This is a review for “BENFEP, a quantitative database of BENthic Foraminifera from surface sediments of the Eastern Pacific” by Diz et al. BENFEP is, as indicated by the title of the paper, a quantitative database of the distribution of recent benthic foraminifera in the Eastern Pacific. The authors compiled a huge amount of datasets, published since the 1950s, into a comprehensive database. The data has been corrected for species synonyms according to the latest taxonomy published in WORMS. The database includes 1072 fully described species plus 391 benthic foraminiferal entities classified at genus level from 2572 sampling locations. This is a huge database and I have the deepest respect for the amount of work that went into such a piece of work. I can imagine a huge amount of scientific possibilities that will be opened by having access to such a database. First of all, it could be used for calibrating proxies, based on benthic foraminifera assemblages. One example, which directly comes into my mind, is a study like Tetard et al., 2021 regarding foraminifera assemblages on the way to a quantitative oxygen proxy. In addition, this dataset can be used to estimate foraminiferal biomass distribution in the sediments, it can be used to upscale individual metabolic rates to total biogeochemical fluxes and much more. For example, it might be possible to estimate foraminiferal respiration rates in sediments.

Unfortunately, I cannot really access the database files on Pangaea. Even, if I log in, I get the message that access is forbidden due to the data moratorium (Error 403). I shortly screened the other reviews and it seemed that they did not experience this issue. So my review is purely based on the background information in the paper and I cannot review in any way the content of the database. I rated the data quality “good” because I had to give a grade in the review system.

In my opinion, based on the background information in the paper, this database will be of very high interest for a large scientific community including micropaleontologists that try to solve taxonomic issues, paleoceanographers that calibrate new proxies and biogeochemical modelers, who try to include biogeochemical flux rates from activities of different organisms into their models. I have some suggestions that may improve the accessibility of the data and some questions about the future of this database and would only suggest minor revisions for the accompanying paper. As mentioned above, at the moment it seems impossible for me to access the database via PANGAEA, so I cannot review the content of the database itself.

Points of revision for the paper:

Line 121-122: “The whole database can be managed using R version 4.2.1 (R Core Team, 2022). It can be uploaded and managed with geographic information system software such as QGIS and ArcGIS after changing the table format from wide to long.”

It would be great, if the database could be managed with geographic information system software such as ArcGIS. If I would have access to the database, I would have tested this myself. I think it would be really helpful, if the authors provide some more information on how to do this. For example, either the database could be uploaded in different file formats. At the moment the file that is uploaded at Pangaea seems to be an .xlsx file. This is of course great for standard users of .xls based software but makes it more complicated for other users (R for example), especially with such a big file. One solution could be to upload the database in different file versions. One text file in long format that could be directly imported into QGis or ArcGIS, another one in wide format for other applications and the original .xslx file. This would be very helpful for users but also would make updating the database in the future very tedious, since all the files would have to be exchanged.

Another possibility would be to publish an R friendly text file that could easily be imported into .xls based software. If this file is in a wide format, it would be easy to provide an R script to convert it into a QGIS/ArcGIS friendly long format, either in the paper or in the supplement. It is also not directly clear

to me, which columns would have to be merged to make the format GIS friendly. Here is an example for such a cookbook (using the gather() argument from the tidyr library):

benfep_wide

#> station lattitude longitude uvigerina_peregrina hoeglundina_elegans cibicides_wuellerstorfi

#> 1 1 10°56’ 178°23’ 17.3 1000.7 0.1

#> 2 2 12°02’ 175°59’ 5273.2 50.0 42

library(tidyr)

 # The arguments to gather():

 # - data: Data object

 # - key: Name of new key column (made from names of data columns)

 # - value: Name of new value column

 # - ...: Names of source columns that contain values

 # - factor_key: Treat the new key column as a factor (instead of character vector)

benfep_long <- gather(benfep_wide, species, population_density, uvigerina_peregrina:cibicides_wuellerstorfi, factor_key=TRUE)

benfep_long

	station	lattitude	longitude	species	population_density
#> 1	1	10°56’	178°23’	uvigerina_peregrina	17.3
#> 2	1	10°56’	178°23’	hoeglundina_elegans	1000.7
#> 3	1	10°56’	178°23’	cibicides_wuellerstorfi	0.1
#> 4	2	12°02’	175°59’	uvigerina_peregrina	5273.2
#> 5	2	12°02’	175°59’	hoeglundina_elegans	50.0
#> 6	2	12°02’	175°59’	cibicides_wuellerstorfi	42

These are of course only optional suggestions on how the accessibility of the manuscript might be improved. My other point is not really a change in the manuscript itself but more a question/suggestion about the future of this database:

Line 338-342: “This database is conceived as a springboard to store future quantitative data of benthic foraminifera in the East Pacific and make them available to the scientific community. It can be enlarged with new records as they are being generated or after the authors request, therefore providing an ongoing live resource. Any changes to add, correct, or update taxonomic categories to an existing

version will be indicated in PANGAEA. We also encourage users of the BENFEP database to quote original data sources."

How do the authors think about the future of the database? Should there be an easy protocol or "cookbook", how to add data of new records as they are being generated? Are there any permanent members of the group that are able to sustain, clean up and update the database? I think for individual members this might be an impossible task but it should be important to avoid misuse of the feature to update the database. One suggestion would be to provide a review system like in WORMS: Uploaded datasets that are not reviewed, yet, will be marked as "not reviewed". Reviewed datasets might be marked as "accepted" or "unaccepted", if there are any issues considered by a reviewer. Reviewers could be volunteer foraminifera taxonomists. I think there might be a big support in the community for such an effort.

Finally, are there any plans to include other ocean basins into the database in the future? These are just some questions/suggestions about the future of the database and of course, this is an own project by itself and some things cannot be directly integrated into the paper and database. For example, the review system for future datasets would need an own platform or deeper collaboration with Pangaea. Though, I think it is worth it to think about the legacy of such a huge project and maybe to integrate some points about the discussion of the future of the database at the end of the paper about "Data availability and future plans".