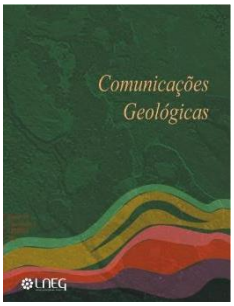
Comunicações Geológicas

Comunicações Geológicas (2023) 110, 1, 25-29
**ISSN:** 0873-948X; **e-ISSN:** 1647-581X

# Portuguese Fossil Database: first phase report




# Base de dados de fósseis portugueses: a primeira fase de construção

P. Fialho[1,4]*, R. da Silva[2], S. Patrocínio[2,3], B. Costa[1], A. Burigo[2,3]

DOI: https://doi.org/10.34637/fkv1-6v97
Recebido em 01/11/2022 / Aceite em 05/04/2023
Publicado online em maio de 2023


Artigo original
Original article

**Abstract**: Knowing what has been previously done before starting a new study is of primary importance when conducting novel scientific research. Portugal has more than two centuries of research in Paleontology; however, this information is currently scattered. The present project aims to compile all fossil occurrences reported in the country by creating an occurrence-based paleobiodiversity database. We collected and compiled three primary data types through an exhaustive bibliographic search and analysis process: (i) references to scientific publications; (ii) listed taxa; (iii) occurrences per taxon. Some of our medium/long-term goals are: (i) the analysis of the taxonomic distribution of Portuguese fossil taxa; (ii) estimation of yearly scientific production indexes; (iii) identification of areas with significant paleontological potential. Presently, we have completed the analysis of 330 scientific publications and recorded a total of 11,011 fossil occurrences. We were able to identify over 380 Portuguese outcrops, with 58,246 fossils attributed to 2,890 taxa.

**Keywords**: Bibliography, review, analysis, taxonomy, occurrences, Portugal.

**Resumo**: Conhecer o passado na ciência, antes de se iniciar um novo estudo, é de importância primordial. Portugal tem mais de dois séculos de investigação em Paleontologia; no entanto, essa informação encontra-se atualmente dispersa. O presente projeto visa compilar todas as ocorrências fósseis registadas no país, através da criação de uma base de dados de paleobiodiversidade baseada em ocorrências. Coletaram-se e compilaram-se três tipos primários de dados através de pesquisa e análise bibliográfica exaustivas: (i) referências a publicações científicas; (ii) taxa listados: (iii) ocorrências por taxon. Alguns dos nossos objectivos a médio/longo prazo são: (i) analisar a repartição taxonómica dos taxa fósseis presentes em território Português; (ii) estimar índices de produção científica anuais; (iii) identificar áreas com elevado potencial de património paleontológico. Até agora, concluímos a análise de 330 publicações, correspondendo a 11 011 ocorrências fósseis. Identificaram-se mais de 380 afloramentos, com 58 246 fósseis atribuídos às 2 890 taxa.

**Palavras-chave**: Bibliografia, revisão, análise, taxonomia, ocorrências, Portugal.

[1] GeoBioTec Research Center, NOVA School of Science and Technology, 2829-516, Caparica, Portugal.
[2] University of Évora, Rua Romão Ramalho 59, 7000-761, Évora, Portugal.
[3] NOVA School of Science and Technology, Campus da Caparica, 2829-516 Caparica, Portugal
[4] Sociedade Portuguesa de Paleontologia, Rua João Luis de Moura, 95, 2530-518, Lourinhã, Portugal.

* Corresponding author / Autor correspondente: prfialho181@gmail.com

## 1. Introduction

Knowledge is defined as the "understanding of or information about a subject acquired by experience or study, either known by one person or by people generally" (Cambridge University Press, 2022). Present-day technology has made knowledge accessible and available to everyone, anywhere, at any time. Tasks that were once herculean feats, such as reviewing the entire bibliography on a given topic, are now relatively more straightforward with the help of digital tools like online search engines, although they are still time-consuming. A database is a type of digital tool that is used to organize in a logical way a collection of digitally stored data, which can then be directly searched and viewed (Meier and Kaufmann, 2019), therefore improving the data search, compilation, and analysis processes. Paleontology greatly benefits from the creation and use of such platforms for data availability, accessibility and sharing with the scientific community and the general public. However, amidst the digitization and organization effort of paleontological data worldwide, we observed a gap regarding the approach of the Portuguese scientific community to the fossil material discovered in the territory. Although there are references to some of these fossils and species in international databases (e.g., The Paleobiology Database), they do not reflect the richness of our geological and paleontological sites.

Therefore, the main objective of the current project is to fill this void by developing an occurrence-based paleobiodiversity database focused on the Portuguese fossil heritage, inspired in the Paleobiology Database structure. With this project, we will also be able to gather information on the paleodiversity recovered in Portugal, and feed accurate data to databases with broader scopes such as: (i) GBIF; (ii) iDigBio; (iii) iDigPaleo (Huang *et al*., 2022). Finally, we plan on contributing to the communication effort of the fossil diversity known so far, while also making advances in the education and scientific research fields.

## 2. Methods

According to Huang *et al*. (2022), most current databases with paleontological data can be divided into the following six categories: (i) taxonomically oriented (authoritative lists of

taxonomy and systematics); (ii) multipurpose biodiversity; (iii) data archives or repositories; (iv) data harvesters or recombiners, which aggregate high volumes of data from external sources; (v) occurrence-based (paleo)biodiversity; (vi) application programming interfaces (APIs). The focus of this project is the design and development of a type (v) database based on regional datasets (extracted from publications), compiled by our team. We drew inspiration from the Paleobiology Database (PBDB), which has been operating since 1998 (Paleobiology Database, 2022).

## 3. Bibliographic search and analysis

Paleodiversity data is available in two primary sources: (i) paleontological collections and (ii) scientific literature. Although we concur with Allmon *et al*. (2018) on the importance of consulting the physical data (*i.e*. fossils) stored in museums and other entities, we consider the bibliographic review process a priority during the first stages of the project. Conducting this bibliographic analysis will not only enhance our previous knowledge of these collections, but also help prepare the project team for future visits to over 30 known public entities in Portugal which house paleontological collections of interest (Mateus, 2020), such as the Geological Museum of the National Laboratory of Energy and Geology (MG-LNEG), the Museum of Lourinhã or the Museum of Natural History and Science of the University of Porto (MHNC-UP).

The bibliographic search has been conducted on data archives, digital search engines, and physical archives (libraries). This way, we can access both recent works that only exist in a digital version and older works that still need to be made available digitally. To guarantee the quality and feasibility of the data collected, we limited the search to scientific publications with some degree of peer review, such as (i) MSc, MRes and PhD dissertations, (ii) congress abstracts, (iii) scientific books, (iv) articles. After identifying these data sources, we applied the methodology described on the workflow present in figure 1.

We extracted three types of primary data from the bibliographic review. The first type comprises the publications of interest taken from the reference list of each work that was analysed, constituting a secondary source of publications for the bibliographic review. The second one includes the listed taxa (overall diversity and new taxa of each publication), and the third refers to the occurrences recorded for each taxon (including geographic coordinates, geochronological data, fossil material, and collections where they were housed). In the framework of this project, we define an occurrence as a record of a particular taxon in a specific geological site and a specific lithostratigraphic unit. Therefore, there can be several occurrences of a taxon in the same location if they originate from different lithostratigraphic layers.

In addition to the primary data types, we found it helpful to keep a record of the data entry date and the name of the reviewer, so that the data input process may be successfully traced back.

The data analysis was conducted on Numbers (version 12.2.1, 2022), the equivalent software to Excel for macOS (version Ventura 13.1, 2022) and comprised simple statistical analysis. The results are here presented in the Preliminary Results section.

## 4. Project phases

Due to the diversity of datasets and the expected high volume of data, we designed a work plan with three major phases. The first phase comprised the concept test, in which the team designed, applied, and refined the methodology represented in figure 1 to maximize data collection. To complete phase 1 of the project, it would be necessary to reach the 10,000 fossil occurrences registered in the database (surpassing the 9,433 Portuguese fossil occurrences recorded in The PaleoBiology Database on 11/05/2022) and analyse 300 scientific publications (at the time 20% of the entire list of references). Phase 1 has now concluded, one year and two months after the beginning of this project.

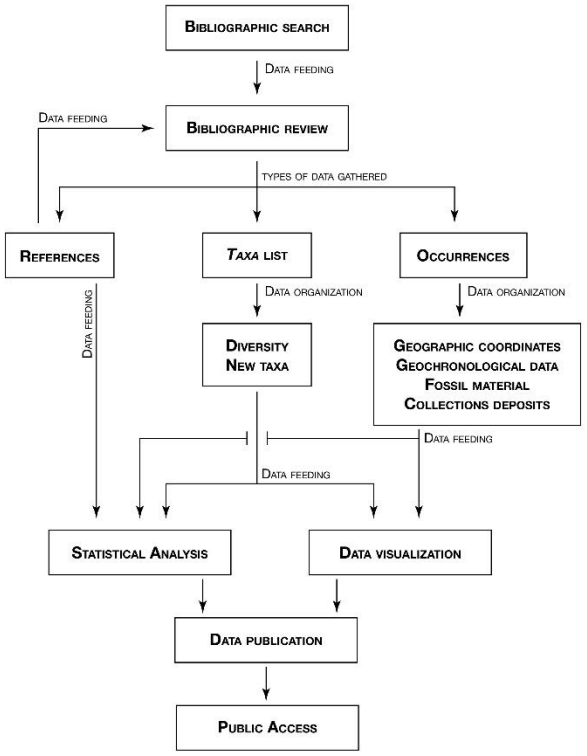



Figure 1. Database building workflow: data gathering, feeding and overall organization processes.

Figura 1. Fluxo de trabalho da construção da base de dados: coleta de dados, povoamento e processos gerais de organização de informação.

The project is currently in the second phase, which consists of scaling the database. We expect this to be the most time-consuming phase, as it encompasses the analysis of the complete list of publications and the data harmonization of all database tables. Although our first estimations pointed to the completion of this phase in three to six years, this is no longer possible due to the high input of new references into the publications list. Presently, the team is able to analyse approximately 0.6 publications per day. With the current total of publications surpassing 1,800, we predict that the bibliographic review will conclude in eight to ten years.

Phase 3 consists in the creation of a platform that will allow the disclosure of and access to all the collected data to the scientific community and the general population, thus concluding this project.

## 5. Data harmonization

According to Huang *et al*. (2022), regional datasets collected by a single user or a group require the establishment of data standards. To achieve data harmonization within the current database and

allow future data integration into other platforms, it is necessary to fill in existent gaps and validate collected data. The literature data includes: (i) the designation conveyed by the authors and the currently accepted designation of a given taxon; (ii) taxonomy; (iii) if the material is fossil or ichnofossil; (iv) GPS coordinates of the studied outcrop; (v) geological dating of the source layer; (vi) the amount of material collected; (vii) legal deposit; (viii) reference of the data source; (ix) input author.

We based the taxonomy data input on division of life in 7 Kingdoms (Animalia, Fungi, Plantae, Chromista, Protozoa, Archaea and Bacteria) in Ruggiero *et al.* (2015a, b) and the classifications provided by the following platforms or digital search tools: (i) The Paleobiology Database (https://paleobiodb.org/#/); (ii) Google Scholar (https://scholar.google.pt/); (iii) Integrated Taxonomic Information System (www.itis.gov); (iv) the Catalog of Life (www.catalogueoflife.org); (v) Mindat (www.mindat.org); (vi) Foraminifera (www.foraminifera.eu); (vii) NOW Database (https://nowdatabase.org/); (viii) WoRMS – World Register of Marine Species (https://www.marinespecies.org/); (ix) Index to Organism Names (ION) (http://www.organismnames.com/); (x) WMSDB – Worldwide Mollusc Species Data Base (https://www.bagniliggia.it/WMSD/WMSDhome.htm); (xi) GBIF – Global Biodiversity Information Facility (https://www.gbif.org/); (xii) Laboratory of Paleobotany (https://paleobotany.ru/); (xiii) FOSSIILID.INFO (https://fossiilid.info/); (xiv) GBD3 Type Fossils (http://www.3d-fossils.ac.uk/home.html); (xv) Encyclopedia of Life (https://eol.org/).

Modern publications often provide the location data needed to accurately pinpoint the location of each fossil occurrence, including GPS coordinates. However, the same cannot be said for older literature. Whenever needed, GPS coordinates of the geological sites were confirmed or obtained through the comparison of the original description of the site with the Portuguese Geological Letters and maps from the following softwares or platforms: (i) Google Earth; (ii) Google Maps; (iii) Mapcarta (https://mapcarta.com/pt/); (iv) Geocaching (https://www.geocaching.com/play); (v) Geossítios (https://geossitios.progeo.pt/); (vi) Mapas On-line by Direção-Geral do Território (http://mapas.dgterritorio.pt/). In the case of dubious descriptions, we opted to consider GPS coordinates of a nearby street or county.

## 6. Data visualization

During preliminary testing, we were able to determine that point clouds would not allow clear visualization of the information due to the vertical superposition of occurrences in geological sites. Therefore, we opted to use density maps for visualizing geographic data concentration. Since one of the cores of this project is the paleobiodiversity geographical occurrences data, we consider that the grid data representation found in ecological studies (Birch *et al*., 2007) is the most adequate to our dataset. According to Birch *et al*. (2007), the rectangular grid is the most frequently used grid type for data density mapping in ecology studies, as seen in the Amphibian and Reptile Atlas of Portugal (Loureiro and Sillero, 2008). However, we opted to use a hexagonal grid since it allows for clear and easy observation of any curvature of the data patterns and thus a more natural representation of the data (ESRI, 2022).

The following software has been used for data visualization within the project's scope: (i) QGIS (version 3.22, 2022); (ii) Adobe Illustrator (version 26.2.1, 2022).

## 7. Preliminary results

Antunes and Balbino (2010) state that Portugal is rich in fossils despite its small geographic extent. With occurrences dating back to the Neoproterozoic (Gonçalves and Palacios, 1984), the geological record of our country encompasses a great fossil diversity, not only in terms of taxonomy but also from a geochronological point of view. The "petrifications" found in Almada mentioned by Bourguet (1742) may represent one of the first references to fossils occurring in Portugal, marking at least 280 years of fossil discoveries in the country.

So far, we were able to compile a list of 1,803 peer-reviewed publications on fossil material collected in Portugal, including MSc, MRes and PhD dissertations, congress abstracts, scientific books and articles. However, we are fully aware that this is only part of the scientific production so far on the topic. It should be noted that the results presented in this publication have a preliminary nature and derive from the bibliographic analysis conducted on 330 of these publications, during the period of May 2021 until October 2022, and thus are not representative of the total fossil diversity and abundance present in the geological record of Portugal. Our goal with sharing these preliminary results is to portray the current status of the project and the information that the data collected in the analysis of the publications mentioned above can provide.

At the moment, we have recorded approximately 58,246 fossils from 11,011 occurrences. These fossils were attributed to 1,927 species and 963 supraspecific identifications (*i.e.*, Genus, Family, Order, Class, Phylum, or Kingdom), and a few occurrences of ichnofossils or fossils classified as incertae sedis. There is a clear bias in the current database due to an initial focus of the bibliographic review effort on specific taxonomic groups of the Animalia kingdom, which totalizes 68.75% of the occurrences recorded so far. Since then, the workflow has been changed to focus on the publication year rather than identified taxa. For the same reason, only 12.02% of the occurrences are ascribed to Plantae and 16.24% to Chromista kingdoms.

Of the 669 fossiliferous sites described in the analysed literature, only 384 were successfully mapped through confirmation or approximation of GPS coordinates (mainland and in the archipelagos of Azores and Madeira). However, due to the absence of precise site descriptions in most of the publications analysed so far, as seen in figure 2, most of the sites pinpointed correspond to geographic approximations (A).

Based on these points, we are able to represent approximately 5,500 fossil occurrences recorded in the database. figure 3 shows a clear difference in the distribution of fossil occurrences, reflecting the authors' intention to study distinct geographic areas of the country. Additionally, we observe a high-density zone of fossil occurrences over the city of Lisbon and Setúbal Peninsula, as a high percentage of analysed publications focus on the Miocene of Lisbon.

The chronological data associated with each fossil occurrence recorded allows us to observe the geographic distribution of geological sites with sediments dated from the Cenozoic (Fig. 4A), Mesozoic (Fig. 4B) and Paleozoic (Fig. 4C). Some of the hexagons represent geological sites that point out occurrences of more recent r older sediments in areas without correspondence in the Geological Map of Portugal, raising uncertainty about their validity. However, these may correspond to outcrops not detectable by the scale of the geological map, sites in areas with "geological accidents", data that was incorrectly recorded, or even data which was incorrect in the original source.

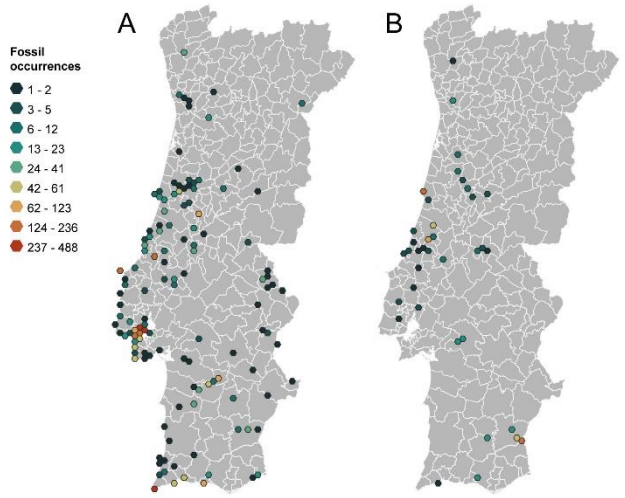



Figure 2. Fossil occurrences geographic mapping: (A) sites without precise GPS coordinates; (B) sites with precise GPS coordinates.

Figura 2. Mapeamento geográfico das ocorrências fósseis: (A) locais sem coordenadas GPS precisas; (B) locais com coordenadas GPS precisas.

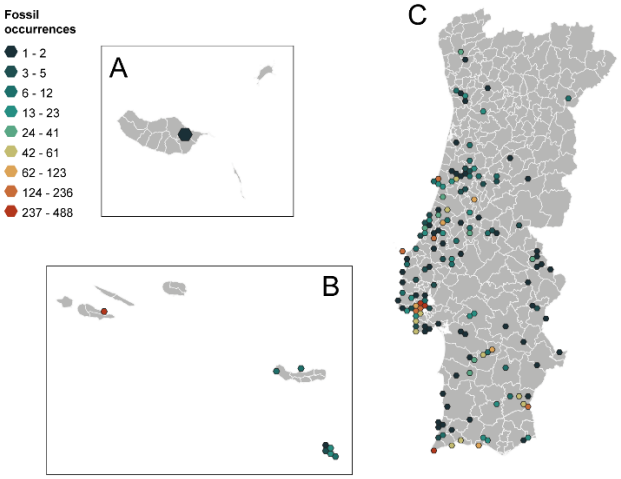



Figure 3. Geographic mapping of the approximately 5,500 fossil occurrences present in the database: (A) Madeira archipelago; (B) Azores archipelago; (C) Portugal mainland.

Figura 3. Mapeamento geográfico das cerca de 5,500 ocorrências fósseis presentes na base de dados: (A) Arquipélago da Madeira; (B) Arquipélago dos Açores; (C) Portugal continental.

## 8. Conclusions

This work represents the early stages of development of an occurrence-based paleobiodiversity database dedicated to the fossil record of Portugal and its associated paleodiversity.

As preliminary results, 58,246 fossils from 11,011 occurrences were accounted for and assigned to 1,927 species and 963 supraspecific taxa. These fossils come from over 660 Portuguese outcrops discovered in mainland Portugal and in the Azores and Madeira archipelagos. However, only 384 have been geographically pinpointed so far. This data corresponds to 330 publications reviewed from 1,808 listed so far. It will still take some time until the relative abundance of each group can be assumed to be representative of the Portuguese fossil record.

As an immediate goal, this project intends to contribute for a better overall knowledge of the Portuguese paleodiversity, namely the total known material and occurrences per taxon, and the species erected in Portugal. As a medium/long-term goals, we aim to progress towards more complex objectives, such as observing the geographic distribution of outcrops, materials, and taxa. This will allow for the identification of distinct paleoenvironments and faunal and botanical associations with an emphasis on less studied groups, and for inferring the areas with significant paleontological potential. The database will also feature information regarding the history of Paleontology in Portugal, with yearly analysis of scientific production and the creation of a checklist of authors, current housing of the Portuguese paleo collections, and type material of species erected in Portugal.

Ultimately, this project will provide accurate and complete data about the Portuguese paleodiversity to broader occurrence-based paleobiodiversity databases and data harvesters/recombiners databases, thus contributing to the improvement of the knowledge of Portuguese Paleontology.

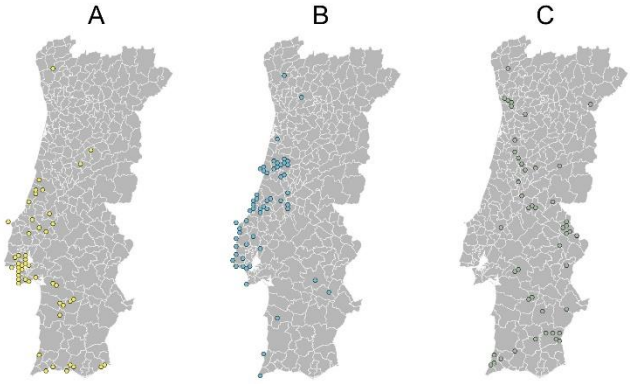



Figure 4. Geochronological distribution of the sites with fossil occurrences: (A) Cenozoic; (B) Mesozoic; (C) Paleozoic.

Figura 4. Distribuição geocronológica dos locais com ocorrências fósseis: (A) Cenozoico; (B) Mesozoico; (C) Paleozoico.

## Acknowledgements

We want to express our appreciation to the SPdP – Sociedade Portuguesa de Paleontologia, which has supported the bibliographic research process since September 2021. We would also like to thank the reviewers for their constructive suggestions.

## References

Allmon, W. A., Dietl, G. P., Hendricks, J. R., Ross, R. M., 2018. Bridging the two fossil records: Paleontology's "big data" future resides in museum collections. *In*: Rosenberg, G. D., Clary, R. M. (Eds.), *Museums at the forefront of the history and philosophy of geology: history made, history in the making*, 535.

Antunes, M. T., Balbino, A. C., 2010. Fósseis de Portugal. *In*: Carvalho, I. S. (Ed.), *Paleontologia*, **3**: 633-659.

Birch, C. P., Oom, S. P., Beecham, J. A., 2007. Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological modelling*, **206**(3-4): 347-359. https://doi.org/10.1016/j.ecolmodel.2007.03.041.

Bourguet, L., 1742. *Traité des petrifications: avec figures*. Paris: Chez Briasson de l'imprimerie de Gissey, 254.

Cambridge University Press (Ed.), *Cambridge Advanced Learner's Dictionary and Thesaurus.* Retrieved May 17th, 2022, from https://dictionary.cambridge.org/dictionary/english/.

ESRI, 2022. *Why hexagons? ArcGIS Pro*. Retrieved October 29, 2022, from https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-whyhexagons.htm.

Gonçalves, F., Palacios, T., 1984. Novos elementos paleontológicos e estratigráficos sobre o Proterozóico português da Zona de Ossa-Morena. *Memórias da Academia das Ciências de Lisboa*, **25**: 225-235.

Huang, H. H. M., Yasuhara, M., Horne, D. J., Perrier, V., Smith, A. J., Brandão, S. N., 2022. Ostracods in databases: State of the art, mobilization and future applications. *Marine Micropaleontology*, 102094. https://doi.org/10.1016/j.marmicro.2022.102094.

Loureiro, A., Sillero, N., 2008. Metodologia. *In*: Loureiro, A., Ferrand de Almeira, N., Carretero, M. A., Paulo, O. S. (Eds.), *Atlas dos Anfíbios e Répteis de Portugal*, 71-80.

Mateus, S. 2020. *Património Paleontológico. O que é, onde está e quais as coleções públicas portuguesas*. Unpublished Ph.D. dissertation, University of Porto, Portugal.

Meier, A., Kaufmann, M., 2019. *SQL and NoSQL databases*. Berlin/Heidelberg, Germany: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-24549-8.

Paleobiology Database, 2022. *Frequently Asked Questions*. Retrieved May 17th, 2022 from https://paleobiodb.org/#/.

Ruggiero, M. A., Gordon, D. P., Orrell, T. M., Bailly, N., Bourgoin, T., Brusca, R. C., Cavalier-Smith, T., Guiry, M. D., Kirk, P. M., 2015a. A Higher Level Classification of All Living Organisms. *PLoS ONE*, **10**. https://doi.org/10.1371/journal.pone.0119248.

Ruggiero, M. A., Gordon, D. P., Orrell, T. M., Bailly, N., Bourgoin, T., Brusca, R. C., Cavalier-Smith, T., Guiry, M. D., Kirk, P. M., 2015b. Correction: A Higher Level Classification of All Living Organisms. *PLoS ONE*, **10**. https://doi.org/10.1371/journal.pone.0130114.