Reusability of data with complex semantic structure

Anne Strack[1], Lukas Jonkers[1], Robert Huber[1,2], and Michal Kucera[1]

[1]MARUM - Center for Marine Environmental Sciences, University of Bremen, Bremen
[2]PANGAEA - Data Publisher for Earth & Environmental Science

July, 7, 2023

This work has been funded by the German Research Foundation (NFDI4Earth, DFG project no. 460036893, https://www.nfdi4earth.de/).

Pilot title	Reusability of data with complex semantic structure
Project Duration	01.08.22 - 31.03.23
Contributors[1]	**Lukas Jonkers:** Software (lead), Validation (equal), Writing – Original Draft Preparation (equal), Writing – Review & Editing (equal). **Anne Strack:** Software (supporting), Validation (equal), Writing – Original Draft Preparation (equal), Writing – Review & Editing (equal), **Robert Huber:** Conceptualization (supporting), Software (supporting), Writing – Original Draft Preparation (supporting), Writing – Review & Editing (supporting). **Michal Kucera:** Conceptualization (lead), Writing – Review & Editing (supporting).
DOI	10.5281/zenodo.8124211
Corresponding author	Lukas Jonkers, MARUM Universität Bremen ✉ ljonkers@marum.de

Abstract

Data on the occurrence and abundance of fossils provide invaluable insights into past climate and biodiversity change. However, lack of common taxonomic standards and associated vocabularies, limit reusability of fossil data and thus global assessments. Inconsistent and variable taxonomy are a common challenge faced in biodiversity research using species occurrence data. This pilot aimed to resolve those semantic barriers for the example of planktonic foraminifera. We designed and developed an R workflow that applies the resolved semantics on legacy data stored in PANGAEA while making use of WoRMS (World Register of Marine Species). Furthermore, we provide community guidelines for new data submissions of species abundance data to generate sustainable ways of combining legacy and new data. As the pilot is closely linked to PANGAEA, we expect that many users will benefit from our workflows and best practice solutions. Since heterogeneous data structures and inadequate ontology support are a common problem for many other geoscientific and biodiversity research communities, we hope that our approach can be transferred on different types of long-tail data.

[1] Based on the CRediT Contributor Roles Taxonomy: https://credit.niso.org/

I. Introduction

The occurrence of living organisms reflects their adaptations to environmental conditions and their fossils preserved in the geological record can therefore be used as a source of information on past environments and climate and on the reaction of past ecosystems to perturbations (Yasuhara et al., 2017). Large amounts of data on the occurrence and abundance of fossils have been collected over decades, but because of the complex semantics, reflecting the intricacies of biological nomenclature and its inconsistent application by users, the resulting treasure trove of paleoecological data is hard to mine for global analyses. This situation can be exemplified by data on the composition of sedimentary assemblages of planktonic foraminifera, prolific marine microplankton with excellent and richly studied fossil record. These data stand for only a small section of the diversity of fossil organisms, but already for this group, the amount of available data exceeds the threshold for manual curation even for data originating from a single time slice (Siccha and Kucera, 2017). Using a semi-automated pipeline to process data lodged in public repositories, Siccha and Kucera (2017) generated a globally consistent resource, which have been used to show how modern plankton ecosystems departed from their pre-industrial state (Jonkers et al., 2019) and how tropical marine diversity may decline under future warming scenarios (Yasuhara et al., 2020). These examples are just scratching the surface of the potential of the existing data: analysing global patterns of biotic response to climate change, such as rates of assemblage turnover and range expansion, changes in the structure of trophic webs, functional consequences of declining biodiversity or the origin of modern communities all require access to syntheses across many taxa and time scales.

Even though a large portion of the micropalaeontological fossil occurrence and abundance data is publicly available in highly organised data repositories, most microfossil assemblage data does not fully comply with the FAIR (findable, accessible, interoperable and reusable; Wilkinson et al., 2016) data principles. Not all raw data is findable and/or accessible and inconsistent formatting often hinders interoperability. However, reusability issues are the biggest challenge to meeting FAIR principles, because of the complexity of taxonomic data and insufficient metadata. The existence of different taxonomic schools and evolving taxonomic insights both render standardisation difficult and many semantic issues arise from the use of synonyms. As a consequence, reusing micropalaeontological data is cumbersome, even when findable, accessible and interoperable, because manual curation and expert knowledge is needed. Moreover, semantic complexity leads to confusion and archiving errors, further complicating data reusability. The same problem applies to many other types of earth science data and as a result, preparing data and ingesting them into analysis platforms (e.g., for statistics and models) consumes a large share of the required resources to analyze the data by

researchers. The lack of community consensus on vocabularies and the lack of a suitable workflow associated with archiving of Earth science data with complex semantic structure hinders effective data ingest and harmonisation that would facilitate reusability. As a result, the treasure trove of paleoecological (and other) data is growing without any improvement in ways to allow automated and objective subsequent analysis.

Within the framework of the German National Research Data Infrastructure (NFDI) we developed processing pipelines and a proposal for best practices to harmonise taxonomic data, using abundances of Quaternary planktonic foraminifera as a model. We report on common problems associated with the standardisation of taxonomic data identified in a large number of micropalaeontological datasets publicly available at PANGAEA. We further developed a workflow to increase the reusability of legacy data using an R pipeline and propose community guidelines for the archiving of new datasets. We hope that this workflow can serve as a model that can be adopted in other user communities who need to access and reuse data without incompatibility or semantic barriers.

II. Results

The results of this pilot can be divided into three main sections. In a first step, we identified the most **common problems and challenges** associated with the standardisation of taxonomic data by analysing a large number of micropalaeontological datasets that are publicly available at PANGAEA (poster The Micropalaeontological Society meeting, https://doi.org/10.5281/zenodo.8123935). A first scanning of these roughly 2,400 data files that contained planktonic foraminifera species yielded 230 different names for extant planktonic foraminifera species. Considering that only approximately 50 extant morphospecies are generally recognized (Brummer and Kučera, 2022), the need for taxonomic harmonisation is obvious. The most common taxonomic issues arise from the **use of synonyms**. In many cases, standardisation can be achieved by one-to-one and many-to one mapping of ontologies. However, one-to-many mapping may lead to ambiguity due to the lack of taxonomic consensus. Context often helps with these problems which are often caused by only some main species complexes. For planktonic foraminifera these are the *Globigerinoides ruber* & *Globigerinoides elongatus* complex, the Neogloboquadrinid complex and the *Trilobatus sacculifer* complex. To solve these complexes, we developed specific decision trees for every species complex that were later scripted into the R pipelines.

The analysis of the 2,400 data files also revealed that semantic complexity in micropalaeontological data often leads to **archiving errors**, further complicating data reusability. For example, PANGAEA already links a considerable amount of AphiaIDs to its parameter names using an automatized parameter annotation service (see also Diepenbroek et al., 2017) whose taxon recognition, however, shows a certain margin of

error and, in particular, also makes incorrect assignments for the mentioned complexes. Unclear taxonomy and grouping of taxa may lead to **duplicated names** as the same species name is being used for different (groups of) taxa. The unnecessary archiving of **grouped taxa** further increases the data complexity and may lead to meaningless species names. These issues can often be resolved by context and the data itself, but they require extra processing steps. Further archiving errors are introduced, because often **relative abundances** rather than absolute abundances are reported. Relative abundance data omit important information about the reliability of the data itself as count statistics are not available and percentages are more sensitive to errors that accumulate over time (Telford, 2019). Small errors arise from rounding issues, whereas more serious errors are often due to double counting of (grouped) taxa, which cannot always be corrected. So, the tendency to report relative abundances rather than absolute abundances and the habit to include counts of both individual and lumped species in the same data set, has led to an enormously high amount of erroneous archived data sets. Of the approximately 43,000 assemblages with relative abundances we analysed, only half of the sums of percentages added up to 100±5%.

Different solutions are needed **for legacy data** that are already published in public repositories **and new data submissions**. So, after identifying the common problems and challenges that occur with taxonomic data harmonization, we used the gained knowledge to develop a **R pipeline** to provide the community with an online tool to harmonize planktonic foraminifera taxonomy in legacy data that are already published in PANGAEA (see **Implemented Solution**).

Almost all of the issues and common problems identified in our analysis can easily be avoided with simple data archiving guidelines that need to be agreed on by the community and communicated and adhered to by data generators and data curators (e.g. PANGAEA). For this, we have compiled a preliminary proposal of **community guidelines** that may be followed for the archiving of new datasets to ensure that data is not only findable, accessible and interoperable in public repositories, but also reusable by others (community consultation in progress; white paper in preparation). The proposed recommendations for data archiving include aspects of the data itself as well as the metadata. When working with taxonomic assemblage data it is crucial to include information about the taxonomic concept that was used and to be clear about exceptions. When uploading the data to public repositories it is always advisable to keep the taxonomic harmonisation in mind. This can be done by providing information on alternative classification or linking the own data to internationally recognized classification schemes (e.g. using the AphiaIDs of the World Register of Marine Species). It is always advised to always report the full genus and species names and no abbreviations and to include subspecies when needed. State whether the full assemblages have been counted and whether “missing” taxa are assumed to be absent from the sample. Further, information about morphological variants and how they map

onto the species can also be included. For the data itself it is important to always report the data at the highest possible taxonomic resolution and to specifically avoid group or lumped taxa. It is also advised to include unidentified species to your data table and report depth rather than age, because it is fixed with the sample and not meant to change. As mentioned before, the use of relative abundances with assemblage data has many disadvantages, so we highly recommend to report absolute counts rather than percentages.

a) Implemented Solution

Considering the most common problems and challenges that we identified in our data analysis, we developed a R pipeline to harmonize planktonic foraminifera taxonomy of legacy data that are already stored in PANGAEA (Figure 1). The current version of the script relies on an external synonym dictionary and the classification scheme of the harmonized taxonomy follows WoRMS (World Register of Marine Species; www.marinespecies.org). By using WoRMS, we were able to assign an AphiaID (a persistent globally unique identifier) to every recognized extant taxon name in the legacy data. In cases where the WoRMS status of the original species name is unaccepted, we then use the AphiaID to correct the original species name to the accepted name.

The script automatically downloads the desired data set and linked metadata directly from the PANGAEA repository using the corresponding persistent identifier (DOI). In a first step the script checks if there are extant planktonic foraminifera species in the data (by comparison with a list of extant species) and, if yes, drops all columns that contain no species abundance data before data harmonization. However, it is important to note that the original data and all original metadata are preserved. Since a lot of problems arise because of the unnecessary archiving of grouped taxa, the script checks if there are any columns in the data that are the sum of two others. Though, this approach only works if the abundance data is numeric and not with non-numeric semi-abundance data (which is often generated for biostratigraphic purposes). The sum-check also includes a threshold value that can be easily adjusted by the user, because the use of relative abundances often produces rounding issues that need to be considered. Summed columns can only be deleted if the two constituent columns are present, that is when there are three columns with the same name or when *Globigerinoides ruber*, *Globigerinoides ruber subsp. albus* and *Globigerinoides ruber subsp. ruber* are present or when morphological variations of *Trilobatus sacculifer* (with and without sac) have different genus names. The sum-check also deals with cases where only a single column is not 0 (no action needed) or where there is one column that exists entirely of 0s (removes one of the remaining two columns when equal). The script further deals with a common case where one species (*Globorotalia menardii*) is often grouped with another species (*Globorotalia*

tumida) and removes the merged case when both species are individually present. Furthermore, the script checks for duplicated species columns that might be redundant and asks the user whether or not the identified column(s) can be removed. An essential part of the script is the sorting of the three main species complexes (as described above), because this is the part where most non-trivial challenges occur (see Figure 1).

If the data contains relative abundances, the script then ends with a small statistic on how many samples deviate from 100% by more than 0.5%. This statistic gives a good estimate on the reliability of the data itself as well as the data harmonization (Telford, 2019).

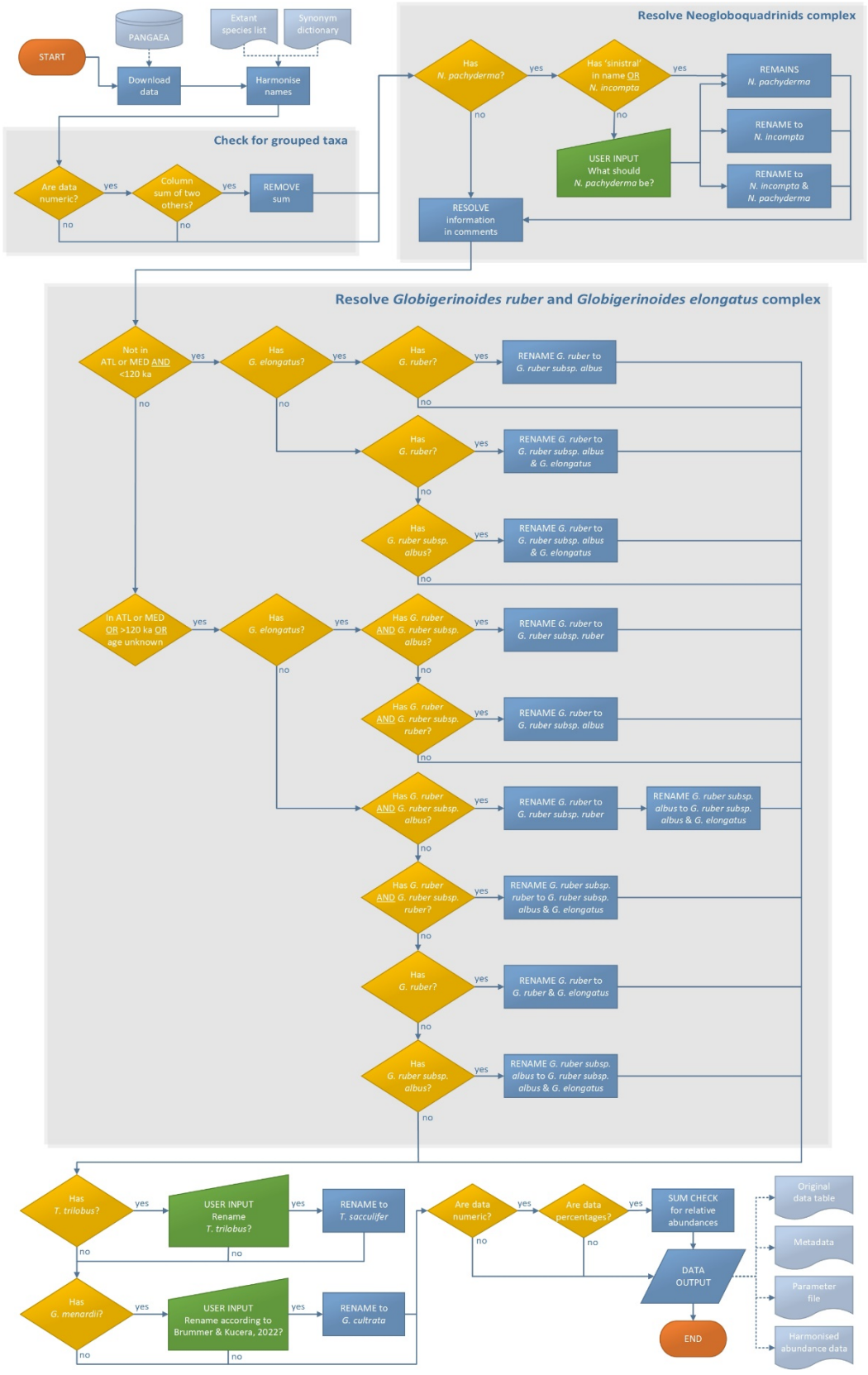

Figure 1: Simplified cross-functional flowchart sketching the functioning of the R script to harmonize planktonic foraminifera abundance data.

The final script output contains the original data table, some metadata (information on the citation, event, URL from which the data were downloaded and license of the data), a parameter file and a harmonized abundance data file. The parameter file lists all parameters of the data with their original name, a column that states whether or not the parameter is used for the harmonization and, if used, also states the harmonized parameter name including the corresponding AphiaID. The harmonized abundance data file is given in long-format and contains all parameters that are set to "included" in the parameter file.

The proposed community guidelines for new data submissions (as described above in more detail) will be written into a white paper. Even though this pilot focuses on a specific example, heterogeneous data structures and inadequate ontology support are a common problem for many other geoscientific and biodiversity research communities. Therefore, we have used our findings to start a community-consultation process to formulate micropalaeontological data standards. It is our intention to expand this community beyond planktonic foraminifera specialists, in order to make the guidelines useful for the entire micropalaeontological research community.

In addition, we have manually reassigned some previously incorrectly assigned annotations within the PANGAEA's metadata and we improved PANGAEA's parameter annotation service (https://ws.pangaea.de/param-annotator/) for which we identified e.g. difficulties to recognise species names which may include optional subgenus names and work towards an integration of RDA I-ADOPT concepts within this service (see: RDA 20 Plenary talk: https://www.rd-alliance.org/plenaries/rda-20th-plenary-meeting-gothenburg-hybrid/practical-implementations-i-adopt-framework-and).

b) Data and Software availability

The script and the used synonym list and extant species list can be accessed via GitHub (https://github.com/lukasjonkers/harmonisePFTaxonomy). Version 1 has been released on Zenodo (https://doi.org/10.5281/zenodo.8124240). The code comes along with sufficient information in the read me to apply it to datafiles with planktonic foraminifera abundance data stored at PANGAEA.

c) Innovation and FAIRness

Paleoecological data holds an invaluable treasure trove of information that could be used to analyse global patterns of, for instance, the biotic response to climate change. However, the lack of common taxonomic standards and associated vocabularies, limit the reusability of these data and hinders us from unlocking this treasure. It has been shown that even for planktonic foraminifera, which only stand for a small section of the diversity of the fossil organisms, the amount of available data already exceeds the threshold for manual curation (Siccha and Kucera, 2017). So, the need for semi-

automated pipelines to process archived data and therefore increase the value of microfossil assemblage data is immanent.

Here, we provide a solution that helps to harmonise legacy taxonomic data of planktonic foraminifera and summarise our recommendations into community guidelines for new data submissions. With this proposed set of measures, we mainly aim to improve the reusability of taxonomic assemblage data. However, since our recommendations also address the full life cycle of research data, we also hope to improve the findability, accessibility and interoperability of taxonomic data. Our proposals are also meant as a starting point for a discussion with the entire community, which hopefully further improves the FAIRness of micropalaeontological data.

III. Challenges and Gaps

Due to incomplete linking of PANGAEA data to the specialised ontology of WoRMS we could not rely completely on external ontologies for i) identifying data containing planktonic foraminifera abundance information and ii) resolving taxonomic inconsistencies. Instead, we used our specialist knowledge to query PANGAEA using keywords that likely resulted in matches with planktonic foraminifera datasets to find the relevant data at PANGAEA. We also manually curated a list of synonyms for taxa not yet linked to WoRMS to resolve taxonomic confusion. With the improvements in the PANGAEA annotation service, the use maintenance of this list will become obsolete.

Due to time constraints we were only able to start the community consultation process to develop FAIR micropalaeontological data standards. We foresee a major step forward in this process during the upcoming FORAMS conference (http://forams2022.it/) in June 2023 (https://doi.org/10.5281/zenodo.8123959). As we remain convinced that this is an essential step forward for our community we aim to finalise the white paper describing the need for standards and laying community-approved standards out towards the end of 2023.

IV. Relevance for the community and NFDI4Earth

The principal beneficiaries of this pilot are the scientific communities (re-)using long-tail data with complex semantic structure, especially micropalaeontologists. However, even though (micro)palaeontological data are particularly complex due the temporally varying pool of species names due to the evolutionary processes of speciation and extinction, our insights could in principle be applied to any kind of taxonomic data, which clearly links this project not only to the NFDI4Earth community, but also to NFDI4Biodiversity. We provide a tool to generate sustainable ways of combining legacy and new data into analysis ready formats. Furthermore, the generated standards and workflows also

benefit data curators by streamlining submission of new data and infrastructure providers by generating a model of community-driven process to resolve complex ontology.

The pilot uses a specific example to address a common problem for many other geoscientific research communities facing the challenge of heterogeneous data structures or inadequate ontology support. Therefore, it may serve as a best practise example on how to integrate the scientific community within the data archiving and publication workflow. While the focus was on the reusability aspect of FAIR data principles, the pilot addresses the full life cycle of research data, thus also improves findability, accessibility and interoperability of taxonomic data. The pilot is closely linked to the information system PANGAEA which is a certified, internationally renowned long-term archive and data publisher. It can be expected that any developments and improvements in the data workflows will engage many users and guide them as best practice solutions.

V. Future Directions

The first and most obvious further development of the processing pipeline we developed during the pilot is to make better use of external specialist ontologies such as WoRMS. Recent work at PANGAEA is aimed at completing the linking to WoRMS and once this process is finished, we can update our pipeline and make it independent of user-defined synonym lists. To increase the user-friendliness of our taxonomic harmonisation pipeline it would be desirable to better integrate it into PANGAEA. In its simplest form, this could be done by linking each micropalaeontological data set to the script on GitHub. Ideally though, the pipeline should be fully integrated with PANGAEA, offering users the possibility to harmonise taxonomic data and perform error checking with the click of a button.

This pilot focused on a time frame and species group where the species pool remained stable. Making the pipeline applicable to data from deep time, where evolutionary processes make harmonisation even more complicated as synonyms vary with time, would be a next step to increase the applicability of our tool. Since age information is often not available in the same datafile as the species abundances, it is essential to better link datasets from the same timeseries. Technically this is already possible by making use of the metadata stored in PANGAEA. In practice there remain challenges related to depth scales and the presence of multiple age-depth models for single time series. Application to deep time also requires information on the stratigraphic range of taxa, information that is currently not available at PANGAEA and which would hence needed to be obtained from external sources, such as mikrotax (www.mikrotax.org). This linking is in principle possible through the WoRMS AphiaIDs.

In our experience a lot of micropalaeontological data remains unfindable and inaccessible. This applies especially to older datasets generated prior to the digitalisation and attention to good data stewardship. To rescue this old and invisible data, more targeted community-wide efforts are needed before the generation of scientists who generated these data retires. A considerable portion of micropalaeontological data used today also remains inaccessible as only derived products are made available. Our challenge thus remains to increase the FAIRness of micropalaeontological data to fully unlock their potential in putting the current biodiversity crisis in a long-term context.

VI. References

Brummer, G.-J. A., & Kučera, M. (2022). Taxonomic review of living planktonic foraminifera. J. Micropalaeontol., 41(1), 29-74. doi:10.5194/jm-41-29-2022

Diepenbroek, M., Schindler, U., Huber, R., Pesant, S., Stocker, M., Felden, J., Buss, M., & Weinrebe, M. (2017). Terminology supported archiving and publication of environmental science data in PANGAEA. J Biotechnol, 261, 177-186. doi:10.1016/j.jbiotec.2017.07.016

Jonkers, L., Hillebrand, H., & Kucera, M. (2019). Global change drives modern plankton communities away from the pre-industrial state. Nature, 570(7761), 372-375. doi:10.1038/s41586-019-1230-3

Siccha, M., & Kucera, M. (2017). ForCenS, a curated database of planktonic foraminifera census counts in marine surface sediment samples. Sci. Data, 4, 170109. doi:10.1038/sdata.2017.109

Telford, R. J. (2019). Tools for identifying unexpectedly low microfossil count sums (Preprint). PaleorXiv. doi:10.31233/osf.io/mvgc3

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. 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., Groth, P., Goble, C., Grethe, J. S., Heringa, J., t Hoen, P. A., 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., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data, 3, 160018. doi:10.1038/sdata.2016.18

Yasuhara, M., Tittensor, D. P., Hillebrand, H., & Worm, B. (2017). Combining marine macroecology and palaeoecology in understanding biodiversity: microfossils as a model. Biol. Rev., 92(1), 199-215. doi:10.1111/brv.12223

Yasuhara, M., Wei, C. L., Kucera, M., Costello, M. J., Tittensor, D. P., Kiessling, W., Bonebrake, T. C., Tabor, C. R., Feng, R., Baselga, A., Kretschmer, K., Kusumoto, B., & Kubota, Y. (2020). Past and future decline of tropical pelagic biodiversity. Proc. Natl. Acad. Sci. USA, 117(23), 12891-12896. doi:10.1073/pnas.1916923117