Who knows *Globorotalia mentum*? – Making micropalaeontological data FAIR

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

NFDI4Earth

PANGAEA
Data Publisher for Earth & Environmental Science

Introduction

Many micropalaeontological studies rely on merging data from different sources. Apart from accessibility issues and a myriad of different data formats, **reusability** is hindered because **micropalaeontological data are semantically complex.** This is because of the existence of different **taxonomic schools** and evolving **taxonomic insights**, which both render standardisation difficult. As a consequence **reusing** micropalaentological data is cumbersome, even **when they are findable, accessible and interoperable**. Moreover, semantic complexity leads to confusion and **archiving errors**, further complicating data reusability.

Within the framework of the German National Research Data Infrastructure (NFDI) we aim to develop tools and processing pipelines to harmonise taxonomic data. Here 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 present **possible solutions** to increase the reusability of legacy data and propose guidelines for archiving of new datasets. This poster is intended as a **starting point for discussion**.

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.

Recommendations for data archiving

Make your science reproducible and your data useful! Deposit them in a public repository to make sure they are findable, accessible and interoperable.

Metadata

- Include information about the taxonomic concept used and be clear about exceptions. Keep taxonomic harmonisation in mind by providing information on alternative classifications. Try to link to internationally recognised classifications using e.g. the World Register of Marine Species.
- Report full genus and species names, include subspecies when needed.
- Include information about morphological variants and how they map onto the species.

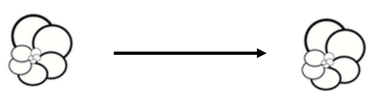
Data

- Report data at the highest possible taxonomic resolution, avoid including grouped taxa.
- Include unidentified specimens.
- Report counts rather than percentages.
- Report depth, not age.

Common challenges

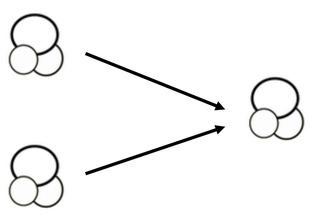
Scanning ~2,300 files with planktonic foraminifera species data archived on PANGAEA yielded **230 different names for extant species, about 180 more than are generally recognised** and thus clearly highlighting the need for harmonisation. Common problems related to the mapping of synonyms can be divided into three categories:

One-to-one



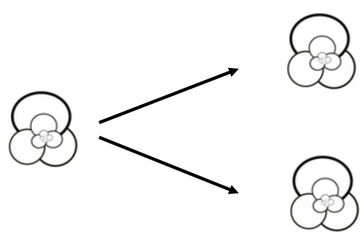
Easy to harmonise

Many-to-one



Easy to harmonise

One-to-many



Hard to harmonise, but context often helps

Images courtesy Peter Kiss

Notorious trouble makers: most problems are caused by a handful of species (complexes). Solving these makes all the difference. For planktonic foraminifera these are *Globigerinoides ruber* - *elongatus*; Neogloboquadrinids and *Trilobatus sacculifer*.

Complexity leads to archiving errors

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.

Context and the data itself can sometimes help to resolve these issues, but they require 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 may lead to meaningless species names.

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 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 solutions to harmonise taxonomic data

Different solutions are needed for **legacy data** that is already in the public domain and **new data** submissions. Any solution needs to be transparent, preserve changes, be future proof and scalable to other groups.

Legacy data

Different solutions to standardise legacy data are possible. We envision the following options:

1

Provide the community with online tools to harmonise taxonomy.

+

Flexible. User-specified taxonomy.

-

Updated taxonomy not preserved. Harmonisation needs to be done every time the data is used.

2

Work under the hood of PANGAEA to provide the data with harmonised taxonomy and provide tools to extract this information.

+

Harmonised taxonomy archived and no need to repeat standardisation every time data are used.

Future updates on taxonomy can be more easily incorporated.

-

Relies on a single classification stored in external library: WoRMS World Register of Marine Species

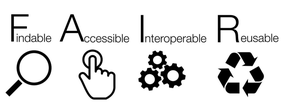
New data

Almost all of the issue 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).

We have formulated a set of guidelines (see above) and would like to hear your opinion about them and how to best disseminate them.

Join the conversation!

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



ljonkers@marum.de