Ten simple rules to follow when cleaning occurrence data in palaeobiology

LEWIS A. JONES[1], CHRISTOPHER D. DEAN[1], BETHANY J. ALLEN[2], HARRIET B. DRAGE[3], JOSEPH T. FLANNERY-SUTHERLAND[4], WILLIAM GEARTY[5], ALFIO ALESSANDRO CHIARENZA[1], ERIN M. DILLON[6], BRUNA M. FARINA[7], and PEDRO L. GODOY[8]

[1]Department of Earth Sciences, University College London, Gower Street, London WC1E 6BT, UK

[2]GFZ Helmholtz Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany

[3]Institute of Earth Sciences, University of Lausanne, Lausanne, 1015 Switzerland

[4]School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT

[5]Open Source Program Office, Syracuse University, Syracuse, NY 13244, USA

[6]Smithsonian Tropical Research Institute, Balboa, Republic of Panama

[7]Department of Biology, University of Fribourg, Fribourg, 1700 Switzerland; Swiss Institute of Bioinformatics, Switzerland

[8]Department of Zoology, Institute of Biosciences, University of São Paulo, São Paulo, 05508-090, Brazil

Corresponding author: Lewis A. Jones (Lewis.Jones@ucl.ac.uk)

ABSTRACT

Large datasets of fossil occurrences, often downloaded from online community-maintained databases, are a vital resource for understanding broad-scale evolutionary patterns, such as how biodiversity has changed through time and space. Such datasets, however, are not infallible and must be 'cleaned' of inaccurate, incomplete, or duplicate data prior to analysis. Researchers must decide upon the extent, feasibility, and value of data cleaning steps to perform, but while guides are available for working with neontological occurrences, there is currently no clear procedure for palaeobiological data despite its unique attributes. Here, we outline ten rules that aim to aid the process of cleaning fossil occurrence data for downstream analysis. These rules cover the major steps involved in processing data prior to analysis, including project setup, data exploration and cleaning, and finalising and reporting work. We provide accompanying examples and a vignette covering the entire data cleaning process to demonstrate the application of each rule. We believe that these rules will serve as a useful guideline to support data cleaning and foster new standards for the palaeobiological community.

Keywords: palaeontology, fossils, biodiversity, reproducibility, data cleaning

INTRODUCTION

Large-scale fossil occurrence datasets have revolutionised our understanding of the evolution of biodiversity on Earth (e.g. Alroy et al., 2008; Alroy, 2010; Close et al., 2020a, 2020b) and enabled a diverse range of studies across palaeobiology, palaeoecology, and conservation (e.g. Powell et al., 2015; Pimiento et al., 2017; Dean et al., 2019; Jones et al., 2019; Allen et al., 2020; Mathes et al., 2021; Boag et al., 2021; Chiarenza et al., 2023). Such datasets provide information about the temporal and spatial distribution of organisms through geological time, along with associated stratigraphic, environmental and biological data (e.g. preservation, palaeoenvironmental information, trait data). Over the last 30 years, palaeobiology has seen the introduction of large-scale collaborative online databases (e.g. Neptune [Lazarus, 1994], the Paleobiology Database [Uhen et al., 2023], Neotoma [Williams et al., 2018]) of fossil occurrences where data are entered (or uploaded) by researchers from around the world with a range of goals, parameters, and collection methods. Using such databases is now commonplace within the field, with the Paleobiology Database (PBDB) and Neotoma both reporting over 500 associated official publications each at time of writing (March, 2025). The scale of these databases has moved palaeontology into the age of 'big data' (Allmon et al., 2018), allowing for the interrogation of Phanerozoic scale patterns that would have been impossible to implement previously.

Despite their value, the use of large-scale databases can be hindered by data quality issues such as variable data curation efforts (e.g. resolving and updating taxonomic opinions, updating geochronological ages), inconsistencies during data entry, general error from those inputting data, ambiguity in the original published documents, and lack of familiarity with the underlying data. Resolving these data issues at the source can be challenging; such databases can contain millions of records but only maintained by a small group of volunteers who lack the necessary resources (e.g. time, funding, or relevant expertise) to identify and resolve incorrect records at pace. These issues can be non-random and consequently lead to bias in downstream analysis (Panter et al., 2020). Unfortunately, issues related to data quality are commonplace within all large datasets (Cai and Zhu, 2015; Isaac and Pocock, 2015), and palaeobiological resources are no exception. A recent estimate based on flowering plants (~19,000 records) from the PBDB suggested at least ~6% of records could be viewed as potentially 'problematic' (Zizka et al., 2019), while another estimate based on fossil occurrences from the Hell Creek Formation suggested an error rate up to 92.6% in taxonomic data (Schroeder et al., 2022). Cleaning occurrence data is therefore critical to ensure accurate, reliable, and up-to-date data analysis. However, it is by no means a trivial task, particularly for complex datasets where values may change over time (e.g. due to updates in taxonomy or nomenclature).

Here, we offer ten simple rules as guidance to follow when cleaning fossil occurrence data in preparation for palaeobiological analysis (Fig. 1). Many of these guidelines are equally applicable for neontological occurrence data and have previously been advocated for by ecologists (e.g. Chapman, 2005; Zizka et al., 2019; Panter et al., 2020; Ribeiro et al., 2022). We expand upon these guidelines and present them within a specifically palaeobiological context. The rules are structured broadly in chronological order to aid in carrying out an individual research project, covering project setup (Rules 1–3), data exploration and cleaning (Rules 4–8), and finalising and reporting work (Rules 9–10). For each rule, we provide guidance on the value of its implementation and, where appropriate, highlight useful resources. Additionally, we demonstrate how each rule can be put into practice within the in-text boxes and in an accompanying vignette on crocodylian biogeography, available within the supplementary material and at https://tenrules.palaeoverse.org/. We hope this guidance acts as a helpful checklist for researchers to follow when cleaning their data, and highlights the extensive skill and knowledge often required to prepare datasets in preparation for palaeobiological analysis. While the rules presented here aim to be of use to the broader community, our intention is to specifically support researchers getting started with analyses using fossil occurrence data. As such, we assume no former knowledge on the subject, and start by defining fossil occurrence data and data cleaning.

WHAT IS FOSSIL OCCURRENCE DATA?

Fossil occurrence data comprise records of the presence of a particular taxon at a unique location in space and geological time. This is distinct from specimen-level data, which provides information about a specific fossil specimen. For example, if three specimens of *Tyrannosaurus rex* are present in the same geological bed at a single location, an occurrence-level dataset would record just one occurrence of *T. rex*. Typically, occurrence data will include information about the observed organisms such as detailed taxonomy (e.g. scientific name and taxonomic affiliation), location (e.g. modern and/or palaeo-geographic coordinates), geological context (e.g. bed, member, formation) and age (e.g. age, epoch, period, era, eon), and may also contain various associated metadata (e.g. references). From a user perspective, fossil occurrence data are most frequently organised as a single wide-format data table (Box 1) where each column represents a unique field and each row represents a unique occurrence record. From a user-perspective this is a common structure, but fossil occurrence data are regularly hosted in online databases as a set of relational data tables, linked through unique identifiers.

Fossil occurrence data can be sourced from a variety of online databases such as the Paleobiology Database (https://paleobiodb.org/#/) (Uhen et al., 2023), Neotoma (https://www.neotomadb.org/) (Williams et al., 2018), Triton (Fenton et al., 2021), Global Biodiversity Information System (https://www.gbif.org/), and the Geobiodiversity Database (http://geobiodiversity.com) (Fan et al., 2013). An exhaustive list of other data sources can be found in Supplementary Table 1 in Dillon et al. (2023).

Table 1: A list of terms used in this article and their respective definitions.

Term	**Definition**
Data cleaning	The process of fixing or removing incorrect, duplicate, or incomplete data present within a dataset (e.g. incomplete locality information, misspellings).
Data filtering	The process of removing data present within a dataset that is beyond the scope of the study (e.g. taxonomically, geographically, temporally, etc.).
Data imputation	The process of replacing missing values within a dataset with modelled values based on the existing observed values.
Data preparation	The process of preparing and transforming raw data so it is suitable for analysis and processing.
Duplicate data	Non-unique data records.
Data outlier	A data record value that notably deviates from other comparable data records.
Inconsistent data	Non-uniform or non-standardised data record values.
Metadata	Structured information that describes, explains, locates, or makes it easier to retrieve, use, or manage data.
Reproducibility	The ability to obtain consistent results using the same data and analyses.
Reusability	The ability to reapply data or code for purposes other than their original purpose.

WHAT IS AND IS NOT DATA CLEANING?

Data cleaning is the process of fixing or removing incorrect, duplicate, or incomplete data present within a dataset (Chapman, 2005). This process typically involves checking that essential fields like taxonomic names, location, and stratigraphic information contain accurate, consistent, and complete information. Common steps for palaeobiological datasets may involve correcting spelling errors in taxonomic names, updating ages of geological formations, or investigating and resolving occurrences suspected to contain inaccurate information. Within our definition of data cleaning, we exclude the use of filtering to remove data outside the scope of the study, whether that be temporally, spatially, environmentally, taxonomically, or by other criteria (see Table 1). For instance, if investigating the evolution of Phanerozoic terrestrial biodiversity, removing marine organisms from the occurrence dataset would constitute data filtering. However, if a fossil occurrence or taxon had been mistakenly coded as a marine organism (e.g. with crocodylomorphs) when it was in fact terrestrial, fixing this issue would constitute data cleaning (e.g. Mannion et al., 2015, 2019).

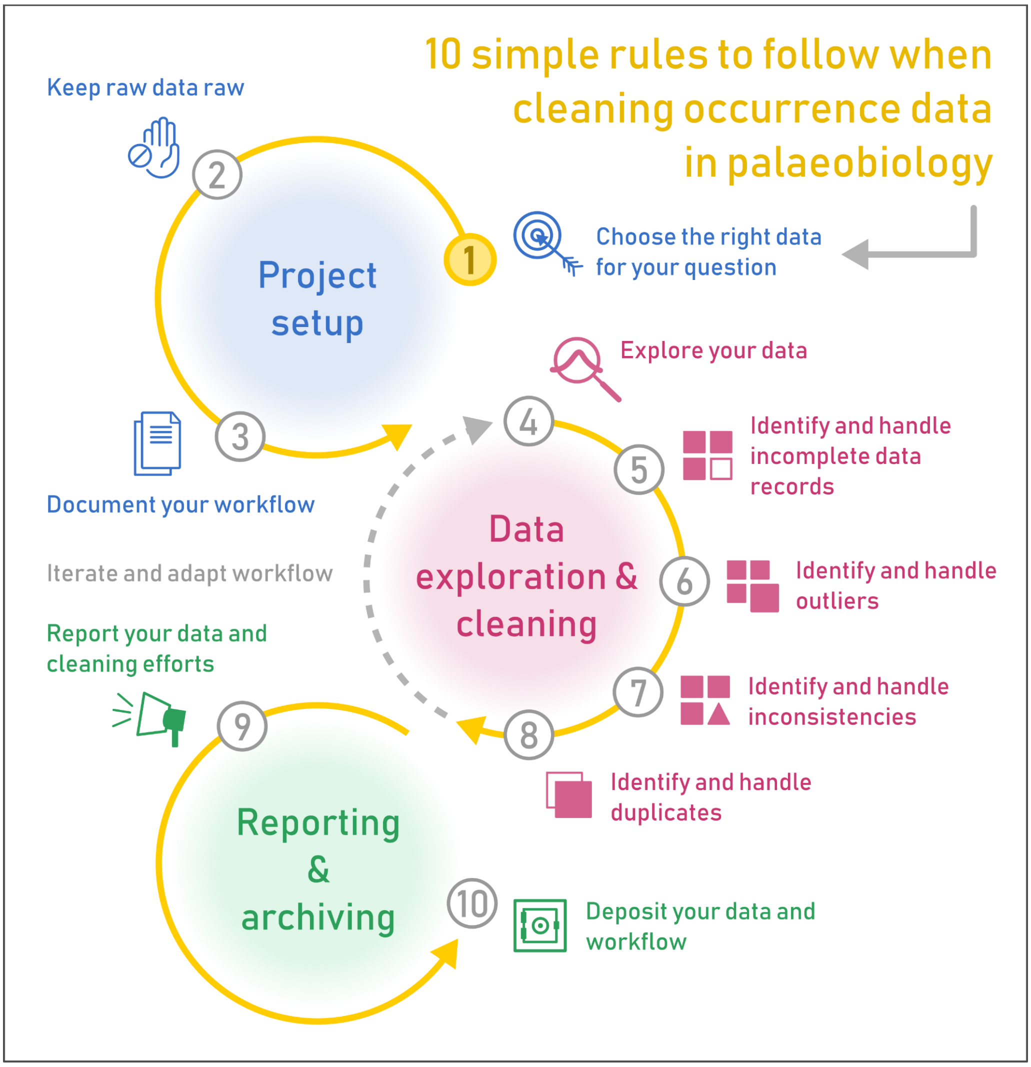
10 simple rules to follow when cleaning occurrence data in palaeobiology
Project setup
1
Choose the right data for your question
2
Keep raw data raw
3
Document your workflow
Data exploration & cleaning
4
Explore your data
5
Identify and handle incomplete data records
6
Identify and handle outliers
7
Identify and handle inconsistencies
8
Identify and handle duplicates
Iterate and adapt workflow
Reporting & archiving
9
Report your data and cleaning efforts
10
Deposit your data and workflow

Figure 1: Graphic summary of the proposed ten rules and steps to follow when cleaning occurrence data for palaeobiological analysis. The rules are grouped within their respective theme: project setup (Rules 1–3); data exploration and cleaning (Rules 4–8); and reporting and archiving (Rules 9 and 10).

RULE 1: CHOOSE THE RIGHT DATA FOR YOUR QUESTION

Selecting the right data is a crucial first step in addressing your research question. Failure to do so can lead to wasted effort in data cleaning, biased results, or misleading conclusions. The data required to address a research question depends on the scope of the study, whether it involves taxonomic diversity, biogeographic patterns, evolutionary rates, ecological reconstructions, or some other thematic area. Before gathering data, whether through fieldwork or using existing databases, researchers must determine what fields, resolution (e.g. taxonomic rank, chronostratigraphic level), and coverage (e.g. temporal, spatial, environmental) are required for their specific inquiry. During this process, researchers should consider whether flexibility related to data resolution and coverage (e.g. taxonomic, temporal, or geographic sampling) may be useful, or introduce unnecessary biases and/or analytical noise. For example, are the same macroevolutionary or ecological trends still identifiable at coarser taxonomic levels or temporal resolutions (e.g. Sepkoski, 1997; Pandolfi, 2001; Hendricks et al., 2014)? Can macroecological trends be reliably reconstructed given the available spatial sampling (e.g. Darroch et al., 2020; Jones et al., 2021; Maidment et al., 2021)? Is sufficient granularity available to determine which environments favour high diversification (e.g. Kiessling et al., 2010)? While data-specific questions are important, defining a research question can be an iterative process and can be refined to meet what data is available, rather than abandoning a project altogether. This refinement may be necessary to ensure analyses are both robust and relevant, as well as to reduce bias and increase the reliability of palaeobiological interpretations.

Many steps exist in identifying the right data to address a research question, and often vary between research questions. Nevertheless, some are shared across palaeobiological studies. The initial steps for data selection often include defining the target group (be that taxonomic, geographical, temporal, etc.) and the level of data resolution required. Including data at inappropriate resolutions can either dilute meaningful signals (if too broad) or introduce unnecessary noise (if too fine-grained), particularly if taxonomic or temporal assignments are uncertain or in flux (e.g. Paterson, 2020). For example, studies on species-specific ecological interactions or evolutionary trends require species-level data resolution (e.g. Kempf et al., 2020; Raja et al., 2021; Godbold et al., 2025), whereas broader macroevolutionary patterns may be addressed at the genus or family level (e.g. Sahney and Benton, 2008; Kiessling and Kocsis, 2015; Mannion et al., 2015; Dimitrijević et al., 2020; Drage and Pates, 2024). This can be dependent on the taxonomic group of choice; for instance, there may be insufficient occurrences identified at the species level to enable analysis at this resolution, such as commonly the case with fossil pollen (e.g. Goring et al., 2013). When considering taxonomic resolution, researchers might also assess whether their study will benefit from incorporating multiple taxonomic groups. While focusing on a single clade may allow for taxon-specific trends to be identified, integrating data from multiple lineages can provide insights into ecosystem-wide responses and provide higher data coverage (e.g. Song et al., 2020). Nevertheless, increasing taxonomic breadth should be done deliberately, as different groups may have distinct preservation biases or ecological niches, complicating direct comparisons (e.g. Fernández-Jalvo et al., 2011; Kiessling and Kocsis, 2015; Dean et al., 2019; Shaw et al., 2020, 2021). Studies conducted at wide taxonomic breadth may therefore provide a large-scale picture of the clade included, but risk averaging across the nuanced trends of the individual subclades within it.

Temporal resolution is equally important as taxonomic resolution. Overly broad temporal bins can obscure evolutionary or ecological signals, while excessively fine bins may introduce sampling noise and/or empty bins if observed fossil occurrences are sparse (Olszewski, 1999; Dean et al., 2020; Fan et al., 2020). For example, analysing faunal turnover leading up to the end-Cretaceous mass extinction within a regional setting requires well-constrained stratigraphic placements, rather than general assignments to the Late Cretaceous (Dean et al., 2020). Consequently, researchers should consider whether increasing temporal precision is truly necessary for their study or whether it will introduce more noise than clarity.

Geographic resolution and coverage should also align with the research question. A global-scale study on biodiversity change must incorporate data from diverse regions rather than being limited to well-sampled areas like North America and Europe (Vilhena and Smith, 2013). If data from key regions are unavailable due to sampling biases (e.g. poor fossil records or insufficient sampling effort), researchers should reconsider whether their question can still be adequately addressed, then explicitly acknowledge this limitation if so. This assessment should be made before cleaning data, ensuring that all necessary regions are included and that limitations are acknowledged in the study design. Failure to do so can result in global signals being obfuscated by regional trends, or highlight apparent ‘global’ trends that are actually sampling artefacts (Allison and

Briggs, 1993; Vilhena and Smith, 2013; Brusatte et al., 2015; Jablonski and Shubin, 2015; Antell et al., 2020; Close et al., 2020b; Flannery-Sutherland et al., 2022b).

If the planned study uses existing data rather than collecting new data (e.g. from a publication or online database), then selecting the right data source is a critical step. Different databases serve different purposes, and the choice depends on the research question and required resolution and coverage. The PBDB is a widely used resource for fossil occurrences, providing broad-scale taxonomic, geographic, and stratigraphic data (Uhen et al., 2023) that is best suited for large-scale palaeobiogeographic and macroevolutionary studies. The Neotoma Paleoecology Database specialises in Quaternary palaeoecological data, including pollen, vertebrates, and geochemistry, making it ideal for studies on more recent environmental changes (Williams et al., 2018). The Geobiodiversity Database (GBDB) is a taxonomic, stratigraphic, and geographic database providing occurrence, collection, and strata data within geological sections (Fan et al., 2013) that is well-suited to high-resolution temporal analyses (Fan et al., 2020). The Global Biodiversity Information Facility (GBIF) and Ocean Biodiversity Information System (OBIS) include modern and fossil occurrences/specimens, which can be leveraged to integrate information from palaeontological and neontological datasets (e.g. Kiessling et al., 2012; Lima-Ribeiro et al., 2017; Jones et al., 2019; Pilotto et al., 2021; Chiarenza et al., 2023; Hodgson et al., 2025). Many other potential data sources exist and a comprehensive list can be found in Supplementary Table 1 in Dillon et al. (2023). Finally, cross-referencing and combining data from multiple databases can be important for enhancing data reliability and completeness, although particular care is needed to ensure datasets and collection approaches are compatible, and that this does not create duplicates. Researchers should consider the full range of data sources available and their data quality, accessibility, resolution and coverage before committing to a dataset.

Box 1. Rule 1: Choose the right data for your question

Robin is starting a project looking at the palaeodiversity of crocodiles through time, assessing their biogeographic patterns during the Paleogene. They decide to download the necessary data from the Paleobiology Database, where Crocodylia are reasonably well represented for this time interval and where relevant information (e.g. taxonomic, geographic, age) are available. When downloading these data, Robin sets the time interval as "Paleogene" and the taxa to include as "Crocodylia", also making sure to only include body fossils in the download and therefore avoiding the potential for ichnotaxa or ootaxa in the dataset. As they are interested in biogeographic patterns, Robin also makes sure to include information related to geographic coordinates, such as both modern and palaeo- latitude and longitude. They also want to assess the association between Crocodylia occurrences and the number of Crocodylia-bearing geologic formations through time, so they make sure that geological information is included within the download.

Table 2: Example occurrence dataframe of "Crocodylia" fossil occurrences from the Paleobiology Database (https://paleobiodb.org/) demonstrating the structure of a wide-format dataframe.

occurrence_no	**collection_no**	**accepted_name**	**max_ma**	**min_ma**	**lng**	**lat**	**…**
40163	3113	Crocodylia	59.2	56	-74.68	39.97	…
40167	3113	Gavialoidea	59.2	56	-74.68	39.97	…
40168	3113	Gavialoidea	59.2	56	-74.68	39.97	…
…	…	…	…	…	…	…	…

RULE 2: KEEP RAW DATA RAW

Once you have identified or collected appropriate occurrence data for the desired research question, a digital copy must be obtained. This digital copy is defined as raw data and remains so if it does not undergo any form of transformation, leaving the structure and composition of its fields and records identical to the data at the point of acquisition. As such, raw data represents the information available to the researcher at that moment in time (see Box 2). Although data cleaning is likely necessary prior to analyses, it is essential to keep a raw copy alongside any cleaned data. Keeping raw data raw is crucial for two reasons. The first is to allow identification of errors inadvertently introduced during data transformation, by ensuring that the original data remains available for cross-reference. The second is to enable scientific reproducibility, by ensuring that exactly the same data that informed an analysis is available for scrutiny and reuse by future researchers.

Raw data is not necessarily primary data. For example, a fossil occurrence dataset sourced from the supplementary information of a published article, or a static data repository (e.g. Zenodo), may constitute first-hand field observations, or a compilation from previous literature (as is usually the case for large online databases). What matters here is that the raw data are new and unedited with respect to the project currently being conducted.

Upon acquisition, raw data files should be immediately stored locally in a dedicated directory using a simple, descriptive file name, and in a format that preserves its structure and integrity (Borer et al., 2009). If a dataset contains entries with non-ASCII-printable text, such as accented characters (e.g. Candelária Formation), then it may also be appropriate to ensure that the file encoding will preserve this text as accurately as possible (e.g. a .csv file with UTF-8 encoding). If compression is required to meet memory restrictions, then a lossless format should also be used to avoid degradation of the raw data (e.g. a zip folder), although this is unlikely to be an issue for fossil occurrence datasets, which are frequently less than 1 GB in size.

Manually opening raw data files should be avoided where possible; different software programs and versions may—and often do—perform automatic formatting upon opening, potentially resulting in mass data alteration (Perkel, 2019). A file may be stored in a read-only format to prevent inadvertent alteration of the raw data (Broman and Woo, 2018), with backups stored in other locations to further guard against future losses or alterations (Wilson et al., 2017). To avoid editing raw data, a researcher can perform manual edits on a working copy of the static file, or by reading the file data into a programming environment where scripted edits can be made to the temporary copy in the computer's memory using a programming language (e.g. R or Python). In the latter case, the script then also functions as a precise log of any alterations to that dataset (see Rule 3; vignette) (Borer et al., 2009).

Understandably, a researcher may wish to make small, practical alterations to the raw data itself (e.g. renaming column headers, manual correction of singular or overwhelmingly rare typographical errors) or performing simple reformatting (e.g. extraction of relevant columns or data sheets) to improve ease of downstream use. In most cases, such procedures can be scripted and manual manipulation of the raw data should still be avoided (Borer et al., 2009). If manual editing of the raw data is essential, this should be kept to the minimum possible, and a comprehensive description of these changes should be documented (e.g. as a plain text file) and kept alongside the static raw data file.

Every effort should be made to ensure that any raw data acquired for analyses remains static and accessible for future users. New data are constantly being added to online community databases (e.g. PBDB and

Neotoma), while existing entries can be revised, merged, or deleted for a range of reasons including—but by no means limited to—human error, changes in taxonomic opinion, and refined age dating. As such, online community databases are not strictly static repositories, as a future user may obtain a different dataset from that of a past user, even with identical download parameters. Some databases provide a service to archive a copy of a raw data download on request (e.g. PBDB; Uhen et al., 2023), and others automatically do so (e.g. GBIF), providing a citable unique digital object identifier (DOI). However, it should not be taken for granted that raw data being archived at the source will always be available, whether that be an online database or the supplementary files of a journal article. Raw data may become unavailable in the future due to the loss of funding and maintainers, file corruption, and journals becoming non-operational. To further guarantee the long-term availability of raw data, raw data should be archived in a suitable open-access repository whenever possible (see Rule 10).

Box 2. Rule 2: Keep raw data raw

Robin downloads the occurrence data as a '.csv' file to their computer, checking the option to "include metadata at the beginning of the output" to preserve information about the download. They then immediately copy the downloaded dataset to a separate raw data folder, and save it as 'read-only' to make sure that it can't be accidentally manipulated. The raw data file has a total of 886 occurrences.

RULE 3: DOCUMENT YOUR WORKFLOW

In almost every data-oriented project, researchers carry out some form of filtering, cleaning, formatting, or other operations to transform raw data into a workable and appropriate state for analysis (see Rules 4–8). Documenting these steps is essential to ensure transparency, reproducibility, and a clear understanding of how data have been processed (Stoudt et al., 2021). Together, these steps can be described as a 'workflow', which represents a sequence of tasks or processes that are systematically organised to achieve a specific purpose (Box 3). In a workflow, each step often depends on the previous one, and tasks are completed in a particular order to maintain efficiency, consistency, and accuracy. Workflows can be simple, involving just a few steps (e.g. restructuring of data), or complex (e.g. data cleaning and imputation), encompassing multiple transformations.

Having a clearly defined workflow can help streamline data processing steps, reduce errors, and enhance reproducibility by providing a clear, repeatable structure for completing work.

Documenting your workflow improves the transparency, reproducibility, and overall value of your research by serving as a reference or guide for repeat, follow-up, or new analyses; whether by the individual who documented the workflow, a collaborator, or any member of the research community. This can be particularly vital when going through the review process or onboarding new team members and collaborators. Documented workflows can also serve as a key avenue for transferring knowledge about data processing decisions through preserving the 'what' (i.e. what data is being transformed), 'why' (i.e. why is the data being transformed), and 'how' (i.e. how is the data being transformed).

Workflows for cleaning occurrence data in palaeobiology fall into two categories that can be used independently or in combination: (1) manual transformation (e.g. hand-typed step-by-step actions in spreadsheet software) and (2) programmatic transformation (e.g. use of automated functions or pipelines within a script of a programming language). Manual manipulation of occurrence data often takes place in spreadsheet software such as Microsoft Excel, Google Sheets, or Apple Numbers, but can also be implemented in text editors. While transforming data in such software can often be more intuitive and user friendly than through programmatic solutions (e.g. in R or Python), the process of documenting the exact steps taken when transforming raw data can be laborious and prone to a lack of clarity. Conversely, programmatic data cleaning provides a clear and traceable workflow, recording the steps taken to clean the data. Through commenting code, additional context for specific data cleaning steps can also be provided to justify decisions made (e.g. taxonomic updates, exclusion of a specific data point), or simply to guide future users. In addition, several formal workflow tools exist that can be leveraged to support data cleaning and workflow documentation (e.g. SnakeMake [Köster and Rahmann, 2012; Mölder et al., 2021] and Galaxy [Giardine et al., 2005; The Galaxy Community, 2024]). To achieve sufficient code proficiency to the extent that a fully programmatic workflow can be developed, however, is not necessarily easy or efficient, and can be a steep learning curve (Brousil et al., 2023). While we generally advocate for a code-based approach to occurrence data cleaning herein, succinctly described manual data cleaning steps can be of equal value and may even be more accessible to the broader community. For researchers with less familiarity with programmatic data transformation (e.g. regex, text parsing), resources are also available for generating a reproducible script of manual data transformation (e.g. OpenRefine). Notably, even in workflows which are entirely code-based, some elements may still require a degree of manual notation. For instance, when acquiring secondary data (e.g. downloading a dataset), it can be important to document the date of download, which may not inherently be obvious within an entirely code-based pipeline. Through the implementation of Rule 2 and Rule 10, the exact data cleaning that has taken place can be inferred through file comparison software (even with manual workflows).

Box 3. Rule 3: Document your workflow

Robin then begins to set up their project. They make a new project in RStudio, which they also link to their GitHub account to ensure that they have version control and therefore a record of all the steps taken when developing their code and assessing their data. They begin to set up their R workflow, making sure to have a clear overarching structure in their project, making use of section labels. Robin also begins to set up their manuscript file, documenting the steps taken so far in the "Methods" section. They will continue to update this with relevant information as they carry out their analysis, and will make sure to add inline comments to the R script explaining what they're doing and why.

RULE 4: EXPLORE YOUR DATA

After obtaining the raw data to address your research question and deciding how to document your workflow (see Rules 1–3), a practical next step is to explore your data. Exploratory data analysis (EDA) involves using graphical tools and basic statistical techniques to better understand the characteristics of your dataset, identify anomalies, and uncover patterns (Tukey, 1977; Quinn and Keough, 2002). This step is important for a variety of reasons. First, EDA can reveal the structure and attributes of your dataset, such as variable types and distributions, numbers of observations, and spatial or temporal dependencies between observations. Second, it can highlight relationships between variables to guide future analyses and maximise statistical insights. Third, EDA can help you select appropriate statistical tools and verify their assumptions to avoid type I (false positive) and II (false negative) errors that might lead to incorrect conclusions (Zuur et al., 2010). In doing so, EDA can illuminate aspects of your data that should be accounted for when constructing models, such as non-normality, collinearity or interactions between covariates, and spurious correlations. EDA can also flag systematic biases (e.g. taphonomic or sampling biases) that warrant careful consideration when interpreting your results. Lastly, EDA can reveal missing values (see Rule 5), outliers (see Rule 6), inconsistencies (see Rule 7), duplication (see Rule 8), and other unusual or erroneous values that require cleaning. Together, EDA is used to assess the quality and completeness of your dataset and gauge whether it can provide a meaningful and representative sample to address your research question. Without this step, you run the risk of applying inappropriate statistical techniques or making faulty inferences.

EDA is a creative and iterative process that is driven by asking questions about your dataset. As such, EDA workflows will inherently be dataset dependent. Nonetheless, the core data exploration steps often include the following: (1) creating data summaries, (2) visualising distributions of individual variables, and (3) visualising relationships between variables. These data exploration steps, together with data cleaning, will often take up the majority of the time you spend analysing your data (Zuur et al., 2010). However, starting simple and being thorough upfront can ultimately produce a more robust and insightful data analysis.

A first step when becoming familiar with your dataset is to produce descriptive summary statistics of the central tendencies and variances of groups in the data. Histograms are typically used to plot the distributions of individual variables, flag outliers, determine whether there are high numbers of zeros, and assess normality (along with QQ-plots and formal tests like Shapiro-Wilk). A combination of scatterplots, correlation matrices, box plots, ordinations (e.g. principal component analysis), and cluster analyses should then be used to visualise bivariate and multivariate relationships between variables, depending on the data types present (see Zuur et al., 2010). These graphical tools can reveal interesting patterns between variables and highlight covariates that might be important to include as predictors in more complex models. This process can also help refine the hypotheses being tested, especially given the observational nature of palaeobiological data, yet caution should be exercised to avoid circularity (Hammer and Harper, 2024). Circular reasoning can arise when the same variable is used to both define *and* test for differences between groups, such that the outcome is guaranteed by the analytical approach (Makin and Orban de Xivry, 2019). For example, you might notice during EDA that your occurrences cluster in a particular way. If you then use those clusters to filter your data and define groups (e.g. clades that either increase or decrease in richness through time), you run into issues if you then examine differences in diversity across those groups because the statistic inference is tied to your grouping criteria; it's a self-fulfilling prophecy. For more in-depth treatment of these tools, Zuur et al. (2010) outlines protocols for EDA in ecology, which can readily be adapted to palaeobiological data (see Birks et al., 2012).

Each of these steps can be scripted in R, other computer programming languages, or even in spreadsheet software, and used to create a transparent and reproducible log of the EDA workflow (see Rule 3), what was discovered, and how these initial inferences shaped the final analysis. To wrangle data and generate basic summary statistics, the *dplyr* (Wickham et al., 2023b) and *tidyr* (Wickham et al., 2024) packages (part of the tidyverse; Wickham et al., 2019) as well as *skimr* (Waring et al., 2022) are particularly helpful. These packages can be used in tandem with *palaeoverse* (Jones et al., 2023), which contains functions designed for working with fossil occurrence data such as temporal or spatial binning, range calculations, identifying unique taxa, and flagging misspellings of taxonomic names. For example, you might want to assess how many bins you have data available for. To visualise relationships between variables, *ggplot2* (Wickham, 2016), *psych* (e.g. `pairs.panels` function; Revelle, 2024), *GGally* (e.g. `ggpairs` function; Schloerke et al., 2024), *corrplot* (Wei and Simko, 2024), and *DataExplorer* (Cui, 2024) offer useful graphical functions. A multitude of online resources exist to help build competency in programming as you explore your data, including *R for Data Science* (Wickham et al., 2023a), *R Graphics Cookbook* (Chang, 2018), and Posit cheat sheets (https://posit.co/resources/cheatsheets/). Importantly, we recommend commenting code and keeping a record of EDA results and visualisations to refer back to as you develop analyses and communicate findings (see Rule 9).

Box 4. Rule 4: Explore your data

To get an idea for how their data is distributed and its various characteristics, Robin first decides to generate some basic summary statistics and plots. As they are interested in assessing palaeodiversity, Robin checks the proportions of the different taxonomic ranks in the dataset. They find that ~28% of the occurrences—about 250 in total—are assigned to the species level, and that a further ~28% are assigned to genera. Because of this, they think it might be wise to carry out palaeodiversity analysis at the rank of genus to ensure that they have enough data to find meaningful patterns. However, they will decide upon this after doing a more thorough assessment of the data. They also look at the geographic distribution of occurrences by looking at their associated country codes, finding that Paleogene crocodiles are found in a total of 46 countries.

However, after sorting these data, they find this number drops to 45 countries. Something odd has happened that they will have to investigate during future data cleaning steps.

RULE 5: IDENTIFY AND HANDLE INCOMPLETE DATA RECORDS

When exploring your dataset by carrying out EDA (see Rule 4), you may encounter ambiguous, incomplete, or missing data entries. These incomplete or missing data records can occur due to various reasons. In some cases, the data truly do not exist or cannot be estimated due to issues relating to taphonomy, collection approaches, or biases in the fossil record (e.g. information derived from highly fragmentary fossils, historical collections without associated geological or chronological information, or underrepresentation of certain taxonomic groups). In other cases, discrepancies may arise because data were collected when definitions or contexts differed, such as shifts in geopolitical boundaries and country names over time (e.g. an occurrence that only has "Czechoslovakia" listed as the country of origin cannot be precisely located today). Additionally, data may be incomplete for some records, but can be inferred through other available data (e.g. inferring country of origin through geographic coordinates). Although an intuitively common issue in palaeobiology given the uneven and incomplete nature of the fossil record (Raup, 1972; Allison and Briggs, 1993; Cherns and Wright, 2000; Vilhena and Smith, 2013; Dean et al., 2019), missing information can bias the results of palaeobiological studies (e.g. Norell and Wheeler, 2003; Kearney and Clark, 2003; Wiens, 2003; Marshall et al., 2018; Jones et al., 2021; Dean and Thompson, 2025). Occurrence data are inherently based on the existence of a particular fossil, but missing data associated with that fossil occurrence can also affect analyses that rely on that associated data (e.g. studies examining environmental associations will be impacted by missing environmental data).

Depending on your research goals and the data required to address your questions, incomplete entries may either be removed through filtering or addressed through imputation techniques. Data imputation approaches can be used to replace missing data with values modelled on the observed data using various methods (Gendre et al., 2024). These can range from simple approaches, like replacing missing values with the mean for continuous variables (e.g. morphometric measurements or associated climatic variables), to more advanced statistical or machine learning techniques (Demirtas, 2018; see Van Buuren, 2018; Haghish, 2022). If you do decide to impute missing data, it is essential that this process and its effects on the dataset are clearly justified and documented (see Rule 3) so that future users of the dataset or analytical results are aware of these decisions. Although missing data can reduce the statistical power of analyses and bias the results, imputing missing values can introduce new biases, potentially also skewing results and interpretations of the examined data (Newman, 2014). Therefore, if a dataset has sufficient data to test the desired hypotheses, or if incomplete data entries cannot be imputed reliably, these entries should be deleted in their entirety during the data cleaning process, while clearly documenting how entries were chosen for exclusion (see Rule 3). Alternatively, some data analyses allow for incomplete data entries (e.g. non-metric multidimensional scaling), and so where these methods are appropriate, you may choose to retain your incomplete data entries as-is.

To decide how to handle missing data, start by identifying the gaps in your dataset, which are often represented by empty entries or 'NA' (meaning "not available" or "not applicable"). For imputing missing values, numerous methods and tools are available in your coding language of choice, such as *missForest* (Stekhoven and Buehlmann, 2012), *mice* (Van Buuren and Groothuis-Oudshoorn, 2011), and *kNN* (Kowarik and Templ, 2016). Additionally, the R packages *TDIP* (Gendre et al., 2024) and *mlim* (Haghish, 2022) integrate various imputation and error identification methods, facilitating method comparison. Many detailed open-access references exist with which to compare the underlying methodologies of imputation approaches, and which provide guidance on the different missing data types and how to choose imputation methods and parameters (e.g. see Van Buuren, 2018).

Removing missing data can be straightforward when working with small datasets. For manual removal, tools such as spreadsheet software can be sufficient (although see Rule 3). In R, built-in functions such as complete.cases() and na.omit() quickly identify and remove missing values. The *tidyr* package also provides the drop_na() function for this purpose (Wickham et al., 2024). However, incomplete data entries can also be of use without imputation or removal; for example, the tax_unique() function from the *palaeoverse* R package (Jones et al., 2023) can flag 'cryptic diversity' that arises due to taxa not assigned to a specific species or genus, but which represent the only appearance of that clade in the geographic region or time period of choice (e.g. Mannion et al., 2011).

Box 5. Rule 5: Handling incomplete data records

Robin next begins to systematically explore their data in more detail, first making sure that the occurrences aren't missing vital information. As they are assessing biogeography, they first find any occurrences that are missing palaeocoordinates and decide to remove them from the dataset rather than trying to estimate new palaeocoordinates using available tools. After removing these data, they check to make sure that all of the occurrences have both modern and palaeo- coordinates, then decide to revisit the issue of missing data within the 'country code' field. They find that there are two occurrences which have a value of 'NA'; normally this would mean missing data, on further checking their geographic position using modern coordinates, Robin finds that they are actually from Namibia (i.e. NA!). It seems R has misconstrued these records!

RULE 6: IDENTIFY AND HANDLE OUTLIERS

Outliers, data points which lie to the extremes of the distribution of all data or otherwise deviate from comparable data points, will become readily apparent when applying EDA to your dataset (see Rule 4). Outliers may arise from a mistake in data entry, or because the value represents a genuine anomaly compared to the other available data. Identifying outliers is therefore doubly useful: it is a way of highlighting potentially suspect data for subsequent checking, and also allows us to better understand the range of values our data holds. Outliers are particularly important when an analysis investigates the maximum and minimum values of a field, or for calculations involving confidence intervals, as unusually small or large values can influence such analyses more strongly than other data points.

Most data types are amenable to some form of outlier analysis. For numerical data, this usually involves identifying the points lying at the extremes of the range of values. A simple example of this is creating a box plot, where typically the 'whiskers' are quantified based on some range of values describing the data, and any points lying outside of this range are plotted as individual outliers. Here, the choice of cut-off is very important, and many different methods exist for setting outlier cut-off points that might be applicable in different situations (Aggarwal 2017). The shape of the distribution of the data also matters. Many methods of generating confidence intervals assume that data are normally distributed, but this is often not the case for real-world biological or palaeobiological datasets, and should be borne in mind when selecting a method. For categorical data, a more appropriate method of identifying outliers might be examining abundance counts for the different categories to identify those with only a few instances. On such topics, we recommend referring to classic textbooks on statistics for (palaeo-)ecologists (e.g. Hammer & Harper 2024).

The types of data commonly present in occurrence datasets can be checked for outliers in a multitude of ways. Checking age data for outliers can be very important: if we wish to quantify the temporal or stratigraphic range of a taxon, then a misplaced data point could falsely prolong our inferred range by millions of years. This is true for both numerical (e.g. '250 Ma') and categorical (e.g. 'Triassic') forms of age data. Collecting tip or node age priors for phylogenetic inference is a common use of such data for which identifying outliers can be particularly important for downstream analyses (Mulvey *et al.* In Press). For such questions, the data resolution at which outliers are quantified should be carefully considered: for example, the age of an occurrence may appear anomalous for a specific species, but not within the context of the wider genus. This difference may alter the appropriate course of action for dealing with such data points. An example of a palaeontology-specific outlier detection method is the "Pacman" method (Lazarus et al. 2012), which uses 'known' age distributions for biostratigraphic markers to identify outliers in numerical stratigraphic data. This approach, and other relevant functions, are available in the *fossilbrush* R package (Flannery-Sutherland *et al.* 2022*b*).

Exploring data to search for taxonomic outliers can also be a helpful way of identifying mistakes. In the case that a collection of fossils is stated to contain nine species of bivalve and one species of shark, it is worth checking that the shark occurrence is correct. Otherwise, for example, it could be that the shark species actually has the same name as a bivalve species and has been miscategorised, or that the shark species is a misspelling (an example of this being the genus *Megalodon*, a bivalve from the Jurassic, being confused with *Otodus megalodon*, the giant shark from the Neogene). For multivariate data (e.g. geographic coordinates), convex hulls can be generated to identify points that form the corners of the hull, and therefore lie at the extremes of the data. The distance of these points from the rest of the data can then be quantified, with those at the greatest distance highlighted for further checking. However, it is worth considering that geographic coordinates are often subject to limits which can artificially create clumpiness in the data. At a global scale, the distribution of the continents serves as a major control on the potential spread of both species and fossil preservation, and an apparently large distance between any two data points may simply represent an area of ocean between two continents. *CoordinateCleaner* (Zizka *et al.* 2019) is an R package designed specifically for cleaning the geographic coordinates of occurrence data, including via outlier detection.

It is also possible to design downstream analytical workflows with outliers in mind, which may be particularly appropriate when it is unclear whether outliers should be removed from a dataset or not. For example, a simple strategy is to calculate and use the 90th or 95th percentile of the data instead of maximum values, or median values over mean values. More complex alternatives include bootstrapping, jackknifing, and related methods implement repeated subsampling of a dataset; this has the overall effect of amplifying the signal of common data values, and diminishing the signal of rare data values (which typically include any outliers). This can reduce the influence of outliers on the results without completely excluding these values from analysis.

Box 6. Rule 6: Identify and handle outliers

Happy that the dataset contains the information needed, Robin sets out to identify potential outliers that might affect the specific variables that relate to their research question. To do this, Robin first plots a map of where crocodiles have been found across the globe to see if any fall in places that we would not expect. They find several occurrences that appear within Antarctica, which is outside the expected climate tolerances of the group. By checking these occurrences against the associated references, it turns out that the collections associated with these anomalous occurrences appear to be legitimate, but the occurrences themselves are only listed as “Crocodylia indet.”. Robin could consider removing these occurrences due to this lack of certainty, but they would have to be consistent in their approach across the data, and make sure that a record of this is documented so that future researchers can follow their approach (see Rule 3).

RULE 7: IDENTIFY AND HANDLE INCONSISTENCIES

When carrying out EDA on your dataset (see Rule 4), it is also likely that inconsistencies will become apparent. Inconsistencies refer to deviations in the format, structure, or definitions of data values in a dataset, and they can occur in all types of variables (e.g. numerical, categorical, etc.). Inconsistencies can represent information that is definitively incorrect (e.g. a taxonomic name spelt both correctly and incorrectly in different records) but can also arise from variation of input into a dataset. This could be due to inconsistencies in standards or unclear definitions of variables (e.g. alternative, but correct, spellings of the same geological formation or different date formats being used in the same column), standards which have changed over time (e.g. a stage being given new age boundaries as a result of increased accuracy of new radiometric dates) or conflicting scientific opinions (e.g. two fossils of the same species input under different taxonomic names by researchers holding differing opinions). Although it is common for inconsistencies to apply across different rows within a single column of variables, they can also apply across multiple related columns. For example, columns for the earliest and latest ages of a fossil occurrence may have different data formats, or there could be a discrepancy between the named chronological interval for an occurrence in one column and its numerical age in a separate column. Inconsistencies may not inherently represent errors in data values, but their inclusion in a dataset can lead to a variety of downstream issues during data analysis, including skewing of summarised values, or the incorrect parsing of data by software. These issues can have serious knock-on effects for the interpretation of results, so it is essential that they are rectified prior to further data analysis. Given the variety of ways that inconsistencies can arise in a dataset, identifying them is challenging and can require high familiarity with the dataset. EDA should therefore be performed iteratively (see Rule 4) to minimise their risk of inclusion.

When searching for inconsistencies in your data, it is essential to first set definitions and standards for the data, which may be different from those associated with the original format of the dataset. This involves ensuring that you have made clear and consistent decisions on value formats, structures, and classes (e.g. are dates listed as DD-MM-YYYY or MM-DD-YYYY?), variable definitions (e.g. the column 'min_ma' is referring to the minimum possible numerical age of the fossil occurrence in millions of years; see Box 1), and the necessary precision of your values (e.g. all measurements in a column will be in centimeters rather than millimetres). When making decisions regarding the formatting of a column, it is always advisable to make edits in a copy of that column to retain the original information (see Rules 2 and 3). Similarly, adding new columns and comments that contextualise your decisions or concerns about a column's accuracy can help avoid the pitfalls of manual workflows (see Rule 3) and aid future users of your data.

Many inconsistencies will become apparent as you familiarise yourself with the spread of data within a particular column (see Rule 4). When using R, the 'table()' function can highlight the frequency of categorical values within a column, which can quickly reveal inconsistent data. Additionally, systematically checking within and between columns for formatting and spelling discrepancies will flag data values which appear problematic. Some inconsistencies may relate to facets of your data that you are less familiar with. This could result in incorrectly identifying values as inconsistencies which are actually separate data points (e.g. close taxonomic spellings, which represent different taxonomic units rather than spelling mistakes. For instance, *Varanops* is a genus of early Permian carnivorous synapsid, whereas *Varanopus* is an ichnogenus of tetrapod footprints also from the Permian), or missing inconsistencies due to a lack of knowledge (e.g. two geological formation names that have now been united under one name). In these cases, we recommend flagging potential issues and obtaining assistance from the literature or other researchers who have expertise in that particular area, rather than making decisions which may result in inaccurate data.

Because inconsistencies are inherently related to the values of the data that you are working on, the ultimate resource for resolving issues is the literature for the corresponding geographic region, taxonomic group or time period of study. Additionally, there are a variety of packages in R that can help identify potential inconsistencies in your dataset. The *fossilbrush* package (Flannery-Sutherland *et al.* 2022*b*) aims to assist with chronostratigraphic and taxonomic harmonisation within a dataset. Similarly, the 'tax_check()' function of the *palaeoverse* package (Jones *et al.* 2023) can help to check for and tally potential spelling variations of the same taxon. The previously mentioned *CoordinateCleaner* package (Zizka *et al.* 2019) is also widely used to automatically and systematically flag common spatial and temporal errors in biological and palaeobiological collection datasets in a way that is systematic, transparent and easily built into personal workflows. However, packages such as these automatically flag records based on predetermined mathematical rules and so are blind to the context of the data that they are assessing. Consequently, such approaches should be used as a complement to, rather than a replacement for, decision making by the researcher.

Box 7. Rule 7: Identify and handle inconsistencies

It's then time for Robin to do a thorough check for inconsistencies in the dataset. They check whether the class types of the fields in the dataset make sense (e.g. the 'max_ma' and 'min_ma' variables are listed as 'numeric'), and makes sure that there aren't inconsistencies between columns in the dataset (e.g. making sure that occurrences with the same value in the 'max_ma' column all have the same value for 'early_interval'). Robin then uses several automatic check functions in different R packages to flag any taxonomic or formation names that might have several different spellings. They quickly find that there are several formations which have suspiciously similar names, one obvious pair being “San Sebastián” and “San Sebastian”. After checking the literature to make sure that these are indeed the same formation, Robin corrects the spelling to ensure consistency across the dataset.

RULE 8: IDENTIFY AND HANDLE DUPLICATES

Duplicate appearances of data entries are also a common issue with occurrence datasets. The identification of duplicate fossil occurrences is an essential step in data cleaning, as neglecting them can directly impact the accuracy of analyses in a non-random way, i.e. by increasing the signal of repeated data points in the dataset (see Rules 6 and 7). There are several ways in which the same occurrence might be recorded in a dataset multiple times. The first is identical duplicates, where the exact same record appears twice or more within a dataset. This is unlikely, as occurrences within large databases are often assigned consecutive unique identifiers and by definition cannot appear twice. However, there are several circumstances where this can occur. For example, when two previously taxonomically unique occurrences are synonymised under the same taxonomic name, when merging occurrences sourced from different databases (e.g. the same fossil specimen could be independently entered into both GBIF and the PBDB), or from user error when manually manipulating a dataset (although this should be minimal if following Rules 2 and 3). A more common form of data duplication is the entry of the same fossil or collection of fossils as two separate occurrences or collections by different contributors to the database in question.

The first step for resolving duplicate occurrences in your dataset is choosing the criteria for identifying duplicates. Identical duplicates should be inherently easy to spot, as they will consist of exactly the same values across all variables (after inconsistencies have been addressed). Duplicate occurrences arising from multiple entries of the same fossil are more challenging, as user variation during data entry will mean that not all variables are likely to be identical. When this is the case, one potential way to identify duplicates is to use columns in the dataset related to the reference (e.g. original descriptive publication) from which the occurrence was acquired; though consideration of what constitutes a duplicate should be established for your specific project (e.g. if we are interested in the total number of localities, multiple references may refer to the same locality and therefore could be defined as duplicates). Multiple occurrences of the same taxon from the same reference might indicate that data duplication has taken place; checking the original reference will help resolve this. Other columns that are likely to have obvious duplicate values include those that tie a data record to a particular geographic or temporal position (e.g. two records with similar/identical geographical coordinates) (Pires *et al.* 2015; Zizka *et al.* 2020; Bonnet-Lebrun *et al.* 2023).

Once the criteria for removing duplicates are established, only one occurrence record should be retained in the processed dataset if multiple share the same taxonomy, geological age, and coordinates. It is ultimately the researcher's decision whether to exclude potential duplicates from the dataset, and the reasons for doing so should be documented (see Rules 3 and 9). However, accidental removal of non-duplicate data can also bias the results of a study, and so it is advisable to be conservative when removing entire occurrence entries. Data duplicates can be more difficult to identify if inconsistencies (see Rule 7) are present in the dataset, such as if the same taxon has an entry for two different ages or geological localities, where the age/location names have been redefined or have different regional names. This means that identification of inconsistencies and duplications (see Rule 8) should often be performed iteratively.

Identification and removal of duplicates can be done manually, but this approach has a high time-cost with large datasets, particularly when identifying them can be challenging in the first place. Alternatively, different softwares can help streamline this process. Duplicates can be removed using Excel by filtering the different columns of your dataset, though this can be too time intensive. In Python, this can be achieved using *Pandas* (McKinney 2011), a library developed specifically for data manipulation. Scripting in R offers quick and effective alternatives; unique() or distinct() from the *dplyr* package (Wickham *et al.* 2023*b*) can be used to return a dataset with any direct duplicates removed. More complex approaches, such as *CoordinateCleaner* (Zizka *et al.* 2019) and *fossilbrush* (Flannery-Sutherland *et al.* 2022*b*), can flag spatial, temporal, and taxonomic errors in occurrence data. As discussed in Rule 7 and above, thorough literature and repository searches, or external expertise on variables/groups you are less familiar with, should also be used in tandem with the above analytical approaches to resolve data duplications.

Box 8. Rule 8: Identify and handle duplicates

For the last step of data cleaning, Robin needs to remove any duplicates that might have crept into the dataset, as these could impact further analyses. Robin makes a new dataset including only the fields ‘collection_no’ and ‘accepted_name’, and then retains only the unique rows. By comparing the number of rows between this dataset and the total dataset, they find that 24 occurrences were absolute duplicates. Robin then double checks these, and removes them from the original dataset. After finishing this step, Robin now has a pretty good idea of how this dataset looks. They therefore decide to go back and re-run their initial summary statistics as well as adding some additional tests, before going back and further refining the dataset.

RULE 9: REPORT YOUR DATA AND CLEANING EFFORTS

After cleaning your data and ensuring that it is fit for purpose, it’s crucial to report on the cleaning steps you took and the overall state of your data. Reporting includes detailing how you carried out the cleaning steps (see Rules 5–8, using the workflow from Rule 3), why these were taken, the impact cleaning had on dataset composition (such as the pre- and post-cleaning occurrence counts; see Rule 4), and dataset summary statistics. Reporting these steps enables reproducibility: without knowing how the data were cleaned, it is impossible to understand the dataset in its processed form or reproduce the downstream analyses. This also increases transparency, such that other researchers will understand how and why the cleaning steps were performed, as well as the time investment on pre-analysis steps that is not otherwise well documented. Reporting on data cleaning also provides a venue for furthering acknowledgement; we can take this space to document other data sources and software (e.g. R packages) that contributed to the dataset in question before or during the cleaning process.

Reporting should involve carefully documenting at minimum: (1) how the data were chosen to be collected (see Rule 1); (2) the data exploration performed (see Rule 4); (3) how outliers, inconsistencies, and duplicates were identified, their counts, and how they were dealt with (e.g. removed, corrected, resampled; see Rules 5–8); and (4) the pre- and post-cleaning dataset summary statistics. The summary statistics should cover, for both the original raw dataset and the final cleaned dataset: the overall counts of occurrences, sampling units, or any other variables of interest; if applicable to the data, aspects like means and standard deviations or ranges of variables of interest; the degree of uncertainty regarding pertinent variables (e.g. how certain are the taxonomic assignments or stratigraphic occurrences, and to what granularity are these recorded?); the impact of any filtering (i.e. *n* occurrences were excluded by cleaning step *n*); and any imputation in the dataset. Reporting your data cleaning should be clearly documented in the methods section, in the supplementary material, or accompanying the dataset (see Rule 3).

Dataset reporting should also cover any cleaning cases specific to your data or difficulties in data processing that would be of interest to future data users or relevant specialists. This might include removing any occurrences of specific taxa due to a debate over synonymisation or higher group assignment, or removing occurrences from specific geographical regions or localities due to uncertain age assignment. For example, a study on global trilobite evolutionary trends might choose to identify and exclude entries in their occurrence dataset of families that recent assignments place within the poorly defined (i.e. 'waste-basket') order 'Ptychopariida' (by following a published taxonomic list, such as Adrain 2011). A global study on Cambrian palaeobiogeography might explain that they chose to time-bin their dataset differently because the Cambrian Stage 10 (Cohen *et al.* 2013) has an as-yet undefined base. In both examples, these data cleaning decisions require direct explanation because they are not obvious to non-specialists (or future researchers) on the taxonomic group or time period, and will have extensive impacts on the analysis results, which might influence how other researchers view or use the data or results in the future.

Several resources exist to aid the reporting process. When downloading raw occurrence data, such as from the PBDB, you can often download a supplementary reference list citing all the contributors to the data you downloaded. These should then be incorporated into publication reference lists (preferably) or supplemental references (see Smith *et al.* 2024 for discussion). If you gathered data from the primary literature, or used literature to verify potentially erroneous entries in your dataset (e.g. Rules 7 or 8), then you should compile a list of references manually or using bibliographic software (e.g. Zotero). Similarly, you can download package version citations in R or Python for those used during cleaning. Additionally, pre-formatted reporting templates exist, such as those by PRISMA (Page *et al.* 2021), which could be included in the supplementary information of an article.

Box 9. Rule 9: Report your data and cleaning

Robin now has a cleaned dataset that they use to run some analyses, and they find some results which are worthy of publication. When Robin writes up their manuscript, they make sure to report all the steps that they took to clean the data in their 'methods' section and in the associated supplementary materials, drawing attention to the decisions that they made on particular occurrences (e.g. what Robin decided to do with the 'Crocodylia indet.' specimens from Antarctica). Robin makes sure their code is clean, structured, and legible, and sufficiently commented such that it can be followed by someone who is less familiar with the approaches that they took.

RULE 10: DEPOSIT YOUR DATA AND WORKFLOW

Once you have documented and reported how you have followed Rules 1–8 (see Rule 9), it is critical that you deposit all of your data and workflow files in a reliable archival repository, preferably prior to review. This enables transparency, data accessibility, and reusability as well as research reproducibility (see Table 1) for the foreseeable future. Further, by uploading your workflow, you allow others to apply your cleaning and filtering steps to their own data, reinforcing standard practices and preventing duplicated effort. At the minimum, your archived files should include your raw data file(s) (see Rule 2) and your data processing documentation (see Rule 3). However, you should aim to archive as much of your entire research workflow as possible (see Rule 9). For example, such an archive would ideally include the scripts that you wrote to perform cleaning and filtering operations (see Rule 3) and/or analysis and visualisation of your cleaned data, including any figures in the accompanying paper (see Rule 4). It should also include modified versions of the data file created before or after manual and/or automated cleaning and filtering steps have been performed, and your reporting on how the data was changed by cleaning (see Rule 9). Finally, in addition to depositing these files (preferably in non-proprietary formats, e.g. .csv or .txt), you should also include a metadata file which describes the attributes of your various files, including their source, purpose, and, in the case of data files, column definitions (Baca 2016). In the case of occurrence data, the standards set forth and resources created by Darwin Core (https://dwc.tdwg.org/) may be useful (see https://fairsharing.org/ for other data and metadata standards). In addition to increasing the accessibility and reusability of your data, accurate and descriptive metadata is also vital for improving the discoverability of your data (Löffler *et al.* 2021).

There are different types of repositories for different purposes. The PBDB and Neotoma serve as ideal repositories for individual occurrence data, and we strongly encourage you to input new occurrence and taxonomic information in these repositories or other appropriate repositories. Nevertheless, these repositories are not intended for storing your individual project materials such as raw data files and scripts. Further, while the ever-growing and dynamic nature of these databases via community crowdsourcing is a clear benefit to our field, this is also the same reason they are inappropriate for storing static versions of your raw data; they may be edited by other users at some point in the future (see Rule 2). Therefore, you'll need to identify a separate repository for your data archive. However, navigating the data repository landscape can be challenging. For example, as of February 2025, the Registry of Research Data Repositories (https://www.re3data.org/; Pampel *et al.* 2013) lists over 2,850 open repositories available for archiving data, with over 85 of them covering 'Geology and Palaeontology'. Commonly used general repositories for occurrence data and associated files include Dryad, Zenodo, FigShare, the Open Science Framework (OSF), and Pangaea (Felden *et al.* 2023). Institutions (e.g. Yale University, University of Vienna) and national bodies (e.g. UK National Geoscience Data Centre) may also offer their own in-house data archival services. When choosing between repository options, you should consider several archival aspects, including longevity, licensing, accessibility, discoverability, citability, version control, cost, and capacity.

First, you should confirm that your chosen repository will be able to store your files for a long time (i.e. decades, at minimum). This information is often listed as 'longevity', 'persistence', or 'retention' within a repository's policies. Most repositories aim to be sustainable and last indefinitely; however, uncertainties around funding, future costs, and technological developments mean this may not hold true. Many repositories will be clear about how much funding they currently have (usually in a number of years; e.g. OSF currently states it has 50 years of funding for hosting data), with the potential for further funding in the future. If a repository does not list a longevity of decades or guarantee permanent hosting, it should probably be avoided (see Lin *et al.* 2020 for further discussion).

Next, your repository should either be clear of what copyright license your files are shared under or provide you with a selection of copyright licenses to choose from. For data, the licenses developed by the Creative Commons should be adequate, covering public domain, attribution, and non-commercial license types. In general, datasets containing only new data are usually published under the CC0 license (“No Rights Reserved”; https://creativecommons.org/public-domain/cc0/), which releases data into the public domain and makes the data easy to reuse for other projects. For example, data in the PBDB are licensed under a CC0 license (Uhen *et al.* 2023). On the other hand, data from the Neotoma database (Williams *et al.* 2018) are licensed under a CC-BY license, meaning the data must be attributed accordingly. For sharing code, there is a wider variety of licenses to choose from, with some of the most popular licenses including the MIT License, Apache License, and GNU General Public License. If you find yourself having a hard time choosing between licenses, you can find handy tools to choose a license from Creative Commons (https://creativecommons.org/choose/) and GitHub (https://choosealicense.com/).

You should also ensure that your repository will make it easy to find and cite your data archive (Wilkinson *et al.* 2016). The most common currency of academic scholarship is citation count, which is often used as one of the determining factors for hiring, promotion, and funding decisions in academia, for better or worse (Ravenscroft *et al.* 2017; Desrochers *et al.* 2018; Smith *et al.* 2024). For a long time, datasets, particularly those of occurrence data, were not citable in the same way in which we cite publications (Payne *et al.* 2012; Silvello 2018). Many repositories, such as Dryad, FigShare, and Zenodo, have introduced the automatic assignment of permanent and unique identification numbers called Digital Object Identifiers (DOIs) to archived datasets (Brown 2021). Theoretically, DOIs have brought data on par with standard publications with regards to citability (although note that other restrictions may remain such as limits to the total number of references imposed by journals [Payne *et al.* 2012] and the lack of inclusion of data citations in many common citation indices [Silvello, 2018; Smith et al., 2024]). Some repositories may not automatically assign DOIs, but may have other ways to provide unique identifiers. For example, GitHub (a common repository for software and data files) does not assign DOIs and is therefore often not a citable repository in journal publications. However, it does allow for integration with Zenodo which will archive each ‘release’ of a public GitHub repository and assign each archive a DOI. This also ensures static versioning of the respective code and data files. Similarly, OSF, which can optionally provide a DOI for a public repository, can be linked to many other storage solutions such as Amazon S3, Dropbox, and OneDrive which are not otherwise citable. In addition to citability, it is also important that the repository provides a way for other researchers to discover your data. For example, Zenodo and FigShare provide simple search interfaces to search for datasets archived with their respective services. Note that Google Scholar historically has explicitly not indexed datasets, but tools such as Google Dataset Search and Science Explorer (https://scixplorer.org/) support finding of archived datasets across the web.

Finally, hosting files costs money, and therefore most repositories have limits to the amount of storage that they provide to individual users or for individual repositories. For example, at the time of writing, free FigShare accounts can only upload up to a total of 20 GB for free, whereas Zenodo and OSF limit each free public repository to 50GB (with no account limits). Dryad similarly offers a storage limit of 50 GB per repository but at a base cost of $150 USD, though this cost can be covered by partnerships with journals or fee waivers. Most repositories will have the option to increase these quotas for a cost. For example, Dryad charges $50 USD for every 10 GB of storage above the base 50 GB, whereas FigShare offers a paid premium service that enables users to archive larger files and repositories with pricing based on the amount of storage required. Fortunately, as mentioned previously, occurrence datasets tend to be relatively small (<1 GB), so these free storage quotas should be sufficient for most occurrence data repositories.

Box 10. Rule 10: Deposit your data and workflow

When Robin submits the finished manuscript to *Palaeontology*, they make sure to upload their raw dataset, the cleaned dataset, and their R scripts to a data repository service. Robin then also makes sure to cite the dataset Digital Object Identifier (DOI) in the manuscript, drawing attention to where the data is kept. They can then sit back and wait for the (hopefully!) positive reviews on the manuscript, knowing that they have done their best to make sure that their research is accurate and easily reproducible.

CONCLUSIONS

Large fossil occurrence datasets have revolutionised the research questions that can be asked of the fossil record. However, a variety of decisions and processes must be carried out prior to conducting analyses that impact these data and subsequent conclusions, including how we set up projects (Rules 1–3), explore and clean data (Rules 4–8), and report our work (Rules 9–10). These steps can be further complicated by the specificities of palaeobiological data, particularly those collected over long time frames where collecting and reporting practices or broader geopolitical shifts may impact the quality and consistency of data being reported.

Consequently, despite data cleaning aiming to be an objective process, it is ultimately the product of researchers who will make decisions based on their professional expertise. In this article, we provide general guidelines to serve as a framework to follow for those working with and cleaning fossil occurrence data. Some of these guidelines may or may not be relevant for individual projects, and they may not always be easy to implement. However, we posit that each rule that can be followed will ultimately provide a clearer understanding of the decisions made to process a dataset prior to analysis. This is an essential step to improve the reproducibility of research; a necessary goal in the face of a broader reproducibility crisis within science (Fidler *et al.* 2017). We hope that, in following these rules, we as a community can produce datasets that not only benefit our own work in the present, but can assist future researchers for many years to come by providing clear and consistent explanations for how we have carried out our work.

DATA ACCESSIBILITY

The data and code generated for this article have been included within a dedicated GitHub repository: https://github.com/palaeoverse/ten-rules. In addition, they have been uploaded to a Zenodo repository through integrated version control: https://doi.org/10.5281/zenodo.14938533.

AUTHORS' CONTRIBUTIONS

L.A.J. conceived the project; all authors contributed to the development of the project; all authors contributed to the writing of the manuscript; C.D.D., H.B.D., and L.A.J. edited the manuscript with contributions from all authors; B.M.F., J.S., and P.G. produced the manuscript figure; A.A.C., B.J.A, E.M.D., and W.G. produced the vignette. All authors approved the final version of the manuscript.

COMPETING INTERESTS

We declare we have no competing interests.

ACKNOWLEDGEMENTS

We are grateful to all our funders for providing the support needed to realise this work. The contributions of L.A.J. were supported by a UK Natural Environment Research Council (NERC) Independent Research

Fellowship (UKRI185) and a Norman Newell Early Career Grant from the Paleontological Society. C.D.D. was supported by a Royal Society grant (RF_ERE_210013, awarded to Philip D. Mannion). B.J.A. was funded by Deutsche Forschungsgemeinschaft Collaborative Research Centre 1211 (grant no. 268236062). H.B.D. was funded under a Swiss National Science Foundation Sinergia grant (198691). A.A.C was supported by a Royal Society Newton International Fellowship (NIF\R1\231802). E.M.D. was supported by an Earl S. Tupper Postdoctoral Fellowship from the Smithsonian Tropical Research Institute. B.M.F. was funded by the Swiss Government Excellence Scholarship (2021.0350). W.G. was funded as part of a grant (#G2023-20946) from the Alfred P. Sloan Foundation. P.L.G. was funded by the São Paulo Research Foundation (FAPESP grant no. 2024/09825-7). This is Paleobiology Database official publication **[TO BE GENERATED UPON ACCEPTANCE]**.

REFERENCES

ADRAIN, J. M. 2011. Class Trilobita Walch, 1771. In: Zhang, Z.-Q.(Ed.) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. *Zootaxa*, **3148**, 104–109.

AGGARWAL, C. C. 2017. *Outlier Analysis*. Springer.

ALLEN, B. J., WIGNALL, P. B., HILL, D. J., SAUPE, E. E. and DUNHILL, A. M. 2020. The latitudinal diversity gradient of tetrapods across the Permo-Triassic mass extinction and recovery interval. *Proceedings of the Royal Society B: Biological Sciences*, **287**, 20201125.

ALLISON, P. A. and BRIGGS, D. E. G. 1993. Paleolatitudinal sampling bias, Phanerozoic species diversity, and the end-Permian extinction. *Geology*, **21**, 65–68.

ALLMON, W. D., DIETL, G. P., HENDRICKS, J. R. and ROSS, R. M. 2018. Bridging the two fossil records: Paleontology's "big data" future resides in museum collections. .

ALROY, J. 2010. The Shifting Balance of Diversity Among Major Marine Animal Groups. *Science*, **329**, 1191–1194.

———, ABERHAN, M., BOTTJER, D. J., FOOTE, M., FÜRSICH, F. T., HARRIES, P. J., HENDY, A. J. W., HOLLAND, S. M., IVANY, L. C., KIESSLING, W., KOSNIK, M. A., MARSHALL, C. R., MCGOWAN, A. J., MILLER, A. I., OLSZEWSKI, T. D., PATZKOWSKY, M. E., PETERS, S. E., VILLIER, L., WAGNER, P. J., BONUSO, N., BORKOW, P. S., BRENNEIS, B., CLAPHAM, M. E., FALL, L. M., FERGUSON, C. A., HANSON, V. L., KRUG, A. Z., LAYOU, K. M., LECKEY, E. H., NÜRNBERG, S., POWERS, C. M., SESSA, J. A., SIMPSON, C., TOMAŠOVÝCH, A. and VISAGGI, C. C. 2008. Phanerozoic Trends in the Global Diversity of Marine Invertebrates. *Science*, **321**, 97–100.

ANTELL, G. S., KIESSLING, W., ABERHAN, M. and SAUPE, E. E. 2020. Marine Biodiversity and Geographic Distributions Are Independent on Large Scales. *Current Biology*, **30**, 115-121.e5.

BACA, M. 2016. *Introduction to metadata*. Getty Publications.

BIRKS, H. J. B., LOTTER, A. F., JUGGINS, S. and SMOL, J. P. 2012. *Tracking environmental change using lake sediments: data handling and numerical techniques*. Vol. 5. Springer Science & Business Media.

BOAG, T. H., GEARTY, W. and STOCKEY, R. G. 2021. Metabolic tradeoffs control biodiversity gradients through geological time. *Current Biology*, **31**, 2906-2913.e3.

BONNET-LEBRUN, A.-S., SWEETLOVE, M., GRIFFITHS, H. J., SUMNER, M., PROVOOST, P., RAYMOND, B., ROPERT-COUDERT, Y. and VAN DE PUTTE, A. P. 2023. Opportunities and limitations of large open biodiversity occurrence databases in the context of a Marine Ecosystem Assessment of the Southern Ocean. *Frontiers in Marine Science*, **10**.

BORER, E. T., SEABLOOM, E. W., JONES, M. B. and SCHILDHAUER, M. 2009. Some Simple Guidelines for Effective Data Management. *The Bulletin of the Ecological Society of America*, **90**, 205–214.

BROMAN, K. W. and WOO, K. H. 2018. Data Organization in Spreadsheets. *The American Statistician*, **72**, 2–10.

BROUSIL, M. R., FILAZZOLA, A., MEYER, M. F., SHARMA, S. and HAMPTON, S. E. 2023. Improving ecological data science with workflow management software. *Methods in Ecology and Evolution*, **14**, 1381–1388.

BROWN, R. F. 2021. The Importance of Data Citation. *BioScience*, **71**, 211.

BRUSATTE, S. L., BUTLER, R. J., BARRETT, P. M., CARRANO, M. T., EVANS, D. C., LLOYD, G. T., MANNION, P. D., NORELL, M. A., PEPPE, D. J., UPCHURCH, P. and WILLIAMSON, T. E. 2015. The extinction of the dinosaurs. *Biological Reviews*, **90**, 628–642.

CAI, L. and ZHU, Y. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, **14**.

CHAMBERLAIN, S., SZOECS, E., FOSTER, Z., ARENDSEE, Z., BOETTIGER, C., RAM, K., BARTOMEUS, I., BAUMGARTNER, J., O'DONNELL, J., OKSANEN, J., TZOVARAS, B. G., MARCHAND, P., TRAN, V., SALMON, M., LI, G. and GRENIÉ, M. 2020. *Taxize: Taxonomic information from around the web*. .

CHANG, W. 2018. *R Graphics Cookbook: Practical Recipes for Visualizing Data*. O'Reilly Media, Inc.

CHAPMAN, A. D. 2005. *Principles and methods of data cleaning*. Global Biodiversity Information Facility.

CHERNS, L. and WRIGHT, V. P. 2000. Missing molluscs as evidence of large-scale, early skeletal aragonite dissolution in a Silurian sea. *Geology*, **28**, 791–794.

CHIARENZA, A. A., WATERSON, A. M., SCHMIDT, D. N., VALDES, P. J., YESSON, C., HOLROYD, P. A., COLLINSON, M. E., FARNSWORTH, A., NICHOLSON, D. B., VARELA, S. and BARRETT, P. M. 2023. 100 million years of turtle paleoniche dynamics enable the prediction of latitudinal range shifts in a warming world. *Current Biology*, **33**, 109-121.e3.

CLOSE, R. A., BENSON, R. B. J., SAUPE, E. E., CLAPHAM, M. E. and BUTLER, R. J. 2020*a*. The spatial structure of Phanerozoic marine animal diversity. *Science*, **368**, 420–424.

CLOSE, R. A., BENSON, R. B. J., ALROY, J., CARRANO, M. T., CLEARY, T. J., DUNNE, E. M., MANNION, P. D., UHEN, M. D. and BUTLER, R. J. 2020*b*. The apparent exponential radiation of Phanerozoic land vertebrates is an artefact of spatial sampling biases. *Proceedings of the Royal Society B: Biological Sciences*, **287**, 1–10.

COHEN, K. M., FINNEY, S. C., GIBBARD, P. L. and FAN, J.-X. 2013. The ICS international chronostratigraphic chart. *Episodes*, **36**, 199–204.

CUI, B. 2024. *DataExplorer: Automate data exploration and treatment*. .

DARROCH, S. A. F., CASEY, M. M., ANTELL, G. T., SWEENEY, A. and SAUPE, E. E. 2020. High Preservation Potential of Paleogeographic Range Size Distributions in Deep Time. *The American Naturalist*, **196**, 454–471.

DEAN, C. D. and THOMPSON, J. R. 2025. Museum 'dark data' show variable impacts on deep-time biogeographic and evolutionary history. *Proceedings of the Royal Society B: Biological Sciences*, **292**, 20242481.

———, CHIARENZA, A. A. and MAIDMENT, S. C. R. 2020. Formation binning: a new method for increased temporal resolution in regional studies, applied to the Late Cretaceous dinosaur fossil record of North America. *Palaeontology*.

DEAN, C. D., ALLISON, P. A., HAMPSON, G. J. and HILL, J. 2019. Aragonite bias exhibits systematic spatial variation in the Late Cretaceous Western Interior Seaway, North America. *Paleobiology*, **45**, 571–597.

DEMIRTAS, H. 2018. Flexible imputation of missing data. *Journal of statistical software*, **85**, 1–5.

DESROCHERS, N., PAUL-HUS, A., HAUSTEIN, S., COSTAS, R., MONGEON, P., QUAN-HAASE, A., BOWMAN, T. D., PECOSKIE, J., TSOU, A. and LARIVIÈRE, V. 2018. Authorship, citations, acknowledgments and visibility in social media: Symbolic capital in the multifaceted reward system of science. *Social Science Information*, **57**, 223–248.

DILLON, E. M., DUNNE, E. M., WOMACK, T. M., KOUVARI, M., LARINA, E., CLAYTOR, J. R., IVKIĆ, A., JUHN, M., CARMONA, P. S. M., ROBSON, S. V., SAHA, A., VILLAFAÑA, J. A. and ZILL, M. E. 2023. Challenges and directions in analytical paleobiology. *Paleobiology*, **49**, 377–393.

DIMITRIJEVIĆ, D., RAJA SCHOOB, N. and KIESSLING, W. 2020. Corallite sizes and their link to extinction risk of scleractinian corals across the triassic-jurassic boundary. .

DRAGE, H. B. and PATES, S. 2024. Distinct causes underlie double-peaked trilobite morphological disparity in cephalic shape. *Communications Biology*, **7**, 1–18.

FAN, J., CHEN, Q., HOU, X., MILLER, A. I., MELCHIN, M. J., SHEN, S., WU, S., GOLDMAN, D., MITCHELL, C. E., YANG, Q., ZHANG, Y., ZHAN, R., WANG, J., LENG, Q., ZHANG, H. and ZHANG, L. 2013. Geobiodiversity Database: a comprehensive section-based integration of stratigraphic and paleontological data. *Newsletters on Stratigraphy*, 111–136.

FAN, J., SHEN, S., ERWIN, D. H., SADLER, P. M., MACLEOD, N., CHENG, Q., HOU, X., YANG, J., WANG, X., WANG, Y., ZHANG, H., CHEN, X., LI, G., ZHANG, Y., SHI, Y., YUAN, D., CHEN, Q., ZHANG, L., LI, C. and ZHAO, Y. 2020. A high-resolution summary of Cambrian to Early Triassic marine invertebrate biodiversity. *Science*, **367**, 272–277.

FELDEN, J., MÖLLER, L., SCHINDLER, U., HUBER, R., SCHUMACHER, S., KOPPE, R., DIEPENBROEK, M. and GLÖCKNER, F. O. 2023. PANGAEA - Data Publisher for Earth & Environmental Science. *Scientific Data*, **10**, 347.

FENTON, I. S., WOODHOUSE, A., AZE, T., LAZARUS, D., RENAUDIE, J., DUNHILL, A. M., YOUNG, J. R. and SAUPE, E. E. 2021. Triton, a new species-level database of Cenozoic planktonic foraminiferal occurrences. *Scientific Data*, **8**, 160.

FERNÁNDEZ-JALVO, Y., SCOTT, L. and ANDREWS, P. 2011. Taphonomy in palaeoecological interpretations. *Quaternary Science Reviews*, **30**, 1296–1302.

FIDLER, F., CHEE, Y. E., WINTLE, B. C., BURGMAN, M. A., MCCARTHY, M. A. and GORDON, A. 2017. Metaresearch for Evaluating Reproducibility in Ecology and Evolution. *BioScience*, **67**, 282–289.

FLANNERY-SUTHERLAND, J. T., SILVESTRO, D. and BENTON, M. J. 2022*a*. Global diversity dynamics in the fossil record are regionally heterogeneous. *Nature Communications*, **13**, 2751.

FLANNERY-SUTHERLAND, J. T., RAJA, N. B., KOCSIS, Á. T. and KIESSLING, W. 2022*b*. fossilbrush: An R package for automated detection and resolution of anomalies in palaeontological occurrence data. *Methods in Ecology and Evolution*, **2022**, 2404–2418.

GENDRE, M., HAUFFE, T., PIMIENTO, C. and SILVESTRO, D. 2024. Benchmarking imputation methods for categorical biological data. *Methods in Ecology and Evolution*, **15**, 1624–1638.

GIARDINE, B., RIEMER, C., HARDISON, R. C., BURHANS, R., ELNITSKI, L., SHAH, P., ZHANG, Y., BLANKENBERG, D., ALBERT, I., TAYLOR, J., MILLER, W., KENT, W. J. and NEKRUTENKO, A. 2005. Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, **15**, 1451–1455.

GODBOLD, A., JAMES, C. C., KIESSLING, W., HOHMANN, N., JAROCHOWSKA, E., CORSETTI, F. A. and BOTTJER, D. J. 2025. Ancient frameworks as modern templates: exploring reef rubble consolidation in an ancient reef system. *Proceedings of the Royal Society B: Biological Sciences*, **292**, 20242123.

GORING, S., LACOURSE, T., PELLATT, M. G. and MATHEWES, R. W. 2013. Pollen assemblage richness does not reflect regional plant species richness: a cautionary tale. *Journal of Ecology*, **101**, 1137–1145.

HAGHISH, E. F. 2022. *Mlim: Single and multiple imputation with automated machine learning*. .

HAMMER, Ø. and HARPER, D. A. 2024. *Paleontological data analysis*. John Wiley & Sons.

HENDRICKS, J. R., SAUPE, E. E., MYERS, C. E., HERMSEN, E. J. and ALLMON, W. D. 2014. The

Generification of the Fossil Record. *Paleobiology*, **40**, 511–528.

HODGSON, E., MCCOY, J., WEBBER, K., NUÑEZ OTAÑO, N., O'KEEFE, J. and POUND, M. 2025. A global dataset of fossil fungi records from the Cenozoic. *Scientific Data*, **12**, 316.

ISAAC, N. J. B. and POCOCK, M. J. O. 2015. Bias and information in biological records. *Biological Journal of the Linnean Society*, **115**, 522–531.

JABLONSKI, D. and SHUBIN, N. H. 2015. The future of the fossil record: Paleontology in the 21st century. *Proceedings of the National Academy of Sciences*, **112**, 4852–4858.

JONES, L. A., DEAN, C. D., MANNION, P. D., FARNSWORTH, A. and ALLISON, P. A. 2021. Spatial sampling heterogeneity limits the detectability of deep time latitudinal biodiversity gradients. *Proceedings of the Royal Society B: Biological Sciences*, **288**, 20202762.

———, MANNION, P. D., FARNSWORTH, A., VALDES, P. J., KELLAND, S.-J. and ALLISON, P. A. 2019. Coupling of palaeontological and neontological reef coral data improves forecasts of biodiversity responses under global climatic change. *Royal Society Open Science*, **6**, 1–13.

JONES, L. A., GEARTY, W., ALLEN, B. J., EICHENSEER, K., DEAN, C. D., GALVÁN, S., KOUVARI, M., GODOY, P. L., NICHOLL, C., BUFFAN, L., FLANNERY-SUTHERLAND, J. T., DILLON, E. M. and CHIARENZA, A. A. 2023. palaeoverse: A community-driven R package to support palaeobiological analysis. *Methods in Ecology and Evolution*, 1–11.

KEARNEY, M. and CLARK, J. M. 2003. Problems due to missing data in phylogenetic analyses including fossils: a critical review. *Journal of Vertebrate Paleontology*, **23**, 263–274.

KEMPF, H. L., CASTRO, I. O., DINEEN, A. A., TYLER, C. L. and ROOPNARINE, P. D. 2020. Comparisons of Late Ordovician ecosystem dynamics before and after the Richmondian invasion reveal consequences of invasive species in benthic marine paleocommunities. *Paleobiology*, **46**, 320–336.

KIESSLING, W. and KOCSIS, Á. T. 2015. Biodiversity dynamics and environmental occupancy of fossil azooxanthellate and zooxanthellate scleractinian corals. *Paleobiology*, **41**, 402–414.

———, SIMPSON, C. and FOOTE, M. 2010. Reefs as cradles of evolution and sources of biodiversity in the Phanerozoic. *Science (New York, N.Y.)*, **327**, 196–198.

———, ———, BECK, B., MEWIS, H. and PANDOLFI, J. M. 2012. Equatorial decline of reef corals during the last Pleistocene interglacial. *Proceedings of the National Academy of Sciences*, **109**, 21378–21383.

KÖSTER, J. and RAHMANN, S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics (Oxford, England)*, **28**, 2520–2522.

KOWARIK, A. and TEMPL, M. 2016. Imputation with the R Package VIM. *Journal of Statistical Software*, **74**, 1–16.

LAZARUS, D. 1994. Neptune: a marine micropaleontology database. *Mathematical Geology*, **26**, 817–832.

LAZARUS, D., WEINKAUF, M. and DIVER, P. 2012. Pacman profiling: a simple procedure to identify stratigraphic outliers in high-density deep-sea microfossil data. *Paleobiology*, **38**, 144–161.

LIMA-RIBEIRO, M. S., MORENO, A. K. M., TERRIBILE, L. C., CATEN, C. T., LOYOLA, R., RANGEL, T. F. and DINIZ-FILHO, J. A. F. 2017. Fossil record improves biodiversity risk assessment under future climate change scenarios. *Diversity and Distributions*, **23**, 922–933.

LIN, D., CRABTREE, J., DILLO, I., DOWNS, R. R., EDMUNDS, R., GIARETTA, D., DE GIUSTI, M., L'HOURS, H., HUGO, W., JENKYNS, R., KHODIYAR, V., MARTONE, M. E., MOKRANE, M., NAVALE, V., PETTERS, J., SIERMAN, B., SOKOLOVA, D. V., STOCKHAUSE, M. and WESTBROOK, J. 2020. The TRUST Principles for digital repositories. *Scientific Data*, **7**, 144.

LÖFFLER, F., WESP, V., KÖNIG-RIES, B. and KLAN, F. 2021. Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs? *PLOS ONE*, **16**, e0246099.

MAIDMENT, S. C. R., DEAN, C. D., MANSERGH, R. I. and BUTLER, R. J. 2021. Deep-time biodiversity patterns and the dinosaurian fossil record of the Late Cretaceous Western Interior, North America. *Proceedings of the Royal Society B: Biological Sciences*, **288**, 20210692.

MAKIN, T. R. and ORBAN DE XIVRY, J.-J. 2019. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife*, **8**, e48175.

MANNION, P. D., UPCHURCH, P., CARRANO, M. T. and BARRETT, P. M. 2011. Testing the effect of the rock record on diversity: a multidisciplinary approach to elucidating the generic richness of sauropodomorph dinosaurs through time. *Biological reviews of the Cambridge Philosophical Society*, **86**.

MANNION, P. D., CHIARENZA, A. A., GODOY, P. L. and CHEAH, Y. N. 2019. Spatiotemporal sampling patterns in the 230 million year fossil record of terrestrial crocodylomorphs and their impact on diversity. *Palaeontology*, **62**, 615–637.

———, BENSON, R. B. J., CARRANO, M. T., TENNANT, J. P., JUDD, J. and BUTLER, R. J. 2015. Climate constrains the evolutionary history and biodiversity of crocodylians. *Nature Communications*, **6**, 1–9.

MARSHALL, C. R., FINNEGAN, S., CLITES, E. C., HOLROYD, P. A., BONUSO, N., CORTEZ, C., DAVIS, E., DIETL, G. P., DRUCKENMILLER, P. S., ENG, R. C., GARCIA, C., ESTES-SMARGIASSI, K., HENDY, A., HOLLIS, K. A., LITTLE, H., NESBITT, E. A., ROOPNARINE, P., SKIBINSKI, L., VENDETTI, J. and WHITE, L. D. 2018. Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. *Biology Letters*, **14**, 1–14.

MATHES, G. H., DIJK, J. van, KIESSLING, W. and STEINBAUER, M. J. 2021. Extinction risk controlled by interaction of long-term and short-term climate change. *Nature Ecology & Evolution*, 1–7.

MCKINNEY, W. 2011. pandas: a foundational Python library for data analysis and statistics. *Python for high performance and scientific computing*, **14**, 1–9.

MÖLDER, F., JABLONSKI, K. P., LETCHER, B., HALL, M. B., TOMKINS-TINCH, C. H., SOCHAT, V., FORSTER, J., LEE, S., TWARDZIOK, S. O., KANITZ, A., and OTHERS. 2021. Sustainable data analysis with Snakemake. *F1000Research*, **10**, 33.

MULVEY, L. P. A., NIKOLIC, M. C., ALLEN, B. J., HEATH, T. A. and WARNOCK, R. C. M. In Press. From fossils to phylogenies: exploring the integration of paleontological data into Bayesian phylogenetic inference. *Paleobiology*.

NEWMAN, D. A. 2014. Missing data: Five practical guidelines. *Organizational research methods*, **17**, 372–411.

NORELL, M. A. and WHEELER, W. C. 2003. Missing Entry Replacement Data Analysis: A Replacement Approach to Dealing with Missing Data in Paleontological and Total Evidence Data Sets. *Journal of Vertebrate Paleontology*, **23**, 275–283.

OLSZEWSKI, T. 1999. Taking advantage of time-averaging. *Paleobiology*, **25**, 226–238.

PAGE, M. J., MCKENZIE, J. E., BOSSUYT, P. M., BOUTRON, I., HOFFMANN, T. C., MULROW, C. D., SHAMSEER, L., TETZLAFF, J. M., AKL, E. A., BRENNAN, S. E., CHOU, R., GLANVILLE, J., GRIMSHAW, J. M., HRÓBJARTSSON, A., LALU, M. M., LI, T., LODER, E. W., MAYO-WILSON, E., MCDONALD, S., MCGUINNESS, L. A., STEWART, L. A., THOMAS, J., TRICCO, A. C., WELCH, V. A., WHITING, P. and MOHER, D. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, **372**, n71.

PAMPEL, H., VIERKANT, P., SCHOLZE, F., BERTELMANN, R., KINDLING, M., KLUMP, J., GOEBELBECKER, H.-J., GUNDLACH, J., SCHIRMBACHER, P. and DIEROLF, U. 2013. Making Research Data Repositories Visible: The re3data.org Registry. *PLOS ONE*, **8**, e78080.

PANDOLFI, J. M. 2001. Numerical and taxonomic scale of analysis in paleoecological data sets: Examples from neo-tropical Pleistocene reef coral communities. *Journal of Paleontology*, **75**, 546–563.

PANTER, C. T., CLEGG, R. L., MOAT, J., BACHMAN, S. P., KLITGÅRD, B. B. and WHITE, R. L. 2020. To clean or not to clean: Cleaning open-source data improves extinction risk assessments for threatened plant species. *Conservation Science and Practice*, **2**, e311.

PATERSON, J. R. 2020. The trouble with trilobites: classification, phylogeny and the cryptogenesis problem. *Geological Magazine*, **157**, 35–46.

PAYNE, J. L., SMITH, F. A., KOWALEWSKI, M., KRAUSE, R. A., Jr., BOYER, A. G., MCCLAIN, C. R., FINNEGAN, S., NOVACK-GOTTSHALL, P. M. and SHEBLE, L. 2012. A Lack of Attribution: Closing the Citation Gap Through a Reform of Citation and Indexing Practices. *TAXON*, **61**, 1349–1351.

PEREIRA, R. C., ABREU, P. H., RODRIGUES, P. P. and FIGUEIREDO, M. A. T. 2024. Imputation of data Missing Not at Random: Artificial generation and benchmark analysis. *Expert Systems with Applications*, **249**, 123654.

PERKEL, J. M. 2019. 11 ways to avert a data-storage disaster. *Nature*, **568**, 131–132.

PILOTTO, F., DYNESIUS, M., LEMDAHL, G., BUCKLAND, P. C. and BUCKLAND, P. I. 2021. The European palaeoecological record of Swedish red-listed beetles. *Biological Conservation*, **260**, 109203.

PIMIENTO, C., GRIFFIN, J. N., CLEMENTS, C. F., SILVESTRO, D., VARELA, S., UHEN, M. D. and JARAMILLO, C. 2017. The Pliocene marine megafauna extinction and its impact on functional diversity. *Nature Ecology & Evolution*, **1**, 1100–1106.

PIRES, M. M., SILVESTRO, D. and QUENTAL, T. B. 2015. Continental faunal exchange and the asymmetrical radiation of carnivores. *Proceedings of the Royal Society B: Biological Sciences*, **282**, 20151952.

POWELL, M. G., MOORE, B. R. and SMITH, T. J. 2015. Origination, extinction, invasion, and extirpation components of the brachiopod latitudinal biodiversity gradient through the Phanerozoic Eon. *Paleobiology*, **41**, 330–341.

QUINN, G. P. and KEOUGH, M. J. 2002. *Experimental design and data analysis for biologists*. Cambridge university press.

RAJA, N. B., LAUCHSTEDT, A., PANDOLFI, J. M., BUDD, A. F., KIESSLING, W. and KIM, S. W. 2021. Morphological traits of reef corals predict extinction risk but not conservation status. 1597–1608.

RAUP, D. M. 1972. Taxonomic diversity during the Phanerozoic. *Science*, **177**, 1065–1071.

RAVENSCROFT, J., LIAKATA, M., CLARE, A. and DUMA, D. 2017. Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. *PLOS ONE*, **12**, e0173152.

REVELLE, W. 2024. *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University, Evanston, Illinois.

RIBEIRO, B. R., VELAZCO, S. J. E., GUIDONI-MARTINS, K., TESSAROLO, G., JARDIM, L., BACHMAN, S. P. and LOYOLA, R. 2022. bdc: A toolkit for standardizing, integrating and cleaning biodiversity data. *Methods in Ecology and Evolution*, **13**, 1421–1428.

RUBIN, D. B. 1976. Inference and missing data. *Biometrika*, **63**, 581–592.

SAHNEY, S. and BENTON, M. J. 2008. Recovery from the most profound mass extinction of all time. *Proceedings of the Royal Society B: Biological Sciences*, **275**, 759–765.

SCHLOERKE, B., COOK, D., LARMARANGE, J., BRIATTE, F., MARBACH, M., THOEN, E., ELBERG, A. and CROWLEY, J. 2024. *GGally: Extension to 'ggplot2'*. .

SCHROEDER, K. M., LYONS, S. K. and SMITH, F. A. 2022. Response to Comment on "The influence of juvenile dinosaurs on community structure and diversity". *Science*, **375**, eabj7383.

SEPKOSKI, J. J. 1997. Biodiversity: Past, Present, and Future. *Journal of Paleontology*, **71**, 533–539.

SHAW, J. O., BRIGGS, D. E. G. and HULL, P. M. 2020. Fossilization potential of marine assemblages and environments. *Geology*, **49**, 258–262.

———, COCO, E., WOOTTON, K., DAEMS, D., GILLREATH-BROWN, A., SWAIN, A. and DUNNE, J. A. 2021. Disentangling ecological and taphonomic signals in ancient food webs. *Paleobiology*, **47**, 385–401.

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

SMITH, J. A., RAJA, N. B., CLEMENTS, T., DIMITRIJEVIĆ, D., DOWDING, E. M., DUNNE, E. M., GEE, B. M., GODOY, P. L., LOMBARDI, E. M., MULVEY, L. P. A., NÄTSCHER, P. S., REDDIN, C. J., SHIRLEY, B., WARNOCK, R. C. M. and KOCSIS, Á. T. 2024. Increasing the equitability of data citation in paleontology: capacity building for the big data future. *Paleobiology*, **50**, 165–176.

SONG, H., HUANG, S., JIA, E., DAI, X., WIGNALL, P. B. and DUNHILL, A. M. 2020. Flat latitudinal diversity gradient caused by the Permian–Triassic mass extinction. *Proceedings of the National Academy of Sciences*, 1–6.

STEKHOVEN, D. J. and BUEHLMANN, P. 2012. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics (Oxford, England)*, **28**, 112–118.

STOUDT, S., VÁSQUEZ, V. N. and MARTINEZ, C. C. 2021. Principles for data analysis workflows. *PLOS Computational Biology*, **17**, e1008770.

THE GALAXY COMMUNITY. 2024. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Research*, **52**, W83–W94.

TUKEY, J. W. 1977. *Exploratory data analysis*. Vol. 1. Springer.

UHEN, M. D., ALLEN, B., BEHBOUDI, N., CLAPHAM, M. E., DUNNE, E., HENDY, A., HOLROYD, P. A., HOPKINS, M., MANNION, P., NOVACK-GOTTSHALL, P., PIMIENTO, C. and WAGNER, P. 2023. Paleobiology Database User Guide Version 1.0. *PaleoBios*, **40**.

VAN BUUREN, S. 2018. *Flexible imputation of missing data*. Chapman & Hall/CRC, Boca Raton,.

——— and GROOTHUIS-OUDSHOORN, K. 2011. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, **45**, 1–67.

VILHENA, D. A. and SMITH, A. B. 2013. Spatial Bias in the Marine Fossil Record. *PLoS ONE*, **8**, 1–7.

WARING, E., QUINN, M., MCNAMARA, A., ARINO DE LA RUBIA, E., ZHU, H. and ELLIS, S. 2022. *Skimr: Compact and flexible summaries of data*. .

WEI, T. and SIMKO, V. 2024. *R package ‘corrplot’: Visualization of a correlation matrix*. .

WICKHAM, H. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. *Use R!*. Springer, Houston, Texas, USA.

———, ÇETINKAYA-RUNDEL, M. and GROLEMUND, G. 2023*a*. *R for data science*. O'Reilly Media, Inc.

———, VAUGHAN, D. and GIRLICH, M. 2024. *Tidyr: Tidy messy data*. .

———, FRANÇOIS, R., HENRY, L., MÜLLER, K. and VAUGHAN, D. 2023*b*. *Dplyr: a grammar of data manipulation*. .

———, AVERICK, M., BRYAN, J., CHANG, W., MCGOWAN, L. D., FRANÇOIS, R., GROLEMUND, G., HAYES, A., HENRY, L., HESTER, J., and OTHERS. 2019. Welcome to the tidyverse. *Journal of open source software*, **4**, 1686.

WIENS, J. J. 2003. Missing Data, Incomplete Taxa, and Phylogenetic Accuracy. *Systematic Biology*, **52**, 528–538.

WILKINSON, M. D., DUMONTIER, M., AALBERSBERG, Ij. J., APPLETON, G., AXTON, M., BAAK, A., BLOMBERG, N., BOITEN, J.-W., DA SILVA SANTOS, L. B., BOURNE, P. E., BOUWMAN, J., BROOKES, A. J., CLARK, T., CROSAS, M., DILLO, I., DUMON, O., EDMUNDS, S., EVELO, C. T., FINKERS, R., GONZALEZ-BELTRAN, A., GRAY, A. J. G., GROTH, P., GOBLE, C., GRETHE, J. S., HERINGA, J., 'T HOEN, P. A. C., HOOFT, R., KUHN, T., KOK, R., KOK, J., LUSHER, S. J., MARTONE, M. E., MONS, A., PACKER, A. L., PERSSON, B., ROCCA-SERRA, P., ROOS, M., VAN SCHAIK, R., SANSONE, S.-A., SCHULTES, E., SENGSTAG, T., SLATER, T., STRAWN, G., SWERTZ, M. A., THOMPSON, M., VAN DER LEI, J., VAN MULLIGEN, E., VELTEROP, J., WAAGMEESTER, A., WITTENBURG, P., WOLSTENCROFT, K., ZHAO, J. and MONS, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018.

WILLIAMS, J. W., GRIMM, E. C., BLOIS, J. L., CHARLES, D. F., DAVIS, E. B., GORING, S. J., GRAHAM, R. W., SMITH, A. J., ANDERSON, M., ARROYO-CABRALES, J., ASHWORTH, A. C., BETANCOURT, J. L., BILLS, B. W., BOOTH, R. K., BUCKLAND, P. I., CURRY, B. B., GIESECKE, T., JACKSON, S. T., LATORRE, C., NICHOLS, J., PURDUM, T., ROTH, R. E., STRYKER, M. and TAKAHARA, H. 2018. The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. *Quaternary Research*, **89**, 156–177.

WILSON, G., BRYAN, J., CRANSTON, K., KITZES, J., NEDERBRAGT, L. and TEAL, T. K. 2017. Good enough practices in scientific computing. *PLOS Computational Biology*, **13**, e1005510.

ZIZKA, A., SILVESTRO, D., ANDERMANN, T., AZEVEDO, J., DUARTE RITTER, C., EDLER, D., FAROOQ, H., HERDEAN, A., ARIZA, M., SCHARN, R., SVANTESSON, S., WENGSTRÖM, N., ZIZKA, V. and ANTONELLI, A. 2019. CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, **10**, 744–751.

———, CARVALHO, F. A., CALVENTE, A., BAEZ-LIZARAZO, M. R., CABRAL, A., COELHO, J. F. R., COLLI-SILVA, M., FANTINATI, M. R., FERNANDES, M. F., FERREIRA-ARAÚJO, T., MOREIRA, F. G. L., SANTOS, N. M. C., SANTOS, T. A. B., SANTOS-COSTA, R. C. dos, SERRANO, F. C., SILVA, A. P. A. da, SOARES, A. de S., SOUZA, P. G. C. de, TOMAZ, E. C., VALE, V. F., VIEIRA, T. L. and ANTONELLI, A. 2020. No one-size-fits-all solution to clean GBIF. *PeerJ*, **8**, e9916.

ŽLIOBAITĖ, I. 2022. Recommender systems for fossil community distribution modelling. *Methods in Ecology and Evolution*, **13**, 1690–1706.

ZUUR, A. F., IENO, E. N. and ELPHICK, C. S. 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, **1**, 3–14.