doi.org/10.22498/pages.26.2.78

EarthLife Consortium: Supporting digital paleobiology

Mark D. Uhen[1], S. Goring[2], J. Jenkins[1] and J.W. Williams[2]

The EarthLife Consortium (ELC) aims to support the accessibility, interoperability, and sustainability of paleobiological data across multiple resources. The new ELC Application Programming Interface (API) allows search and retrieval across several databases, and is readily extensible to others.

Paleobiology is a classic example of a 'long-tail' discipline, with the large majority of paleobiological data collected by individuals organized into tight guilds of specialists. Most paleobiologists have a domain of expertise centered on a particular set of organisms (or even on particular fossilized body parts within organisms), a geographic region, and a time period or timescale. For example, one paleobiologist might be an expert on leaves and seeds from the Paleogene of North America (leaving the fossil pollen and other microfossils to other specialists) (e.g. Wing et al. 2009), another might specialize in stable isotope measurements from bones and teeth (e.g. DeSantis et al. 2009), while a third might be a specialist in marine foraminifera, working with ocean-sediment cores collected from across the world (e.g. Barker et al. 2005). These scientists also pursue varied research agendas, both as individuals and research teams.

There is widespread recognition that the whole of the fossil record is greater than the sum of its parts. Many of our discipline's foundational advances – e.g. recognizing five major extinctions in Earth's history; studying speciation and extinction processes during and after extinction events (Raup and Sepkoski 1984; Sepkoski 1997; Peters and Foote 2001); demonstrating the relationship of diversity with climate and productivity variations (Marx and Uhen 2010); demonstrating that species abundances and ranges closely, but individualistically, track climate variations at timescales of 10^2 to 10^5 years during past glacial-interglacial cycles (Huntley and Birks 1983; Webb 1987) – have been made possible by the painstaking synthesis of many individual fossil occurrences into regional- to global-scale databases. Many paleobiological databases exist, some begun and maintained by individual investigators and others that have matured into open, community-curated data resources (CCDRs), with data contributed and stewarded by a broad cross section of the paleobiological community (Uhen et al. 2013; Williams et al. 2018).

The history of cyberinfrastructure development in paleobiology has been "bottom-up", with the attendant advantages and disadvantages. There has been broad and deep participation by paleobiologists in building community-supported cyberinfrastructure.

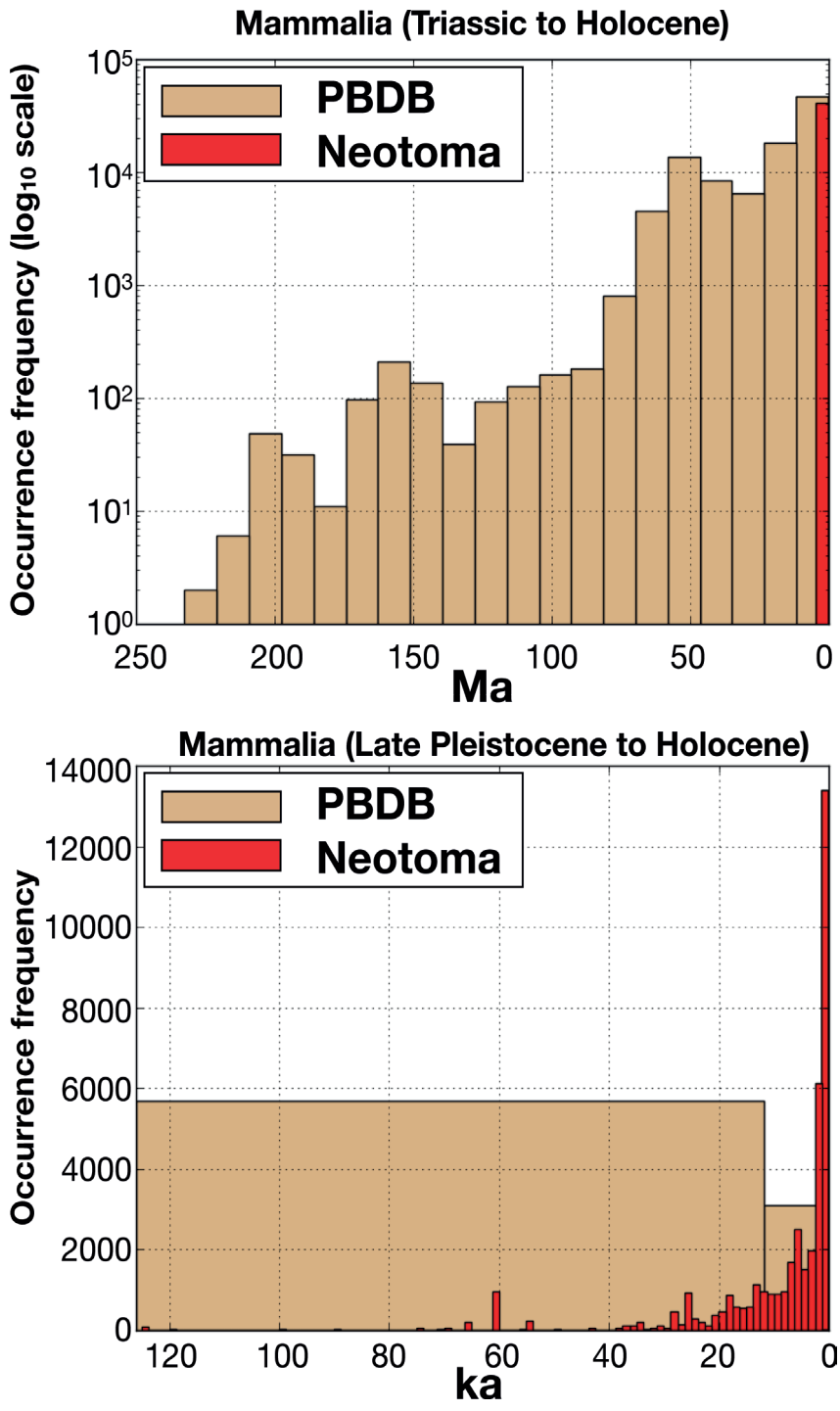

Figure 1: Comparison of the temporal distributions of occurrences of Mammalia in PBDB and Neotoma over **(A)** 250 million years (Ma); and **(B)** 120 thousand years (ka). Note that PBDB has much greater time depth, but that Neotoma has much greater time resolution over a shorter time scale.

Many hard-won lessons have been learned, and well-developed data models have been created to describe paleobiological data in geological contexts. There has also been a proliferation of many small-scale paleobiological resources, with idiosyncratic data and metadata standards and concerns about long-term sustainability of smaller resources.

ELC goals and methods

The ELC project (earthlifeconsortium.org) aims to leverage the long-tail paleobiological data to address large-scale paleobiological questions. Specifically, ELC aims to: improve and expand the interoperability of cyberinfrastructure within the paleobiosciences; promote sharing and use of paleobiological data within paleobioscience and with closely allied geoscience and bioscience disciplines; enhance the sustainability of paleobiological cyberinfrastructure by consolidating smaller resources into larger community-supported repositories; and establish a 4D framework (geography + depth + geologic time) for life and its physical environments that spans all timescales and extends back to the earliest beginnings of the fossil record.

We have advanced towards these goals with the ELC Application Programming Interface (ELC API), which returns data from Neotoma Paleoecology Database (Neotoma, neotomadb.org), which includes paleoecological and co-located paleoenvironmental data at fine temporal grains in the near past, and Paleobiology Database (PBDB, paleobiodb.org), which includes data on all fossil organisms from all of geologic time at coarser temporal grain (Fig. 1). The ELC API is fully documented on Swagger and GitHub, with the capability for extension to other related databases. ELC has already expanded to include occurrence data from the Strategic Environmental Archaeology Database (SEAD; sead.se), demonstrating the ease of database addition to the system. In doing so, we have also established a common data-interchange standard between these resources and contemporary biodiversity databases by adopting the Darwin Core format (Wieczorek et al. 2012) and further extending it for use with additional paleobiological data elements. The ELC project has also supported the incorporation of several smaller databases into Neotoma (Grimm et al. this issue).

Data from the ELC API can be returned either in comma separated value (.csv) text files, or in JSON files for further processing, display, or analysis. We have crafted eight separate endpoints for the API that return datasets based on what data the user is querying. The primary endpoints are: ***Locale***, an intersection of spatial coordinates and geologic time; ***Mobile***, which pre-packages a "light" data set on fossil occurrences for use in mobile applications such as Flyover Country (Myrbo and Loeffler, this issue); ***Occurrence***, which returns a list of occurrences of a given taxon in a specific place and time, including the subtaxa of that taxon (e.g. occurrences of fossils of all species of *Canis*, if given only the genus *Canis*); and ***Taxonomy***, which returns the metadata associated with any given taxon (e.g. ecology, time range, original author, etc.).

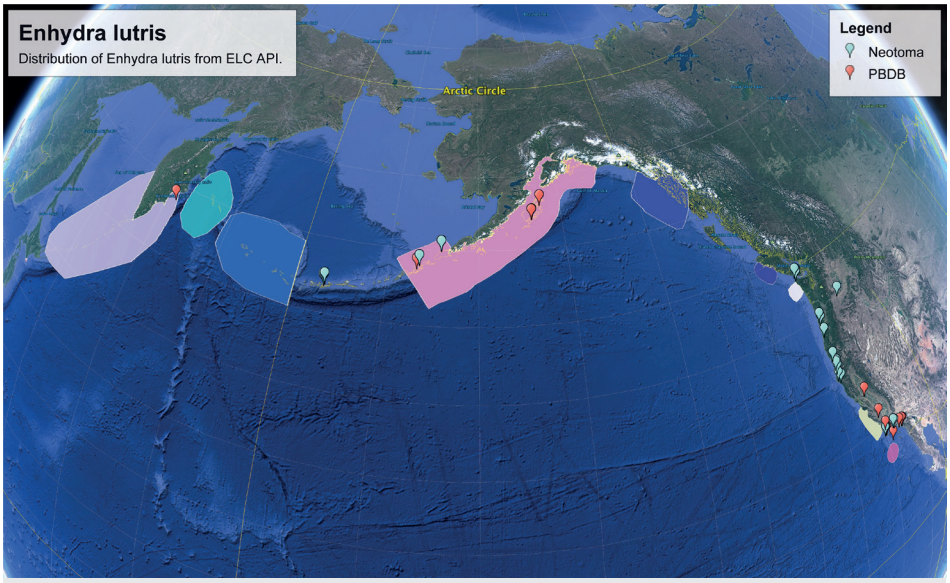

Figure 2: Distribution of the sea otter, *Enhydra lutris*, from the ELC API (All occurrences; Pleistocene-Holocene). Blue points are from Neotoma (n=82), while red points are from PBDB (n=16). Neither database has a complete picture of the distribution of fossil *E. lutris*, but the combined data from ELC more closely resembles the modern distribution of *E. lutris* (from the IUCN Red List), shown in colored polygons representing sub-populations of *E. lutris*, with fossil occurrences demonstrating presences outside the modern range. Base map from Google Earth.

Using these parameters, users can craft queries to answer many questions regarding the distribution and paleoecology of organisms through time and space, from deep geologic time scales, through glacial-interglacial time scales, into the early Anthropocene. For example, the sea otter, *Enydra lutris*, is represented in both PBDB and Neotoma, but neither has a comprehensive view of its distribution in the North Pacific fossil record. Figure 2 shows the occurrences of *Enydra lutris* derived from the ELC API which clearly shows some occurrences from both databases, yielding a much more comprehensive view of its past distribution. While the ELC API returns a limited set of data about each occurrence, end users are able to get further, richer datasets from each constituent database using provided metadata.

ELC Foundation

The Earth Life Consortium Foundation (ELC Foundation) is a non-profit organization currently in its formative stages. The ELC Foundation's missions are to provide easy, free, and global access to scientific data in paleontology, paleoenvironmental studies, and related fields and support the access, development, and sustainability of the community-curated scientific data resources that are the foundation of modern paleobiodiversity science. How best to sustain, develop, and grow these community data resources remains a persistent challenge for the paleogeosciences (Williams et al. 2017). In earlier centuries, professional societies launched peer-reviewed journals as modes of sharing data and knowledge among international networks of scientists. The time may be ripe to extend the mission of professional societies to include the support of high-quality, community-curated scientific data resources. As a starting point, the Paleontological Society and Society for Vertebrate Paleontology have contributed funds to launch the ELC Foundation.

EarthLife Consortium outlook

We welcome the participation by other paleobiological databases and societies in the ELC mission of global access to the full universe of paleobiological data. Others can also participate by joining one of the ELC participating databases, and adding data to these systems which will automatically propagate to ELC. More data in the systems will result in better-supported answers to a wider variety of questions about the history of life on Earth.

ACKNOWLEDGEMENTS

This project was supported by a grant from the US National Science Foundation, ICER 1540997.

AFFILIATIONS

[1]Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, USA
[2]Department of Geography, University of Wisconsin, Madison, USA

CONTACT

Mark D. Uhen: muhen@gmu.edu

REFERENCES

Barker S et al. (2005) Quat Sci Rev 24: 821-834

DeSantis LR et al. (2009) PLoS One 4: e5750

Huntley B, Birks HJB (1983) An atlas of past and present pollen maps for Europe: 0-13000 years ago. Cambridge University Press, 667 pp

Marx FG, Uhen MD (2010) Science 327: 993-996

Peters SE, Foote M (2001) Paleobiology 27: 583-601

Raup DM, Sepkoski JJ Jr. (1984) PNAS 81: 801-805

Sepkoski JJ Jr. (1997) J Paleontol 71: 533-539

Uhen MD et al. (2013) J Vert Paleontol 33: 13-28

Webb T III (1987) Vegetatio 69: 177-187

Wieczorek J et al. (2012) PLoS One 7: e29715

Williams JW et al. (2017) Authorea: 165940

Williams JW et al. (2018) Quatern Res 89: 156-177

Wing SL et al. (2009) PNAS 106: 18627-18632