Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-164-RC2, 2020

Earth System Science Data Discussions

Interactive comment on "Retrospect and prospect of a section-based stratigraphic and palaeontological database – Geobiodiversity Database" *by* Hong-He Xu et al.

Richard Butler (Referee)

r.butler.1@bham.ac.uk

Received and published: 20 September 2020

GENERAL COMMENTS This article has the potential to provide a useful and welcome introduction to the history, structure, content, functionality and analytical tools, and proposed future of the Geobiodiversity Database, a database with huge research potential that remains much less well known and used within the international geoscience community than the US-based Paleobiology Database. The current version of the manuscript does not, however, fully achieve its goals. It is lacking in details in some parts, particularly those dealing with historical aspects, and at times difficult to follow and somewhat repetitive. In this review I first provide some major comments on

three particular sections of the manuscript, and then more detailed comments on other issues with the text.

The history section (lines 35–80) is rather opaque and should provide more details. It would be useful to know (lines 36–45), at which institutions, by which researchers, and using what sources of funding the GBDB was established, how the database was managed, where the data enterers were based etc. In lines 38–39 a "large palaeobiology database" is mentioned, but it is not made explicitly clear that the authors are referring to (I believe) the US-based Paleobiology Database. I am not sure that the statement that the PBDB "temporarily ignored" Chinese data is correct: such data may have been underrepresented, but the PBDB currently includes >13,000 Chinese collections, and several thousand of these were entered prior to 2007. On line 44 there is mention of aligning data entry with standards of "international researchers", but it is not clear what those standards are. On lines 51–54 a number of different statistical and visualisation tools are mentioned by name, but these should be explained in more detail. On lines 56–59, stratigraphic correlation tools (CONOP, SinoCor) are mentioned, but not explained with sufficient detail for readers who are not already familiar with them.

In the section on the data of the GBDB, some aspects of the structure of the database are described with insufficient detail to be understandable. For example:

- Lines 90–91: it is not clear what a 'virtual section' is. This needs expanding and explaining with further detail. It is unclear to me why a fossil without any detailed associated stratigraphic section and a borehole (which presumably has a detailed record of changes in sedimentology) both represent 'virtual sections' as they seem like quite fundamentally different kinds of data.

- Lines 95–96. It is very unclear how the palaeontological data in the GBDB are linked to sections, and how this relates to occurrence-based datasets. Some of the statements made about this seem internally inconsistent.

- Lines 100–101. It is unclear what "opinion data" are here, and how taxonomic opin-

ions are treated by the GBDB. Are these taxonomic opinions, and reflect changes in the identification of fossils from particular sections? Does the change in the taxonomic identity of a fossil from a particular section reflected in other sections at all? Is there an overarching taxonomic framework, similar to the dynamic taxonomy present in the PBDB?

- Do the GSSPs included in the GBDB only include those from China, or is this global?

- Lines 105–107. This collaboration with BGS is very incompletely described. Is data compiled in the same way as in the rest of the database? Is the data accessible to other researchers?

- Lines 108–109. Are these borehole data from oil companies publicly accessible, despite potential commercial sensitivities? Should be made explicit.

The section on newly-added data in the GBDB contains comments that are questionable in parts. For example, although it is undoubtedly true that marine invertebrates have long been the core focus of Phanerozoic diversity studies, it is not true to imply that there have been almost no studies of terrestrial diversity, as is done here where only two local studies of Chinese Palaeozoic diversity are cited. For example, considerable work has been conducted on global land plant diversity, starting with the work of Andrew Knoll, Karl Niklas and Bruce Tiffney in the late 1970s through early 1990s, and continued today by other researchers such as Borja CascalesâĂŘMinana. There is a long track record of studies of terrestrial tetrapod diversity, beginning with the work of Mike Benton on global patterns numerous studies by John Alroy and others on Cenozoic mammal diversity, and then a huge number of papers over the last 15 years on diversity patterns in individual clades, from dinosaurs to hominids. Many of these have used PBDB data and are listed on the official publication list of the database. Finally, there is also a long history of studies of global insect diversity, going back nearly 30 years. This section should be revised in light of this extensive history of terrestrial research.

SPECIFIC COMMENTS The title of the paper is somewhat unclear in its meaning, and I would suggest changing it for something simpler and more accessible. Perhaps something like “The past and future of the Geobiodiversity Database: a section-based stratigraphic and palaeontological database”

There is much use of the term “big data”. But the size of the datasets contained within the Geobiodiversity Database and other comparable databases (PBDB etc.) would generally not qualify as big data under most definitions. A slightly more neutral term, such as “large data sets” might be more appropriate.

Line 12: should be “Here, a thorough introduction is given to the Geobiodiversity Database”

Line 13: in various places you use the word “serial” (here “serial of scientific studies”) when “series” would be correct

Line 14: “Nevertheless, the existing problems of the GBDB limited the using of its data”. This phrasing is problematic in the abstract because the “existing problems” have not been described. I would suggest combining this and the following sentence into something like “Nevertheless, limited use of the GBDB by the wider palaeontological community led to reorganisation and improvements beginning in 2019”.

Line 18: “Further collaborations are proposed” – this is not really developed in the paper, and it would be good to know if, for example, more definite discussions on collaboration have been had with other database leadership teams.

Lines 19–20: This statement on the availability of datasets – should it be in the abstract?

Lines 22–25: These opening lines are quite repetitive. Would suggest rewording to reduce repetition e.g. “Palaeontology and stratigraphy have become increasingly quantitative branches of geoscience in recent decades (REFERENCES). Quantitative analyses of large datasets of fossil and stratum records have become more common in

studies of. . .

Lines 25–27: rather a brief selection of papers are cited as evidence of the increasingly quantitative nature of palaeontological/stratigraphic research, and they are fairly biased towards studies linked to the GBDB. A broader range of citations would be useful here.

Line 27: I would suggest “academic databases”, rather than “professional databases”

Line 30: I would suggest using the term “user-friendly” rather than “friendly” when talking about the accessibility of the database for users.

Line 55: would suggest “unique” rather than “exclusive”

Line 158: opinions in the PBDB are taxonomic opinions, not palaeobiological opinions.

Line 170: the relationship of taxon occurrences to sections needs to be explained in more detail in this manuscript. Do taxa occur associated with distinct horizons within a section (in which case this would quite closely approximate an occurrence-based dataset)? Or are occurrences clumped within sections in some way?

Lines 193–194: It is unclear what is meant here. Do you mean that the GBDB was not being backed up and that its use was hazardous because of the potential for data loss? Needs some more clarity.

Lines 195–197: More information is needed on what a “safe data bank” is.

Lines 198–200: The description of the data entry process is unclear. Who are registered authorizers, and how are they selected? Can anyone enter data, but it has to be checked by a registered authorizer?

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-164, 2020.