Supplementary main text

**Potential issues of the GBDB data and framework and future works**

**1. Geographical bias**

GBDB data are largely limited to Chinese blocks and Paleozoic sections. The low fossil abundance in the post-Paleozoic may largely be attributed to the fact that non-marine strata prevail in the Chinese blocks during this time. Compared to the marine facies, fossil abundance in non-marine strata is generally low in the geological record.

**2. Subjective definition of geological unit**

Geological units are subjectively defined but follow stringent criteria. When a detailed lithological description is provided in a reference, a refined geological unit has been entered (Fig. S1). Detailed descriptions are provided in most Chinese stratigraphic books. If only a general lithological description is provided with respect to lithostratigraphy, a geologic unit is normally assigned to a lithostratigraphic unit, such as a bed, a member, a formation, or a group. In GBDB system, a sedimentary unit measures the smallest lithological part in a section column, and contains fossil collections if there is any. The recognition of a sedimentary unit in a section column is theoretically determined by lithological changes but also depends on arbitrary decisions of scientists. For example, when a section column is featured as limestone but differs in some part as nodular limestone, micrite limestone, then a unit can either be divided based on these changes in lithology or combined as one since it is limestone throughout. The identification of a geological formation or member in a section is usually where there are big changes in lithology and this is normally stated/labeled clearly in the literature. Sometimes, the data enterer will just use formation as unit, so lithostratigraphic units can be coarse. But mostly, the GBDB does encourage enterers to distinguish units according to minor lithological changes, that is, members and event beds may be distinguished.

**3. Taxonomic problems**

Taxonomic data are currently not well organized and poorly vetted in the GBDB (Table S2). Unlike the PBDB, the GBDB does not have a comprehensive framework for taxonomic authority and opinion data. For the current dataset used in this study, only some graptolites have been under critical evaluation by experts at NIGPAS (Nanjing Institute of Geology and Palaeontology) (Fan *et al*. 2014). Careful taxonomic reviews of the rest of data are in process via the volunteers and potential GBDB users. Simultaneously, we are working on cleaning taxonomic data by establishing an automated link to the PBDB.

Supplementary tables and figures

Table S1. The geographic distribution of occurrences in the GBDB. Note that 86% of all occurrences are from China.

|  |  |
| --- | --- |
| **Country** | **Freq.** |
| China | 149306 |
| Iran | 8713 |
| Estonia | 5130 |
| UK and Northern Ireland | 3677 |
| Belgium | 1621 |
| Russian Federation | 1311 |
| Norway | 858 |
| United States of America | 649 |
| Canada | 584 |
| Argentina | 460 |
| Austria | 449 |
| Portugal | 295 |
| Italy | 152 |
| Others | 540 |

Table S2. Occurrence distribution among fossil groups in GBDB.

|  |  |
| --- | --- |
| **Fossil group** | **Number of occurrences** |
| Foraminifera | 38152 |
| Conodonta | 37734 |
| Graptolithina (Hemichordata) | 37531 |
| Brachiopoda | 30258 |
| Trilobitomorpha (Arthropoda) | 18840 |
| Bivalvia (Mollusca) | 11093 |
| Anthozoa (Cnidaria) | 10218 |
| Cephalopoda (Mollusca) | 7731 |
| Crustacea (Arthropoda) | 5411 |
| Sporites | 3287 |
| Cycadophyta | 2468 |
| Acritarchs | 2006 |
| Gastropoda (Mollusca) | 1997 |
| Hyolitha (Mollusca) | 1891 |
| Equisetophyta | 1750 |
| Chitinozoa | 1712 |
| Magnoliophyta | 1601 |
| Bryozoa/Ectoprocta | 1307 |
| Coniferophyta | 1260 |
| Dinoflagellata | 1176 |
| Lycopodiophyta | 775 |
| Radiolaria | 772 |
| Porifera | 543 |
| Pollenites | 484 |
| Filicophyta | 441 |
| Charophyta | 428 |
| Trace fossils | 407 |
| Tentaculita (Mollusca) | 392 |
| Ginkgophyta | 338 |
| Annelida | 301 |
| Monoplacophora (Mollusca) | 268 |
| Archaeocyatha | 224 |
| Hydrozoa (Cnidaria) | 215 |
| Crinozoa (Echinodermata) | 209 |
| Rotifera | 208 |
| Chlorophyta | 193 |
| Placodermi (Vertebrata) | 160 |
| Ciliophora | 140 |
| Rhodophyta | 104 |
| Others | 968 |
| unknown | 4072 |



Figure S1. A simplified example GBDB database structure. The core item of the database is a stratigraphic section. A section can be divided into formations, and a formation can be split into bed-by-bed units. If a unit contains fossil collection(s), then fossil occurrences are listed.



Figure S2. The time series of sampled genus richness and all sample-related variables over the Phanerozoic based on the formations that span longer than one, two, and three geological stage. Note that all curves in each plot mirror with each other. D, raw genus diversity (sampled-in-bin). NC, number of collections. NF, number of formations. NU, number of sedimentary units. SA, sedimentary area. SV, sedimentary volume. Dashed lines denote five mass extinction events during the Phanerozoic. Ma, millions of years ago. Cm, Cambrian; O, Ordovician; S, Silurian; D, Devonian; C, Carboniferous; P, Permian; Tr, Triassic; J, Jurassic; K, Cretaceous; Pg, Paleogene; Ng, Neogene.