Towards FAIRer micropalaeontological data

Lukas Jonkers (ljonkers@marum.de), Robert Huber, Anne Strack and Michal Kucera

Introduction

Sedimentary microfossil assemblage data can provide a crucial baseline of biodiversity prior to human influence, reveal species turnover dynamics across time scales inaccessible using direct observations and are essential for quantitative palaeoclimatology.

Many micropalaeontological studies rely on merging data from different sources. This merging requires that the data are **findable**, **accessible**, **interoperable** and **reusable**, i.e. comply with the FAIR data principles. As a community we are in the favourable position that data sharing through established repositories is common practice. However, challenges remain to make micropalaeontological data truly fair: data sets need additional information (metadata) in order to streamline reusability.

In addition, the biggest challenge to **reusability** is the **semantically complex** nature of species assemblage data**.** This is because of the existence of different **taxonomic schools** and evolving **taxonomic insights**, which both render standardisation difficult. As a consequence **reusing** micropalaeontological data is cumbersome, even **when they are findable, accessible and interoperable**. Moreover, semantic complexity leads to confusion and **archiving errors**, further hampering data reusability. Thus, to make micropalaeontological data FAIRer, we, as a community, need data standards to increase the value of our data and to make our science reproducible.

Within the framework of the German National Research Data Infrastructure (NFDI) we are developing tools and processing pipelines to harmonise taxonomic data. At the same time we are starting a community engagement process to collectively define micropalaeontological data requirements. To this end we invite you to take part in a **survey**.

As an illustration of why we need data standards, we here report on common problems associated with the standardisation of taxonomic data identified in a large number of micropalaeontological datasets publicly available at PANGAEA. Our assessment focussed on planktonic foraminifera because of the relatively simple taxonomy of this group. However, we explicitly invite **feedback** from all data generators and data users from the micropalaeontological community to discuss solutions that work for everyone and can be applied across different taxonomic groups and different research fields.

Have your say about data requirements

We have designed a survey to assess requirements for microfossil species abundance data sets. Please take the time to fill it out and join us making micropalaeontological data FAIRer.

The goal of this survey is to define a community-endorsed checklist of properties of data that are important to ensure their reusability. The survey will be distributed among the entire micropalaeontological research community.

tinyurl.com/mipastandards

The results of the survey will be disseminated through a (white) paper. At the end of the poll is an invitation to participate in the writing.

Challenges to interoperability and reusability

We have scanned ~2,300 files with planktonic foraminifera species abundance data archived on PANGAEA.

To fully make use of those data desired metadata (e.g. size fraction, methodology) was often missing and needed to be scraped from literature.

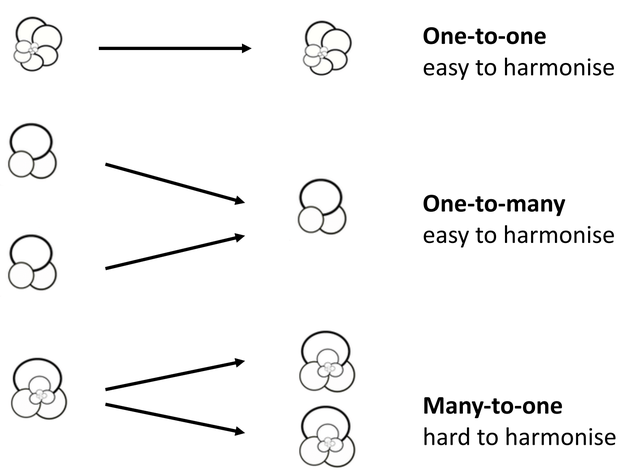
The ~2,300 files with planktonic foraminifera species data yielded **230 different names for extant species, about 180 more than are generally recognised** and thus clearly highlighting the need for harmonisation.

Most taxonomic issues occurred within a small group of species that have subspecies, variants or where the taxonomy has changed (e.g. *Neogloboquadrina pachyderma*, *Globigerinoides ruber* or *Trilobatus sacculifer*).

Virtually all datasets lacked an explanation of the taxonomic concept, a description of how variants map onto each other, or how taxa were lumped. Together, this led to ambiguity and unnecessary errors in the data.

Assessment of taxonomic completeness of the data was challenging as taxa with zero abundance and unidentified specimens were inconsistently reported.

Common taxonomic mapping issues



One-to-one
easy to harmonise

One-to-many
easy to harmonise

Many-to-one
hard to harmonise

Images courtesy Peter Kiss

Taxonomic complexity leads to archiving

Duplicated names

24	Neogloboquadrina pachyderma sinistral	N. pachyderma s	%
25	Neogloboquadrina pachyderma dextral and dutertrei integrade	P/D int	%
43	Neogloboquadrina pachyderma dextral and dutertrei integrade	P/D int	%
44	Neogloboquadrina pachyderma dextral	N. pachyderma d	%

Unclear taxonomy and grouping of taxa leads to the same name being used for different (groups of) taxa. We found at least 400 files (i.e. 17 %) with non-unique taxon names.

Context and the data itself can sometimes help to resolve these issues, but they always require avoidable extra processing steps.

Grouped taxa and compound names

4	Globigerinoides ruber pink	G. ruber p	%
5	Globigerinoides ruber white	G. ruber w	%
6	Globigerinoides ruber	G. ruber	%
7	Globigerinoides tenellus	G. tenellus	%
8	Globigerinoides sacculifer wo sac	G. sacculifer wo sac	%
9	Globigerinoides sacculifer sac	G. sacculifer sac	%
10	Globigerinoides sacculifer	G. sacculifer	%
34	Globorotalia anfracta	G. anfracta	%
35	Globorotalia menardii	G. menardii	%
36	Globorotalia tumida	G. tumida	%
37	Globorotalia menardii flexuosa	G. menardii flexuosa	%
38	Globorotalia mentum	G. mentum	%
39	Candeina nitida	C. nitida	%

Unnecessary archiving of grouped taxa increases data complexity and requires extra processing steps.

Grouping of taxa lead to meaningless taxon names (kudos for those who know the taxon *Globorotalia mentum*).

Percentages = trouble

tus [%]	4 G. ruber p [%]	5 G. ruber w [%]	6 G. ruber [%]	7 G. tenellus [%]	8 G. sacculifer wo sac [%]	9 G. sacculifer sac [%]	10 G. sacculifer [%]	11 S. dehiscens [%]	Sum
0.1	2.5	18.937	21.484	0.2	12.514	3.32	15.84	0.00	99.84
1.3	0.1	18.612	18.759	0.0	9.897	1.18	11.08	0.00	99.78
0.4	0.8	19.284	20.080	1.0	6.958	3.38	10.34	0.00	100.14
0.7	1.4	17.477	18.919	0.9	5.045	3.60	8.65	0.00	100.05
0.4	0.6	19.368	20.000	0.8	7.368	1.90	9.26	0.00	99.62
0.6	0.4	18.939	19.318	0.4	7.765	0.38	8.14	0.19	100.02
0.5	0.4	22.064	22.420	1.6	4.804	1.96	6.76	0.00	100.40

Relative abundance data omit important information about reliability as count statistics are not always available.

Percentage data are more sensitive to errors that accumulate over time and cannot always be corrected. Of the ~43,000 assemblages with relative abundances we assessed only half have sums that add up to 100±5%.

Small errors arise from rounding, more serious errors are due to double counting of taxa due to grouping. In many cases the cause of the errors is unclear.

Towards defining data requirements

Almost all of the issues and errors identified here can easily be avoided with simple data archiving guidelines that need to be communicated and adhered to by data generators and data curators (e.g. PANGAEA).

In order to define such guidelines we have prepared a survey (**tinyurl.com/mipastandards**). Participants are asked to decide for a pre-defined selection of data properties whether they are desired, recommended or essential for the reusability of the data. The survey will be distributed among the entire micropalaeontological research community, starting here at FORAMS.

This ranking of the data properties will be used to define the guidelines, which will be communicated through a publication in a peer-reviewed journal. All participants of the survey are invited to contribute to the writing.

The survey covers the following data requirements:

- Description of the dataset (e.g. goal of the analysis)
- Site information (position, gear)
- Sample preparation (sieving, staining, chemical treatment)
- Counting method (microscope, magnification)
- Taxonomic information (concept, lumping, completeness)
- Attribution (source, contributor, funder)
- Sample characteristics (size, state, dissolution)
- Data requirements (depth, thickness, counts)

Have your say, fill out the survey!

Get in touch if you want to join us making micropalaeontological data

Findable Accessible Interoperable Reusable

ljonkers@marum.de