Paleobiology

cambridge.org/pab

On The Record

*These authors contributed equally.

Cite this article: Smith JA *et al* (2024). Increasing the equitability of data citation in paleontology: capacity building for the big data future. *Paleobiology* **50**, 165–176. https://doi.org/10.1017/pab.2023.33

Received: 29 June 2023
Revised: 16 October 2023
Accepted: 25 October 2023

Keywords:
Biodiversity; Open science; Paleobiology Database; Specimen-based; Taxonomy

Corresponding author:
Jansen A. Smith;
Email: jansen.smith@fau.de

PALEOBIOLOGY A PUBLICATION OF THE Paleontological SOCIETY

CAMBRIDGE UNIVERSITY PRESS

Increasing the equitability of data citation in paleontology: capacity building for the big data future

Jansen A. Smith[1,2,3,4*], Nussaïbah B. Raja[1,*], Thomas Clements[1], Danijela Dimitrijević[1], Elizabeth M. Dowding[1], Emma M. Dunne[1], Bryan M. Gee[5], Pedro L. Godoy[6,7], Elizabeth M. Lombardi[3], Laura P. A. Mulvey[1], Paulina S. Nätscher[1], Carl J. Reddin[1,8], Bryan Shirley[1,9], Rachel C. M. Warnock[1] and Ádám T. Kocsis[1,10,*]

[1]GeoZentrum Nordbayern, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Bayern 91054, Germany
[2]Paleontological Research Institution, Ithaca, New York 14850, U.S.A.
[3]Department of Biology, University of New Mexico, Albuquerque, New Mexico 87110, U.S.A.
[4]Department of Earth and Environmental Sciences, University of Minnesota Duluth, Duluth, Minnesota 55812, U.S.A.
[5]Burke Museum and Department of Biology, University of Washington, Seattle, Washington 98195, U.S.A.
[6]Department of Zoology, Institute of Biosciences, University of São Paulo, São Paulo, SP 04263, Brazil
[7]Department of Anatomical Sciences, Stony Brook University, Stony Brook, New York 11794, U.S.A.
[8]Museum für Naturkunde, Berlin, Bayern, 10115, Germany
[9]Department of Earth Sciences, Faculty of Geosciences, Utrecht University, Utrecht, 3584 CB, The Netherlands
[10]MTA-MTM-ELTE Research Group for Paleontology, Budapest 1431, Hungary

Non-technical Summary

Researchers often use large databases to conduct their studies; however, they do not always provide credit, through citations, to the people who produced the data in the databases. In the field of paleontology, researchers use a large database called the Paleobiology Database (PBDB) to study global patterns and processes over millions of years. These studies use data from the PBDB and typically receive a greater number of citations than the original data-producing papers. This creates a situation where the hard work of collecting the data is not credited and rewarded in a fair way, even though this work is equally important to the field of paleontology. By fixing this issue and giving proper credit to data-producing papers, paleontology itself can be strengthened by increasing the incentives for producing data and at the same time creating more high-quality data for everyone to use.

Abstract

Data compilations expand the scope of research; however, data citation practice lags behind advances in data use. It remains uncommon for data users to credit data producers in professionally meaningful ways. In paleontology, databases like the Paleobiology Database (PBDB) enable assessment of patterns and processes spanning millions of years, up to global scale. The status quo for data citation creates an imbalance wherein publications drawing data from the PBDB receive significantly more citations (median: 4.3 ± 3.5 citations/year) than the publications producing the data (1.4 ± 1.3 citations/year). By accounting for data reuse where citations were neglected, the projected citation rate for data-provisioning publications approached parity (4.2 ± 2.2 citations/year) and the impact factor of paleontological journals ($n = 55$) increased by an average of 13.4% (maximum increase = 57.8%) in 2019. Without rebalancing the distribution of scientific credit, emerging "big data" research in paleontology—and science in general—is at risk of undercutting itself through a systematic devaluation of the work that is foundational to the discipline.

Introduction

> "Both data collectors and data crunchers are important, and certainly the latter would not exist without the former."
>
> MacRoberts and MacRoberts (2018: p. 476)

Large data compilations allow new questions to be asked at previously unreachable scales and provide greater certainty in answering old questions. Across the biological sciences, large compilations have led to theoretical and practical advances, ranging from new insights on biodiversity dynamics in conservation science (Dornelas et al. 2014), to biomedical advances based on compiled genetic data (Benson et al. 2013), to the identification of mass extinctions in

paleontology (Raup and Sepkoski 1982). Compilations of data at these scales, and the associated advances, are accompanied by a need to establish a standard set of protocols for the curation, management, and citation of the underlying data (Altman et al. 2015; Cousijn et al. 2018; Kaufman et al. 2018; Marwick and Birch 2018; Lammey 2019). Recognizing the reliance of large data compilations on the many data-producing studies from which they are built, there is a growing consensus that data users should credit data producers in a way that is on par with the credit attributed to traditionally recognized outputs, like peer-reviewed publications (Piwowar and Vision 2013; Altman et al. 2015; Penev et al. 2017; Cousijn et al. 2018; Kaufman et al. 2018; Silvello 2018; Zhao et al. 2018; Lammey 2019; Pierce et al. 2019; Dosso and Silvello 2020).

Despite this emerging consensus, the scientific community at large has been slow to adopt the practice of citing data sources, and a common procedure for data citation—used inclusively here to refer to any attribution to data provisioners by data users (e.g., Penev et al. 2017; Cousijn et al. 2018; Hood and Sutherland 2021; and see "Balancing Data Use and Citation in Paleontology")—remains elusive (Ingwersen and Chavan 2011; Marwick and Birch 2018; Zhao et al. 2018; Cousijn et al. 2019; Tomaszewski 2019; Silveira et al. 2020; Suhr et al. 2020). For example, in a review of 600 papers from 12 disciplines (e.g., biology, earth sciences, ecology, and environmental sciences; Zhao et al. 2018), when authors used a new or existing dataset in their analysis ($n = 312$), data attribution was variable: 6% included data citations, 9% used unique identifiers for the data (e.g., DOI), 24% mentioned data with only a database name, and 60% referenced their data using a URL—an imperfect citation, given that URLs can expire. Relatedly, 88% of studies ($n = 100$, randomly drawn from 4533 studies) using data compiled by the Global Biodiversity Information Facility failed to appropriately cite the sources of the data they used (Escribano et al. 2018). It has become clear that making data citation a standardized practice will require changes at all stages of academic research—from funders, publishers, editorial boards, data repositories, authors submitting analyses of compiled data, researchers producing the data, scientists evaluating each other's work, and all other persons involved in research production (Kaufman et al. 2018; Marwick and Birch 2018; Cousijn et al. 2019; Colavizza et al. 2020; Silveira et al. 2020).

Like many other scientific disciplines, paleontology has much room for improvement in how data are cited (Payne et al. 2012; Kaufman et al. 2018; Fig. 1). Paleontology has historically been a descriptive field wherein accumulations of fossils are documented when they are found in rocks and sediments. Most basically, individual fossils—alongside information on their location, stratigraphy, and taxonomy—are the raw data of paleontology (Johnson et al. 2005; Allmon et al. 2018). It is typically these records of taxa at a given place and time that are compiled for larger-scale analyses. The analysis of data compilations has deep roots in paleontology (e.g., Phillips 1860; Newell 1952, 1967; Harland 1967; Sepkoski et al. 1981; Sepkoski 1984), and the development of online databases (e.g., ART [Raja et al. 2022a]; BioDeepTime [Smith et al. 2023b]; Geobiodiversity Database [Fan et al. 2013]; Neotoma [Williams et al. 2018]; Neptune Sandbox Berlin [Renaudie et al. 2020]; Paleobiology Database, https://paleobiodb.org; PARED [Kiessling and Krause 2022]; Triton [Fenton et al. 2021]) in the last two decades has helped make these types of analyses a cornerstone of modern paleontology (Supplementary Fig. S1). Paleontologists now routinely analyze compiled data at local to global scales across temporal ranges of hundreds of millions of years, greatly expanding the ambition of the hypotheses and questions that can be addressed about the history of life on Earth (e.g., Kiessling 2005; Payne and Finnegan 2007; Alroy et al. 2008). However, the use of compiled data in paleontology has moved at a faster pace than the development of protocols for best practices in data citation (Payne et al. 2012; Kaufman et al. 2018), which has contributed to a decrease in the number of taxonomic experts in paleontology (e.g., Payne et al. 2012), much as it has in overlapping disciplines (e.g., archaeology [Marwick and Birch 2018], biodiversity research [Escribano et al. 2018; Mandeville et al. 2021], ecology and evolution [Hood and Sutherland 2021], taxonomy [Agnarsson and Kuntner 2007; Engel et al. 2021; Benichou et al. 2022]). As paleontology and related disciplines move toward a FAIR (Findability, Accessibility, Interoperability, and Reuse; Wilkinson et al. 2016; and see https://www.go-fair.org/fair-principles) infrastructure for digital assets in the long-term future, a short-term solution is needed to ensure the continuance of the specimen-based work that is foundational to each of the areas of research.

Here we quantify the extent to which scientific contributions of data-provisioning publications are unseen and uncredited and discuss present and future consequences of this imbalance. We do so by estimating the number of neglected citations, defined here as citations that were not attributed to these studies despite the data being used, in peer-reviewed publications based on analyses of the Paleobiology Database (PBDB; hereafter, "PBDB publications," including only those listed as "official publications"). The PBDB was selected as it is one of the oldest, largest, and most widely used paleontological databases and maintains a list of publications that make use of the database (i.e., "official publications"). We transform the raw estimates of neglected citations into an annual citation rate that enables us to standardize comparison of citations across PBDB publications and the underlying data-provisioning publications and capture an estimate of neglected citations. To demonstrate the effect of neglected citations beyond individual publications, we also estimate changes to the impact factors of paleontological journals (e.g., *Acta Palaeontologica Polonica*, *Journal of Paleontology*, *Palaeontology*) that often publish specimen-based work (used inclusively for taxonomy, systematics, morphology, and other areas associated with data provisioning). Leveraging these comparisons, we advocate for the proper citation of specimen-based work in paleontology and present a strategy for more equitable data citation.

Methods

The data used to produce this study were drawn from published studies based on data from the PBDB (https://paleobiodb.org), bibliometric data from Google Scholar (https://scholar.google.com), and Journal Citation Reports generated by Clarivate (https://jcr.clarivate.com/jcr/home). These data were used to estimate the extent to which data-provisioning publications have been undervalued through a lack of citation when their data have been reused in publications drawing from the PBDB. We estimated how citation metrics for data-provisioning publications would change if they were cited in all instances where their data outputs were reused, as well as the effect this would have on the impact factors of discipline-specific paleontological journals. All analyses were carried out using R 4.1.2.

https://doi.org/10.1017/pab.2023.33 Published online by Cambridge University Press

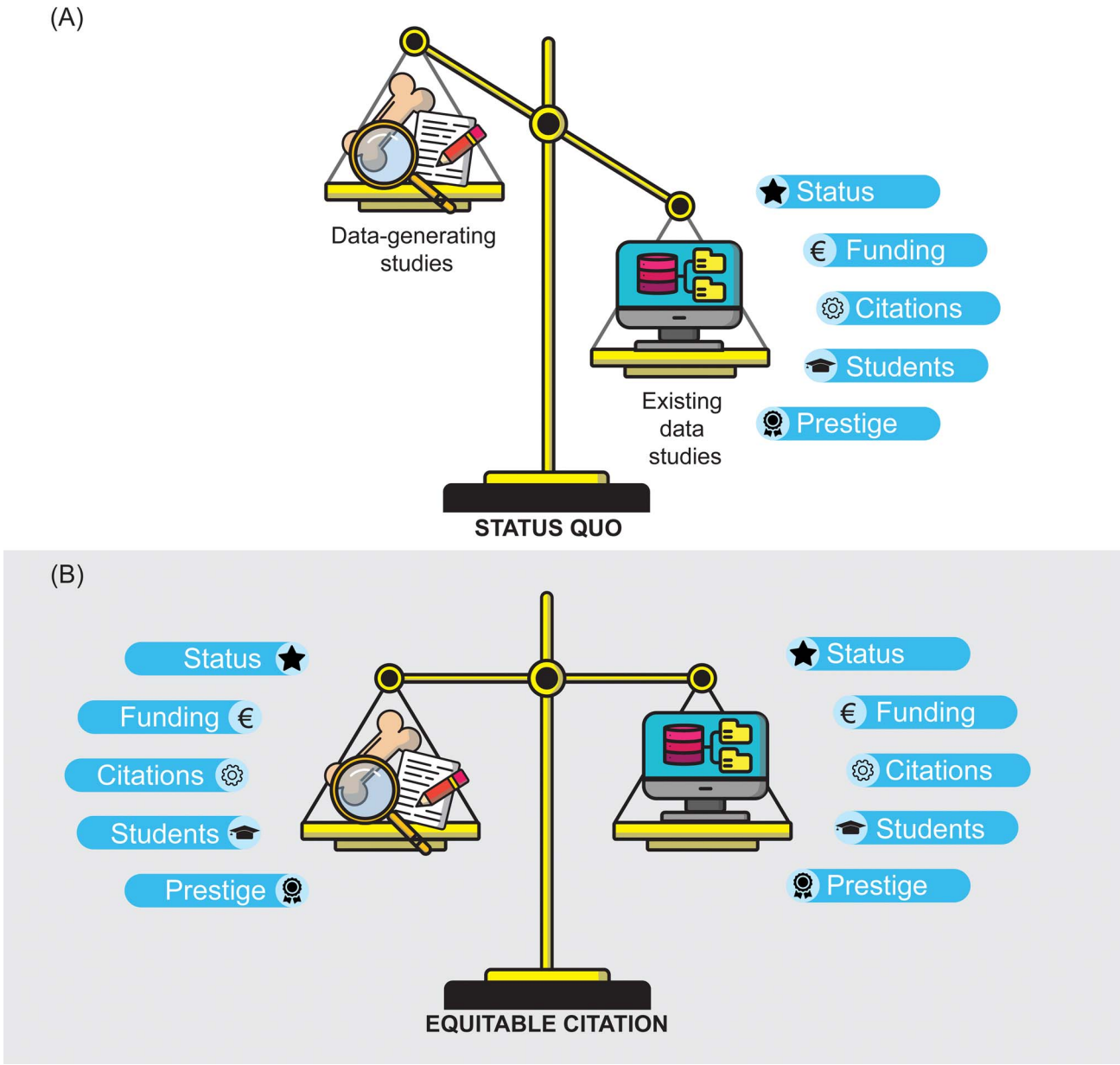

Figure 1. The current balance of credit distribution in paleontology (A) and a reimagined dynamic in which data-provisioning publications are equitably cited (B).

Data Collection on Paleontological Data Reuse

In this study, we focused on the Paleobiology Database (https://paleobiodb.org). The PBDB is among the most commonly used large fossil occurrence databases and is widely used in large-scale temporal and spatial analyses of biodiversity in the fossil record. The PBDB records a list of "official publications," which are publications that use data from the database and have requested an official publication number (see: https://paleobiodb.org/#/publications)—this list is maintained to demonstrate the importance and utility of the database to funding agencies. We compiled this publication list into a dataset on May 6, 2021, and at that time, the list included 396 publications spanning the years 2001–2021.

As our study required the scientometric information for the original publications that contributed the data reused in the PBDB publications, we extracted the raw datasets associated with each PBDB publication whenever they were available (e.g., those uploaded to a data repository linked to the manuscript). It was important to have the raw datasets, because these are downloaded from the PBDB directly and contain the reference information for the data-provisioning publications. We assumed that all data listed in these exported dataset files were used in the subsequent study, and we gave equal weight to a study provisioning 1 or 100 data points (see Dosso and Silvello [2020] for an alternative approach to data credit distribution). When these datasets were not available online, we sent a personalized template email (see Supplementary Material) to the lead or corresponding author(s) of PBDB publications asking for the dataset. If no response was received after 2 weeks, we contacted authors again with a follow-up email. Within a few days (median = 1 day, mean = 5.5 days), 50% of the 167 responses provided either a file or a link to the requested data, 17% of responses indicated

https://doi.org/10.1017/pab.2023.33 Published online by Cambridge University Press

that the files had been lost, 9% of responses indicated only simple use of the PBDB that required no download, and 23% of responses indicated the publication did not use PBDB data. In some cases, authors provided us with the parameters they used to extract their data from the PBDB; however, as the PBDB is a dynamic database, the data produced by these queries change over time and could not be incorporated. We did not receive a response from authors for 25% of the PBDB publications (68/268 requests). In total, we were able to extract the needed information from 151 PBDB publications, accounting for 38% of PBDB publications (total = 396) within the temporal scope of our data collection phase (see Smith et al. 2023a).

Existing and Neglected Citations for Data-provisioning Publications

Using the combined data from the 151 datasets available to us from PBDB publications, we compiled each instance of unique citation information, yielding a list of 49,999 data-provisioning publications. To quantify the magnitude of neglected citations attributed to these references, we first needed to extract the existing number of citations for each publication. This was done by scraping citation data of each data-provisioning publication from Google Scholar in June–August 2021. Google Scholar is detached from academic publishers and other metadata aggregators (e.g., CrossRef, Scopus) and has less transparency than some of these other tools; however, it continues to be commonly used by the academic community and is readily accessed, making it a suitable choice for the objectives of this study. The process of scraping citations was complicated by several factors, including incomplete citation information in some PBDB datasets and issues with Google Scholar not retrieving the correct publication associated with a citation. Consequently, 9816 references required non-automated data extraction by members of the authorship team between August 2021 and April 2022—it is possible that some publications received additional citations during this period, and we assume the overall effect was negligible (e.g., as the median citation rate was relatively low, this is substantiated in our data). Overall, this process produced citation information for 47,122 of the 49,999 (94.2%) data-provisioning publications. Citation data were also extracted for all 396 PBDB publications to enable comparisons between citations of the two publication types.

We tabulated the number of times data from each data-provisioning paper were reused. Although the number of neglected citations is informative on its own (Supplementary Figs. S2, S3), we standardized citations to an annual rate to enable comparison between data-provisioning and PBDB publications. Likewise, we focused on publications from the period of 2001–2021, as this encompasses the period during which PBDB publications have existed and rates of citation are likely influenced by the time period being considered (e.g., more citations and publications in more recent times). Annual citation rates were calculated for data-provisioning publications in three scenarios, using: (1) only existing citations; (2) instances of data reuse in the 151 PBDB publications with data available, in addition to existing citations; and (3) extrapolating to potential neglected citations in the entire dataset of 396 PBDB publications for which we sought data in this study (assuming rates of data reuse in this larger dataset would be similar to those in our smaller set; see Supplementary Material for discussion of assumptions). Citation rate of PBDB publications was calculated solely for existing citations and used as a basis of comparison to approximate the relative seen and unseen contributions of data-provisioning publications to paleontology.

Rates of citation for data-provisioning publications in each of the three scenarios were compared statistically to the citation rate for PBDB publications using median and harmonic mean. Comparison of median citation rates was conducted using a Wilcoxon rank sum test with continuity correction. Harmonic mean was also evaluated to account for outliers in the dataset that might have biased comparisons (e.g., a publication with an exceptionally high citation rate). As the results were similar when using median and harmonic mean, only the results using the median are reported in the main text (see Supplementary Material for results with harmonic means).

Estimating Effects on Paleontological Journal Impact Factors

We estimated the effect that citation of data-provisioning publications in past PBDB publications would have on paleontological journals, using impact factor as a metric. For this analysis, we evaluated changes to the journal impact factor (JIF) of all journals categorized by Clarivate as "Paleontology." As with citations to data-provisioning publications, we first compiled the data used to calculate JIF for the period of 1997–2021 (see Smith et al. [2023a] for raw data). These data included number of citable items published in each journal every year, the number of citations of those citable items every year, and the resulting impact factor, which is calculated as, for example:

$$\frac{\text{Citations in 2021 to items published in 2019 and 2020}}{\text{Number of citable items in 2019 and 2020}} \quad (1)$$

All necessary data are compiled annually by Clarivate and published as Journal Citation Reports (https://jcr.clarivate.com/jcr/home), which we accessed between November 18, 2022, and February 19, 2023, for use in calculating adjusted impact factors. To calculate new impact factors, we tabulated the number of neglected citations to data-provisioning publications, aggregated by journal on an annual basis. These neglected citations were added to the number of citations of citable items for each paleontological journal, and impact factor was recalculated based on these new citation counts. Changes in JIF were converted to percent differences to standardize the results. To contextualize these changes within the scope of publishing in paleontology, in general, we also tabulated and plotted the total number of citable items and citations to those items each year.

Results and Discussion

Balancing Data Use and Citation in Paleontology

PBDB publications were cited at a median rate of 4.28 times each year (median absolute deviation: 3.47), a significantly greater rate than annual citations for data-provisioning publications from the same period of time (1.35/year, median absolute deviation: 1.26; Wilcoxon rank sum test, p-value < 0.0001; Fig. 2; data available in Smith et al. [2023a]). When a citation was credited to each data-provisioning publication within the available subset of data-using publications (151 out of 396 PBDB publications with available data), the citation rate increased to 2.44 each year (median absolute deviation: 1.49). Assuming these 151 publications are a representative sample of all PBDB publications (see Supplementary Material for discussion of assumptions),

https://doi.org/10.1017/pab.2023.33 Published online by Cambridge University Press

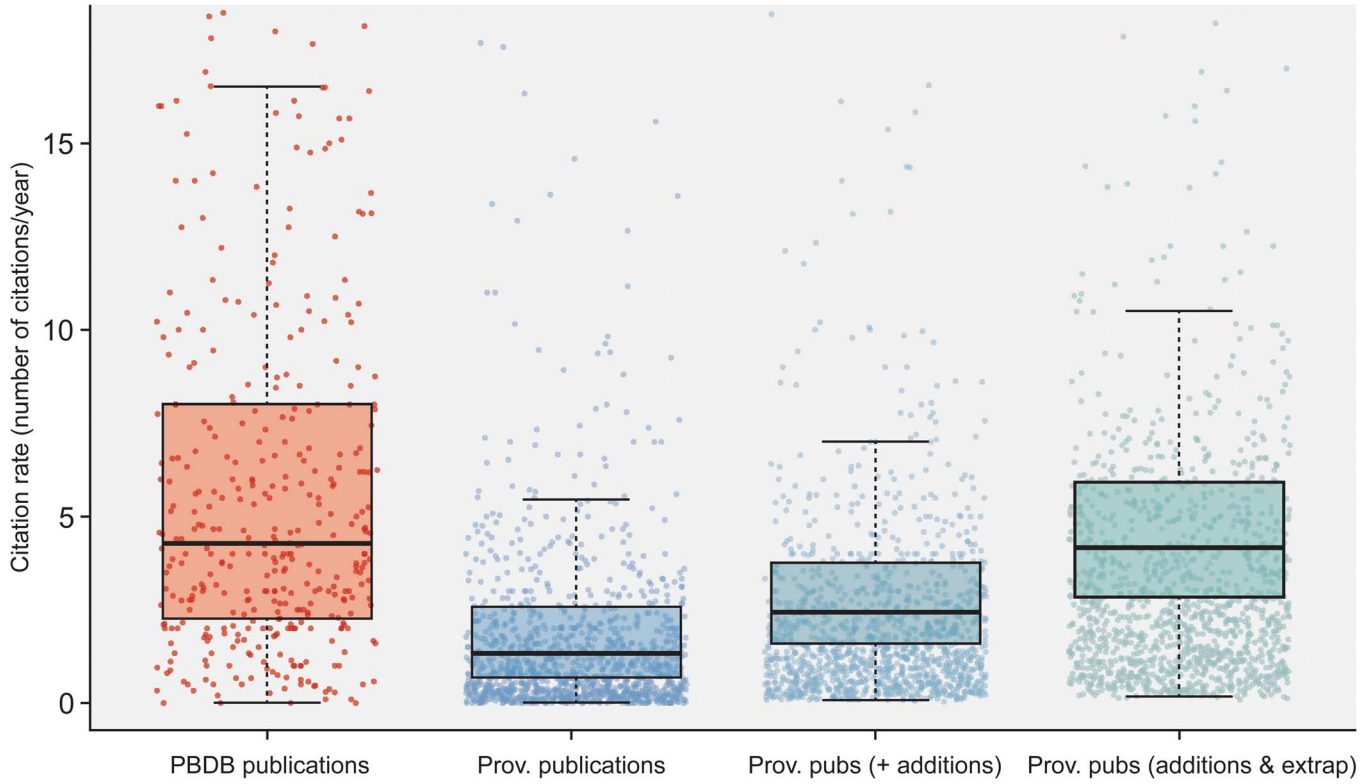

Figure 2. Citation rates for official Paleobiology Database (PBDB) publications and the data-provisioning publications used in those PBDB publications. Only data-provisioning publications from the same time frame (since 2001) as PBDB publications are included to standardize for temporal effects. Citations to data-provisioning publications (i.e., primary literature) are presented as the current rate (i.e., no additions for neglected citations), the projected rate when including citations from PBDB publications where data were available (k = 112; i.e., additions), and the projected rate when making those additions and extrapolating to the entire set of PBDB publications (k = 396; i.e., additions and extrapolated).

extrapolating to the entire set of 396 PBDB publications increased the median citation rate for data-provisioning publications to 4.16 annual citations (median absolute deviation: 2.22), statistically indistinguishable from the median rate for PBDB papers (Wilcoxon rank sum test, p-value = 0.2103; Fig. 2)—using the harmonic mean as the summary statistic to downweight outliers (e.g., publications with extraordinarily high citation rates) resulted in a similar pattern (see "Supporting Analyses" in the Supplementary Material). These results suggest data-provisioning publications should be cited at a rate equal to that for the PBDB publications that reuse their data.

It is clear that the status quo (Fig. 1A)—where the professional reward for PBDB publications is three times greater than for data-provisioning publications—does not give adequate recognition to the importance of data-provisioning publications and the effort required to produce them. At a minimum, the outputs of data-provisioning publications represent intellectual input by their authors and, in many cases, represent dozens or hundreds of hours of work and large financial investment (Agnarsson and Kuntner 2007; Ebach et al. 2011; Baker and Mayernik 2020; Melville et al. 2021). As has been broadly recognized in the literature, data producers deserve credit for their foundational work (Agnarsson and Kuntner 2007; Payne et al. 2012; Penev et al. 2017; Cousijn et al. 2018, 2019; Kaufman et al. 2018; Marwick and Birch 2018; Silvello 2018; Zhao et al. 2018; Jones et al. 2019; Lammey 2019; Pierce et al. 2019; Tomaszewski 2019; Colavizza et al. 2020; Dosso and Silvello 2020; Dorta-González et al. 2021; Hood and Sutherland 2021). In a hypercompetitive academic environment where many aspects of an individual's career (e.g., reputation, career prospects, funding) are influenced by citation counts, this status quo for citation practice is neither fair nor sustainable (Agnarsson and Kuntner 2007; Neylon and Wu 2009; Payne et al. 2012; Piwowar and Vision 2013; Tang et al. 2017; Curry 2018; Gingras and Khelfaoui 2018; MacRoberts and MacRoberts 2018; Silvello 2018; Pierce et al. 2019; Stern and O'Shea 2019; Colavizza et al. 2020; Dosso and Silvello 2020; Raja and Dunne 2022). Without rebalancing the credit distribution (Fig. 1B), emerging "big data" research—considered here as research using large amounts of data (e.g., > 1 TB) including environmental data, images, stratigraphic information, taxonomic records, and more (Leonelli 2014; Allmon et al. 2018)—in paleontology is at risk of undercutting itself by contributing to a systematic devaluation of the specimen-based work that is foundational to the discipline itself.

The estimated citation rate for data-provisioning papers after the addition of neglected citations demonstrates the fundamental and underappreciated value of specimen-based work in paleontology. One way to acknowledge its value and to incentivize future specimen-based work is to cite the data in a formal way when they are used (Piwowar and Vision 2013; Penev et al. 2017; Cousijn et al. 2018, 2019; Kaufman et al. 2018; Silvello 2018; Zhao et al. 2018; Pierce et al. 2019; Dosso and Silvello 2020; Dorta-González et al. 2021; Hood and Sutherland 2021). Although citations are inherently flawed as a metric and subject to biases (e.g., Gingras and Khelfaoui 2018; MacRoberts and MacRoberts 2018; Davies et al. 2021; Hood and Sutherland

https://doi.org/10.1017/pab.2023.33 Published online by Cambridge University Press

2021; Raja and Dunne 2022), citations in one form or another are likely to continue being used to evaluate researchers (see Hicks et al. [2015] for cautionary guidelines and Wilkinson et al. [2016] for discussion of FAIR principles). Citing data producers may be another step toward increased transparency and reproducibility in the pipeline from data production to digital upload and reuse (Wilkinson et al. 2016; Escribano et al. 2018; Hood and Sutherland 2021; see also Supplementary Table S1 and "Additional Contributions to the Paleobiology Database" in the Supplementary Material). Consequently, the development of a clear protocol for citing data can set a community-wide standard that preempts many of the shortcomings reported for traditional text citations (e.g., Gingras and Khelfaoui 2018; MacRoberts and MacRoberts 2018; Davies et al. 2021). The recommended best practices for data citation from the broader literature, from which paleontology can draw (Payne et al. 2012; Kaufman et al. 2018), include two general themes: (1) credit data provisioning by citing the publication in which the data were initially reported, or (2) use new metrics specifically developed for data citation.

Conceptually, the most straightforward way to credit data producers is to cite the publication from which the data were originally reported when the data are reused (Penev et al. 2017; Cousijn et al. 2018; Hood and Sutherland 2021; and as recommended by some databases, e.g., BioTIME, https://biotime.st-andrews.ac.uk/usageGuidelines.php; Neotoma, https://www.neotomadb.org/data/data-use-and-embargo-policy). As the most basic option, this strategy carries the simplifying assumption that all authors participated in data production and credits them equally on this basis (see Pierce et al. [2019] for a counterargument). By virtue of its simplicity, this strategy for citing original publications facilitates ease of use through rapid integration into existing citation metrics, circumventing the need for an independent data citation tool. A prerequisite for using many data citation tools is a unique identifier for datasets (e.g., DOI), which is not available for many past publications (Hood and Sutherland 2021) and, in recent publications, continues to be a shortcoming driven by poor adherence to data-sharing recommendations (Gabelica et al. 2022; see Agosti et al. [2022] for recommendations on use of identifiers). Many of the datasets included in the PBDB do not have unique identifiers, making the application of more complex data citation tools intractable. In alignment with our objective, citing original publications upon data reuse allowed for the most intuitive comparison between data-provisioning and data-using publications with a metric already familiar to academics.

It also bears stating that citing the database itself—in this case study, the PBDB—is necessary but not sufficient. As a consequence of being secondary sources of data, databases indirectly can create a barrier to citation of data-provisioning publications by masking the original data sources. Reflecting this issue, several databases (e.g., BioTIME, Neotoma) provide guidance on citing original data sources and, in the PBDB itself, recommendations toward this end have been made (https://paleobiodb.org/#/faq/how-should-the-paleobiology-database-data-be-cited-; see also Uhen et al. [2023] for a current user guide).

Citation practice is developing rapidly, as a host of data citation tools have been proposed—including the Data Citation Index (Clarivate 2023), SageCite (Lyon 2010), Data Usage Index (Ingwersen and Chavan 2011), and Data Credit Distribution (Dosso and Silvello 2020)—and multiple working groups have been convened on this topic (e.g., Scholix, Data Usage Metrics, Data Citation Synthesis Group). One of the driving principles behind the development of these metrics is the idea that data use is complex and therefore requires a tool that captures the nuances of data (Data Citation Synthesis Group 2014; Cousijn et al. 2019; Dosso and Silvello 2020; Hood and Sutherland 2021). As a scientist's value cannot be distilled to a single metric, using several of these tools in combination with other measures of a person's contributions to science and society may be a more equitable option for evaluating scientists in the future (Neylon and Wu 2009; Curry 2018; Ewers et al. 2019; Stern and O'Shea 2019; Davies et al. 2021; Hood and Sutherland 2021; Westoby et al. 2021). Given the attention to citation practice and alternative metrics in the recent literature and the progress made by working groups on the topic, it may only be a matter of time before data citation metrics become mainstream (Data Citation Synthesis Group 2014; Kaufman et al. 2018; Cousijn et al. 2019; Hood and Sutherland 2021).

Regardless of the citation approach, attributing credit to data provisioning (and all individuals involved in the process of making data available in digital compilations; see Supplementary Table S1; see also Escribano et al. 2018; Benichou et al. 2022) in a professionally meaningful way represents a shift in citation practice and credit distribution in paleontology (Payne et al. 2012; Kaufman et al. 2018; Fig. 1). With the rise of quantitative paleontology and the associated shift away from paleontology's traditional descriptive roots, it is imperative we find reasonable and equitable ways to improve our data citation practices. For example, a single publication might draw data from thousands of primary sources (Supplemental Fig. S4) and, particularly for journals with strict page limits or length-based page charges, it is often not feasible to include citations for each of the data-provisioning publications. Though it will continue to be impractical to cite thousands of papers in printed format, the growing awareness of the importance of data citation and improving digital infrastructure provide a path forward. As a starting point, online archives and preprint servers (e.g., BioArXiv, EarthArXiv, Open Science Framework) can accommodate the long list of references required to cite all data-provisioning publications used by a publication based on the PBDB or another database. These online archives and preprint servers are routinely indexed by aggregators (e.g., Google Scholar, Web of Science). Publication of a reference list with an online repository or preprint server can increase the likelihood that citations are attributed to data-provisioning publications but, critically, the references must be included with the main text in the references section, not placed in the supplementary material. Current processes for aggregating citations do not find references in supplementary material. Raja et al. (2022a) illustrated the feasibility of this approach by publishing their database references in a preprint hosted at Open Science Framework (Raja et al. 2022b). To alleviate the burden and facilitate consistency of compiling these large reference lists, future authors can use the R package refer (https://github.com/adamkocsis/refer). This package offers tools to generate a formatted document containing the metadata and reference list that can be used to upload to the aforementioned online archives. The user is required to provide either a text file containing formatted references or a BibTeX file containing the references in their data-using publication along with other generic information (e.g., title of the publication, author affiliations). The package also includes a ready-made template for the formatting of the document, and experienced users can provide their own templates.

Still, this solution is a stopgap measure, and it would be preferable for journals to implement policies and technical changes

https://doi.org/10.1017/pab.2023.33 Published online by Cambridge University Press

on their platforms to encourage more equitable citation practices. Although many journals still print hard copies, essentially all journals have online versions, and many journals are now published online-only. Even so, many online publishers have retained strict manuscript length policies, thereby limiting the number of references allowed. Rather than relegating data reference information to the depths of supplementary material where they will not be included in citation counts, online journals can, as a first step, omit the reference list from their imposed page limits. Another option for journals is to require authors to submit a list of references for data-provisioning publications as an appendix and to publish this list with the main text references in the online version (e.g., McGill et al. 2016)—a printed issue could still include only the references cited in the main text. Encouragingly, some journals (e.g., *Global Ecology and Biogeography* [McGill et al. 2016], *Scientific Data* [personal experience, e.g., Raja et al. 2022a]) have already made these changes, allowing authors to fully cite their data sources. Admittedly, these changes will be somewhat onerous, as they require managing and formatting thousands of references; however, AI tools and the refer package presented here are viable options for streamlining this process. Whether these changes are adopted more broadly will depend on demand from the community.

Broader Considerations for Paleontology as a Discipline

Improving data citation practice will also have a positive effect on paleontological journals, especially those that publish specimen-based work (e.g., *Acta Palaeontologica Polonica*, *Journal of Paleontology*, *Journal of Vertebrate Paleontology*). Whereas higher-profile outlets (e.g., *Science*, *Nature*) tend to publish paleontological articles on charismatic and unusual specimens (e.g., dinosaurs, fossils in amber) or on large data compilations (e.g., latitudinal diversity gradients, extinction), most paleontological papers are published in discipline-specific journals (Raja and Dunne 2022). As might be expected, publications in these journals traditionally receive fewer citations, and the journals have lower impact factors.

Just as the citation rate for data-provisioning publications increased after accounting for neglected data citations (Fig. 2), the JIF—another flawed but commonly used evaluative metric (e.g., Neylon and Wu 2009; Stephan et al. 2017; Curry 2018; Stern and O'Shea 2019)—increases substantially for paleontological journals (Fig. 3A). Combining our tabulated neglected citations with currently attributed citations used to calculate JIF by Clarivate (https://jcr.clarivate.com/jcr/home), we found that in the last decade (2010–2019), the JIF reported for a journal in a given year (e.g., *Journal of Paleontology* in 2015) would increase on average by ~0.1, or 5.08%. This is a conservative estimate, as it only includes neglected citations from the 151 PBDB publications for which data were available. Extrapolating to the entire dataset of 396 PBDB publications suggests that any of the 55 journals categorized by Clarivate as a paleontological journal would see an increase in JIF by ~0.2, or 13.3% (see "7_paleo_journal_JIFcalculation.csv" in Smith et al. [2023a] for raw data for all 55 paleontological journals from 1997 to 2021 and additional information on language and country of publishing). The change in JIF from neglected citations was not, however, uniform across journals or through time. Whereas some journals had no ($n = 10$; e.g., *Micropaleontology*, *Paleoceanography and Paleoclimatology*, *Stratigraphy*) or few (e.g., *GFF*, *Palaios*, *Zootaxa*) neglected citations in our dataset and limited associated changes to JIF, other journals would have substantial increases in JIF in one or more years (e.g., *Palaeontologia Electronica*, *Palaeontology*, *PalZ*). Furthermore, for those journals with large JIF changes, there is a notable increase in the effect of adding neglected citations in more recent years (Fig. 3B). Neglected citations rarely contributed to JIF in the early part of the decade (2010 onward); however, at the end of the decade, the average JIF for 10 highly impacted journals increased by 27% in 2018 and 36% in 2019 when recalculated to include neglected citations. These differences through time are a consequence of the formula for calculating JIF and publishing trends in paleontology (Fig. 3C, D). Because JIF for a given year (e.g., 2020) is based on citations of research published in the preceding 2 years (e.g., 2018 and 2019), a relatively short turnaround time is needed between publication of a data-provisioning study and subsequent PBDB publication using those data. Consequently, many instances of data reuse cannot be incorporated into this metric because of the limited look-back period. Alternatives to the 2-year JIF do exist (e.g., 5-year JIF); however, analyses comparing 2- and 5-year JIFs show minimal differences between the two (e.g., Campanario 2011; Dorta-González and Dorta-González 2013). Though it continues to be a widely used metric across many branches of science, JIF performs poorly when capturing reuse of data and undervalues journals where data-provisioning studies are published. Accelerating publication rates in paleontology (Fig. 3C,D) and the shift from printed to online publication appear to have reduced the time between initial data publication and data reuse. Moreover, the number of paleontological journals published in 2021 was 55, more than double the 24 published in 1997 when Clarivate began compiling Journal Citation Reports. As the number of publications and citations in paleontology continues to grow (Fig. 3C,D), so too will the consequences of neglected data citations (Fig. 3A,B). A more equitable future in paleontology requires rapid correction to citation practice.

JIF influences more than how journals rank in comparison to one another; it also influences how authors and the work they publish in those journals are regarded and rewarded professionally (Neylon and Wu 2009; Stephan et al. 2017; Curry 2018; Stern and O'Shea 2019). Despite the poor performance of JIF as an indicator of quality, JIF continues to influence an author's choice of publication venue and contributes to the perceived importance of the papers published in the journal and, more broadly, the discipline (Neylon and Wu 2009; Curry 2018; Stern and O'Shea 2019). In lieu of systematic changes in publication practices in science (e.g., Kravitz and Baker 2011; Curry 2018; Davies et al. 2021), increasing the prestige of discipline-specific journals is imperative for increasing the profile of paleontology and will benefit all in the discipline.

Data sharing—particularly when credited appropriately—is an equally important component in any effort to strengthen the field of paleontology. As with data citation, the issue of data sharing is commonly discussed in the literature, and there is a consensus that it is incumbent upon authors to share the data they use to produce their results (e.g., Piwowar and Vision 2013; Kaufman et al. 2018; Marwick and Birch 2018; Jones et al. 2019; Lammey 2019; Mandeville et al. 2021). Even so, data sharing is not practiced consistently (Stuart et al. 2018; Gabelica et al. 2022; Roche et al. 2022). Several large publishers (e.g., Elsevier, Springer, Taylor and Francis, Wiley) have data availability policies with multiple tiers, ranging from written recommendations to strict requirements for publishing data, but it remains at the discretion of journals to enact and enforce these policies (Jones et al. 2019)—

https://doi.org/10.1017/pab.2023.33 Published online by Cambridge University Press

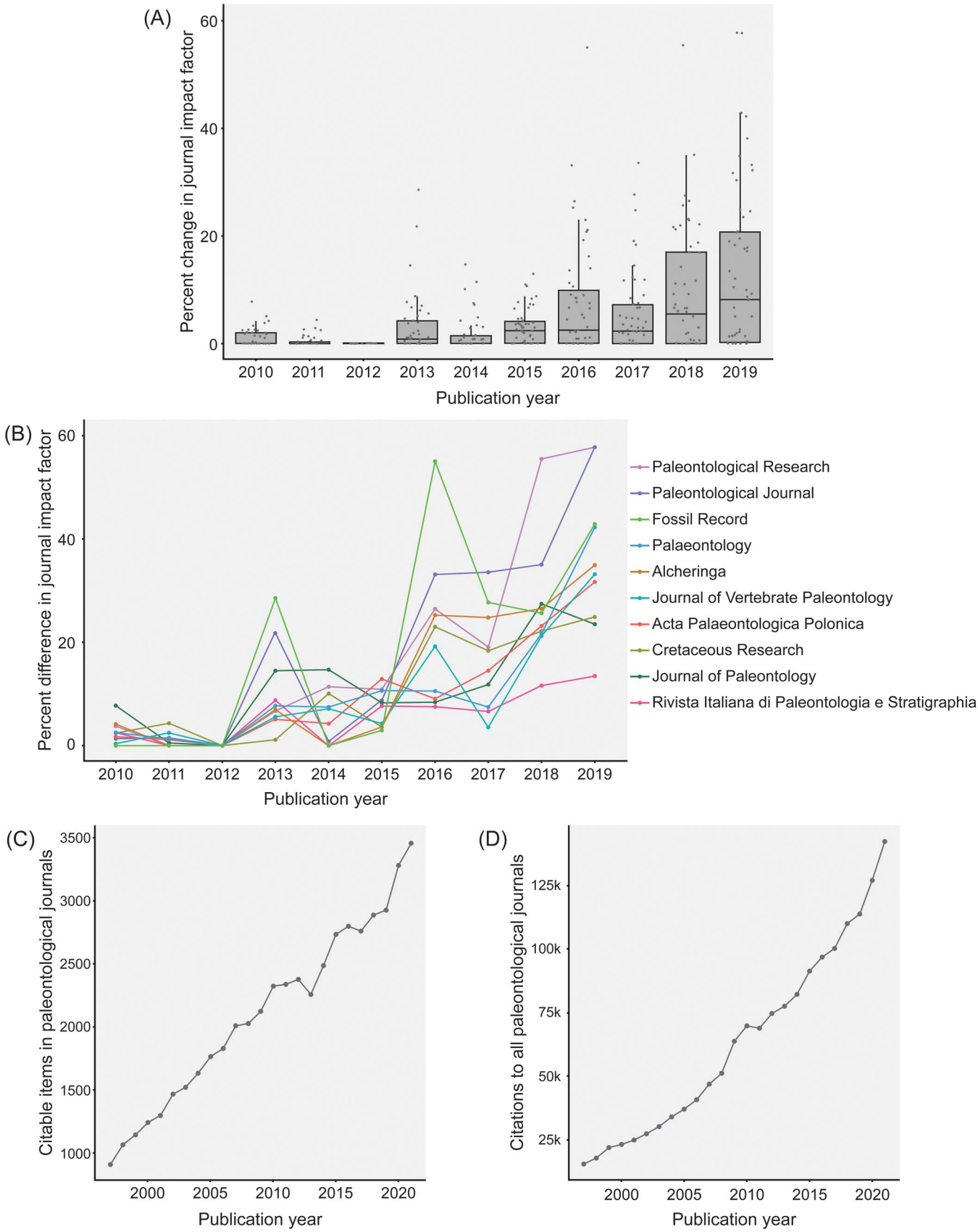

Figure 3. The effects of adding neglected citations from data reuse on journal impact factor (JIF; A, B) and general patterns in publishing trends in paleontology (C, D). A, The increase in JIF for the 55 journals categorized to paleontology by Clarivate, for the period of 2010 to 2019. Note, an outlier value of 172% in 2018 for *PalZ* was not plotted. B, Increases in JIF for the 10 paleontological journals most affected by neglected citations, only including those with complete data for the duration of 2010 to 2019. For raw data for all 55 paleontological journals from 1997 to 2021, see "7_paleo_journal_JIFcalculation.csv" in Smith et al. (2023a). C, The number of citable items published in paleontological journals each year. D, The number of citations to items published in paleontological journals each year.

https://doi.org/10.1017/pab.2023.33 Published online by Cambridge University Press

this may change in the United States, however, with a new mandate for public availability of data produced in federally funded research, beginning in 2023 (National Science and Technology Council 2022). As demonstrated by Gabelica et al. (2022), who found that data were only available for 6.8% ($n = 3,556$) of publications in their review of 333 open access journals from BioMed Central, many data-sharing policies are not effective in practice. Data availability was considerably better in the present study, with data accessible for 32% ($n = 128$) of PBDB publications—note that though data were available, they were not always usable for the analysis conducted here. Encouragingly, when data from PBDB publications were not readily available online, 167 of the 268 (68%) authors who were contacted were responsive, and approximately half ($n = 84$) of these responses included the requested data. Still, we were unable to recover data for 21% ($n = 83$) of PBDB publications for myriad reasons. With pushes toward big data science in paleontology and related disciplines, it will be up to the community to influence journal policies toward required sharing rather than relying on unenforceable recommendations (Payne et al. 2012; Kaufman et al. 2018; Jones et al. 2019).

Improved data sharing requires buy-in from individuals, who may themselves benefit from the practice and enhance the quality of science in paleontology. As reviewed by Marwick and Birch (2018), there are many reasons to share data (e.g., reciprocal data sharing by others; reproducibility of research; enabling others to ask new questions) and some associated costs (e.g., time required to clean data; data use without citation). One of the incentives is that data sharing is associated with increased citation of the publication where the data were initially published (Sears 2011; Piwowar and Vision 2013; Tomaszewski 2019; Colavizza et al. 2020; Dorta-González et al. 2021). For example, Colavizza et al. (2020) reported that when publications included data availability statements with the associated data publicly accessible, those publications saw a 25% increase in their citations compared with publications without available data. As demonstrated here (Fig. 2), the potential citation benefit may be even larger in a discipline like paleontology, where publications on data compilations have become mainstream. Changes to the format on funding proposals, for example, inclusion of a "research outcomes" section that includes datasets by the Deutsche Forschungsgemeinschaft (i.e., German Research Foundation) and a non-publication section in National Science Foundation grant reports, can further encourage data sharing. Of perhaps greater importance, data sharing ensures the reproducibility of scientific results (Piwowar and Vision 2013; Altman et al. 2015; Marwick and Birch 2018). As has been demonstrated to the detriment of many fields of study (e.g., behavioral ecology [Viglione 2020], food science [van der Zee et al. 2017], paleontology [Price 2022], psychology [John et al. 2012]), some researchers have been guilty of misrepresenting their data. Data sharing provides a means to uphold academic integrity and establishes an ethical and practical standard that encourages scientific advancement (Marwick and Birch 2018; Raja and Dunne 2022).

Paleontology has not yet crossed the threshold to become a big data discipline (Allmon et al. 2018) but has the potential to do so in the near future. Realizing this potential will expand research horizons in paleontology but, to be done effectively and equitably (e.g., Raja et al. 2022c), it requires a stable foundation in specimen-based work and reckoning with structural biases. Large paleontological databases, including the PBDB, are far from complete. For example, in examining the collections at nine paleontological museums in the United States, Marshall et al. (2018) estimated that there were 23 times the number of unique localities in only those nine collections than were in the PBDB at the time. Paleontologists should be wary of assuming our databases are comprehensive, as "having a lot of data is not the same as having all of them; and cultivating such an illusion of completeness is a very risky and potentially misleading strategy" (Leonelli 2014: p. 7). Activating the extensive data held in museum collections (e.g., unpublished specimens; "extended specimen" data; Webster 2017; Allmon et al. 2018; Marshall et al. 2018), will require support for the infrastructure sustaining collections and recognition of the importance of specimen-based work that often goes wanting in paleontology and related disciplines (Johnson et al. 2005; Agnarsson and Kuntner 2007; Payne et al. 2012; Allmon et al. 2018; Marshall et al. 2018; Engel et al. 2021; Benichou et al. 2022). A critical component to realizing a big data future in paleontology will be increased funding to support museum collections and data repositories, with respect to both maintaining existing materials and to obtaining and curating new materials. Illustrating the scope of the need, Allmon et al. (2018) estimated that it costs US$1 to digitize each specimen, and digitizing only the currently identified specimens in U.S. collections (as of 2018) would require an investment of US$35 million. That figure increased to US$75 million after including all fossils, not just those with existing taxonomic identifications. Investment at this scale represents a massive increase in funding, as the budget for this type of work in the United States was only US$10 million at the time (Allmon et al. 2018). These monetary estimates also do not account for the costs of data storage and maintenance of data repositories (whether museum-based or external) that provide access to other researchers and the public. Particularly, as complex data become more commonplace (e.g., CT scans, images), infrastructure requirements will be critical to ensuring a big data future in paleontology. Without funding for this fundamental work, growth and advances in paleontology will be slow at best.

The illusion of completeness is elucidated further when considering biases in where data recorded in paleontological databases originated, what organisms are preferentially studied, who contributes to compiling data in databases, and who conducts the research (e.g., Raja and Dunne 2022; Raja et al. 2022c). Compilations of modern biodiversity data show a clear association between data production and wealthier, more resource-rich countries, particularly those in western Europe and North America (Amano and Sutherland 2013; Hughes et al. 2017). The same is true for compilations of paleontological data; a recent study examining data recorded in the PBDB found that 97% of fossil occurrence data were produced by researchers based in high- or upper middle-income countries (Raja et al. 2022c). The same study found a direct link between paleontological data production and socioeconomic factors, such as greater wealth, education level, and political stability (Raja et al. 2022c). These patterns clearly illustrate a global knowledge and power imbalance in paleontological research that can only be rectified by changes to how paleontological research is conducted (Cisneros et al. 2022; Monarrez et al. 2022; Raja et al. 2022c).

Conclusion

The scientific value of large-scale analyses in paleontology is undeniable, and the scope and quality of insights produced in such analyses will only increase with the inclusion of more data.

https://doi.org/10.1017/pab.2023.33 Published online by Cambridge University Press

Databases like the PBDB have been instrumental in making these research directions possible and, with a community initiative to improve data citation and sharing practices, can continue to unlock new discoveries about life on Earth. Although we focus here on paleontology, similar imbalances affect related and overlapping disciplines (e.g., archaeology [Marwick and Birch 2018], biodiversity research [Escribano et al. 2018; Mandeville et al. 2021], ecology and evolution [Hood and Sutherland 2021], taxonomy [Agnarsson and Kuntner 2007; Engel et al. 2021; Benichou et al. 2022]), all of which can benefit from similar structural improvements. Whether citations are attributed to the data or to the original publication, there are potentially large implications for how research and researchers are credited and valued, and how journals are perceived. Our objective here is not to devalue papers examining large-scale trends relying on data compilations drawn from other scientists' work, but rather to ensure it remains feasible for taxonomists, systematists, and other specimen-based workers, and those conducting the equally important work on stratigraphy, lithology, and depositional environments, to publish research and be credited in a way that acknowledges their critical importance to paleontology and all life sciences. Citation counts and the metrics derived from them continue to influence most aspects of a scientific career. When people producing data receive proper credit, the community data pool will increase in availability and quality. At the same time, the profile and prestige of paleontological journals will improve. As a unified science, paleontology will benefit and grow.

Acknowledgments. We thank the many authors of the official PBDB papers who shared their raw data with us, and those responsible for maintaining the PBDB as the excellent community resource that it is. We also thank M. Patzkowsky, G. Jones, M. Hopkins (editor), and P. Monarrez and P. Novack-Gottshall (reviewers) for their comments that improved an earlier version of this article. This work was supported in part by the Paleosynthesis Project, with funding from the Volkswagen Stiftung, and by the TERSANE project, with funding from the Deutsche Forschungsgemeinschaft (FOR 2332; grant nos. KI 806/17–1 (N.B.R., D.D.), BA 5148/1-2 to K. De Baets (P.S.N.), AB 109/11-1 to M. Aberhan (C.J.R.), and Ko 5382/2-1 (Á.T.K.). P.L.G. was supported by the São Paulo Research Foundation (FAPESP 2022/05697-9). B.M.G. was supported by the National Science Foundation (ANT-1947094 to C. Sidor). B.S. was supported by the Deutsche Forschungsgemeinschaft (JA 2718/3-1) and the Netherlands Earth System Science Centre (NESSC).

Author Contributions. J.A.S., N.B.R., and Á.T.K. contributed equally to this work. J.A.S. led manuscript drafting. J.A.S., N.B.R., Á.T.K., L.P.A.M, C.J.R., and B.S. conceived of and designed the study. D.D., J.A.S., N.B.R., E. M. Dunne, E.M.L., L.P.A.M., P.S.N., C.J.R., B.S., and Á.T.K. contributed to data collection. N.B.R., Á.T.K., and J.A.S. generated code to extract and manipulate data. T.C., D.D., and J.A.S. led figure development. J.A.S. and Á.T.K. conducted analyses. All authors edited, reviewed, and approved the submitted manuscript.

Competing Interest. The authors declare that they have no competing interests.

Data Availability Statement. All data and supplementary material are available on Zenodo at https://doi.org/10.5281/zenodo.7881567.

Code Availability Statement. All code used to extract, manipulate, and visualize data for this manuscript are available at https://doi.org/10.5281/zenodo.7881567. The code for the R package refer is available at https://github.com/adamkocsis/refer.

Literature Cited

Agnarsson, I., and M. Kuntner. 2007. Taxonomy in a changing world: seeking solutions for a science in crisis. *Systematic Biology* **56**:531–539.

Agosti, D., L. Benichou, W. Addink, C. Arvanitidis, T. Catapano, G. Cochrane, M. Dillen, *et al.* 2022. Recommendations for use of annotations and persistent identifiers in taxonomy and biodiversity publishing. *Research Ideas and Outcomes* **8**:e97374.

Allmon, W. A., G. P. Dietl, J. R. Hendricks, and R. M. Ross. 2018. Bridging the two fossil records: paleontology's "big data" future resides in museum collections. *In* G. D. Rosenberg and R. M. Clary, eds. *Museums at the forefront of the history and philosophy of geology: history made, history in the making. Geological Society of America Special Paper* **535**:35–44.

Alroy, J., M. Aberhan, D. J. Bottjer, M. Foote, F. T. Fürsich, P. J. Harries, A. J. Hendy, S. M. Holland, L. C. Ivany, and W. Kiessling. 2008. Phanerozoic trends in the global diversity of marine invertebrates. *Science* **321**:97–100.

Altman, M., C. Borgman, M. Crosas, and M. Matone. 2015. An introduction to the joint principles for data citation. *Bulletin of the Association for Information Science and Technology* **41**:43–45.

Amano, T., and W. J. Sutherland. 2013. Four barriers to the global understanding of biodiversity conservation: wealth, language, geographical location and security. *Proceedings of the Royal Society of London B* **280**:20122649.

Baker, K. S., and M. S. Mayernik. 2020. Disentangling knowledge production and data production. *Ecosphere* **11**:e03191.

Benichou, L., J. Buschbom, M. Campbell, E. Hermann, J. Kvaček, P. Mergen, L. Mitchell, C. Rinaldo, and D. Agosti. 2022. Joint statement on best practices for the citation of authorities of scientific names in taxonomy by CETAF, SPNHC and BHL. *Research Ideas and Outcomes* **8**:e94338.

Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. 2013. GenBank. *Nucleic Acids Research* **41**:D36–D42.

Campanario, J. M. 2011. Empirical study of journal impact factors obtained using the classical two-year citation window versus a five-year citation window. *Scientometrics* **87**:189–204.

Cisneros, J. C., N. B. Raja, A. M. Ghilardi, E. M. Dunne, F. L. Pinheiro, O. R. Regalado Fernández, M. A. F. Sales, *et al.* 2022. Digging deeper into colonial palaeontological practices in modern day Mexico and Brazil. *Royal Society Open Science* **9**:210898.

Clarivate. 2023. Data Citation Index. https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/data-citation-index.

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

Cousijn, H., P. Feeney, D. Lowenberg, E. Presani, and N. Simons. 2019. Bringing citations and usage metrics together to make data count. *Data Science Journal* **18**:9.

Cousijn, H., A. Kenall, E. Ganley, M. Harrison, D. Kernohan, T. Lemberger, F. Murphy, *et al.* 2018. A data citation roadmap for scientific publishers. *Scientific Data* **5**:180259.

Curry, S. 2018. Let's move beyond the rhetoric: it's time to change how we judge research. *Nature* **554**:147–147.

Data Citation Synthesis Group. 2014. Joint declaration of data citation principles. M. Martone, ed. San Diego, Calif. FORCE11. https://doi.org/10.25490/a97f-egyk.

Davies, S. W., H. M. Putnam, T. Ainsworth, J. K. Baum, C. B. Bove, S. C. Crosby, I. M. Côté, *et al.* 2021. Promoting inclusive metrics of success and impact to dismantle a discriminatory reward system in science. *PLoS Biology* **19**:e3001282.

Dornelas, M., N. J. Gotelli, B. McGill, H. Shimadzu, F. Moyes, C. Sievers, and A. E. Magurran. 2014. Assemblage time series reveal biodiversity change but not systematic loss. *Science* **344**:296–299.

Dorta-Gonzalez, P., and M. I. Dorta-González. 2013. Impact maturity times and citation time windows: the 2-year maximum journal impact factor. *Journal of Informetrics* **7**:593–602.

Dorta-González, P., S. M. González-Betancor, and M. I. Dorta-González. 2021. To what extent is researchers' data-sharing motivated by formal mechanisms of recognition and credit? *Scientometrics* **126**:2209–2225.

Dosso, D., and G. Silvello. 2020. Data credit distribution: a new method to estimate databases impact. *Journal of Informetrics* **14**:101080.

https://doi.org/10.1017/pab.2023.33 Published online by Cambridge University Press

Ebach, M. C., A. G. Valdecasas, and Q. D. Wheeler. 2011. Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. *Cladistics* **27**:550–557.

Engel, M. S., L. M. P. Ceríaco, G. M. Daniel, P. M. Dellapé, I. Löbl, M. Marinov, R. E. Reis, *et al.* 2021. The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches. *Zoological Journal of the Linnean Society* **193**:381–387.

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

Ewers, R. M., J. Barlow, C. Banks-Leite, and C. Rahbek. 2019. Separate authorship categories to recognize data collectors and code developers. *Nature Ecology and Evolution* **3**:1610–1610.

Fan, J., Q. Chen, X. Hou, A. I. Miller, M. J. Melchin, S. Shen, S. Wu, *et al.* 2013. Geobiodiversity Database: a comprehensive section-based integration of stratigraphic and paleontological data. *Newsletters on Stratigraphy* **46**:111–136.

Fenton, I. S., A. Woodhouse, T. Aze, D. Lazarus, J. Renaudie, A. M. Dunhill, J. R. Young, and E. E. Saupe. 2021. Triton, a new species-level database of Cenozoic planktonic foraminiferal occurrences. *Scientific Data* **8**:1–9.

Gabelica, M., R. Bojčić, and L. Puljak. 2022. Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *Journal of Clinical Epidemiology* **150**:33–41.

Gingras, Y., and M. Khelfaoui. 2018. Assessing the effect of the United States' "citation advantage" on other countries' scientific impact as measured in the Web of Science (WoS) database. *Scientometrics* **114**:517–532.

Harland, W. B. 1967. *The Fossil Record: A Symposium with Documentation, Jointly Sponsored by the Geological Society of London and the Palaeontological Association*. Geological Society of London, London.

Hicks, D., P. Wouters, L. Waltman, S. de Rijcke, and I. Rafols. 2015. Bibliometrics: the Leiden Manifesto for research metrics. *Nature* **520**:429–431.

Hood, A. S. C., and W. J. Sutherland. 2021. The data-index: an author-level metric that values impactful data and incentivizes data sharing. *Ecology and Evolution* **11**:14344–14350.

Hughes, B. B., R. Beas-Luna, A. K. Barner, K. Brewitt, D. R. Brumbaugh, E. B. Cerny-Chipman, S. L. Close, *et al.* 2017. Long-term studies contribute disproportionately to ecology and policy. *BioScience* **67**:271–281.

Ingwersen, P., and V. Chavan. 2011. Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. *BMC Bioinformatics* **12**:1–10.

John, L. K., G. Loewenstein, and D. Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* **23**:524–532.

Johnson, K. G., H. F. Filkorn, and M. Stecheson. 2005. Into focus: paleontology collections on the World Wide Web: the missing link. *Palaeontologia Electronica* **8**. https://paleo.carleton.ca/toc8_2.htm.

Jones, L., R. Grant, and I. Hrynaszkiewicz. 2019. Implementing publisher policies that inform, support and encourage authors to share data: two case studies. *Insights* **32**:11.

Kaufman, D. S., N. Abram, M. Evans, P. Francus, H. Goosse, H. Linderholm, M.-F. Loutre, *et al.* 2018. Technical note: open-paleo-data implementation pilot—the PAGES 2k special issue. *Climate of the Past* **14**:593–600.

Kiessling, W. 2005. Long-term relationships between ecological stability and biodiversity in Phanerozoic reefs. *Nature* **433**:410–413.

Kiessling, W., and M. C. Krause. 2022. PARED—an online database of Phanerozoic reefs. https://www.paleo-reefs.pal.uni-erlangen.de.

Kravitz, D., and C. Baker. 2011. Toward a new model of scientific publishing: discussion and a proposal. *Frontiers in Computational Neuroscience* **5**. https://doi.org/10.3389/fncom.2011.00055.

Lammey, R. 2019. Data sharing and data citation: join the movement. *European Science Editing* **45**:58–59.

Leonelli, S. 2014. What difference does quantity make? On the epistemology of Big Data in biology. *Big Data and Society* **1**:2053951714534395.

Lyon, L. 2010. *UK Digital Curation Centre: Enabling Research Data Management at the Coalface*. Microsoft Research Faculty Summit 2010, Redmond, Wash.

MacRoberts, M. H., and B. R. MacRoberts. 2018. The mismeasure of science: citation analysis. *Journal of the Association for Information Science and Technology* **69**:474–482.

Mandeville, C. P., W. Koch, E. B. Nilsen, and A. G. Finstad. 2021. Open data practices among users of primary biodiversity data. *BioScience* **71**:1128–1147.

Marshall, C. R., S. Finnegan, E. C. Clites, P. A. Holroyd, N. Bonuso, C. Cortez, E. Davis, G. P. Dietl, P. S. Druckenmiller, and R. C. Eng. 2018. Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. *Biology Letters* **14**:20180431.

Marwick, B., and S. E. P. Birch. 2018. A standard for the scholarly citation of archaeological data as an incentive to data sharing. *Advances in Archaeological Practice* **6**:125–143.

McGill, B. J., M. Dornelas, and R. Field. 2016. A new year with a new leadership team at GEB—or how to guarantee your paper gets into GEB. *Global Ecology and Biogeography* **25**:1–2.

Melville, J., D. G. Chapple, J. S. Keogh, J. Sumner, A. Amey, P. Bowles, I. G. Brennan, *et al.* 2021. A return-on-investment approach for prioritization of rigorous taxonomic research needed to inform responses to the biodiversity crisis. *PLoS Biology* **19**:e3001210.

Monarrez, P. M., J. B. Zimmt, A. M. Clement, W. Gearty, J. J. Jacisin, K. M. Jenkins, K. M. Kusnerik, *et al.* 2022. Our past creates our present: a brief overview of racism and colonialism in Western paleontology. *Paleobiology* **48**:173–185.

National Science and Technology Council. 2022. *Desirable characteristics of data repositories for federally funded research*. https://doi.org/10.5479/10088/113528.

Newell, N. D. 1952. Periodicity in invertebrate evolution. *Journal of Paleontology* **26**:371–385.

Newell, N. D. 1967. Revolutions in the history of life. *In* C. C. Albritton, ed. *Uniformity and simplicity: a symposium on the principles of the uniformity of nature*. Geological Society of America Special Paper **89**:63–92.

Neylon, C., and S. Wu. 2009. Article-level metrics and the evolution of scientific impact. *PLoS Biology* **7**:e1000242.

Payne, J. L., and S. Finnegan. 2007. The effect of geographic range on extinction risk during background and mass extinction. *Proceedings of the National Academy of Sciences USA* **104**:10506–10511.

Payne, J. L., F. A. Smith, M. Kowalewski, R. A. Krause Jr., A. G. Boyer, C. R. McClain, S. Finnegan, P. M. Novack-Gottshall, and L. Sheble. 2012. A lack of attribution: closing the citation gap through a reform of citation and indexing practices. *TAXON* **61**:1349–1351.

Penev, L., D. Mietchen, V. Chavan, G. Hagedorn, V. Smith, D. Shotton, É. Ó. Tuama, *et al.* 2017. Strategies and guidelines for scholarly publishing of biodiversity data. *Research Ideas and Outcomes* **3**:e12431.

Phillips, J. 1860. *Life on the Earth: its origin and succession*. Macmillan and Company, Cambridge.

Pierce, H. H., A. Dev, E. Statham, and B. E. Bierer. 2019. Credit data generators for data reuse. *Nature* **570**:30–32.

Piwowar, H. A., and T. J. Vision. 2013. Data reuse and the open data citation advantage. *PeerJ* **1**:e175.

Price, M. 2022. Paleontologist accused of faking data in dino-killing asteroid paper. *Science*. https://www.science.org/content/article/paleontologist-accused-faking-data-dino-killing-asteroid-paper.

Raja, N. B., D. Dimitrijević, M. C. Krause, and W. Kiessling. 2022a. Ancient Reef Traits, a database of trait information for reef-building organisms over the Phanerozoic. *Scientific Data* **9**:425.

Raja, N. B., D. Dimitrijević, M. C. Krause, and W. Kiessling. 2022b. Database references for "Ancient Reef Traits, a database of trait information for reef-building organisms over the Phanerozoic." https://doi.org/10.31219/osf.io/sxq7m.

Raja, N. B., and E. M. Dunne. 2022. Publication pressure threatens the integrity of palaeontological research. *Geological Curator* **11**:407–418.

Raja, N. B., E. M. Dunne, A. Matiwane, T. M. Khan, P. S. Nätscher, A. M. Ghilardi, and D. Chattopadhyay. 2022c. Colonial history and global economics distort our understanding of deep-time biodiversity. *Nature Ecology and Evolution* **6**:145–154.

Raup, D. M., and J. J. Sepkoski Jr. 1982. Mass extinctions in the marine fossil record. *Science* **215**:1501–1503.

https://doi.org/10.1017/pab.2023.33 Published online by Cambridge University Press

Renaudie, J., D. Lazarus, and P. Diver. 2020. NSB (Neptune Sandbox Berlin): an expanded and improved database of marine planktonic microfossil data and deep-sea stratigraphy. *Palaeontologia Electronica* **23**:a11.

Roche, D. G., I. Berberi, F. Dhane, F. Lauzon, S. Soeharjono, R. Dakin, and S. A. Binning. 2022. Slow improvement to the archiving quality of open datasets shared by researchers in ecology and evolution. *Proceedings of the Royal Society of London B* **289**:20212780.

Sears, J. R. L. 2011. Data sharing effect on article citation rate in paleoceanography. *AGU Fall Meeting Abstracts* **1**:1628.

Sepkoski, J. J. 1984. A kinetic model of Phanerozoic taxonomic diversity. III. Post-Paleozoic families and mass extinctions. *Paleobiology* **10**:246–267.

Sepkoski, J. J., R. K. Bambach, D. M. Raup, and J. W. Valentine. 1981. Phanerozoic marine diversity and the fossil record. *Nature* **293**:435–437.

Silveira, L. da, A. D. Barbosa, M. K. Ferreira, and S. E. Caregnato. 2020. Citação de dados científicos: scoping review. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação* **25**:01–31.

Silvello, G. 2018. Theory and practice of data citation. *Journal of the Association for Information Science and Technology* **69**:6–20.

Smith, J., N. B. Raja, T. Clements, D. Dimitrijević, E. M. Dowding, E. M. Dunne, B. M. Gee, *et al.* 2023a. Code and data: increasing the equitability of data citation in paleontology: capacity building for the big data future. Zenodo. https://doi.org/10.5281/zenodo.7881567.

Smith, J. A., M. C. Rillo, Á. T. Kocsis, M. Dornelas, D. Fastovich, H.-H. M. Huang, L. Jonkers, *et al.* 2023b. BioDeepTime: a database of biodiversity time series for modern and fossil assemblages. *Global Ecology and Biogeography* **32**:1680–1689.

Stephan, P., R. Veugelers, and J. Wang. 2017. Reviewers are blinkered by bibliometrics. *Nature* **544**:411–412.

Stern, B. M., and E. K. O'Shea. 2019. A proposal for the future of scientific publishing in the life sciences. *PLoS Biology* **17**:e3000116.

Stuart, D., G. Baynes, I. Hrynaszkiewicz, K. Allin, D. Penny, M. Lucraft, and M. Astell. 2018. Practical challenges for researchers in data sharing. https://apo.org.au/node/224476.

Suhr, B., J. Dungl, and A. Stocker. 2020. Search, reuse and sharing of research data in materials science and engineering—a qualitative interview study. *PLoS ONE* **15**:e0239216.

Tang, M., J. D. Bever, and F.-H. Yu. 2017. Open access increases citations of papers in ecology. *Ecosphere* **8**:e01887.

Tomaszewski, R. 2019. Citations to chemical databases in scholarly articles: to cite or not to cite? *Journal of Documentation* **75**:1317–1332.

Uhen, M. D., B. Allen, N. Behboudi, M. E. Clapham, E. Dunne, A. Hendy, P. A. Holroyd, *et al.* 2023. Paleobiology Database User Guide Version 1. *PaleoBios* **40**(11):1–56.

van der Zee, T., J. Anaya, and N. J. L. Brown. 2017. Statistical heartburn: an attempt to digest four pizza publications from the Cornell Food and Brand Lab. *BMC Nutrition* **3**:54.

Viglione, G. 2020. "Avalanche" of spider-paper retractions shakes behavioural-ecology community. *Nature* **578**:199–200.

Webster, M. S. 2017. *The extended specimen: emerging frontiers in collections-based ornithological research*. CRC Press, Boca Raton, Fla.

Westoby, M., D. S. Falster, and J. Schrader. 2021. Motivating data contributions via a distinct career currency. *Proceedings of the Royal Society of London B* **288**:20202830.

Wilkinson, M. D., M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, *et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**:160018.

Williams, J. W., E. C. Grimm, J. L. Blois, D. F. Charles, E. B. Davis, S. J. Goring, R. W. Graham, A. J. Smith, M. Anderson, and J. Arroyo-Cabrales. 2018. The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. *Quaternary Research* **89**:156–177.

Zhao, M., E. Yan, and K. Li. 2018. Data set mentions and citations: a content analysis of full-text publications. *Journal of the Association for Information Science and Technology* **69**:32–46.

https://doi.org/10.1017/pab.2023.33 Published online by Cambridge University Press