

# Marine micropalaeontological data requirements

\* Indicates required question

---

## Introduction

The sedimentary record of marine microfossil species assemblages offers the unique opportunity to obtain a long-term view of changes in biodiversity and indirectly, of climate.

As a community we are in the fortunate position that large amounts of microfossil data are publicly available in data repositories. This renders data Findable and Accessible. However, the remaining two aspects of FAIR data, Interoperable and Reusable remain challenging because we lack community-defined (meta)data standards and requirements. This reduces the reproducibility of research, renders synthesis cumbersome and has led to many errors in archived data sets.

The goal of this survey is to assess what we as a community deem necessary to increase the interoperability and reusability of marine microfossil assemblage data. The product will be a checklist of (meta)data requirements that can be used by data generators when making their data (publicly) available. At the same time this checklist can be used by data repositories and data stewards to assess if datasets align to community standards.

The survey touches only slightly on the topic of the data format. At this stage, the focus is on what information is needed to ensure reusability of marine microfossil assemblage data.

In this survey you'll be guided through a series of data characteristics and asked if you think they are "desired", "recommended" or "essential". By default all information is desired. Essential implies that the data cannot be reused without this information. And recommended means that this information would be good to have and would increase the value and reusability of the data. When answering the questions, please keep in mind what information you would (like to) use when searching for certain datasets or when filtering a synthesis of microfossil assemblage data. None of the questions is compulsory, so if a question has no relevance for your work or if you don't want to answer it, feel free to skip it.

We plan to write a white paper to disseminate the results among the community. You will be asked if you want to participate in the writing process at the end of the survey.

Thank you for your time!

## Note

Please assume that all information is available when filling out this survey: e.g. when asked "Are details about the chronology optional, desired or essential?", answer this question assuming that a chronology is available. Don't tick "optional" because this information is not always available.

## About

This survey is part of an [NFDI4Earth](#) Pilot project funded by the German Science Foundation DFG on handling data with complex semantic structure. The project is carried out by Anne Strack, Lukas Jonkers and Michal Kucera at [MARUM Center for Marine Environmental Sciences](#) at the University of Bremen and Robert Huber at [PANGAEA - the data publisher for Earth and Environmental Science](#).

We are grateful to Paula Diz, Ivan Hernandez-Almeida, Oscar Romero, Tracy Aze, Deborah Khider, Adam Woodhouse and Karin Zonneveld for the feedback we received during design of this survey.

Please get in touch with [Lukas Jonkers](#) if you have questions about this survey or about the project.

## Description

1. How important is

*Mark only one oval per row.*

	Desired	Recommended	Essential
<b>a description of the goal of the analyses</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>a summary of the analysis/abstract</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

## Information about the sediment core

2. How do you rank the importance of the following data?

*Mark only one oval per row.*

	Desired	Recommended	Essential
<b>Site name</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Site location (longitude, latitude, water depth)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Sampling campaign/cruise</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Sampling/coring method</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Collection date</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Details about chronology when available</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Description of environmental setting</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Description of depositional setting</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Links to ancillary data when available</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Location (repository) where the core is stored</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### Information about sample handling and preparation

The data characteristics considered in the following four sections (sample handling, counting method, taxonomy and attribution) are likely to pertain to all samples in a (sedimentary) sequence. After these sections, questions follow about data characteristics that are likely to vary from sample to sample (e.g. sample-specific information and information related directly related to the count data).

#### 3. How important is the following information about sample handling?

*Mark only one oval per row.*

	Desired	Recommended	Essential
<b>Description of sample storage/handling prior to analyses</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Description of sample preparation (use of chemicals, boiling, centrifugation, settling method etc)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Initial sieve size (and unit)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Description of the mounting medium</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>The size of cover slip used</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### Information about counting methods

4. How do you rate the description of the following?

Mark only one oval per row.

	Desired	Recommended	Essential
<b>Counting method (e.g. light microscope, SEM, AI)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Counting magnification</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Counting marker (e.g. lycopodium)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>When a counting marker is used, information about the batch and amount of the marker</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Taxonomic information**

Taxonomic ambiguities often present the biggest hurdles to reusing microfossil abundance data and have caused many existing datasets that are in the public domain to contain errors. Below is a list of requirements that would reduce taxonomic confusion that we would like to get your feedback on.

**Taxonomic concept**

Different taxonomic schools exist and taxonomic concepts evolve over time. In order to be able to harmonise data counted by researchers following different taxonomic concepts it is important to include this concept in the metadata.

Example: *We considered all 50 species that are recognised as extant by Brummer and Kucera (2022).*

5. How important do you think providing information about the taxonomic concept is?

*Mark only one oval.*

- Desired
- Recommended
- Essential

### Lumped and confusing taxa

In some cases particular taxa are not or cannot be differentiated. If the presence and meaning of such lumped categories is clear, reusability of species count data is improved.

For instance: "*In the census counts, we did not differentiate G. elongatus, which is counted together with G. ruber albus.*"

6. How important is a description of how confusing taxa were treated?

*Mark only one oval.*

- Desired
- Recommended
- Essential

### Variants and subspecies

Some taxa contain subspecies or variants that may be counted separately. Even though in many cases the mapping of subspecies or variants onto the parent taxon is clear from the name (e.g. *Globigerinoides ruber ruber* and *Globigerinoides ruber alba*) a description of this mapping helps to avoid taxonomic confusion and renders data better reusable.

7. How do you rank an explanation of the subspecies and variants mapping?

*Mark only one oval.*

- Desired
- Recommended
- Essential

## Summing of taxa

When variants or subspecies are counted separately, or when species are grouped for other reasons, data files often contain the sum of lumped or grouped taxa *in addition* to the individual ones (e.g. the abundance of subspecies is provided together with the abundance of species, see figure below).

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

This practice leads to confusion and archiving errors even though lumping only involves simple summing of taxa. Moreover, the higher taxonomic resolution may get lost along the way when data sets are merged and only lumped taxa are retained. Finally, data sets with summed taxa make (machine) reading unnecessarily complicated.

8. Do you agree that species assemblage data should be archived at the highest taxonomic resolution and exclude summed taxa when the constituent taxa have been counted separately?

Mark only one oval.

Yes

No

## Taxon names

Taxon names need to be unique and unambiguous to avoid confusion.

9. How do rank providing unique and complete taxon names up the highest level distinguished in full, i.e. genus, species, subspecies/variant?

*Mark only one oval.*

- Desired  
 Recommended  
 Essential

### **Resting stages**

For some microfossil groups, e.g. dinocysts, parallel taxonomies exist for the motile and resting (cyst) stages and cysts can be named following either.

10. How important are the following?

*Mark only one oval per row.*

	Desired	Recommended	Essential
<b>Information about whether cyst or motile taxonomy was used.</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Information on how cyst and motile taxonomy are linked</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### **Link to external ontology**

Taxon names change with progressing insights, often necessitating taxonomic harmonisation even when data sets are perfectly described. This process can be facilitated through by linking taxon names to an external database, such as the [world register of marine species](#) (WoRMS). Use of such external ontologies allows or automated harmonising and updating of the taxonomy as WoRMS contains complete classification, status of the species and provides a list of synonymised names (see an example [here](#)). Some data repositories (e.g. PANGAEA) already use the unique aphia ID that WoRMS assigns to taxa under the hood, but inclusion would be more straightforward and more transparent if provided together with the (count) data by the scientists themselves.

11. How do you rate the inclusion of an AphiaIDs for each taxon in datasets of marine microfossil abundance?

*Mark only one oval.*

- Desired
- Recommended
- Essential

### **Zero abundances**

The absence of certain species can be very informative, yet for many data sets one needs to assume that species that were not reported were absent.

12. How do you rank the inclusion of species with zero abundance?

*Mark only one oval.*

- Desired
- Recommended
- Essential

### **Unidentified specimens**

Taxonomic completeness of species data is necessary for many applications (e.g. determining biodiversity, quantitative palaeo-environmental reconstructions etc), yet it is often difficult to assess from the data itself.

13. How do you rank the following aspects related to taxonomic completeness of species counts?

*Mark only one oval per row.*

	Desired	Recommended	Essential
<b>Include counts of unidentified specimens</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Indication of taxonomic completeness in the metadata</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### **Reworked specimens**

Sometimes specimens can be recognised as being reworked

14. How do you rank the inclusion of counts of reworked specimens?

*Mark only one oval.*

- Desired
- Recommended
- Essential

### **Attribution**

Tracking the source of the data is important for enquiries and to give credit.

15. How important are the following

*Mark only one oval per row.*

	Desired	Recommended	Essential
<b>Source (DOI/publication)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Contributor</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Institution</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Project</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Funder</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Additional remarks about site or sampling**

The questions above should be applicable to all samples. Sample and data specific questions follow after this question. However, do you at this stage, have any other remarks about (meta)data requirements?

16. Please write down other things you deem optional, desirable or essential to increase the FAIRness

---

---

---

---

---

**Sample characteristics**

The previous questions were applicable to all samples from a single site. The data properties that follow are, or can be, sample specific.

17. How do you rank the following sample characteristics?

*Mark only one oval per row.*

	Desired	Recommended	Essential
<b>Size of fraction that was used for counting (minimum, maximum, unit)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>State of the specimens in the sample, when available (e.g. stained, with cytoplasm, cell content)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>An estimate of sample preservation (quantitative or qualitative)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Sample mass used for counting (including a unit)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Dry bulk density of the sample (including a unit)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

18. Is there something else about the sample characteristics that you think needs to be included? Please indicate whether you think these should be optional, desired or essential.

---

---

---

---

---

### Data information

These are characteristics that vary among the samples.

19. How important do you think are the following

*Mark only one oval per row.*

	Desired	Recommended	Essential
<b>Depth, either top and bottom, or mid and thickness. With unit</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Sedimentary unit</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Sample ID when available (e.g. IODP, IGSN)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Age and age unit (when available)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

## Raw data

Many microfossil data sets only contain relative species abundances. However, count statistics cannot be evaluated from relative abundance data, preventing direct quality assessment (e.g. rarefaction). Even when count sums are reported, back calculation of raw data is associated with uncertainty because of rounding issues. Moreover, percentage data have proven too be extremely prone to archiving errors (in a large part of publicly archived datasets, the relative abundances do not sum to 100%). Many of these issue can be easily addressed and avoided by archiving the data as they were counted (integers).

20. Do you agree that raw, rather than derived, data should be archived?

*Mark only one oval.*

Yes

No

21. If you do not agree that raw data should be archived, can you please explain why?

---

---

---

---

---

## Qualitative data

Some microfossil data are qualitative only. Reuse of such data can be challenging if the categories (e.g. absent, rare, frequent, abundant) are not clearly defined in the metadata or not spelled out clearly and unambiguously.

22. How important is a description of the classification system for qualitative data?

*Mark only one oval.*

Desired

Recommended

Essential

23. Do you have other remarks about data requirements?

---

---

---

---

---

Final remarks, questions or suggestions

24. Do you have any final remarks, questions or suggestions that you would like to share?

---

---

---

---

---

**Demography**

Please help us understand the results of this survey better by providing a few details about yourself.

25. Are you an early career researcher? Within 5 years of active research since obtaining a PhD. \*

*Mark only one oval.*

- Yes
- No

26. Which is the microfossil group you mostly work with? \*

*Mark only one oval.*

- Benthic foraminifera
- Coccolithophores
- Diatoms
- Dinocysts
- Ostracods
- Planktonic foraminifera
- Radiolaria
- Other...

27. In which country are you based (working)? \*

*Mark only one oval.*

- Afghanistan
- Akrotiri
- Albania
- Algeria
- American Samoa
- Andorra
- Angola
- Anguilla
- Antarctica
- Antigua and Barbuda
- Argentina
- Armenia
- Aruba
- Ashmore and Cartier Islands
- Australia
- Austria
- Azerbaijan
- Bahamas, The
- Bahrain
- Bangladesh
- Barbados
- Bassas da India
- Belarus
- Belgium
- Belize
- Benin
- Bermuda
- Bhutan
- Bolivia
- Bosnia and Herzegovina

- Botswana
- Bouvet Island
- Brazil
- British Indian Ocean Territory
- British Virgin Islands
- Brunei
- Bulgaria
- Burkina Faso
- Burma
- Burundi
- Cambodia
- Cameroon
- Canada
- Cape Verde
- Cayman Islands
- Central African Republic
- Chad
- Chile
- China
- Christmas Island
- Clipperton Island
- Cocos (Keeling) Islands
- Colombia
- Comoros
- Congo, Democratic Republic of the
- Congo, Republic of the
- Cook Islands
- Coral Sea Islands
- Costa Rica
- Cote d'Ivoire
- Croatia
- Cuba
- Cyprus

- Czech Republic
- Denmark
- Dhekelia
- Djibouti
- Dominica
- Dominican Republic
- Ecuador
- Egypt
- El Salvador
- Equatorial Guinea
- Eritrea
- Estonia
- Ethiopia
- Europa Island
- Falkland Islands (Islas Malvinas)
- Faroe Islands
- Fiji
- Finland
- France
- French Guiana
- French Polynesia
- French Southern and Antarctic Lands
- Gabon
- Gambia, The
- Gaza Strip
- Georgia
- Germany
- Ghana
- Gibraltar
- Glorioso Islands
- Greece
- Greenland
- Grenada

- Guadeloupe
- Guam
- Guatemala
- Guernsey
- Guinea
- Guinea-Bissau
- Guyana
- Haiti
- Heard Island and McDonald Islands
- Holy See (Vatican City)
- Honduras
- Hong Kong
- Hungary
- Iceland
- India
- Indonesia
- Iran
- Iraq
- Ireland
- Isle of Man
- Israel
- Italy
- Jamaica
- Jan Mayen
- Japan
- Jersey
- Jordan
- Juan de Nova Island
- Kazakhstan
- Kenya
- Kiribati
- Korea, North
- Korea, South

- Kuwait
- Kyrgyzstan
- Laos
- Latvia
- Lebanon
- Lesotho
- Liberia
- Libya
- Liechtenstein
- Lithuania
- Luxembourg
- Macau
- Macedonia
- Madagascar
- Malawi
- Malaysia
- Maldives
- Mali
- Malta
- Marshall Islands
- Martinique
- Mauritania
- Mauritius
- Mayotte
- Mexico
- Micronesia, Federated States of
- Moldova
- Monaco
- Mongolia
- Montserrat
- Morocco
- Mozambique
- Namibia

- Nauru
- Navassa Island
- Nepal
- Netherlands
- Netherlands Antilles
- New Caledonia
- New Zealand
- Nicaragua
- Niger
- Nigeria
- Niue
- Norfolk Island
- Northern Mariana Islands
- Norway
- Oman
- Pakistan
- Palau
- Panama
- Papua New Guinea
- Paracel Islands
- Paraguay
- Peru
- Philippines
- Pitcairn Islands
- Poland
- Portugal
- Puerto Rico
- Qatar
- Reunion
- Romania
- Russia
- Rwanda
- Saint Helena

- Saint Kitts and Nevis
- Saint Lucia
- Saint Pierre and Miquelon
- Saint Vincent and the Grenadines
- Samoa
- San Marino
- Sao Tome and Principe
- Saudi Arabia
- Senegal
- Serbia and Montenegro
- Seychelles
- Sierra Leone
- Singapore
- Slovakia
- Slovenia
- Solomon Islands
- Somalia
- South Africa
- South Georgia and the South Sandwich Islands
- Spain
- Spratly Islands
- Sri Lanka
- Sudan
- Suriname
- Svalbard
- Swaziland
- Sweden
- Switzerland
- Syria
- Taiwan
- Tajikistan
- Tanzania
- Thailand

- Timor-Leste
- Togo
- Tokelau
- Tonga
- Trinidad and Tobago
- Tromelin Island
- Tunisia
- Turkey
- Turkmenistan
- Turks and Caicos Islands
- Tuvalu
- Uganda
- Ukraine
- United Arab Emirates
- United Kingdom
- United States
- Uruguay
- Uzbekistan
- Vanuatu
- Venezuela
- Vietnam
- Virgin Islands
- Wake Island
- Wallis and Futuna
- West Bank
- Western Sahara
- Yemen
- Zambia
- Zimbabwe

### Contact details (optional)

Please indicate your name and email address if we may contact you in case of questions or if you would like to participate in writing a manuscript disseminating the results of this survey to the community.

28. First name

---

29. Last name

---

30. What is your email?

---

---

This content is neither created nor endorsed by Google.

Google Forms

