Identifying, cross-referencing, and extracting dark data using GeoDeepDive

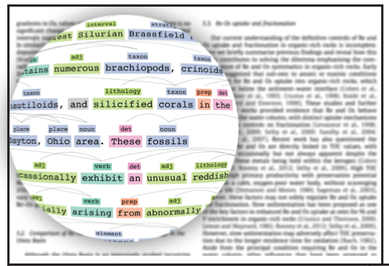
Erika T. Ito, Andrew A. Zaffos, Valerie J. Syverson, Ian A. Ross, and Shanan E. Peters

What is GeoDeepDive?

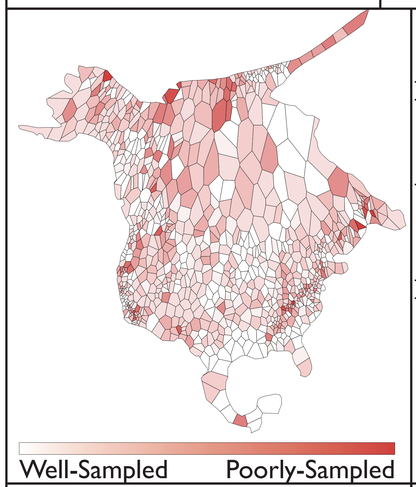
GeoDeepDive is a library of over 3 million machine-readable scientific documents (e.g., journal articles, monographs). It downloads PDFs from partnered publishers (e.g., Elsevier, Wiley, Taylor & Francis), and turns them into machine-readable products using natural language processing, optical character recognition, and other techniques. This processing enables intelligent text-mining and grammatical analysis of scientific documents.

Identifying Dark Data

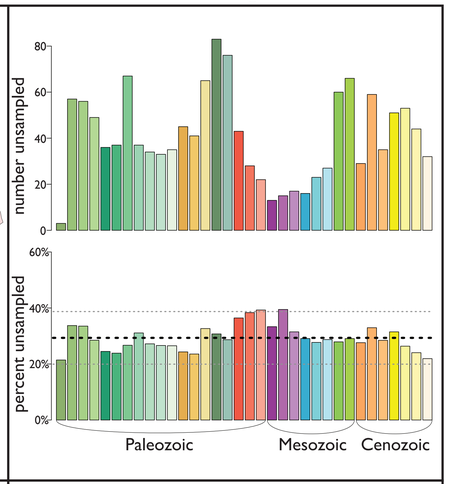
The Paleobiology Database is the largest public repository of global fossil occurrences. However, there are ongoing concerns that certain time-periods, geographic regions, or paleo-environments are systematically understudied. How can we identify understudied areas and test for systematic biases?



1. Identify formal geologic formations not described in the Paleobiology Database.
2. Search documents for the co-occurrence of these formations and keywords indicating they are fossil bearing - e.g., "fossiliferous".
3. Create rules to sift out bad matches, such as using geographic location to verify the correct formation is being referenced.



A map of spatial sampling coverage in the Paleobiology Database relative to the scientific literature.



A barplot of temporal sampling coverage in the Paleobiology Database relative to the scientific literature.

Acknowledgements

Research funded by NSF ICER 1343760 and EAR-1150082

GeoDeepDive is made possible by our partnered publishers: Elsevier, Wiley, Taylor & Francis, SEPM, Canadian Science Publishing, PubMed Central, the United States Geological Survey, the Geological Society of America, the Public Library of Science, and the Treatise of Invertebrate Paleontology.

Cross-Referencing Dark Data


```
"type": "article",
"title": "New pelecinid wasps (Hymenoptera: Pelecinidae) from Upper Cretaceous Myanmar Amber",
"volume": "67",
"journal": "Cretaceous Research",
"link": [
    {"url": "http://www.sciencedirect.com/science/article/pii/S0195667116301409",
     "type": "publisher"},
    {"url": "/data/downloads/Cretaceous Research/S0195667116301409.pdf",
     "type": "filepath"}
  ],
"publisher": "Elsevier",
"author": [
    {"name": "Guo, Lichao"},
    {"name": "Shih, Chungkun"},
    {"name": "Li, Longfeng"},
    {"name": "Ren, Dong"}
  ],
"pages": "84--90",
"number": "1",
"identifier": [
    {"type": "doi",
     "id": "10.1016/j.cretres.2016.07.003"}
  ],
"year": "2016"
```


What if we want to link scientific literature to occurrences in online databases like the Paleobiology Database? Once accomplished, we can enhance existing database entries with additional information extracted from the primary literature.

1. Download bibliographic metadata of target articles matching some initial criterion. For example, all articles mentioning a formal taxonomic name.
2. Build a predictive model and extract assumed model predictors from both the article metadata and the Paleobiology Database - e.g., author, publication year, publication name, article title, or taxonomic names.
3. Fit the model to a training set of known correct and false matches to constrain the correct model coefficients.
4. Use the model to predict the likelihood that a fossil occurrence in the Paleobiology Database refers to a specific article in the GeoDeepDive library.

Classic

The first fossil of Lindsaeaceae (Polypodiales) from the Cretaceous amber forest of Myanmar [doi: 0.1016/j.cretres.2016.12.003]

Collembola (Arthropoda, Hexapoda) from the mid Cretaceous of Myanmar (Burma) [doi: 10.1016/j.cretres.2005.07.003]

New beaded lacewings (Neuroptera: Berothidae) from Upper Cretaceous Myanmar amber [doi: 10.1016/j.cretres.2016.08.007]

The earliest Attagenus species (Coleoptera: Dermestidae: Attageninae) from Upper Cretaceous Burmese amber [doi: 0.1016/0022-2011(68)90160-2]

A new earwig (Dermaptera: Pygidicranidae) from the Upper Cretaceous Myanmar amber [doi: 0.1016/j.cretres.2017.02.012]

The first record of Ichneumonidae (Insecta: Hymenoptera) from the Upper Cretaceous of Myanmar [doi:10.1016/j.cretres.2016.11.001]

Suggest papers with similar attributes in GeoDeepDive.

A new genus with a new species, *Brachypelecinus euthyntus* gen. et sp. nov., and two new species, *Abropelecinus lythus* sp. nov. and *Zoropelecinus perlosus* sp. nov., are described and figured from three exquisitely preserved pelecinid wasps in the Upper Cretaceous Myanmar (Burma) amber.

In this paper, we report a new genus and species, *Brachypelecinus euthyntus* gen. et sp. nov., and two new species, *Abropelecinus lythus* sp. nov. and *Zoropelecinus perlosus* sp. nov., based on three Myanmar amber specimens.

Abropelecinus lythus sp. nov., holotype, CNU-HYM-MA2016002.

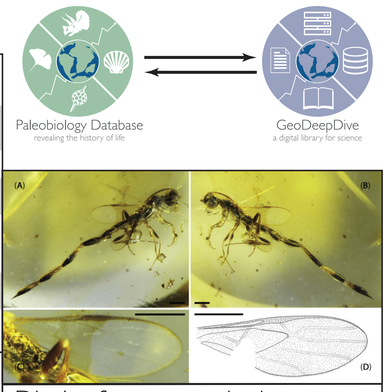
The forewing of *Abropelecinus lythus* sp. nov. (male, body length 4.4 mm) (Fig. 5B) has hyaline membrane with only two tubular veins present (C and R).

For example, the body lengths of amber species range from 4.4 mm in a male specimen of *Abropelecinus lythus* sp. nov. to 12.7 mm in a female specimen *Zoropelecinus zigrasi* Engel & Grimaldi, 2013, while the body lengths of compression fossils range from 3.8 mm in a female specimen *Eopelecinus minutus* Zhang & Rasnitsyn, 2004 to 50.9 mm in a female specimen *Megapelecinus changi*.

A, *Brachypelecinus euthyntus* gen. et sp. nov. B, *Abropelecinus lythus* sp. nov. C, *Zoropelecinus perlosus* sp. nov. D, Zoropelecinus zigrasi. E, *Pelecinopteron tubuliforme*.

The forewing of *Abropelecinus lythus* sp. nov.with hyaline membratene has only two tubular veins present (C and R) , which is the first amber male reported with this type of reduced venation.

Present sentences mentioning each taxon in the text.

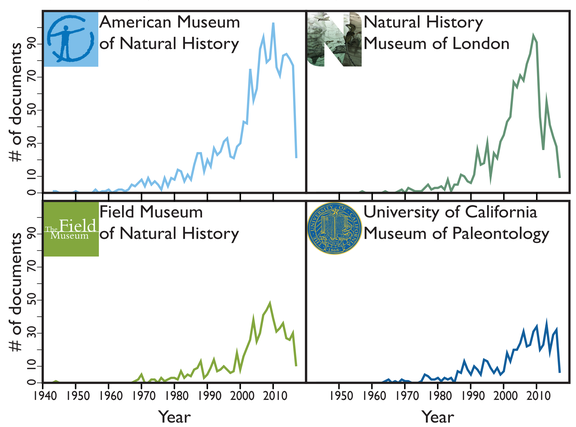


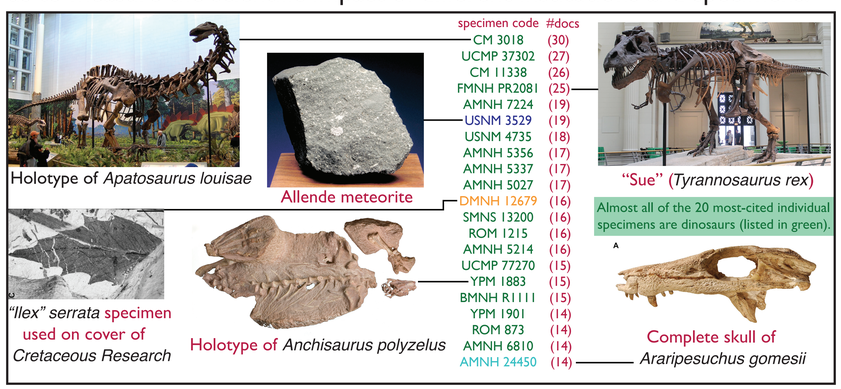
Display figures mentioning a taxon in the text.

Extracting Dark Data

What if we want to build an entirely new database by extracting information from the literature? We illustrate how to build a database of museum specimens. Such a database allows us to measure the scientific impact of different institutions and specimens.

1. Begin with a list of institution names and codes.
2. Index articles containing codes or institution names.
3. Index articles containing taxonomic names.
4. Find the intersect of institution, taxonomic names, and numeric digits within a sentence and evaluate the probability it is a museum specimen code.





www.geodeepdive.org

Outcomes

1. A machine executed assesment of how thoroughly scientific literature is collated by paleontologists and a high-priority list for future data acquisition.
2. A join table linking different databases or database entries to the primary scientific literature.
3. An evaluation fo the scientific impact of different musuem specimens and scientific institutions.