Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-164-RC1, 2020

Open Access 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.

Peter Sadler (Referee)

sadler@ucr.edu

Received and published: 11 August 2020

GENERAL COMMENTS ON THE TEXT: It is good to learn that the GBDB (Geobiodiversity Database) may be fully back-in-business. As a section-based database, it had enjoyed a useful position among taxon-based and collection-based information hubs. It has been the foundation of groundbreaking publications in macroevolution and a test bed for correlation tools. Speculation about its fate had begun when access became more challenging and updates ceased. There followed stories to the effect that the trusted original team had quit for untold reasons and there were rumors of a law suit. My comments are intended to improve the positive impact of this article on potential

users of the resuscitated GBDB.

Some research paleobiologists and stratigraphers trust public databases only as a way to find the primary literature. They like to do their own quality control and are hard to convince that a database management team can be as thorough as themselves. Database managers must be very careful not to undermine the credibility of their database in the minds of potential users. User-trust is a fragile commodity, easily lost. This retrospective article risks undermining the credibility of the GBDB in some ways, especially with its limited historical overview and veiled criticism of the former management. A full update summary that describes the details of new and tested functionality would do more to restore the trust of professional users.

Retrospection is a strange approach to an update. Effective product updates look forward, optimistically, not backward; they should not waste time criticizing prior functionality. Given the speculations about the change of management of the GBDB, dwelling on the criticisms could seem more political than intellectual. Previous user criticisms can be presented positively as wishes fulfilled.

Retrospection is, at best, a chance to display depth of knowledge about the longer history of numerical stratigraphy and the databases that came to support it. Here the article falls short, especially in the introduction. It is widely known that the GBDB came late to the suite of community databases that was already led in paleontology by the PBDB (PaleoBiology DataBase). Insightfully advised, the GBDB made a new niche for itself – a section-based structure and the promise of access to Chinese locality data. The longer-established PBDB was founded on the database compiled by Jack Sepkoski, a member of the Chicago school, under Dave Raup, that had pioneered numerical paleobiology. The same school trained the founder of the Macrostrat database. The introduction does not properly acknowledge this deep history and offers nothing on the "Fossil Record" database, a long-standing collaboration of museums, universities and government agencies in New Zealand. Nor is Neptune mentioned; that database was built for the Deep Sea Drilling Project and its successors. Other geosciences, no-

tably seismology, were substantially ahead in pioneering and funding big community databases, but even here the separately funded organizations have recently shown some fragility.

The introduction also attempts to cover numerical correlation tools. These too have a much deeper and richer history than is evident in the article, and the portrayal is not entirely accurate about those connected to the GBDB. In very different ways, SinoCor and CONOP built on two-dimensional graphical correlation, which was published by Alan Shaw in 1964. Like GraphCor before it, SinoCor is computer-assisted graphical correlation. Given the enormous hardware advances, SinoCor was able to present users with a more elegantly straightforward graphical user interface, which helps students understand graphic correlation. CONOP, by contrast, is multidimensional. It takes Alan Shaw's principle of "Economy-of-Fit" to a wider array of stratigraphic data types and separates the consideration of sequencing from spacing. In this regard, CONOP resembles Lucy Edwards' "No-Space Graphs" from the 1970's. Another important biostratigraphic sequencing tool, Biograph, supports important macro-evolutionary studies, but is not mentioned. I would hope to hear of plans to incorporate time series tools that are well developed for the correlation of geochemical data and in astrochronology. For greater efficiency in sequencing huge datasets, which might be called big-data among paleobiologists, the original GBDB developers have since modified CONOP to CONOPSAGA. That development supports the most recent of the papers cited here and on the GBDB website, but the enabling software development is not mentioned.

Two other numerical sequencing tools that appeared years ago deserve mention: Jean Guex' Unitary Associations (published in book form as early as 1991, and the basis of Biograph) and John Alroy's CONJUNCT (also Appearance Event Ordination). They are especially relevant because they rely upon co-occurrences of taxa, the kind of information that is readily extracted from the collection-based PBDB. Guex and Alroy have separately made the case that co-occurrence is more trustworthy stratigraphic information than the order of all first- and last-occurrences in a single stratigraphic section.

These are arguments for including collections that lack the context of a stratigraphic section together with the section-based data. A collection is a stratigraphic section of restricted extent; the limiting case, if you will. If the collection has been assembled across many beds, however, it is no longer evidence of coexistence – a problem more easily avoided or recognized in section-based data. It is unfortunate to call these collections "virtual" data. It is very odd to call borehole records "virtual" sections (line 92). The term "virtual" is better reserved for composite sections that have been built by combining data from multiple locations, as in graphic correlation. For these, the familiar image of a stratigraphic section serves to hold all compiled information about sequence and spacing, but unlike cores, has no physical existence.

The authors (line 58) imply that SinoCor has few users because its file format is unique, and yet their own GBDB is correctly claimed (line 55-56) to readily export data for the software tool. Perhaps the authors overestimate the market for any of these tools. No matter, tool development has moved on. The GBDB needs to consider new developments like CONOPSAGA, Dynamic Programming (with its portfolios of alternative hiatus treatments), AstroChron and its potential hybrid AstroConop.

The authors' remarks about fossil terrestrial organisms are surprising. They recognize John Alroy as a key contributor to PBDB-based publications on biodiversity. John's much earlier dissertation and the publications that established his expertise were exhaustive big-data(?) analyses of all Cenozoic land mammal fossils of North America. John is another product of the Chicago school; he writes paleontological database management scripts and sequencing tools that have been integral to the PBDB. Vertebrate paleontologists might be disappointed to learn that the GBDB does not yet extend to collections younger than Eocene (line 162). Most professional paleontologists will know that section-based data are inherently more sparse for non-marine than marine organisms.

The article repeatedly emphasizes the section-based structure of the GBDB, and yet risks undermining this essential quality. The authors seem to be critical that "for a long

time the GBDB focused on stratigraphic records instead of only fossils" (lines 169-171). And yet the leaders of the project were all paleontologists. The impressive record of their publications from the GBDB, as evident from this article and the GBDB web site, is overwhelmingly paleontological. The key to better resolving power in the geologic history of taxon richness is unambiguous evidence of superposition and sequence. That evidence is more richly contained in sections with multiple, ordered collections of fossils, than in individual collections. Lithologic records hold clues to hiatuses and habitat biases. This is the one of the distinguishing advantages that the GBDB gave to paleobiologists.

The authors and GBDB managers need to distinguish those uses that are essentially paleobiologic from those that are biostratigraphic. There are studies that legitimately build from collections. There are those that are concerned with high-resolution sequencing. The latter are the vital constituency of the GBDB and it is these publications that will drive success. Some of these potential customers build their own databases and manage their own quality control, not relying on either the GBDB or the PBDB except as an aid in finding publications. The PBDB began that way. Paleozoic ammonoid experts, for example, have built their own database (GONIAT.ORG), which some find more open and user-friendly than either the GBDB or the PBDB. For Mesozoic and Cenozoic micro-paleontologists, there is MICROTAX.ORG. Other open paleontologic databases cover Baltic cores, Ocean drilling data and Cretaceous sections, for example. These databases and others like them have the advantage of narrower scope than the GBDB and PBDB. They set a high standard for completeness, expert authentication and community involvement

The article provides two numbered lists: the first presents customer dissatisfaction with the GBDB (no statistics, dates or sources given); and the second presents implemented and planned improvements. Neither list is particularly specific. It would place the GBDB in more positive light to emphasize the upgrades already in place and to show that these are responses to customer wishes. Some of the upgrades concern

flashy look-and-feel. The serious potential users, whose papers draw envious attention to the GBDB, want better control of downloads, versatile analytical tools (downloadable and on-line web services), and an assurance of quality control. It is also reassuring to see that the managers of a database trust its quality for their own high-profile, collaborative publications. Quality control has two levels: 1) accurate transcription of published information; and 2) expert cleaning and updating of legacy information. The PBDB, for example, seems to promise users an updating of the higher taxonomic assignment of species and a modern evaluation of synonymy. The breadth and expertise of the data compilers and authenticators may impress users more than the number of data transcribers who may not be accomplished taxonomists. The GBDB opinions are potentially an attractive feature. In the past, some users found that they could know that there was an opinion, but had difficulty accessing the source or the content of the opinion. It is not clear whether this has changed. Expert cleaning can be more than perfect matching to the published source. Those who mine data from this primary literature notice some obvious mistakes in measurements, numbers and names; some are straightforward to correct; so too are some types of obsolescence.

"Big data" is a popular jargon phrase. Paleontologists have, indeed, been compiling large datasets for the past 2-3 decades. These are small data, however, compared with those that come from geoscientific instruments that continuously monitor seismic-waves, GPS satellites, weather stations or tide gauges, for example. By comparison, it seems to be an overstatement to use the term "big data" for a data set whose essence that can be downloaded to a single Excel flat file and readily manipulated on an old Core i5 laptop to select subsets of data and sources or generate a world map of localities. Presumably the GBDB itself is a true relational database with a far more versatile data structure.

The article concludes with a relatively weak return to what has been true in the past and to statements that are too generalized. That dissipates the excitement of reading that the GBDB is alive again. The three databases mentioned (there are others,

of course) still complement one another and still aim to be readily accessible. Nobody should doubt that. Readers need more specific information about the updates and reassurance that the database is still managed by experienced geobiologists and stratigraphers. The rate at which GBDB employees can enter data is very impressive. One advantage of the PBDB, in the minds of many users however, is that data entry is a community effort involving authorized users, each with proven and vetted expertise. Published data include errors and obsolescence. Many of us trust the PBDB authorizers to improve the quality of the information during data entry.

One of the most difficult aspects of published paleobiological data is to continually improve upon the initial estimates of the numerical age of fossil occurrences. The GBDB offered an advantage over the PBDB, because the ease of recovering superpositional order of events from sections enabled users to recalibrate their composite data sets with the most recent geochronometry. I trust that the GBDB will continue to build on this research advantage.

Paleobiological databases of the future ought to be able to continually sequence global and regional virtual sections. When new data are entered, the database ought to be able to respond immediately whenever the new data would be the oldest or youngest find of a taxon or a previously unproven co-existence; that is, data that potentially change the global range of a taxon should be flagged as soon as possible after entry. That is just one example of a deeper intellectual advantage of summary graphics for the whole database. In other words, automated and continual sequencing of the fossil record could become part of the validation and cleaning stage in the GBDB schema. CONOP- or SinoCor-generated graphics of richness and extinction through time should become a routine splash-page of the database, just like paleogeographic summaries. Like the maps of the spatial distribution of data, these time-series graphics could indicate how data-support and uncertainty (rarefaction, perhaps) vary through time. Both could be incentives to contribute new data where coverage is weak.

DATA UNIQUENESS: Although all these data have been previously published in nu-

merous papers, the uniqueness of the PBDB database is that so much information is all compiled in one place, and in one format. Most of the primary sources are in Chinese. As compiled here, the data are uniquely accessible to a much wider audience of potential users. The majority of the data were compiled and available through the GBDB years before this article was written. The article compares the total holdings of the GBDB and PBDB and contrasts the geographic distribution of their primary sources. From this we may assume that the overlap is not substantial, but there is no systematic analysis of duplication between the databases - another possible approach to quality control.

DATA USEFULNESS: The data have already proven their usefulness in several major publications about macro-evolution and mass extinction. Their usefulness is primarily via the tools of the GBDB. The Excel file made available here might improve the usefulness for some researchers, but the data management tools of the GBDB are almost certainly superior.

DATA COMPLETENESS: The data are surely the most comprehensive of their kind for localities in China. The entire data set has been posted in the flat Excel file.

DATA QUALITY: The GBDB makes a good faith effort to ensure that the data are true to the original publications. Corrections for error and obsolescence are handled by "opinions." There is no guarantee that the opinion process amounts to a complete assurance of quality. The best assurance of the quality of these data is that they are not only already included in a permanent and managed repository, but they have also been used in major research articles that have stood the test of peer review and publication.

PRESENTATION QUALITY: The article is an abbreviated retrospective on the history of the GBDB. It is less about the data themselves than an advertisement for on-going and future changes in the GBDB platform that makes these data public. There is nothing in the article that would be at odds with the data themselves, but there are statements that might undermine readers' confidence that the database management team has a

sufficient breadth of expertise. Perhaps the need for a brief article has produced a condensed history that readers may find deficient and not always accurate. This problem for readers' confidence in the future of the GBDB is compounded by a tendency toward general rather than specific wording and all the usual misunderstandings that arise in translation between the Chinese and English languages. These linguistic flaws would be relatively easy for a native English-speaker to correct.

ACCESSING THE DATA: I navigated to the Xenodo site from the PDF abstract and was able to download the GBDB "all section" data as a Microsoft Excel flat-file with 764545 rows. Each row appears to be one taxon assigned to a collection with lithological information. From this file I easily generated my own world map of the data locations, using Excel functions. The location symbols produce a recognizable world map, as expected, with dense coverage for China, sparse patterns recognizable as Europe and North America.

The abstract posted at the Xenodo site provides a web address for the GBDB. [https://www.geobiodiversity.com/] This address could not be reached in several attempts; it finally responded slowly with the new statistical splash page for the GBDB. The abstract submitted for review does not end with this url.

The new website presents users with some problems and disappointments. The "fossil ontology" label is intriguing, but clicking it does not work. Not all buttons in the bar at the bottom left do anything noticeable and not all have a call-out explanation. The paleogeographic maps that summarize section locations are fascinating to play with via the circular geologic time scale. The popup words to explain the eye icons are covered by the cursor and cannot be read. Occurrences is mistyped as "Occureences." The publications have urls, but these are not clickable. The site is clearly new, more impressive than the old one in some ways, but still not fully fledged.

Google-searching for the GBDB on line yielded only the old Geobiodiversity Database site "Powered by Junxuan Fan Copyright 2006-2017 Released version: 1001130." The

newest posting on that site is from January 2018. Manuscript first-author Xu Honghe is listed as the contact. Perhaps that site should help potential users by linking them tothe new site. The new site does include a link to the old; this may be welcomed by some long-time users.

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