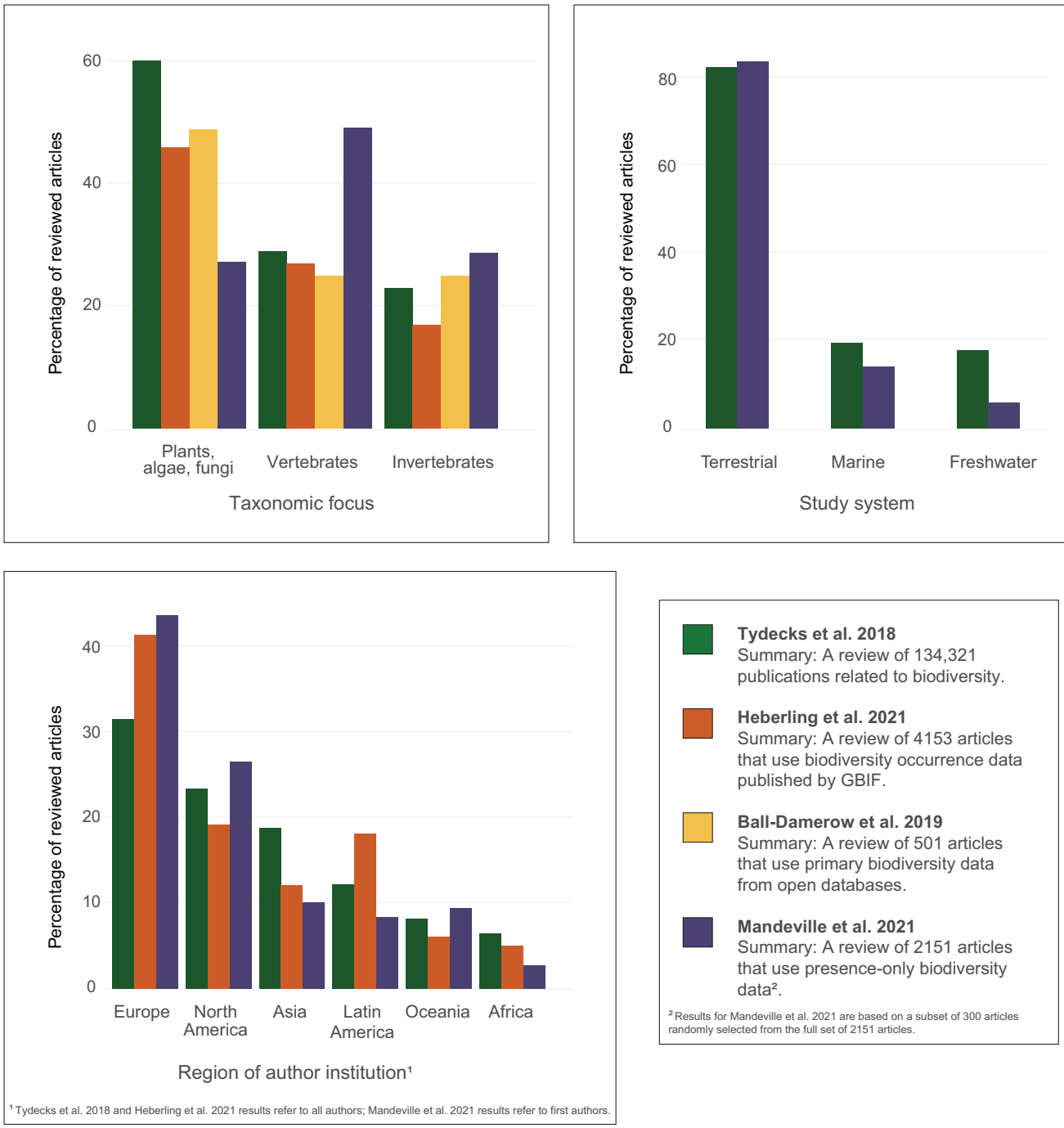



***Figure 4. A comparison of trends in taxonomic focus, study system, and geographic region of the biodiversity literature identified by this review and three complementary reviews covering different aspects of the biodiversity literature. See each cited paper for specific methods and results, because the methods of defining and measuring each trend may differ slightly between articles.***

movement; it will be essential to ensure that open sharing of data is supportive rather than exploitative of Global South researchers (Serwadda et al. 2018, Eichhorn et al. 2020, Pettorelli et al. 2021, Trisos et al. 2021). One example of an approach to this issue from within the biodiversity data community is the ongoing effort to repatriate biodiversity data that have been collected within a historically exploited region but stored and managed elsewhere, in order to transfer primary data custody and decision-making power back to the communities from which the data were collected (Dias et al. 2017, Eichhorn et al. 2020, Heberling et al. 2021).

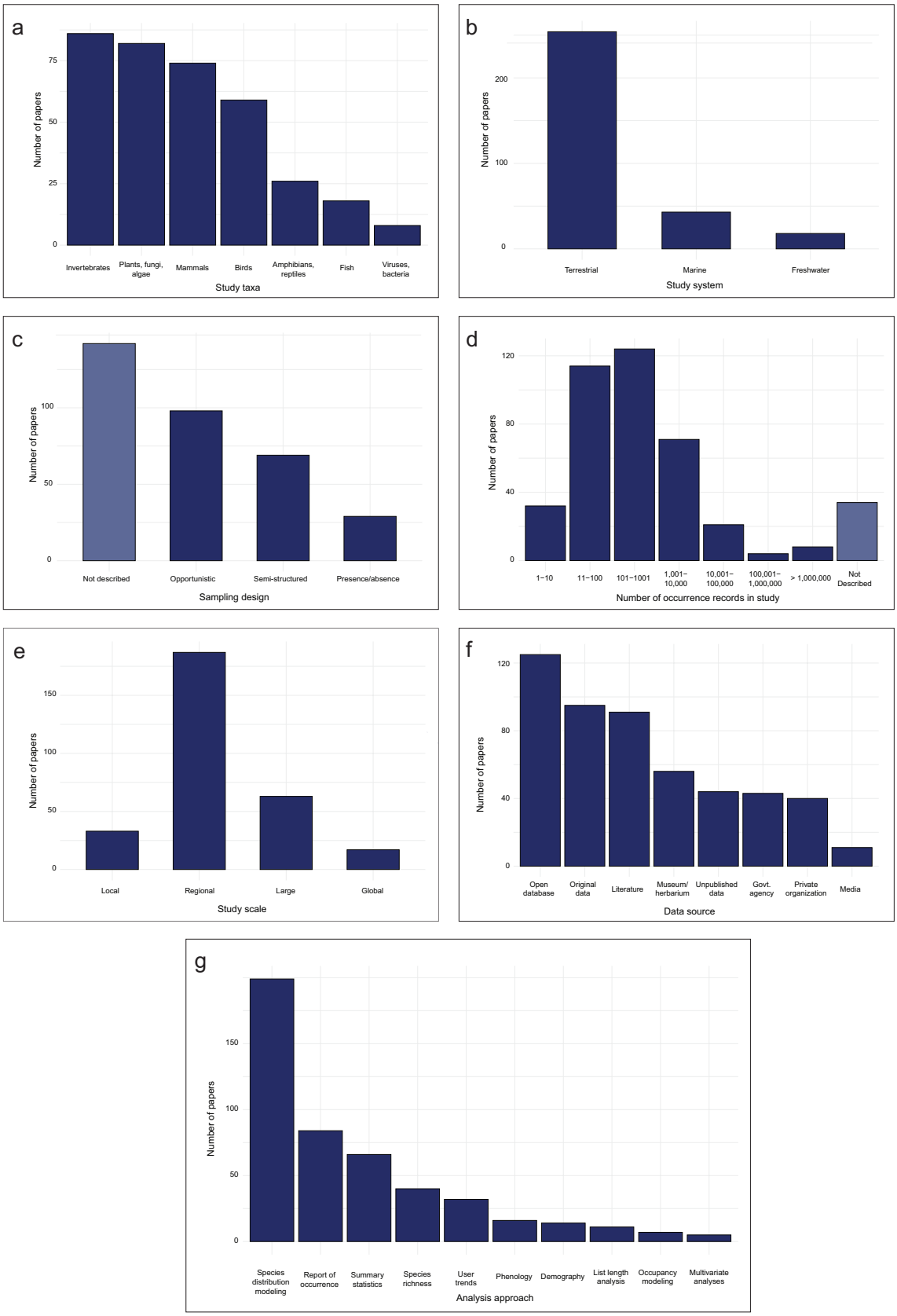



***Figure 5. The frequency of characteristics among the subset of 300 randomly selected articles: (a) study taxa, (b) study system, (c) sampling design, (d) sample size, (e) study scale, (f) direct data source, and (g) analysis approach. Characteristics are not mutually exclusive; multiple responses per characteristic can apply to an article.***

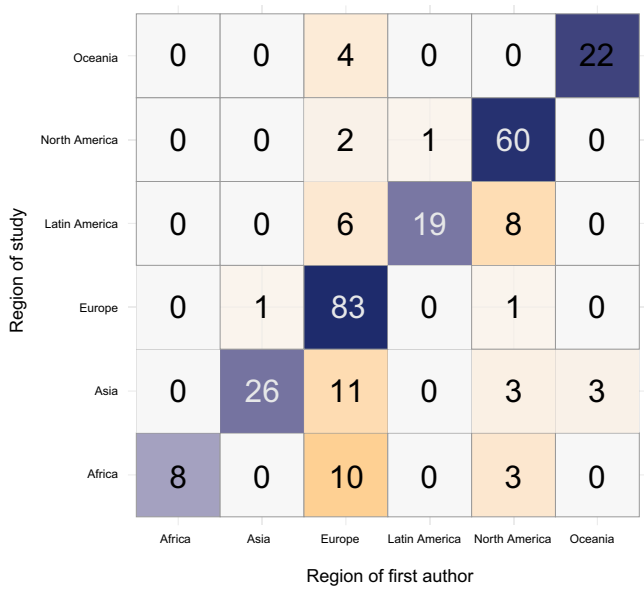



***Figure 6. The study regions of the subset of 300 articles are indicated on the* y*-axis and the region of the first author of each article, defined by institutional affiliation, is indicated on the* x*-axis. The number in each cell indicates the number of articles written about the region on the* y*-axis by a first author based in the corresponding region on the* x*-axis.***

**Presence-only data: A lens into current trends in the access, analysis, and publishing of openly accessible biodiversity data.** As the biodiversity research literature continues to grow, the open sharing of biodiversity data is increasingly recognized as necessary and is quickly becoming normalized (Peterson et al. 2018, Ball-Damerow et al. 2019, Heberling et al. 2021). Presence-only biodiversity data are relatively representative of broad taxonomic and geographic trends associated with the field of biodiversity as a whole, but they differ in the ease with which they can be shared in accordance with currently recognized best practices (König et al. 2019, Anderson et al. 2020, Wüest et al. 2020, Gadelha et al. 2021). Therefore, as practices continue to be developed to facilitate the sharing of a wide range of data types (Anderson et al. 2020), presence-only data can serve as an early indicator to illustrate the progress, challenges, and limitations to the adoption of biodiversity data sharing practices. The work of recent reviews focused on presence-only data from open databases (e.g., Ball-Damerow et al. 2019 and the GBIF Science Review series) makes it clear that open data infrastructure actively supports a large body of research. But to understand the extent to which biodiversity research in the traditional peer-reviewed literature serves to facilitate or slow the progress toward open data, it is necessary to consider presence-only data from a wider range of sources.

In the sections that follow, we focus on three aspects of the presence-only biodiversity data literature indexed in the Web of Science Core Collection, with an emphasis on open data practices. We first consider the sources of presence-only data in this body of literature. Next, we consider how presence-only data are analyzed and whether these analyses are supported by well-documented metadata. Finally, we characterize the data publication practices associated with the presence-only biodiversity data in this set of literature. Our objective is to delineate the current state of data sharing practices and to identify areas for growth, many of which will apply to both presence-only data and also more generally to a range of biodiversity data types.

## Sources of presence-only biodiversity data

Openly accessible databases—that is, searchable online repositories in which biodiversity data from many original sources are aggregated—make billions of biodiversity data points freely available for anyone to access and use (Peterson et al. 2018, Ball-Damerow et al. 2019). Researchers may choose to access data from openly accessible databases for many reasons: to avoid duplicating research effort that has been undertaken in the past, to access data on a larger temporal and spatial scale than could be collected through original field work, to synthesize data from disparate sources, or to replicate or build on a previous study. So it is unsurprising that openly accessible databases were the most common direct data source in our review, accessed by 42% of the reviewed articles. However, only 19% of the reviewed articles used data exclusively from open databases; the vast majority accessed some or all of their data from sources other than open databases. Other common data sources include original fieldwork, the literature, and museums and herbaria (figure 5). Ball-Damerow and colleagues (2019) identified these same three sources of occurrence data as the most commonly integrated with occurrence data accessed from open databases.

In many cases, it is likely that researchers choose to collect new data or compile data from a variety of original sources because the data they need are not available in an openly accessible database (Troudet et al. 2017, Ball-Damerow et al. 2019). For instance, articles in our review were substantially more likely to address vertebrate species than in reviews in which all articles rely at least partially on open data (figure 4). In particular, a large percentage of the articles in our review addressed mammals (figure 5). Although mammals are considered overrepresented in open databases on a per-species basis, they make up a relatively small portion of the total volume of data available from open databases, likely because of many mammal species' lower detection probability, wider-ranging distributions, and relatively lower dedicated citizen science interest than some other taxa (Troudet et al. 2017, Parsons et al. 2018). This may explain why articles that addressed mammal species were relatively unlikely to obtain data from an open database and more likely to obtain data from government agencies, private organizations, and through original data collection. Overall, the relatively small percentage of articles based on open presence-only data corroborates a growing sentiment from

the literature: Although the volume of openly accessible biodiversity data continues to grow, there are substantial taxonomic and spatial gaps for which there is minimal open data (Pino-Del-Carpio et al. 2014, Chambers et al. 2017, Troudet et al. 2017, Ondei et al. 2018, Wetzel et al. 2018, Ball-Damerow et al. 2019, Hochkirch et al. 2020). Our results corroborate the many studies that have identified gaps in biodiversity data, making it clear that the majority of researchers who conduct presence-only analyses do not find the data they need in open databases. This highlights the need for the biodiversity research community to continue ongoing efforts to identify and fill critical taxonomic and spatial knowledge gaps in open databases.

Data gaps can be filled through both novel data collection and mobilization of existing data that are not yet openly accessible. Many large pools of data exist outside the open data infrastructure—for example, in government agencies and private organizations (Stephenson et al. 2017, Wetzel et al. 2018, Cretois et al. 2020). Identifying these sources of data, supporting policies and infrastructure that facilitate their access and reuse, and incentivizing data sharing at an institutional level is needed to facilitate more open access to these data (Vořišek et al. 2018). This is critical for establishing the long-term records that are essential for studying trends across space and time and informing conservation interventions in the face of global change (Wetzel et al. 2018). Opening existing data for reuse is also necessary to avoid duplication of data collection effort and research waste, freeing research resources to target true data gaps (Grainger et al. 2020). Consider, for example, that 13% of the articles in our review accessed data from 10 or more nonopen sources, some accessing well over one thousand distinct sources. The collation of data from multiple sources represents an extensive research effort that will likely need to be repeated by future researchers if the data are not made more openly accessible. Reducing inefficiencies by supporting the access and reuse of data will allow researchers to prioritize generation of data that will fill gaps in the available knowledge. To achieve this, efforts to build relationships between data aggregators and the research community will continue to be essential.

In other cases, openly accessible data may be available to replace or supplement data from other sources but authors may neglect to use it, either because they are not aware of it or because they do not trust its quality (Faith et al. 2013). Even when data are aggregated in an open database, some researchers may choose to access the data from their original sources rather than from the open database (Singer et al. 2020). In some cases, researchers may be aware of open data but believe they lack the skills to access and use it effectively (Poisot et al. 2019). Indeed, a broad survey of researchers found that the perceived value and efficiency of reusing open data were major factors in whether researchers chose to access open data (Curty et al. 2017). Finally, it is also important to note that inequities in technological infrastructure, competence, and training mean that access to digital platforms is also inequitable (Johnson et al. 2021). Finding solutions to the barriers that keep researchers from accessing open biodiversity data should be a goal of the biodiversity research community.

**Practices for accessing and citing open data vary widely.** Among open databases, data sources varied widely. We identified 117 open databases that were used to access presence-only occurrence data (see supplemental file S6). We classified nine of these as large open databases, defined as relatively well known, established databases that contain data covering a very large geographic range, a wide range of taxa, or both. The most commonly accessed was the GBIF, which was accessed by 37 articles, followed by eBird (9 articles). The remaining 108 open databases, classified as small databases, had a narrower geographic or disciplinary scope and were each accessed by an average of 1.2 articles. Of the articles that accessed open data from at least one source, 55% accessed a large database and 65% accessed a small database. Two thirds directly accessed just one database, whereas the remaining third accessed between two and 10 distinct open databases. Of course, because many open data sources serve to aggregate many smaller databases, data users that accessed just one database may still have obtained data from a wide range of original sources. These results are similar to the findings of Ball-Damerow and colleagues (2019), who also found that a small number of open data sources were cited by many articles, whereas a large number of open data sources were cited very few times.

The frequent reliance on small open databases is probably due in large part to the prevalence of small databases within specific research areas (Costello and Wieczorek 2014, Ball-Damerow et al. 2019, Singer et al. 2020) and may also be partially explained by a lack of familiarity with or trust in large databases (Faith et al. 2013). We recognize many values of small databases, including responsiveness to specific disciplinary requirements (Franz and Sterner 2018) and the cultivation of strong relationships between data curators and communities of data users (Blair et al. 2020, Monfils et al. 2020). However, small open databases may lack the standardization and interoperability that are built into larger data aggregators (Poisot et al. 2019), they may lack consistent leadership to maintain growing content and keep up with developing best practices (Costello et al. 2013), and they are more likely to become technologically obsolete, rendering the data inaccessible (Vines et al. 2014, Tessarolo et al. 2017, Ball-Damerow et al. 2019, Blair et al. 2020).

We attempted to access all of the databases referred to in our reviewed articles and found that we could not locate or access 9% of the small databases from which articles in our review had obtained data. In a few other cases, the database website could be accessed, but it was not clear that the data were still accessible; for example, data could be visualized but the link to download data was broken, or it was requested that visitors contact the database managers to request access. Although still concerning, it is perhaps a cause for cautious optimism that the proportion of

inaccessible databases in our review is considerably lower than the 26% of databases found to be inaccessible by Ball-Damerow and colleagues (2019), who reviewed articles published through April 2017. An additional 15% of the small databases had been consolidated into a different database but were still accessible. All nine large databases remained accessible. Because of the important role played by small databases, we do not intend to suggest that authors avoid them; rather, we caution the biodiversity data community to be cognizant that these small databases are strongly relied on and to be proactive about supporting them over time (Costello and Wieczorek 2014). The true reliance on small databases is likely to be even higher than identified in our study because small regional databases may be cited more frequently by articles published in regional journals and gray literature, which may not be indexed by the Web of Science and so may have been underrepresented in our search (Calver et al. 2017).

The proliferation of open data aggregators, along with the rapidly evolving best practices for their use, has resulted in an uneven landscape of how such data are cited in the literature (Escribano et al. 2018, Ball-Damerow et al. 2019, Luo et al. 2021). Citation of a digital object identifier (DOI) that is uniquely connected to the full data set analyzed in an article has emerged as the best practice in this area (Brown 2021, Heberling et al. 2021); this practice enables the data set to be clearly replicated and all original sources to be credited (Escribano et al. 2018, Luo et al. 2021). But not all researchers are yet aware of this best practice, because it is relatively new. Furthermore, not all open databases have a clear mechanism for producing a citable DOI (Altman and Crosas 2014, Penev et al. 2017). We found a great deal of variation in how open databases were cited among the articles in our review. The vast majority of articles simply listed the names of the databases from which they obtained data, sometimes accompanied by a brief description of the type of original sources from which the data were aggregated. Only 4% of the data sets accessed from an open database were cited with a DOI, and another 3% were not cited but, instead, were described in the text of the article with a direct link to the full data set or other thorough directions that would enable a reader to replicate the data retrieval process. Interestingly, the proportion of articles in our review that included a database citation with a URL or DOI was much lower than the 34% observed by Ball-Damerow and colleagues (2019). This may reflect a difference in search strategy; the search terms used by Ball-Damerow and colleagues (2019) ensured that all reviewed articles at least mentioned the type of database accessed, whereas our search terms required only that articles mentioned the type of data. The differing results obtained by these two searches suggest that the use of appropriate citation practices may be correlated with authors' use of specific terminology to refer to open databases, perhaps signaling their perception of their work as related to the open data movement.

A small number of authors in our review found alternative ways to recognize original providers of data even when there was no mechanism to do so through the open database—for example, by listing all original data sources in the supplemental material. Giving credit to the original providers of open data is critical for incentivizing data sharing to researchers, institutions, and funders (Escribano et al. 2018, Ball-Damerow et al. 2019, Groom et al. 2020) and for recognizing and supporting the diverse landscape of organizations and institutions that engage in biodiversity monitoring (Kühl et al. 2020). This may be especially true when data were collected through public involvement in citizen science. Thirty-four percent of the articles in our review identified citizen science as the original source of some or all of their data, although the true percentage of articles that derived data from citizen science is likely higher because citizen science data are frequently reused without their source being clearly described (Cooper et al. 2014). Citizen science plays an important role in biodiversity data collection but long-term funding and support for many citizen science programs may be dependent on the demonstrated impact, so appropriate citation is critical (Chandler et al. 2017, Pearce-Higgins et al. 2018, MacPhail and Colla 2020, Mandeville and Finstad 2021).

### Analysis and reporting of presence-only biodiversity data and associated metadata

The growth of interest in presence-only data in the mid-2000s was paralleled by innovation in species distribution modeling approaches tailored to this data type (Vaz et al. 2015, Araújo et al. 2019, Ball-Damerow et al. 2019), so it is unsurprising that species distribution modeling was the dominant analysis approach in our review (figure 5). These methods have become increasingly sophisticated and widely popular (Hao et al. 2019, Norberg et al. 2019, Zurell et al. 2020). A large review of articles that use GBIF data found a similar prevalence of species distribution modeling and identified a recent transition in focus from methodological developments to widespread application similar to that seen in our overall set of reviewed articles (Heberling et al. 2021). Although the initial development of species distribution modeling approaches for presence-only data was at least partially a response to the increased availability of the data type, we suggest that their subsequent wide adoption has created a positive feedback effect whereby researchers, driven by the growing ease of analyzing presence-only data, have increasingly begun to seek out presence-only data from a wider range of sources.

Despite its prevalence, however, species distribution modeling is far from the only analysis method applicable to presence-only data. Our results illustrate a wide range of analysis approaches, including both inferential statistics and a variety of descriptive statistics. Presence-only data are also occasionally used indirectly—for example, to validate the results of another analysis or to inform a sampling design. Methodological innovation in inferential approaches is ongoing, and since 2012, a number of articles have applied a variety of less common inferential approaches, including phenology analyses, demography analyses, list length

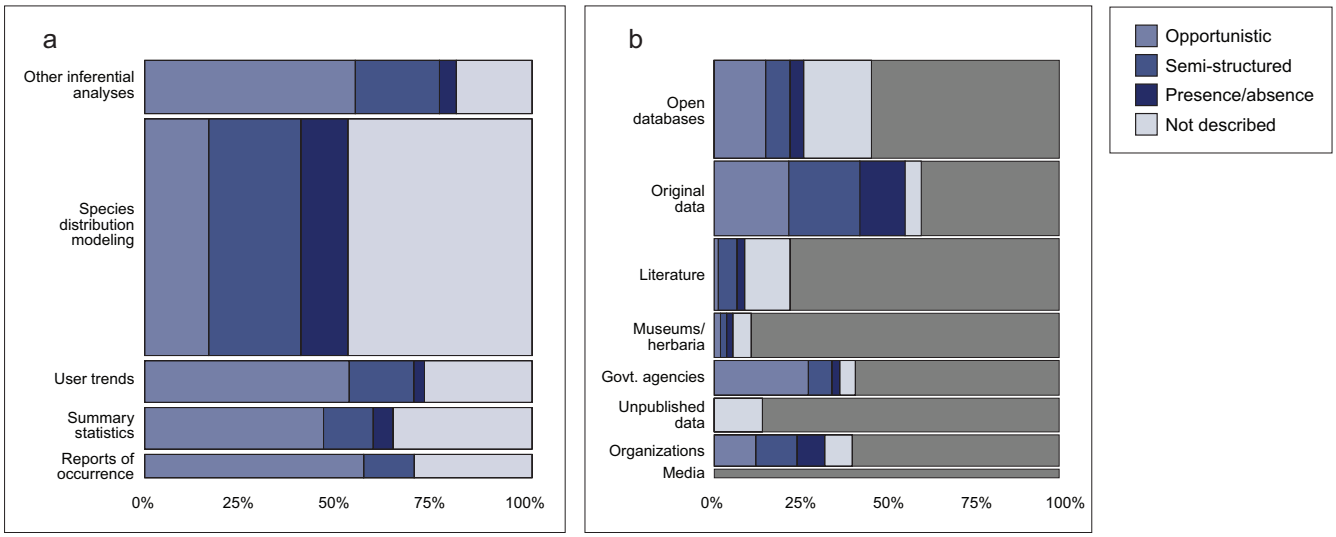



***Figure 7. The percentage of the 300-article subset that is associated with each type of data structure, as a function of (a) analysis approach and (b) direct data source accessed by study authors. In panel (a), the* y*-axis categories represent all articles for which the indicated analysis approach was the most complex approach applied (with the exception of "user trends," in which case all articles using this approach are represented). The bar widths indicate the number of articles in the 300-article subset within each category. In panel (b), the* y*-axis categories represent all articles that use data from the indicated data source. The bar widths indicate the overall proportion of the 300-article subset that used each data type. The gray portions of the bars represent articles that integrated data from the indicated source with data from other sources; because of the confounding effect of data integration on metadata reporting, metadata reporting trends are not reported for these articles. The portions of the bars shaded according to the legend represent articles for which the indicated source was the only source accessed by the article.***

analysis, occupancy modeling, and multivariate statistics (figure 5). In particular, the integration of presence-only data with other types of biodiversity data is of growing interest in the literature (Pacifici et al. 2017, Fletcher et al. 2019, Miller et al. 2019, Isaac et al. 2020, Simmonds et al. 2020, Zipkin et al. 2021). In our review, articles that integrated presence-only data with other types of biodiversity data were nearly three times as likely to employ an uncommon inferential analysis approach as the articles that used only presence-only data, indicating that data integration can open a wider range of analysis options for presence-only data.

Clearly documented metadata, particularly an explicit description of the data structure and original sampling design, also enable a wider range of analytical approaches, including data integration (Isaac et al. 2014, Araújo et al. 2019, Dobson et al. 2020). This trend is reflected in our results, with articles that employed more complex analysis approaches being correspondingly more likely to describe the underlying data structure (figure 7). Articles that employ species distribution modeling are the major exception to this trend; despite the relative statistical complexity of species distribution modeling, articles that modeled species distributions were the least likely to document data structure (figure 7). This likely reflects the growing accessibility of species distribution modeling approaches, which have become increasingly straightforward to implement through user-friendly platforms that can be implemented as a black box by researchers without a clear understanding of the method (Joppa et al. 2013, Merow et al. 2013, Kass et al. 2018). Although the growing accessibility of species distribution modeling offers great potential for research and conservation (Rapacciuolo 2019, Sofaer et al. 2019), we caution that it is still essential to share metadata whenever possible to aid in interpretation and evaluation of results (Soranno et al. 2020, Zurell et al. 2020, Muscatello et al. 2021, Sillero and Barbosa 2021, Foster et al. 2021). Relatedly, it is important to check for and correct data quality errors in data and metadata, particularly when data are obtained from open databases or collated from several sources (Ball-Damerow et al. 2019). In addition to supporting data interpretation and analysis, the reporting of high quality metadata facilitates a wide range of potential future data uses.

**Reporting of metadata is inconsistent.** Despite the value of clear metadata, around half of the articles that we reviewed did not explicitly describe the structure or sampling design of all of their data, corroborating previously reported trends (figure 5; Kervin et al. 2013, Roche et al. 2015). Of course, researchers can only report metadata if they have access to this information, and researchers reusing data may simply not have information on the original data structure. For instance, 118 articles obtained data from museums,

herbaria, and the literature and 77% of these did not report the structure of their data; in the vast majority of these cases, metadata on the original sampling design were likely unavailable. Users of open data also have inconsistent access to metadata, and around half of the articles that obtained data exclusively from open sources did not describe data structure (figure 7). Although many openly accessible databases enable and encourage metadata standardization and sharing, most prominently through the Darwin Core standard (Wieczorek et al. 2012), many data available through open databases have been digitized from historical records, for which such metadata may be unavailable or may have been lost over time (Specht et al. 2018). Articles that rely on data collected by government agencies and private organizations describe data structure more frequently (figure 7). In the instances in which the structure of data from these sources is not described, it may be due to the loss of information that occurs when complete information was not passed from the data owners to the data users. Standardizing the methods used by governmental and private institutions to share data with researchers may reduce instances of data loss associated with more informal sharing of data (Kühl et al. 2020). Unsurprisingly, articles exclusively based on original field work were most consistent in documenting data structure (figure 7). The combination of data from multiple sources is an additional barrier to describing presence-only data because of practical challenges associated with describing a large number of separate sampling schemes. For each additional source accessed by an article in our review, the likelihood of data structure being described decreased by 12%. Although authors may have little recourse when working with data sets for which metadata are unavailable or with large data sets for which it may be impractical to describe a large number of separate sampling schemes, improving data citation practices may provide a partial solution by making it possible to trace data to its original source to gather any available metadata.

Of articles that described the structure of their data, most described one or more data source as opportunistic (i.e., collected with no predefined sampling design), followed by semistructured (*sensu* Dobson et al. 2020), and finally a smaller percentage used presence or absence data and discarded the absence records before analysis. Of the articles that converted presence or absence data to presence-only format before analysis, one third did this for the purpose of comparing different modeling approaches. The remaining two thirds discarded the absence data and conducted analyses exclusively in a presence-only framework. Previous authors have cautioned that it is not advisable to analyze presence or absence data in a presence-only framework (Yackulic et al. 2013), so it is concerning that some articles in our review took this approach. In some cases researchers may be motivated to convert presence or absence data to presence-only to facilitate merging presence or absence and presence-only data sets, but many recent studies suggest approaches for integrating various data types without reducing data structure (Pacifici et al. 2017, Fletcher et al. 2019, Miller et al. 2019, Isaac et al. 2020, Zipkin et al. 2021).

The articles in our review were more consistent in reporting the scope of their presence-only data set, in terms of both sample size and study scale. The sample size varied considerably between articles, but the majority of studies were small to mid-size (figure 5). The studies' geographic scale followed a similar trend, with the majority addressing a regional scale (figure 5). The small number of articles that did not explicitly state a sample size tended to involve several separate analyses of a large number of species and stated a total sample size and total number of species rather than the sample size for each analysis. The tendency toward mid-size studies has remained relatively consistent over time, with the exception of studies with a sample size of over one hundred thousand occurrence records. These very large studies were absent from our reviewed articles until 2014. This recent increase in large studies likely reflects growing infrastructure for and interest in big data macroecology (Hampton et al. 2013, Wüest et al. 2020). Such large studies are more likely to rely on open data than studies with a smaller scope.

### How often are presence-only data made available for reuse?

Our results suggest that the majority of data used in presence-only analyses are not made available after the analyses are published, although there is a recent trend toward increased data sharing. To characterize trends in data sharing, we excluded the 19% of articles that were based entirely on data accessed from open sources. Of the remaining articles that used data from at least one source other than an open database, just 21% made all data used in the study openly available on publication of the article. Of these, 18% published their data in an openly accessible online database, whereas the rest used a different form of publication, such as supplementary material or an online repository (figure 8). The most common means of sharing data was to directly include it in the article, either the main text or the supplemental material. Data formats varied from those that facilitate reuse relatively easily (e.g., CSV files, spatial data files) to those that pose challenges for reuse (e.g., PDF files). Online repositories, including Dryad, Figshare, and GitHub, were also used by a small number of articles to share data. Only nine articles indicated that their data sets had been shared in an openly accessible database, although it is possible that the authors of some articles in our review published their data to an open database but neglected to mention this in the article. Of course, the data analyzed in the 19% of reviewed articles that obtained data exclusively from open databases remained openly available as long as the databases from which the authors accessed their data were still accessible.

To maximize their research value, data must be published in a way that is both searchable and persistent (Wilkinson et al. 2016, Bishop et al. 2019). Therefore, publication of data in aggregated databases is preferable to publication in supplemental material. In particular, larger

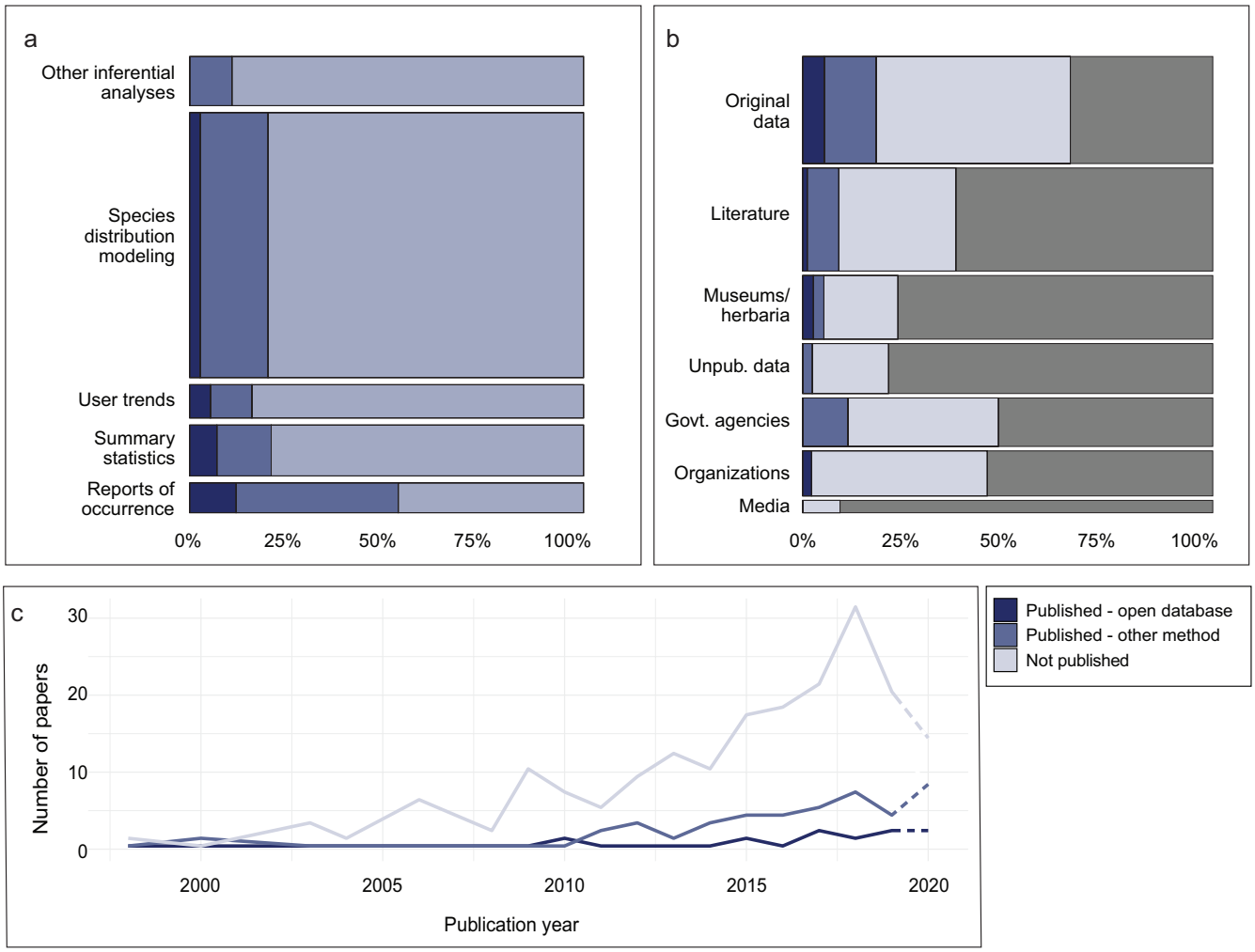



***Figure 8. The percentage of the 300-article subset that is associated with the three levels of data availability as a function of (a) analysis approach and (b) direct data source accessed by study authors. For all panels of this figure, articles based entirely on data accessed from open databases have been excluded, leaving a subset of 242 articles that access data from at least one source other than an open database. In panel (a), the y-axis categories represent all articles for which the indicated analysis approach was the most complex approach applied (with the exception of "user trends," in which case all articles using this approach are represented). The bar widths indicate the total number of articles within each category. In panel (b), the y-axis categories represent all articles in which the indicated direct data source was accessed. The bar widths indicate the overall proportion of the 242-article subset that used each data type. The portions of the bars shaded according to the legend represent articles for which the indicated source was the only source accessed by the article or which integrated the indicated source with open data. The gray portions of the bars represent articles that integrated data from the indicated source with data from other sources; because of the confounding effect of data integration on data sharing, data sharing trends are not reported for these articles. Panel (c) indicates trends in data availability over time. 2020 is indicated with dashed lines because the results for 2020 may be less complete than those for other years; although the set of articles was obtained with a search on 4 January 2021, some articles with a 2020 publication date may not yet have been indexed by journals or the Web of Science.***

databases are more likely to have greater longevity, stability, and infrastructure to maintain current best practices for data management in this rapidly developing field (Costello and Wieczorek 2014, Poisot et al. 2019). Much like small open databases, it has been demonstrated that data in supplementary material often become inaccessible over time (Vines et al. 2014, Stodden et al. 2018). We attempted to access all data shared by our reviewed articles and found that it was largely, but not entirely, still accessible: 7% of the data sets shared in journal supplementary materials were no longer available, and 22% of the data sets shared in an open database were no longer available. The inaccessible data from open databases were exclusively shared in small databases.

Although the overall accessibility of openly available presence-only data has increased dramatically in recent years, our results make it clear that the traditional peer-reviewed literature still largely serves as a sink for presence-only biodiversity data rather than facilitating its sharing and reuse. Making presence-only data more accessible should be a clear priority. Because strong infrastructure and clear best practices already exist for sharing presence-only occurrence data (Costello and Wieczorek 2014, Peterson et al. 2018, Hackett et al. 2019, Anderson et al. 2020) this should be achievable. However, several barriers can stand in the way of data sharing, including researchers' lack of incentive and ability, data ownership, and data set complexity. The strategies for overcoming these barriers will differ on the basis of the original source, ownership, and structure of the data.

**Data sharing considerations for different types of presence-only data.** The most straightforward type of presence-only data to target for increased data sharing are likely those collected by the study authors. Our results do indicate that original data are the most frequently shared, but the sharing rate is still low, at just 27% (figure 8). The publishing rate of original data collected with citizen science was somewhat higher than average, although still fewer than half of the articles based on original citizen science published their data. This is problematic, because studies have shown that citizen science participants generally expect and want their data to be made available for research, conservation, and policymaking (Chandler et al. 2017, Ganzevoort et al. 2017, Groom et al. 2017, Fox et al. 2019, Larson et al. 2020). Further integration of citizen science with open biodiversity data aggregators should therefore be a priority.

We anticipated lower rates of data publication from articles that compiled data from third party data owners, including the literature and museums and herbaria, and our results indicated rates of data publication that were just slightly lower than that of original data (figure 8). We suggest two major reasons why authors may not share data they have collated from other data owners. First, they may lack (or perceive that they lack) the permission to do so. And second, they may perceive that data sharing is unnecessary, assuming that readers wishing to reproduce their data set could retrace the data acquisition methods described in the paper to reassemble the data set from its original sources. Although this may sometimes be true, collating data from multiple sources takes a great deal of time and effort, so it is not a trivial process for a reader to reassemble a data set following a process described in the literature. And even if original data sources are well documented and still accessible, it cannot be assumed that a reader will be able to replicate the steps taken to collect data; literature is often behind paywalls, and access to institutional databases may be limited. Therefore, researchers working with data compiled from museums, herbaria, and journal articles should strive to provide as thorough a description as possible of their exact process of compiling their data set or, better yet, publish their complete data set whenever possible (Cousijn et al. 2018). Widespread progress on this issue will depend in part on the support of institutions: Institutions that host data should institute mechanisms to generate citations when data are accessed, making data easier to cite (Mooney and Newton 2012, Fenner et al. 2019, Powers and Hampton 2019), and journals that publish research should outline clear policies that support and facilitate data sharing and citation (Hrynaszkiewicz et al. 2020).

Finally, there are circumstances in which researchers may be unable to share data because of its proprietary or sensitive nature. We expect that this issue is most relevant to data obtained from private organizations or government agencies; in the present review, articles that accessed data primarily from one of these sources were characterized by low rates of data publication (figure 8). This is a complex issue, but we would encourage owners of sensitive data to use existing decision tools and prioritization schemes to consider whether there is a suitable way to make these data available for reuse, even in a more limited format (Clements et al. 2018, Tulloch et al. 2018, Chapman 2020). Because 37% of reviewed articles derive at least a portion of their data from sources that are assumed to generally be nonopen (e.g., data provided by government agencies, private organizations, or personal communications), and 41% derive some or all of their data from sources that are potentially accessible but cannot be assumed to be available to all readers (e.g., museums, literature, media), it is clear that a large portion of the presence-only biodiversity literature relies on data that are not accessible, hampering the replicability of these studies and the reusability of the data on which they are based.

A separate but related issue concerns data ethics and ownership. Issues of data ownership and governance are inherently related to social governance, and it is essential that the ethics of data sharing be held in the forefront at all stages of data management (Carroll et al. 2021, Rubert-Nason et al. 2021, Trisos et al. 2021). Data relevant to local communities must be made accessible to community members and must not be used in ways that are counter to community priorities (Johnson et al. 2021). This is particularly essential when it comes to Indigenous data; the CARE Principles for Indigenous Data Governance are a critical framework for ensuring Indigenous peoples' rights to the control of Indigenous data (GIDA 2019, Carroll et al. 2021). In addition, when data are collected by community members, as with citizen science, it is important to understand and respect volunteers' motivations for and concerns about the use of data they have contributed (Ganzevoort et al. 2017, Lynn et al. 2019, Tengö et al. 2021). The continued normalization of open data sharing must center scholarship and practice that respects ethical data governance, stewardship, and access.

**The future of presence-only biodiversity data sharing.** Data sharing practices in the presence-only biodiversity literature have until recently remained relatively constant over time, but the

proportion of reviewed articles that publish their data has increased somewhat since 2016 (figure 8). This is cause for optimism and continued efforts to normalize open sharing of biodiversity data. Recent studies document overwhelmingly positive attitudes to data sharing (Tenopir et al. 2020, Soeharjono and Roche 2021), so if practical barriers can be overcome, there is a high likelihood that data sharing will continue to increase. Increased sharing of biodiversity data may even produce a ripple effect across disciplines; biodiversity research has historically exhibited a higher rate of open data sharing than closely related scientific disciplines such as ecology and conservation science (Michener 2015, Osawa 2019, Shin et al. 2020), but given the broad and growing application of presence-only biodiversity data across many related scientific disciplines (Ball-Damerow et al. 2019, Heberling et al. 2021), continued improvements in open sharing of presence-only biodiversity data may serve to spread awareness of open data practices across disciplines.

Past studies have indicated that the majority of biodiversity researchers support data sharing but may be held back by lack of sufficient incentive, lack of familiarity with data aggregators, lack of information on data set structure or ownership, and lack of trust in public databases (Huang et al. 2012, Tenopir et al. 2020). We compared articles that did and did not publish their data to examine the relative impact of some potential barriers to data sharing. First, we anticipated that two measures of data set complexity might negatively correlate with data sharing: first, the number of data sources accessed to compile a data set and, second, whether the original sampling design was reported. We expected that authors might be held back from sharing data by the complexity of crediting multiple original sources or by their own lack of complete information on data structure. However, we did not find either of these relationships in our results. This finding suggests that data set complexity may not be the primary factor prohibiting researchers from publishing their data sets. It is a concern but is more likely secondary to other barriers. Because lack of familiarity with open databases has also been cited as a barrier to data sharing, we expected that authors' familiarity with open data, as has been demonstrated by the integration of data from open databases with presence-only data from other sources, would correlate with greater rates of data publication. This was not the case: Of the articles that integrated data from open databases and other sources, 76% did not publish the data that were not already open.

These findings suggest that other concerns, including lack of researcher incentive and concern about receiving appropriate credit for shared data, may be more serious barriers to data sharing (Escribano et al. 2018, Tenopir et al. 2020). Some developments have begun to address the issue of researcher incentive: Data sharing is increasingly incentivized through journal policies, funding agency requirements, and the promotion of data citations (Mills et al. 2015, Colavizza et al. 2020, Walters 2020). Continuing to normalize these incentives may help overcome existing barriers to data sharing, especially in situations in which data users are the original data owners (Chavan and Penev 2011, Mooney and Newton 2012, Kattge et al. 2014, Escribano et al. 2018). Furthermore, researchers are increasingly taking ownership over the process of data sharing, establishing grassroots collaborations that organize specific research communities to engage with open data infrastructure and practices (Aubin et al. 2020). This integration of open data practices into local networks of biodiversity researchers has great potential to incentivize open data sharing by establishing it as a key component of network building and collaboration within specific research areas. As open data sharing becomes increasingly normalized, it will be essential that practitioners of open science maintain a supportive, rather than critical, approach to encouraging researchers who are taking their first steps into open data sharing. Researchers do not all have equal access to the resources, training, technical capacity, and institutional support to fully engage in open data practices, and small steps toward open data sharing must be welcomed while the field as a whole shifts to become more equitably supportive of open data practices (Bahlai et al. 2019, Chawinga and Zinn 2019, Powers and Hampton 2019, Soeharjono and Roche 2021).

## Conclusions

Open access to high quality biodiversity occurrence data is key to many emerging themes in biodiversity research and conservation, including development and implementation of international biodiversity assessments and targets (Hochkirch et al. 2020), research synthesis for conservation decision-making (Nakagawa et al. 2020), and near-term ecological forecasting of species abundance in space and time (Callaghan et al. 2021), so continued efforts to increase the open sharing of biodiversity data will be critical. This will require increased incentivization, institutional support, ongoing shifts in cultural norms, and a growing emphasis on an ethical, equitable framework for data sharing. Recent trends toward increased sharing of presence-only biodiversity data are a cause for optimism, but there is still a great deal of work to be done in normalizing the use of best practices in data access, documentation, citation, and sharing. Still, we see evidence in the trends reported in the present article for an often-reported survey result: Researchers generally feel positively toward reusing and sharing data, despite persistent uncertainty about best practices and concern about credit and incentives (Ross-Hellauer et al. 2017, Tenopir et al. 2020, Soeharjono and Roche 2021). Such evidence includes the recent increase in the proportion of articles that produce open data, the efforts taken by some authors to credit original data providers even when no clear mechanism had yet been developed to do so, and the above-average sharing rate for citizen science data.

For researchers looking to begin or continue their journey into reuse and sharing of open biodiversity data, there are many excellent resources that offer an entry point into accessing and sharing open data; we particularly

point such researchers to Hampton and colleagues (2015), Wilkinson and colleagues (2016), Boland and colleagues (2017), Alston and Rick (2021), and to guides such as the FAIR Principles (GO FAIR 2021), the CARE Principles of Indigenous Data Governance (GIDA 2019), and the Quick Guide to Publishing Data Through GBIF.org (GBIF 2021). To those beginning to engage with open data, we echo the wisdom of Bahlai and colleagues (2019), Alston and Rick (2021), and others in encouraging researchers to begin with any first steps, however small, that are feasible given their circumstances. Increased open data sharing will rely on both the progressive adoption of data sharing practices by individual researchers and ultimately on broad cultural shifts within biodiversity and related fields (Chawinga and Zinn 2019). This shift to a culture of ethical open data sharing will be essential to meet challenges associated with the growing biodiversity crisis and to support a growing need for biodiversity assessment, monitoring, and conservation.

## Acknowledgments

This work is part of the Norwegian University of Science and Technology's Transforming Citizen Science for Biodiversity project, and we thank the other project members for their valuable feedback on earlier stages of this work. We also greatly appreciate the feedback of three anonymous reviewers and an anonymous editor. Funding for CPM was provided by the Transforming Citizen Science for Biodiversity program within the Digital Transformation Initiative of the Norwegian University of Science and Technology. WK was funded by the Norwegian Research Council (grant no. 272947) and the Norwegian Biodiversity Information Centre. EBN was funded by the Norwegian Institute for Nature Research.

## Supplemental material

Supplemental data are available at *BIOSCI* online.

## References cited

Alston JM, Rick JA. 2021. A beginner's guide to Conducting reproducible research. Bulletin of the Ecological Society of America 102: e01801.

Altman M, Crosas M. 2014. The evolution of data citation: From principles to implementation. IASSIST Quarterly 37: 62.

Amano T, Lamming JDL, Sutherland WJ. 2016. Spatial Gaps in global biodiversity information and the role of citizen science. BioScience 66: 393–400.

Anderson RP, Araújo MB, Guisan A, Lobo JM, Martínez-Meyer E, Peterson AT, Soberón JM. 2020. Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society. Frontiers of Biogeography 12.

Araújo MB et al. 2019. Standards for distribution models in biodiversity assessments. Science Advances 5: eaat4858.

Aria M, Cuccurullo C. 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics 11: 959–975.

Ariño AH. 2010. Approaches to estimating the universe of natural history collections data. Biodiversity Informatics 7: 81–92.

Asmussen CB, Møller C. 2019. Smart literature review: A practical topic modelling approach to exploratory literature review. Journal of Big Data 6: 93.

Aubin I, Cardou F, Boisvert-Marsh L, Garnier E, Strukelj M, Munson AD. 2020. Managing data locally to answer questions globally: The role of collaborative science in ecology. Journal of Vegetation Science 31: 509–517.

Bahlai CA, Bartlett LJ, Burgio KR, Fournier AMV, Keiser CN, Poisot T, Whitney KS. 2019. Open science isn't always open to all scientists. American Scientist 107: 78–82.

Ball-Damerow JE, Brenskelle L, Barve N, Soltis PS, Sierwald P, Bieler R, LaFrance R, Ariño AH, Guralnick RP. 2019. Research applications of primary biodiversity databases in the digital age. PLOS ONE 14: e0215794.

Bayraktarov E, Ehmke G, O'Connor J, Burns EL, Nguyen HA, McRae L, Possingham HP, Lindenmayer DB. 2019. Do big unstructured biodiversity data mean more knowledge? Frontiers in Ecology and Evolution 6: 239.

Bishop BW, Hank C, Webster J, Howard R. 2019. Scientists' data discovery and reuse behavior: (Meta)data fitness for use and the FAIR data principles. Proceedings of the Association for Information Science and Technology 56: 21–31.

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.

Boland MR, Karczewski KJ, Tatonetti NP. 2017. Ten simple rules to enable multi-site collaborations through data sharing. PLOS Computational Biology 13: e1005278.

Boshoff N. 2009. Neo-colonialism and research collaboration in Central Africa. Scientometrics 81: 413–434.

Brown RF. 2021. The importance of data citation. BioScience 71: 211–211.

Callaghan CT, Poore AGB, Mesaglio T, Moles AT, Nakagawa S, Roberts C, Rowley JJL, Vergés A, Wilshire JH, Cornwell WK. 2021. Three frontiers for the future of biodiversity research using citizen science data. BioScience 71: 55–63.

Calver MC, Goldman B, Hutchings PA, Kingsford RT. 2017. Why discrepancies in searching the conservation biology literature matter. Biological Conservation 213: 19–26.

Carroll SR, Herczog E, Hudson M, Russell K, Stall S. 2021. Operationalizing the CARE and FAIR Principles for Indigenous data futures. Scientific Data 8: 108.

Chambers LE, Barnard P, Poloczanska ES, Hobday AJ, Keatley MR, Allsopp N, Underhill LG. 2017. Southern Hemisphere biodiversity and global change: Data gaps and strategies. Austral Ecology 42: 20–30.

Chandler M et al. 2017. Contribution of citizen science towards international biodiversity monitoring. Biological Conservation 213: 280–294.

Chapman AD. 2020. Current Best Practices for Generalizing Sensitive Species Occurrence Data. Global Biodiversity Information Facility.

Chavan V, Penev L. 2011. The data paper: A mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics 12: S2.

Chawinga WD, Zinn S. 2019. Global perspectives of research data sharing: A systematic literature review. Library Information Science Research 41: 109–122.

Clements HS, Selinske MJ, Archibald CL, Cooke B, Fitzsimons JA, Groce JE, Torabi N, Hardy MJ. 2018. Fairness and transparency are required for the inclusion of privately protected areas in publicly accessible conservation databases. Land 7: 96.

Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B. 2020. The citation advantage of linking publications to research data. PLOS ONE 15: e0230416.

Cooper CB, Shirk J, Zuckerberg B. 2014. The invisible prevalence of citizen science in global research: Migratory birds and climate change. PLOS ONE 9.

Costello MJ, Wieczorek J. 2014. Best practice for biodiversity data management and publication. Biological Conservation 173: 68–73.

Costello MJ, Michener WK, Gahegan M, Zhang Z-Q, Bourne PE. 2013. Biodiversity data should be published, cited, and peer reviewed. Trends in Ecology Evolution 28: 454–461.

Cousijn H et al. 2018. A data citation roadmap for scientific publishers. Scientific Data 5: 180259.

Cretois B, Linnell JDC, Grainger M, Nilsen EB, Rød JK. 2020. Hunters as citizen scientists: Contributions to biodiversity monitoring in Europe. Global Ecology and Conservation 23: e01077.

Curty RG, Crowston K, Specht A, Grant BW, Dalton ED. 2017. Attitudes and norms affecting scientists' data reuse. PLOS ONE 12: e0189288.

Di Marco M et al. 2017. Changing trends and persisting biases in three decades of conservation science. Global Ecology and Conservation 10: 32–42.

Dias D, Baringo Fonseca C, Correa L, Soto N, Portela A, Juarez K, Tumolo Neto RJ, Ferro M, Gonçalves J, Junior J. 2017. Repatriation data: More than two million species occurrence records added to the Brazilian Biodiversity Information Facility Repository (SiBBr). Biodiversity Data Journal 2017: e12012.

Dobson ADM et al. 2020. Making messy data work for conservation. One Earth 2: 455–465.

Eichhorn MP, Baker K, Griffiths M. 2020. Steps towards decolonising biogeography. Frontiers of Biogeography 12: e44795e.

Escribano N, Ariño AH, Galicia D. 2016. Biodiversity data obsolescence and land uses changes. PeerJ 4: e2743.

Escribano N, Galicia D, Ariño AH. 2018. The tragedy of the biodiversity data commons: A data impediment creeping nigher? Database 2018: bay033.

Faith D, Collen B, Ariño A, Koleff PKP, Guinotte J, Kerr J, Chavan V. 2013. Bridging the biodiversity data gaps: Recommendations to meet users' data needs. Biodiversity Informatics 8.

Fanelli D, Larivière V. 2016. Researchers' individual publication rate has not increased in a century. PLOS ONE 11: e0149504.

Farley SS, Dawson A, Goring SJ, Williams JW. 2018. Situating ecology as a big-data science: Current advances, challenges, and solutions. BioScience 68: 563–576.

Fenner M et al. 2019. A data citation roadmap for scholarly data repositories. Scientific Data 6: 28.

Fletcher RJ, Hefley TJ, Robertson EP, Zuckerberg B, McCleery RA, Dorazio RM. 2019. A practical guide for combining data to model species distributions. Ecology 100: e02710.

Foster SD, Vanhatalo J, Trenkel VM, Schulz T, Lawrence E, Przeslawski R, Hosack GR. 2021. Effects of ignoring survey design information for data reuse. Ecological Applications: e2360.

Fox R, Bourn NAD, Dennis EB, Heafield RT, Maclean IMD, Wilson RJ. 2019. Opinions of citizen scientists on open access to UK butterfly and moth occurrence data. Biodiversity and Conservation 28: 3321–3341.

Franz NM, Sterner BW. 2018. To increase trust, change the social design behind aggregated biodiversity data. Database 2018: bax100.

Gadelha LMR et al. 2021. A survey of biodiversity informatics: Concepts, practices, and challenges. WIREs Data Mining and Knowledge Discovery 11: e1394.

Ganzevoort W, Van den Born RJG, Halffman W, Turnhout S. 2017. Sharing biodiversity data: Citizen scientists' concerns and motivations. Biodiversity and Conservation 26: 2821–2837.

[GBIF] Global Biodiversity Information Facility. 2019. GBIF Science Review 2019. GBIF. https://doi.org/10.15468/QXXG-7K93.

[GBIF] Global Biodiversity Information Facility. 2021. Quick Guide to Publishing Data Through GBIF.org. GBIF. www.gbif.org/publishing-data.

Gelfand AE, Shirota S. 2019. Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. Ecological Monographs 89: E01372.

[GIDA] Global Indigenous Data Alliance. 2019. CARE Principles for Indigenous Data Governance. GIDA. GIDA-global.org.

GO FAIR. 2021. FAIR Principles. GO FAIR. www.go-fair.org/fair-principles.

Grainger MJ, Bolam FC, Stewart GB, Nilsen EB. 2020. Evidence synthesis for tackling research waste. Nature Ecology and Evolution 4: 495–497.

Grimmett L, Whitsed R, Horta A. 2020. Presence-only species distribution models are sensitive to sample prevalence: Evaluating models using spatial prediction stability and accuracy metrics. Ecological Modelling 431: 109194.

Groom Q, Weatherdon L, Geijzendorffer IR. 2017. Is citizen science an open science in the case of biodiversity observations? Journal of Applied Ecology 54: 612–617.

Groom Q, Güntsch A, Huybrechts P, Kearney N, Leachman S, Nicolson N, Page RDM, Shorthouse DP, Thessen AE, Haston E. 2020. People are essential to linking biodiversity data. Database 2020: baaa072.

Guisan A et al. 2013. Predicting species distributions for conservation decisions. Ecology Letters 16: 1424–1435.

Habel JC et al. 2014. Towards more equal footing in north–south biodiversity research: European and sub-Saharan viewpoints. Biodiversity and Conservation 23: 3143–3148.

Hackett RA, Belitz MW, Gilbert EE, Monfils AK. 2019. A data management workflow of biodiversity data from the field to data users. Applications in Plant Sciences 7: e11310.

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH. 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment 11: 156–162.

Hampton SE, et al. 2015. The Tao of open science for ecology. Ecosphere 6: 120.

Hao T, Elith J, Guillera-Arroita G, Lahoz-Monfort JJ. 2019. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. Diversity and Distributions 25: 839–852.

Heberling JM, Miller JT, Noesgaard D, Weingart SB, Schigel D. 2021. Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences 118: e2018093118.

Hedrick BP et al. 2020. Digitization and the future of natural history collections. BioScience 70: 243–251.

Hochkirch A et al. 2020. A strategy for the next decade to address data deficiency in neglected biodiversity. Conservation Biology 35: 502–509. doi:10.1111/cobi.13589

Hrynaszkiewicz I, Simons N, Hussain A, Grant R, Goudie S. 2020. Developing a research data policy framework for all journals and publishers. Data Science Journal 19: 5.

Huang J, Frimpong EA. 2015. Using historical atlas data to develop high-resolution distribution models of freshwater fishes. PLOS ONE 10: e0129995.

Huang X, Hawkins BA, Lei F, Miller GL, Favret C, Zhang R, Qiao G. 2012. Willing or unwilling to share primary biodiversity data: Results and implications of an international survey. Conservation Letters 5: 399–406.

Huettmann F. 2009. The global need for, and appreciation of, high-quality metadata in biodiversity database work. Pages 25–28 in Spehn EM, Korner C, eds. Data Mining for Global Trends in Mountain Biodiversity. Taylor and Francis.

Isaac NJB et al. 2020. Data integration for large-scale models of species distributions. Trends in Ecology and Evolution 35: 56–67.

Isaac NJB, Strien AJ van, August TA, Zeeuw MP de, Roy DB. 2014. Statistics for citizen science: Extracting signals of change from noisy ecological data. Methods in Ecology and Evolution 5: 1052–1060.

James SA, Soltis PS, Belbin L, Chapman AD, Nelson G, Paul DL, Collins M. 2018. Herbarium data: Global biodiversity and societal botanical needs for novel research. Applications in Plant Sciences 6: e1024.

Johnson N, Druckenmiller ML, Danielsen F, Pulsifer PL. 2021. The use of digital platforms for community-based monitoring. BioScience 71: 452–466.

Johnston A, Hochachka WM, Strimas-Mackey ME, Gutierrez VR, Robinson OJ, Miller ET, Auer T, Kelling ST, Fink D. 2021. Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. Diversity and Distributions 27: 1265–1277.

Joppa LN, McInerny G, Harper R, Salido L, Takeda K, O'Hara K, Gavaghan D, Emmott S. 2013. Troubling trends in scientific software use. Science 340: 814–815.

Kattge J, Díaz S, Wirth C. 2014. Of carrots and sticks. Nature Geoscience 7: 778–779.

Kass JM, Vilela B, Aiello-Lammens ME, Muscarella R, Merow C, Anderson RP. 2018. Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. Methods in Ecology and Evolution 9: 1151–1156.

Kays R, McShea WJ, Wikelski M. 2020. Born-digital biodiversity data: Millions and billions. Diversity and Distributions 26: 644–648.

Kelling S et al. 2019. Using semistructured surveys to improve citizen science data for monitoring biodiversity. BioScience 69: 170–179.
Kervin K, Michener W, Cook R. 2013. Common errors in ecological data sharing. Journal of eScience Librarianship 2: 1024.
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.
Kühl HS et al. 2020. Effective biodiversity monitoring needs a culture of integration. One Earth 3: 462–474.
Larson LR, Cooper CB, Futch S, Singh D, Shipley NJ, Dale K, LeBaron GS, Takekawa JY. 2020. The diverse motivations of citizen scientists: Does conservation emphasis grow as volunteer participation progresses? Biological Conservation 242: 108428.
Li K, Greenberg J, Dunic J. 2020. Data objects and documenting scientific processes: An analysis of data events in biodiversity data papers. Journal of the Association for Information Science and Technology 71: 172–182.
Luo M, Xu Z, Hirsch T, Aung TS, Xu W, Ji L, Qin H, Ma K. 2021. The use of Global Biodiversity Information Facility (GBIF)-mediated data in publications written in Chinese. Global Ecology and Conservation 25: e01406.
Lynn SJ, Kaplan N, Newman S, Scarpino R, Newman G. 2019. Designing a platform for ethical citizen science: A case study of CitSci.org. Citizen Science: Theory and Practice 4: 14.
MacPhail VJ, Colla SR. 2020. Power of the people: A review of citizen science programs for conservation. Biological Conservation 249: 108739.
Mandeville CP. 2021. Open data practices among users of primary biodiversity data. Open Science Framework. https://osf.io/jueqc. doi:10.17605/OSF.IO/JUEQC
Mandeville CP, Finstad AG. 2021. Community science supports research on protected area resilience. Conservation Science and Practice 2021: e442.
Merow C, Smith MJ, Silander JA. 2013. A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. Ecography 36: 1058–1069.
Mesibov R. 2018. An audit of some processing effects in aggregated occurrence records. ZooKeys 751: 129–146.
Michener WK. 2015. Ten simple rules for creating a good data management plan. PLOS Computational Biology 11: e1004525.
Miller DAW, Pacifici K, Sanderlin JS, Reich BJ. 2019. The recent past and promising future for data integration methods to estimate species' distributions. Methods in Ecology and Evolution 10: 22–37.
Miller SE et al. 2020. Building natural history collections for the twenty-first century and beyond. BioScience 70: 674–687.
Mills JA et al. 2015. Archiving primary data: Solutions for long-term studies. Trends in Ecology and Evolution 30: 581–589.
Monfils AK et al. 2020. Regional collections are an essential component of biodiversity research infrastructure. BioScience 70: 1045–1047.
Mooney H, Newton MP. 2012. The anatomy of a data citation: Discovery, reuse, and credit. Scholarly Communication 1: eP1035–eP1035.
Muscatello A, Elith J, Kujala H. 2021. How decisions about fitting species distribution models affect conservation outcomes. Conservation Biology: 10.1111/cobi.13669.
Nakagawa S et al. 2020. A new ecosystem for evidence synthesis. Nature Ecology and Evolution 4: 498–501.
Nelson G, Ellis S. 2019. The history and impact of digitization and digital data mobilization on biodiversity research. Philosophical Transactions of the Royal Society B 374: 20170391.
Norberg A et al. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecological Monographs 89: e01370.
Nuñez MA, Barlow J, Cadotte M, Lucas K, Newton E, Pettorelli N, Stephens PA. 2019. Assessing the uneven global distribution of readership, submissions and publications in applied ecology: Obvious problems without obvious solutions. Journal of Applied Ecology 56: 4–9.
Ondei S, Brook BW, Buettel JC. 2018. Nature's told stories: An overview on the availability and type of on-line data on long-term biodiversity monitoring. Biodiversity and Conservation 27: 2971–2987.
Osawa T. 2019. Perspectives on biodiversity informatics for ecology. Ecological Research 34: 446–456.
Pacifici K, Reich BJ, Miller DAW, Gardner B, Stauffer G, Singh S, McKerrow A, Collazo JA. 2017. Integrating multiple data sources in species distribution modeling: A framework for data fusion. Ecology 98: 840–850.
Parsons AW, Goforth C, Costello R, Kays R. 2018. The value of citizen science for ecological monitoring of mammals. PeerJ 6: e4536.
Pearce-Higgins JW, et al. 2018. Overcoming the challenges of public data archiving for citizen science biodiversity recording and monitoring schemes. Journal of Applied Ecology 55: 2544–2551.
Pelayo-Villamil P, Guisande C, Manjarrés-Hernández A, Jiménez LF, Granado-Lorencio C, García-Roselló E, González-Dacosta J, Heine J, González-Vilas L, Lobo JM. 2018. Completeness of national freshwater fish species inventories around the world. Biodiversity and Conservation 27: 3807–3817.
Penev L et al. 2017. Strategies and guidelines for scholarly publishing of biodiversity data. Research Ideas and Outcomes 3: e12431.
Petersen TK, Speed JDM, Grøtan V, Austrheim G. 2021. Species data for understanding biodiversity dynamics: The what, where and when of species occurrence data collection. Ecological Solutions and Evidence 2: e12048.
Peterson AT, Asase A, Canhos D, Souza S de, Wieczorek J. 2018. Data leakage and loss in biodiversity informatics. Biodiversity Data Journal 6: e26826.
Pettorelli N, Barlow J, Nuñez MA, Rader R, Stephens PA, Pinfield T, Newton E. 2021. How international journals can support ecology from the Global South. Journal of Applied Ecology 58: 4–8.
Pino-Del-Carpio A, Ariño AH, Villarroya A, Puig J, Miranda R. 2014. The biodiversity data knowledge gap: Assessing information loss in the management of Biosphere Reserves. Biological Conservation 173: 74–79.
Poisot T, Bruneau A, Gonzalez A, Gravel D, Peres-Neto P. 2019. Ecological data should not be so hard to find and reuse. Trends in Ecology and Evolution 34: 494–496.
Powers SM, Hampton SE. 2019. Open science, reproducibility, and transparency in ecology. Ecological Applications 29: e01822.
R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing. www.R-project.org.
Ramirez KS et al. 2018. The future of ecology is collaborative, inclusive and deconstructs biases. Nature Ecology and Evolution 2: 200–200.
Rapacciuolo G. 2019. Strengthening the contribution of macroecological models to conservation practice. Global Ecology and Biogeography 28: 54–60.
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.
Roche DG, Kruuk LEB, Lanfear R, Binning SA. 2015. Public data archiving in ecology and evolution: How well are we doing? PLOS Biology 13: e1002295.
Ross-Hellauer T, Deppe A, Schmidt B. 2017. Survey on open peer review: Attitudes and experience among editors, authors and reviewers. PLOS ONE 12: e0189311.
Rubert-Nason K, Casper AA, Jurjonas M, Mandeville C, Potter R, Schwarz K. 2021. Ecologist engagement in translational science is imperative for building resilience to global change threats. Rethinking Ecology 6: 65–92.
Serra-Diaz JM, Enquist BJ, Maitner B, Merow C, Svenning J-C. 2017. Big data of tree species distributions: How big and how good? Forest Ecosystems 4: 30.
Serwadda D, Ndebele P, Grabowski MK, Bajunirwe F, Wanyenze RK. 2018. Open data sharing and the Global South: Who benefits? Science 359: 642–643.
Shin N, Shibata H, Osawa T, Yamakita T, Nakamura M, Kenta T. 2020. Toward more data publication of long-term ecological observations. Ecological Research 35: 700–707.

Sicacha-Parada J, Steinsland I, Cretois B, Borgelt J. 2020. Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway. Spatial Statistics 100446.

Sillero N, Barbosa AM. 2021. Common mistakes in ecological niche models. International Journal of Geographical Information Science 35: 213–226.

Simmonds EG, Jarvis SG, Henrys PA, Isaac NJB, O'Hara RB. 2020. Is more data always better? A simulation study of benefits and limitations of integrated distribution models. Ecography 43: 1413–1422.

Singer RA, Ellis S, Page LM. 2020. Awareness and use of biodiversity collections by fish biologists. Journal of Fish Biology 96: 297–306.

Soeharjono S, Roche DG. 2021. Reported individual costs and benefits of sharing open data among Canadian Academic Faculty in ecology and evolution. BioScience. https://doi.org/10.1093/biosci/biab024.

Sofaer HR et al. 2019. Development and delivery of species distribution models to inform decision-making. BioScience 69: 544–557.

Soranno PA et al. 2020. Ecological prediction at macroscales using big data: Does sampling design matter? Ecological Applications 30: e02123.

Specht A, Bolton MP, Kingsford B, Specht RL, Belbin L. 2018. A story of data won, data lost and data re-found: The realities of ecological data preservation. Biodiversity Data Journal 6: e28073.

Speed JDM, Bendiksby M, Finstad AG, Hassel K, Kolstad AL, Prestø T. 2018. Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data. PLOS ONE 13: e0196417.

Stephenson PJ et al. 2017. Unblocking the flow of biodiversity data for decision-making in Africa. Biological Conservation 213: 335–340.

Støa B, Halvorsen R, Mazzoni S, Gusarov VI. 2018. Sampling bias in presence-only data used for species distribution modelling: Theory and methods for detecting sample bias and its effects on models. Sommerfeltia 38: 1–53.

Stodden V, Seiler J, Ma Z. 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. Proceedings of the National Academy of Sciences 115: 2584–2589.

Stork H, Astrin JJ. 2014. Trends in biodiversity research—a bibliometric assessment. Open Journal of Ecology 4: 354–370.

Sullivan BL et al. 2017. Using open access observational data for conservation action: A case study for birds. Biological Conservation 208: 5–14.

Tengö M, Austin BJ, Danielsen F, Fernández-Llamazares Á. 2021. Creating synergies between citizen science and indigenous and local knowledge. BioScience 71: 503–518.

Tenopir C, Rice NM, Allard S, Baird L, Borycz J, Christian L, Grant B, Olendorf R, Sandusky RJ. 2020. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. PLOS ONE 15: e0229003.

Tessarolo G, Ladle R, Rangel T, Hortal J. 2017. Temporal degradation of data limits biodiversity research. Ecology and Evolution 7: 6863–6870.

Theobald EJ et al. 2015. Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. Biological Conservation 181: 236–244.

Trisos CH, Auerbach J, Katti M. 2021. Decoloniality and anti-oppressive practices for a more ethical ecology. Nature Ecology and Evolution 2021: s41559-021-01460-w.

Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F. 2017. Taxonomic bias in biodiversity data and societal preferences. Scientific Reports 7: 9132.

Tulloch AIT, Possingham HP, Joseph LN, Szabo J, Martin TG. 2013. Realising the full potential of citizen science monitoring programs. Biological Conservation 165: 128–138.

Tulloch AIT et al. 2018. A decision tree for assessing the risks and benefits of publishing biodiversity data. Nature Ecology and Evolution 2: 1209–1217.

Tydecks L, Jeschke JM, Wolf M, Singer G, Tockner K. 2018. Spatial and topical imbalances in biodiversity research. PLOS ONE 13: e0199327.

Vaz UL, Cunha HF, Nabout JC, Vaz UL, Cunha HF, Nabout JC. 2015. Trends and biases in global scientific literature about ecological niche models. Brazilian Journal of Biology 75: 17–24.

Vines TH, Albert AYK, Andrew RL, Débarre F, Bock DG, Franklin MT, Gilbert KJ, Moore J-S, Renaut S, Rennison DJ. 2014. The availability of research data declines rapidly with article age. Current Biology 24: 94–97.

Voříšek P, Gregory RD, Keller V, Herrando S, Lindström Å, Nagy S, Burfield IJ, Noble D, Ramírez I, Foppen RPB. 2018. Wetzel et al. fail to identify the real gaps in European bird monitoring. Biological Conservation 225: 245–246.

Walters WH. 2020. Data journals: Incentivizing data access and documentation within the scholarly communication system. Insights 33: 18.

Westgate MJ. 2019. revtools: An R package to support article screening for evidence synthesis. Research Synthesis Methods 10: 606–614.

Wetzel FT, et al. 2018. Unlocking biodiversity data: Prioritization and filling the gaps in biodiversity observation data in Europe. Biological Conservation 221: 78–85.

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.

Wilkinson MD et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018.

Wüest RO et al. 2020. Macroecology in the age of Big Data: Where to go from here? Journal of Biogeography 47: 1–12.

Yackulic CB, Chandler R, Zipkin EF, Royle JA, Nichols JD, Grant EHC, Veran S. 2013. Presence-only modelling using MAXENT: When can we trust the inferences? Methods in Ecology and Evolution 4: 236–243.

Zipkin EF, Zylstra ER, Wright AD, Saunders SP, Finley AO, Dietze MC, Itter MS, Tingley MW. 2021. Addressing data integration challenges to link ecological processes across scales. Frontiers in Ecology and the Environment 19: 30–38.

Zurell D et al. 2020. A standard protocol for reporting species distribution models. Ecography 43: 1261–1277.

*Caitlin P. Mandeville (caitlin.mandeville@ntnu.no) is a PhD candidate in the Centre for Biodiversity Dynamics and Department of Natural History at the Norwegian University of Science and Technology, in Trondheim, Norway. Wouter Koch is a senior advisor of biodiversity informatics at the Norwegian Biodiversity Information Centre, as well as a PhD candidate in the Centre for Biodiversity Dynamics and Department of Natural History at the Norwegian University of Science and Technology, in Trondheim, Norway. Erlend B. Nilsen is a senior research scientist at the Norwegian Institute for Nature Research in Trondheim, Norway, as well as a professor of ecology at the Faculty of Biosciences and Aquaculture at Nord University, in Steinkjer, Norway. Anders G. Finstad is a professor in the Centre for Biodiversity Dynamics and Department of Natural History at the Norwegian University of Science and Technology, in Trondheim, Norway.*