

UC Merced

Frontiers of Biogeography

Title

The EarthLife Consortium API: an extensible, open-source service for accessing fossil data and taxonomies from multiple community paleodata resources

Permalink

<https://escholarship.org/uc/item/7gs0h1w1>

Journal

Frontiers of Biogeography, 13(2)

Authors

Uhen, Mark D.
Buckland, Philip I.
Goring, Simon J.
et al.

Publication Date

2021

DOI

10.21425/F5FBG50711






Copyright Information

Copyright 2021 by the author(s). This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed



The EarthLife Consortium API: an extensible, open-source service for accessing fossil data and taxonomies from multiple community paleodata resources

Mark D. Uhen^{1*} , Philip I. Buckland² , Simon J. Goring^{3,4} ,
Julian P. Jenkins¹  and John W. Williams^{3,4} 

¹ Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, VA 22030 USA; ² The Environmental Archaeology Lab, Umeå University, 901 87 Umeå, Sweden; ³ Department of Geography, University of Wisconsin-Madison, Madison, WI 53706 USA; ⁴ Center for Climatic Research, University of Wisconsin-Madison, Madison, WI 53706 USA. *Corresponding author: Mark D. Uhen, muhen@gmu.edu

Abstract

Paleobiologists and paleoecologists interested in studying biodiversity dynamics over broad spatial and temporal scales have built multiple community-curated data resources, each emphasizing a particular spatial domain, timescale, or taxonomic group(s). This multiplicity of data resources is understandable, given the enormous diversity of life across Earth's history, but creates a barrier to achieving a truly global understanding of the diversity and distribution of life across time. Here we present the Earth Life Consortium Application Programming Interface (ELC API), a lightweight data service designed to search and retrieve fossil occurrence and taxonomic information from across multiple paleobiological resources. Key endpoints include Occurrences (returns spatiotemporal locations of fossils for selected taxa), Locales (returns information about sites with fossil data), References (returns bibliographic information), and Taxonomy (returns names of subtaxa associated with selected taxa). Data objects are returned as JSON or CSV format. The ELC API supports tectonic-driven shifts in geographic position back to 580 Ma using services from Macrostrat and GPLates. The ELC API has been implemented first for the Paleobiology Database and Neotoma Paleocology Database, with a test extension to the Strategic Environmental Archaeology Database. The ELC API is designed to be readily extensible to other paleobiological data resources, with all endpoints fully documented and following open-source standards (e.g., Swagger, OGC). The broader goal is to help build an interlinked and federated ecosystem of paleobiological and paleoenvironmental data resources, which together provide paleobiologists, macroecologists, biogeographers, and other interested scientists with full coverage of the diversity and distribution of life across time.

Highlights

- Paleodiversity data are essential to understanding the ecology, evolution, and biogeography of life over Earth's entire history.
- The Earth Life Consortium Application Programming Interface (ELC API) provides an easy-to-use framework for the automated search and retrieval of paleobiodiversity data across several existing community-curated data resources (CCDRs).
- These tools can currently access the Paleobiology Database and Neotoma Paleocology Database, with a test extension to the Strategic Environmental Archaeology Database, and have the capability to be expanded to other existing and future paleobiodiversity resources.
- All ELC APIs follow open-source standards and are fully documented, supporting their use by scientists seeking to find data and developers seeking to add their resources to the ELC API umbrella.

Keywords: community-curated data resource (CCDR), cyberinfrastructure, database, informatics, paleobiodiversity, paleobiogeography, paleoecology, taxonomy.

Introduction

Study of the patterns and processes governing the diversity of life on earth at long timescales and broad spatial scales requires the assembly of many individual fossil occurrences into larger, open, community-curated data resources (CCDRs; Williams et al. 2018a) such as the Paleobiology Database (PBDB), the Neotoma Paleoecology Database, and others (Uhen et al. 2013). In an era of global change, when stewarding biodiversity is urgent (IPBES 2019), conservation biologists, global change ecologists, paleontologists and other Earth system scientists use the geological record to study biodiversity dynamics during large and rapid transitions (National Research Council 2005, 2011, 2012, Willis and Birks 2006, Dietl and Flessa 2011, Willis and MacDonald 2011, Fritz et al. 2013, Kidwell 2015, Fordham et al. 2020). For example, large paleodata syntheses are used to understand how contemporary ecological systems are shaped by historical legacies of slow-acting processes (e.g., Whittaker et al. 2001, Jablonski 2008), test the ecological forecasting models used to project and prepare for the impacts of 21st-century climate change (e.g., Veloz et al. 2012, Blois et al. 2013), assess the patterns and causes of abrupt ecological and environmental change (Williams et al. 2010, Shanahan et al. 2015, Shuman et al. 2019), constrain phylogenetic models of species divergence and rates of evolution (Muller & Reisz 2005), assess the novelty of contemporary ecosystems relative to historic or deeper-time baselines (Jackson & Williams 2004, Radeloff 2015), and understand the fundamental processes that generate, maintain, and rebuild biodiversity (Crame 2001, Jablonski et al. 2013).

These open paleodata resources also make paleobiological data accessible to scientists from allied disciplines, powering the next generation of convergent research. For example, the fossil record is used by sedimentologists and economic geologists studying facies relationships and employing biostratigraphic controls for correlating rock strata (Metcalf & Nicoll 2007), structural geologists and geophysicists seeking biogeographic constraints on reconstructions of former tectonic plate positions (Chaloner & Creber 1988, Wright et al. 2013), paleoclimatologists building proxy-based reconstructions of past climates (Bartlein et al. 2011, Marsicek et al. 2018, Routson et al. 2019), and archaeologists seeking to understand how past societies shaped and were shaped by their environment (e.g., O'Regan et al. 2011, Kohler et al. 2018).

In response, many paleoecological and paleobiological data resources have emerged over the years, of varying size and scope, some begun and maintained by individual investigators and others maturing into publicly available, community curated data resources (Williams et al. 2018b), with data contributed and curated by a broad cross-section of the paleobiological community (Uhen et al. 2013, Williams et al. 2018a). The PBDB, launched in 1998 to study global biodiversity dynamics across the history of life, is a global-scale data resource with holdings across the Phanerozoic to present and a temporal grain on the order of 10^6 years. The Neotoma Paleoecology

Database, a coalition of constituent databases that use a common database platform (Grimm et al. 2018, Williams et al. 2018), emphasizes records from the Late Neogene to present and temporal grains of 10^1 to 10^3 years and has multiple origins, often linked to efforts to reconstruct past climates, test climate models, and map species responses to environmental change (e.g., COHMAP 1988, FAUNMAP 1996, Harrison et al. 2013, Grimm et al. 2018). The PBDB and Neotoma together have been cited over 50,000 times, with H-factors of 94 and 79, respectively. Many other paleobiological data resources exist, of varying size and scope, including the New and Old World Database (NOW), the Strategic Environmental Archaeology Database (SEAD), Neptune, and others (Uhen et al. 2013).

The next stage of evolution is to consolidate or federate paleodata resources. Consolidation, in which data from one data resource are added to another, is a good solution for data resources with simpler data models or that are unlikely to persist on their own, e.g., if the lead investigator(s) retire or move to other projects. Both the PBDB and Neotoma have grown in part through consolidation. PBDB has incorporated data from the Evolution of Terrestrial Ecosystems (ETE) project, and several research projects have housed their data in the PBDB from the start instead of creating standalone databases. In addition, PBDB has incorporated several large datasets that had been stored in various off line formats and made them available to all. Many of these are now downloadable as PBDB Data Archives (<https://www.paleobiodb.org/classic/app/archive/list>). Constituent Databases within Neotoma include FAUNMAP (Graham et al. 1996), the European Pollen Database (Fyfe et al. 2009), the Neotoma Ostracode Database (Curry et al. 2013), the Diatom Database of the Academy of Natural Sciences (Sullivan & Charles, 1994), Neotoma Testate Amoebae Database (Amesbury et al. 2018), and others (Williams et al. 2018b).

However, for some data resources, consolidation may not be feasible if data models have reached a level of complexity that precludes simple merging of semantic and ontological schema, if sustained funding requires maintaining a standalone identity (e.g., for national-scale data resources), or if consolidation would disrupt the linkage between a CCDR and the community that it supports. The last consideration is perhaps the most critical; the ultimate guarantor of sustainability of all CCDRs is close engagement with and support of their networks of data contributors and stewards.

This persistent multiplicity of paleodata resources, although understandable, presents a challenge for macroecologists, biogeographers, and other scientists seeking broad-scale, integrative understanding of the diversity and distribution of life. Simply discovering all pertinent paleodata resources is a challenge, and each has its own data schema, which hinders integration and understanding. Because paleodata resources often focus on particular spatiotemporal domains and have been assembled by different networks of researchers,

fossil occurrences for given taxonomic groups may be distributed across paleodata resources, with poorly characterized gaps and overlaps.

Here we describe a new resource, developed in partnership by the PBDB and Neotoma and open to all, called the Earth-Life Consortium Application Programming Interface (ELC API). The API is designed to be a common lightweight data standard and associated web services for discovering and obtaining fossil occurrence data from across multiple paleodata resources. A series of API endpoints enable retrieval of different kinds of data. The project is completely open source and ELC API code is designed to be readily extensible to other paleoecological and paleobiological resources.

The ELC API: Overview and Design Process

The ELC API is a composite API that generates and dispatches queries to multiple paleobiological data resources, via subqueries directed to the native API for each resource. Basic operation of the ELC API is illustrated in Fig. 1. The resulting data returns are processed and reformatted by the ELC API to provide the end user with comparable data objects. The ELC API is intentionally designed to be lightweight, with a fairly small number of endpoints and expected parameters. This design supports the goal of searching and returning data from multiple paleodata resources, each with its own particular data model, semantics, and ontology. Results are returned as aggregated

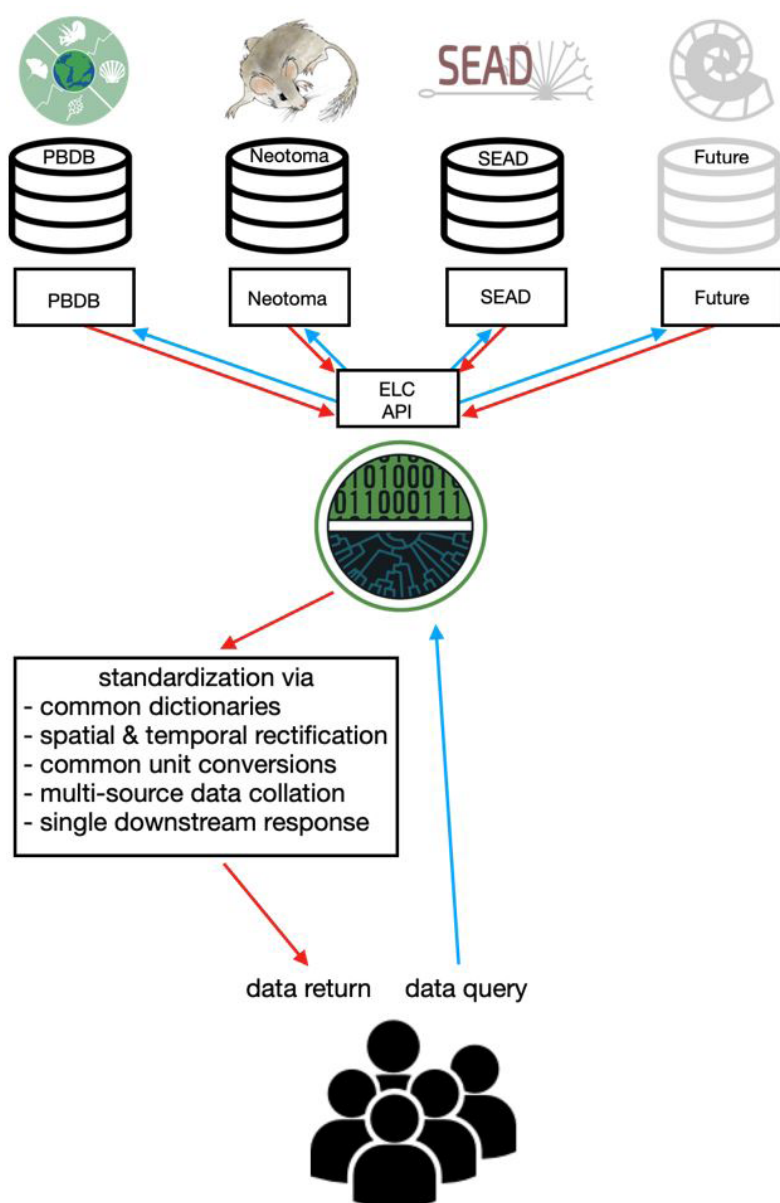


Figure 1. Basic operation of the ELC API. In the diagram, data queries are shown in blue, while data returns are shown in red. The ELC API takes a single query from an end user, and sends it to constituent database APIs. Then, it takes the data returns and standardizes them into a single data return to the end user.

JSON or CSV objects. Currently, the ELC API allows users to query data from the [Paleobiology Database \(PBDB\)](#): [Neotoma Paleocology Database \(Neotoma\)](#): and [Strategic Environmental Archaeology Database \(SEAD\)](#), but the ELC API is easily extensible to additional databases (<http://earthlifeconsortium.org/docs/api-extend.html>).

Design and development of the ELC API followed a user-centered and “API first” development process, that emphasizes careful consideration of how to robustly represent and access information before application development. This approach consists of the following steps:

- 1) Developers and paleobiologists from the Neotoma and PBDB teams met to review the data models of the existing data resources and native APIs, identify semantic commonalities, and points of divergence. Common research queries were identified by paleobiologists and translated by the developers into sketches of API endpoints, each consisting of parameters to be passed to the endpoint and the structure of data objects to be returned by the endpoint.
- 2) The interface schema for the ELC APIs was then established and designed to be compliant with the OpenAPI standard v2.0 (<https://swagger.io/specification/v2/>). Endpoints were generated for key data routes (https://github.com/EarthLifeConsortium/elc_api/blob/master/swagger_server7/swagger/swagger.yaml), and routing of API queries was automatically handled by a Swagger-compliant interface library.
- 3) Server backend code was developed using Python and Flask for each generated endpoint.
- 4) The OpenAPI Python library flask-swagger (<https://pypi.org/project/flask-swagger/>) was generated for browsing the API documentation and testing parameter calls.

This API-first process enabled the developers and scientists to stay closely engaged throughout the development process and for scientific users to quickly test and suggest modifications to APIs. Changes to the API often were made in the schema during development and the changes pushed down through the frontend to backend code.

ELC API Endpoints

Overview

Each ELC API endpoint returns a specific suite of formatted data from all participating data resources, with users having the option to choose subsets of participating resources. API versioning is supported through the use of ‘api_v1’ in the directory path; all descriptions here are for Version 1.0 of the API. All API endpoints and parameters are fully documented at (<http://earthlifeconsortium.org/docs/api-docs.html>) and an interactive ‘sandbox’

is available for testing and designing API queries (http://earthlifeconsortium.org/api_v1/ui/). Here we briefly summarize the features of each endpoint likely to be of particular interest to biogeographers and macroecologists.

Occurrences (occ)

Base Path: http://earthlifeconsortium.org/api_v1/occ?

Parameters: `taxon` (name or comma-separated list of names), `bbox` (wkt polygon), `agerange`, `ageunits` (‘ybp’, ‘ka’, ‘ma’), `coordtype` (‘modern’, ‘paleo’), `includelower` (true,false), `output` (json, csv), `limit`, `offset`, `show` ([‘full’, ‘idx’, ‘poll’]), `run`

Notes:

Occurrences are the individual instances of fossils in time and space. Occurrences of taxa can be specified at any taxonomic level, bounded by any units of time and spatial delimitation. Taxon names can use wild cards (%). Geographic search parameters are described using the well-known text (WKT) standard for describing polygons developed by the Open Geospatial Consortium. The WKT implementation is compliant with ISO 19125-1 and 19125-2 standards. The Open Street Map Playground (<https://clydedacruz.github.io/openstreetmap-wkt-playground>) graphical interface can be used to create and define search polygons. For all ELC APIs, geographic coordinates are expected to be decimal degrees, ranging from -90 (S) to 90 (N) and -180 (W) to 180 (E).

The ELC API supports spatial searches by both modern spatial coordinates and paleogeographic coordinates, using the `coordtype` parameter. Temporal searches can employ named geologic ages using definitions from the International Commission on Stratigraphy (ICS, Gradstein et al. 2012) or minimum/maximum age ranges, using units of ybp, ka, or Ma (years before present, thousands of years (kiloanna) before present, or millions of years (megaanna) before present). For example, a user could request all fossil occurrences of the Holocene, all occurrences from the Eocene to Pliocene, all occurrences from 12 to 15 ka, or from 1 Ma through the Miocene. Other parameters include `run` (to choose which data resources are queried), `limit` and `offset` (to limit the number of data objects returned and enable serial queries), and `show` (to allow full dataset returns, ID numbers only, or summary statistics only).

Note that the seemingly simple concept of ‘occurrence’ is a foundational point of semantic divergence between the PBDB and Neotoma that required special handling when building data returns. In the PBDB, unique identifiers are assigned to individual fossil occurrences, because the PBDB was originally designed a store of species occurrences in the stratigraphic record, extracted from the literature. In Neotoma, unique identifiers are assigned to samples, based on a data model in which samples are collected from cores and stratigraphic sections (Williams et al. 2018b). Each sample can consist of one fossil specimen (e.g., a single canid femur) or multiple taxa, e.g., separate counts of individual micropaleontological taxa. The ELC API resolves this difference by returning the unique occurrence ID for each PBDB fossil occurrence and a composited unique identifier for Neotoma occurrence that combines the sample identifier and taxon identifier.

At this time, the ELC API returns presence-only information for each fossil occurrence and does not return information about the abundance of individual fossil occurrences; users desiring abundance information should use the native APIs.

Examples:

http://earthlifeconsortium.org/api_v1/occ?taxon=canis+dirus (retrieve all occurrences of dire wolf across all participating data resources)

http://www.earthlifeconsortium.org/api_v1/occ?taxon=pinus&agerange=15000%2C10000&ageunits=ybp&includelower=false&limit=10 (retrieve the first 10 instances of pine between 15,000 and 10,000 years ago)

Locales (*loc*)

Base Path: http://earthlifeconsortium.org/api_v1/loc?

Parameters: *idlist*, *bbox*, *agerange*, *ageunits*, *coordtype*, *limit*, *offset*, *output*, *show*, *run*

Notes:

The Locales endpoint returns information about sites or locations that contain fossil samples. Locales can be searched for using polygons or age ranges, as described for occurrences. The *idlist* parameter also allows locales to be found for lists of collection IDs (PBDB) or dataset IDs (Neotoma), using the format `[database]:[datatype]:id_number,` All other parameters follow the format for the occurrences endpoint.

Examples:

[http://earthlifeconsortium.org/api_v1/loc?bbox=POLYGON\(\(139.8%20-33.7,%20150.1%20-33.7,%20150.1%20-39.1,%20139.8%20-39.1,%20139.8%20-33.7\)\)](http://earthlifeconsortium.org/api_v1/loc?bbox=POLYGON((139.8%20-33.7,%20150.1%20-33.7,%20150.1%20-39.1,%20139.8%20-39.1,%20139.8%20-33.7)))

(retrieve all locales within a polygon roughly covering the Australian state of Victoria)

http://earthlifeconsortium.org/api_v1/loc?idlist=pbdb:col:9191,neot:dst:998

(return the location metadata for PBDB collection ID 9191 (Davis Fm, Elvins Group, MO) and Neotoma dataset ID 998 (Gytgykai Lake). Note that in this example, '%20' is an escape code used to indicate a space break. Resources such as the ELC API Sandbox and Open Street Map Playground will automatically insert these escape codes where needed.

References (*ref*)

Base Path: http://earthlifeconsortium.org/api_v1/ref?

Parameters: *idlist*, *output* (*json*, *csv*, *bibjson*, *ris*), *show*, *run*

Examples:

http://earthlifeconsortium.org/api_v1/ref?idlist=pbdb:col:60676,neot:dst:5664,neot:dst:5562,pbdb:col:105101

(Provide a list of four reference IDs, each prefaced by a string representing the holding data resource; references will be returned in default BIBJSON format)

Notes:

References returns bibliographic information about the publications stored in paleobiological data resources that are linked to fossil records. References

can be returned in JSON, CSV, BIBJSON, or RIS format (specified in output parameter) with BIBJSON. The *idlist* parameter format follows that for the locales endpoint. Because the *idlist* parameter requires knowledge of database-specific ID numbers for publications, the references API is usually called programmatically, given knowledge of publication IDs provided by the occurrence data return.

Taxonomy (*tax*)

Base Path: http://earthlifeconsortium.org/api_v1/tax?

Parameters: *taxon* (name or comma-separated list of names), *idlist*, *includelower* (*true*, *false*), *output* (*json*, *csv*), *show* (*all*, *poll*, *idx*), *run* (all or list of database names)

Notes:

The Taxonomy endpoint reveals the taxonomic names and hierarchies stored inside paleobiological data resources. Taxonomic name requests can be sent as either as a list of taxon IDs (using *idlist*) or as one or more taxonomic names (*taxon*). If *includelower* is set to *true* (default is *false*), Taxonomy will return all species or sub-species names within a named taxon (PBDB) or will run wildcard searches (Neotoma). If the taxonomic name does not occur in the given data resource, the ELC API will not return any data from that database. Taxonomic concepts may differ among included data resources.

The PBDB and Neotoma handle taxonomies and taxon names differently. The PBDB, which draws taxonomic names directly from the published literature, allows multiple taxonomic names to be stored for the same taxon and is dynamically updated by data authorizers (Peters & McClennen, 2016). The most recent name entered for a given name is used as the current taxon name. The PBDB also employs a rank-ordering system with taxonomic names assigned to levels of species, genus, etc. Neotoma uses defined vocabularies of taxonomic names in which Data Stewards can propose the addition or modification of taxonomic names and designated Taxonomic Experts approve these additions and modifications taxonomic names (Williams et al. 2018b). Taxonomic names in Neotoma can include information about fossil morphology or taxonomic uncertainty, e.g., Poaceae (<50µm), *Odocoileus* cf. *O. virginianus*, or *Ambrosia*-type. Because most taxa in Neotoma are still extant, Neotoma attempts when possible to link to current taxonomic authorities, e.g., using phylogenetic-based classification for plants (e.g., Cantino et al. 2007). Taxonomic names in Neotoma currently are not placed into rank-order designations of species, genus, etc.

Examples:

http://earthlifeconsortium.org/api_v1/tax?taxon=Carcharocles&includelower=true

(Return all taxon names that are children of the taxon *Carcharocles*)

Miscellaneous API Endpoints

The ELC API also supports a number of miscellaneous utility endpoints. These are all located within the misc pathway.

Paleocoordinates (*paleocoord*)

Base Path: http://earthlifeconsortium.org/api_v1/misc/paleocoords?

Parameters: *coords* (latitude, longitude), *age*, *ageunits* (ybp,ka,ma)

Notes:

Converts modern day cartesian coordinates into paleocoordinates using the GPlates (Wright et al. 2013) tectonics reconstruction model. Operationally this model is hosted by the Macrostrat Project (Peters et al. 2018) at the University of Wisconsin. As above, geographic coordinates are expected to be decimal degrees, ranging from -90 (S) to 90 (N) and -180 (W) to 180 (E). Age can be either an integer, representing a single age, or a named geological age.

Examples:

http://www.earthlifeconsortium.org/api_v1/misc/paleocoords?coords=21%2C80&age=66&ageunits=Ma

Timebound

Base Path: http://earthlifeconsortium.org/api_v1/misc/timebound?

Parameters: *agerange*, *ageunits* (ybp,ka,ma)

Notes:

Return the oldest and youngest ages spanning the specified range. Age range requests can be passed as individual or pairs of geologic ages or numeric values, or as a single geological age. Geologic ages are resolved according to ICS definitions (Gradstein et al. 2012).

Examples:

http://earthlifeconsortium.org/api_v1/misc/timebound?agerange=late+cretaceous,selandian

(return the earliest and latest age bounds from the Late Cretaceous epoch to the Selandian stage of the Paleocene, using the default age units of Ma)

http://earthlifeconsortium.org/api_v1/misc/timebound?agerange=triassic&ageunits=ybp

(return the earliest and late age bounds of just the Triassic, in years before present)

http://earthlifeconsortium.org/api_v1/misc/timebound?agerange=2,miocene

(return age bounds from the beginning of the Miocene to 2 Ma).

Subtaxa

Base Path: http://earthlifeconsortium.org/api_v1/misc/subtaxa?

Parameters: *taxon* (name or comma-separated list of names), *synonyms* (true, false)

Notes:

Return a list of all taxonomic names hierarchically below the specified taxon, optionally including synonyms. As with occurrence, a single name or lists of taxon names can be passed in via *taxon*, with the % wildcard also allowed. Subtaxa defaults to returning synonyms.

Examples:

http://earthlifeconsortium.org/api_v1/misc/subtaxa?taxon=tyrannosauridae&synonyms=false

(retrieve all subtaxa of Tyrannosauridae excluding synonyms.)

Mobile

Base Path: http://earthlifeconsortium.org/api_v1/misc/mobile?

Parameters: *taxon* (name or comma-separated list of names), *bbox*

Notes:

This is a custom lightweight endpoint designed for use with Flyover Country (Loeffler 2018, Myrbo et al. 2018) and other mobile apps. Mobile only requires two parameters (*taxon* name(s) and geographic polygon) and returns a combination of occurrence data with associated taxonomic and select environmental details. As above, wildcard operators are permitted. The response is nested JSON with a highly compact vocabulary.

Examples:

[http://earthlifeconsortium.org/api_v1/misc/mobile?taxon=Dasyprocta,Borreria&bbox=POLYGON\(\(-73.037109375%20-6.839169626342809%2C-64.423828125%202.635788574166611%2C-51.76757812500001%202.1088986592431382%2C-36.298828125000014%20-6.31529853833004%2C-52.29492187499999%20-32.546813173515154%2C-57.56835937500001%20-16.636191878397653%2C-73.037109375%20-6.839169626342809\)\)](http://earthlifeconsortium.org/api_v1/misc/mobile?taxon=Dasyprocta,Borreria&bbox=POLYGON((-73.037109375%20-6.839169626342809%2C-64.423828125%202.635788574166611%2C-51.76757812500001%202.1088986592431382%2C-36.298828125000014%20-6.31529853833004%2C-52.29492187499999%20-32.546813173515154%2C-57.56835937500001%20-16.636191878397653%2C-73.037109375%20-6.839169626342809)))

(retrieve all occurrences of the extant plant *Borreria* and the rodent *Poaceae* in Brazil).

Use Case Examples

The best use cases available to demonstrate the utility of the ELC API are those where both databases have significant numbers of occurrences of the same fossil taxon. For Neotoma and the PBDB, the area of greatest shared holdings is for terrestrial vertebrates. Given that Neotoma has a heavy emphasis on the Late Neogene (and particularly the Quaternary), and PBDB covers all of deep time, the Pleistocene is the areas of greatest temporal overlap. As other CCDRs join the ELC API, additional taxonomic, temporal, and spatial parameters will be key to producing data sets blended from several resources.

The sea otter *Enhydra lutris* was used by Uhen et al. (2018) to demonstrate contributions from both Neotoma and PBDB. Here we show a similar return for the polar bear, *Ursus maritimus* using the ELC API. Fig. 2 shows a map of modern *U. maritimus* distribution from OBIS compared to the distribution from Neotoma and PBDB downloaded with the ELC API. Notice that the fossil distribution, primarily from the Pleistocene, shows *U. maritimus* much farther to the south, particularly in Europe and Asia, where they are unknown today. Also note that neither Neotoma, nor PBDB has full coverage of the fossil range of polar bears, but together, a much fuller and clearer picture



Figure 2. Distribution of modern and fossil polar bears, *Ursus maritimus*. Neotoma occurrences are shown as light blue pins, PBDB occurrences are in red, while modern occurrences are in green Xs. Data on modern occurrences were downloaded from OBIS on May 28, 2019. Note how much farther south many fossil occurrences are compared to modern occurrences. Neither Neotoma nor PBDB fully capture the entire fossil distribution of *U. maritimus*, but together they can. Base map from Google Earth. API Query: http://earthlifeconsortium.org/api_v1/occ?taxon=Ursus%20maritimus&output=csv

of the broad distribution of this currently threatened species is available.

Another use case that demonstrates the complementarity of Neotoma and PBDB data in geologic time is shown in Fig. 3, which shows the distribution of occurrences of the Family Canidae. Neotoma is strong in shallow time, while PBDB is strong in deep time. To understand the full distribution of this taxon, both sets of data are necessary, and the ELC API gives access to both.

Discussion

Benefits to scientists and other end users

Paleobiological data are hard-won, requiring substantial field and lab time and deep taxonomic expertise in the identification of fossil specimens. The paleobiological data and knowledge gathered by CCDRs such as Neotoma and the PBDB represents decades to centuries of accumulated data and knowledge and hundreds of millions of dollars of

scientific investment. Hence, these data resources are foundational infrastructure for the paleobiosciences.

The ELC API improves and expands the interoperability of cyberinfrastructure within the paleobiosciences. It also promotes the sharing and use of paleobiological data within and outside the discipline but especially within closely allied geoscience and bioscience disciplines (Uhen et al. 2018, Williams et al. 2018a). This multiplies the usefulness of hard-won fossil occurrence data that has been accumulated by paleontologists for decades. The ELC API also builds interoperability between this paleobioscience CI and current and emergent CI in the biosciences, particularly with respect to networks of biodiversity and ecological databases. Finally, it helps to establish a 4D framework for life and its physical environments at all scales of time.

Extending the ELC API to Other Data Resources

The ELC API is designed to be readily extensible to other data resources, and its strength grows

with the addition of other paleoecological and biological databases. In turn, joining the ELC API helps increase the discovery and citation of participating paleodata resources. All ELC API code makes use of open-source standards (e.g., Swagger, OGC) and is available on GitHub. Research groups and developers are encouraged to fork the GitHub ELC repository (https://github.com/EarthLifeConsortium/elc_api) and follow the steps described below to add new paleodata resources to the ELC API.

The ELC API is an intentionally lightweight data service, because keeping the number of endpoints and codebase small facilitates adoption by other paleobiological data resources. However, this design philosophy means that some kinds of searches may not be able to be performed by the ELC API or some kinds of data returns may not be available. In these cases, users desiring more detailed data returns can use the native Neotoma and PBDB API endpoints, R packages, and other associated resources.

It is possible to link other data resources to the ELC API by adding a customized “handler” file (https://github.com/EarthLifeConsortium/elc_api/tree/master/swagger_server/handlers) that serves to translate the output of the data resource’s own API, and return results in the format expected by the ELC API for each individual endpoint. The use of handlers means that onboarding new resources can happen relatively easily, without changing the underlying framework.

As an initial test of the extensibility of the ELC API, the Occurrence endpoint has been extended to retrieve data from the [Strategic Environmental Archaeology Database](#) (SEAD, Buckland et al. 2018).

The ELC API is denoted by major version within the base path itself and all sub-versions of the API will not break parameter, route or response compatibility. If, in the future, the API is expanded or modified in such a way that backwards compatibility is untenable (which is not anticipated at this time), a new major version will be launched. Earlier versions will remain available.

Extension of the ELC API to other paleobiological data resources requires that they have their own native APIs for accessing internal data holdings. Not all yet do. In these cases, the ELC APIs may offer a useful design template and starting schema for developers of other paleobiological data resources.

We plan to add access to museum collection data from [iDigBio](#) via the [ePANDDA](#) API in the near future. This will allow users to query both published and unpublished museum data sets with single queries.

The ELC API does not attempt to automatically test whether duplicate fossil occurrences are returned by data resources. At present, this is not a major concern because the PBDB and Neotoma have largely non-overlapping data holdings, with the possible exception of some overlap in Miocene and Pleistocene terrestrial vertebrates (e.g., Fig. 2). One of the likely uses of the ELC API, however, is as a starting point for identifying and resolving potential duplicate data holdings.

Conclusions

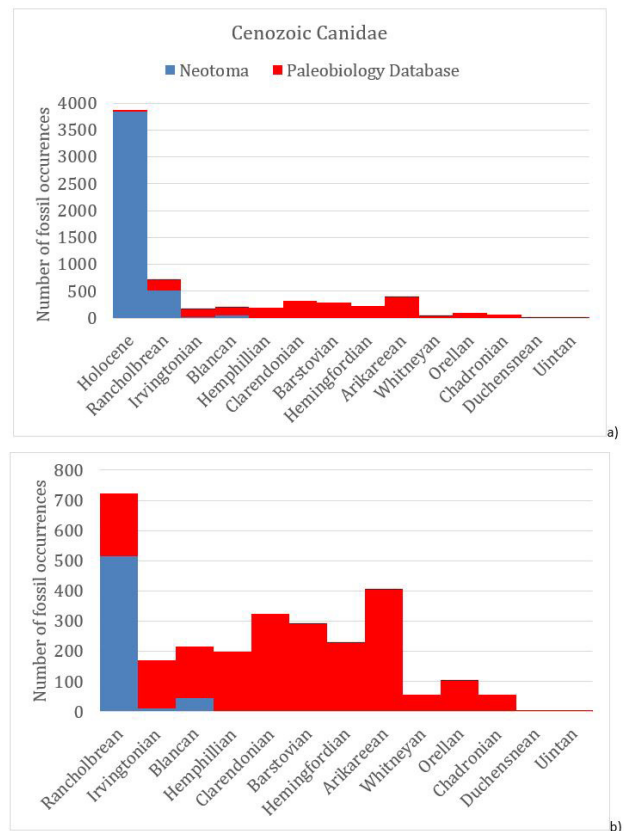


Figure 3. Distribution of Cenozoic Canidae occurrences over time. Occurrences are binned by North American Land Mammal Age, with Holocene occurrences shown (Fig. 3a; first column) or omitted (Fig. 3b). Neotoma data holdings are particularly strong for the two most recent North American Land Mammal ages (Ranchoabrean and Holocene) and extend back to Blancan (4.75 Ma), while PBDB occurrences span the full history of canid presence in the evolutionary record. These analyses have not attempted to detect and remove duplicate occurrences. Source API URL: [http://earthlifeconsortium.org/api/v1/occ?taxon=Canidae&bbox=POLYGON\(\(-84.95521545410158%207.251452594070756%2C-79.05555725097658%207.936835699063607%2C-76.67441531722721%2012.985746596062143%2C-62.268453183970614%2019.003818604968203%2C-46.939284882363154%2049.81620262683731%2C-19.739685058593707%2068.27853709923588%2C-9.096679687499986%2075.47857450805824%2C-4.792785644531217%2082.85133180485437%2C-23.302001953124957%20&includelower=true&output=csv](http://earthlifeconsortium.org/api/v1/occ?taxon=Canidae&bbox=POLYGON((-84.95521545410158%207.251452594070756%2C-79.05555725097658%207.936835699063607%2C-76.67441531722721%2012.985746596062143%2C-62.268453183970614%2019.003818604968203%2C-46.939284882363154%2049.81620262683731%2C-19.739685058593707%2068.27853709923588%2C-9.096679687499986%2075.47857450805824%2C-4.792785644531217%2082.85133180485437%2C-23.302001953124957%20&includelower=true&output=csv)

The ELC API brings together paleobiological databases so that they share a single set of Internet-based commands by which researchers and the public can easily access fossil records from all of Earth history. Key endpoints include Occurrences, Locales, References, and Taxonomy. All code is open-source and the ELC API standard is designed to be readily extensible to new paleobiological data resources. By

coordinating with other emerging efforts in geological and biological data sharing, best practices, and protocols, we ensure that data will be freely available to all, enabling new scientific syntheses and discovery, more powerful educational opportunities, and general exploration of the history of life on Earth.

Acknowledgements

We thank all of the members of the ELC development team including co-PIs Shanan Peters (University of Wisconsin), Russell Graham (Pennsylvania State University); Jessica Blois (University of California, Merced), Alison Smith (Kent State University), Edward Davis (University of Oregon). Development assistance was provided by Brian Bills (Penn State) and Michael McClennen, lead PBDB developer. We thank Phil Buckland and Roger Mähler at Umeå University for pilot testing the extension of the ELC APIs to SEAD. Support for the ELC API came from NSF EarthCube (ICER 1540997, 1541002). We thank participants in the Data Demo Derby workshop, sponsored by the above NSF awards, the ePANDDA project, and the EarthRates RCN, for initial testing of the ELC API endpoints.

References

- Amesbury, M.J., Booth, R.K., Roland, T.P., et al. (2018) Towards a Holarctic synthesis of peatland testate amoeba ecology: development of a new continental-scale palaeohydrological transfer function for North America and comparison to European data. *Quaternary Science Reviews*, 201, 483-500.
- Bartlein, P.J., Harrison, S.P., Brewer, S., et al. (2011) Pollen-based continental climate reconstructions at 6 and 21 ka: a global synthesis. *Climate Dynamics*, 37, 775-802.
- Blois, J.L., Williams, J.W., Fitzpatrick, M.C., Ferrier, S., Veloz, S.D., He, F., Liu, Z., Manion, G. & Otto-Bliesner, B. (2013) Modeling the climatic drivers of spatial patterns in vegetation composition since the Last Glacial Maximum. *Ecography*, 36, 460-473.
- Buckland, P.I., Sjölander, M. & Eriksson, E.J. (2018) Strategic Environmental Archaeology Database (SEAD). In: *Encyclopedia of Global Archaeology* (ed. by C. Smith). Springer, Cham. https://doi.org/10.1007/978-3-319-51726-1_833-2
- Cantino, P.D., Doyle, J.A., Graham, S.W., Judd, W.S., Olmstead, R.G., Soltis, D.E., Soltis, P.S. & Donoghue, M.J. (2007) Towards a phylogenetic nomenclature of *Tracheophyta*. *Taxon*, 56, E1-E44.
- Chaloner, W.G. & Creber, G.T. (1988) Fossil plants as indicators of late Palaeozoic plate positions. Geological Society, London, Special Publications, 37, 201-210.
- COHMAP Members (1988) Climatic changes of the last 18,000 years: observations and model simulations. *Science*, 24, 1043-1052.
- Crame, J.A. (2001) Taxonomic diversity gradients through geological time. *Diversity and Distributions*, 7, 175-189.
- Curry, B.B., Delorme, L.D., Smith, A.J., Palmer, D.F. & Stiff, B.J. (2013) The biogeography and physiochemical characteristics of aquatic habitats of freshwater ostracodes in Canada and the United States. In: *Ostracoda as Proxies for Quaternary Climate Change* (ed. by D. Horne, J. Holmes, J. Rodriguez-Lasaro & F. Viehberg), pp. 85-115. Elsevier Science Publishing.
- Dietl, G.P. & Flessa, K.W. (2011) Conservation paleoecology: putting the dead to work. *Trends in Ecology and Evolution*, 26, 30-37.
- FAUNMAP Working Group (1996) Spatial response of mammals to late Quaternary environmental fluctuations. *Science*, 272, 1601-1606.
- Fordham, D.A., Jackson, S.T., Brown, S.C., Huntley, B., Brook, B.W., Dahl-Jensen, D., Gilbert, M.T.P., Otto-Bliesner, B.L., Svensson, A., Theodoridis, S., et al. (2020) Using paleo-archives to safeguard biodiversity under climate change. *Science* 369, eabc5654.
- Fritz, S.A., Schnitzler, J., Eronen, J.T., Hof, C., Böhning-Gaese, K. & Graham, C.H. (2013) Diversity in time and space: wanted dead and alive. *Trends in Ecology and Evolution*, 28, 509-516.
- Fyfe, R.M., de Beaulieu, J.-L., Binney, H., et al. (2009) The European Pollen Database: past efforts and current activities. *Vegetation History and Archaeobotany*, 18, 417-424.
- Gradstein, F.M., Ogg, J.G., Schmitz, M.D. & Ogg, G.M. (2012) A geologic time scale 2012. Elsevier, Amsterdam.
- Graham, R.W., Lundelius, E. L. Jr, Graham, M.A., et al. (1996) Spatial response of mammals to Late Quaternary environmental fluctuations. *Science*, 272, 1601-1606.
- Grimm, E.C., Blois, J.L., Giesecke, R., Graham, R., Smith, A.J. & Williams, J.W. (2018) Constituent databases and data stewards in the Neotoma Paleocology Database: history, growth, and

- new directions. *Past Global Changes Magazine*, 26, 64-65.
- Harrison, S.P., Bartlein, P.J., Brewer, S., Prentice, I.C., Boyd, M., Hessler, I., Holmgren, K., Izumi, K., & Willis, K. (2013) Climate model benchmarking with glacial and mid-Holocene climates. *Climate Dynamics*, 43, 671-688.
- IPBES (2019) Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. IPBES, United Nations.
- Jablonski, D. (2008) Extinction and the spatial dynamics of biodiversity. *Proceedings of the National Academy of Sciences USA*, 105, 11528-11535.
- Jablonski, D., Belanger, C.L., Berke, S.K., Huang, S., Krug, A.Z., Roy, K., Tomasovych, A. & Valentine, J.W. (2013) Out of the tropics, but how? Fossils, bridge species, and thermal ranges in the dynamics of the marine latitudinal diversity gradient. *Proceedings of the National Academy of Sciences USA*, 110, 10487-10494.
- Jackson, S.T. & Williams, J.W. (2004) Modern analogs in Quaternary paleoecology: here today, gone yesterday, gone tomorrow? *Annual Review of Earth and Planetary Sciences*, 32, 495-537.
- Kidwell, S.M. (2015) Biology in the Anthropocene: challenges and insights from young fossil records. *Proceedings of the National Academy of Sciences USA*, 112, 4922-4929.
- Kohler, T.A., Buckland, P.I., Kintigh, K.W., Bocinsky, R.K., Brin, A., Gillreath-Brown, A., Ludäscher, B., McPhillips, T.M., Opitz, R. & Terstriep, J. (2018) Paleodata for and from archaeology. *Past Global Changes Magazine*, 26, 68-69.
- Loeffler, S. (2018) Improving geoscience data access and interoperability through the Flyover Country mobile app. University of Minnesota, Minneapolis, MN.
- Marsicek, J., Shuman, B.N., Bartlein, P.J., Shafer, S.L., & Brewer, S. (2018) Reconciling divergent trends and millennial variations in Holocene temperatures. *Nature*, 554, 92.
- Metcalfe, I. & Nicoll, R.S. (2007) Conodont biostratigraphic control on transitional marine to non-marine Permian-Triassic boundary sequences in Yunnan-Guizhou, China. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 252, 56-65.
- Muller, J. & Reisz, R. (2005) Four well-constrained calibration points from the vertebrate fossil record for molecular clock estimates. *BioEssays*, 27, 1069-1075.
- Myrbo, A., Loeffler, S., Shinneman, D. & McEwan, R. (2018) Outreach and educational opportunities created by open data resources. *Past Global Changes Magazine*, 26, 74-75.
- National Research Council (2005) The geological record of ecological dynamics: understanding the biotic effects of future. The National Academies Press, Washington, D.C.
- National Research Council (2011) Understanding Earth's deep past: lessons for our climate future. The National Academies Press, Washington, D.C.
- National Research Council (2012) New research opportunities in the Earth sciences. The National Academies Press, Washington, D.C.
- O'Regan, H.J., Turner, A., Bishop, L.C., Elton, S. & Lamb, A.L. (2011) Hominins without fellow travellers? First appearances and inferred dispersals of Afro-Eurasian large-mammals in the Plio-Pleistocene. *Quaternary Science Reviews*, 30, 1343-1352.
- Peters, S.E., Husson, J.M. & Czaplewski, J.J. (2018) Macrostrat: a platform for geological data integration and deep-time Earth crust research. *Geochemistry, Geophysics, Geosystems* 19, 1393-1409.
- Peters, S.E. & McClennen, M. (2016) The Paleobiology Database application programming interface. *Paleobiology*, 42, 1-7
- Radeloff, V.C., Williams, J.W., Bateman, B.L., et al. (2015) The rise of novelty in ecosystems. *Ecological Applications*, 25, 2051-2068.
- Routson, C.C., McKay, N.P., Kaufman, D.S., Erb, M.P., Goosse, H., Shuman, B.N., Rodysill, J.R. & Ault, T. (2019) Mid-latitude net precipitation decreased with Arctic warming during the Holocene. *Nature*, 568, 83-87.
- Shanahan, T.M., McKay, N.P., Hughen, K.A., Overpeck, J.T., Otto-Bliesner, B., Heil, C.W., King, J., Scholz, C.A. & Peck, J. (2015) The time-transgressive termination of the African Humid Period. *Nature Geoscience*, 8, 140-144.
- Shuman, B.N., Marsicek, J., Oswald, W.W. & Foster, D.R. (2019) Predictable hydrological and ecological responses to Holocene North Atlantic variability. *Proceedings of the*

- National Academy of Sciences USA, 116, 5985-5990.
- Sullivan, T.J. & Charles, D.F. (1994) The feasibility and utility of a paleolimnology/paleoclimate data cooperative for North America. *Journal of Paleolimnology*, 10, 265-273.
- Uhen, M.D., Barnosky, A.D., Bills, B., et al. (2013) From card catalogs to computers: databases in vertebrate paleontology. *Journal of Vertebrate Paleontology*, 33, 13-28.
- Uhen, M.D., Goring, S., Jenkins, J. & Williams, J.W. (2018) Earthlife Consortium: supporting digital paleobiology. *Past Global Changes Magazine*, 26, 78-79.
- Veloz, S.D., Williams, J.W., Blois, J.L., He, F., Otto-Bliesner, B. & Liu, Z. (2012) No-analog climates and shifting realized niches during the late Quaternary: implications for 21st-century predictions by species distribution models. *Global Change Biology*, 18, 1698-1713.
- Whittaker, R.J., Willis, K.J. & Field, R. (2001) Scale and species richness: towards a general, hierarchical theory of species diversity. *Journal of Biogeography* 28, 453-470.
- Williams, J.W., Grimm, E.C., Blois, J.L., et al. (2018a) The Neotoma Paleocology Database, a multiproxy, international, community-curated data resource. *Quaternary Research*, 89, 156-177.
- Williams, J.W., Kaufman, D., Newton, A. & Von Gunten, L. (2018b) Special Issue: building and harnessing open paleodata. *Past Global Changes Magazine*, 26, 49.
- Williams, J.W., Shuman, B., Bartlein, P.J., Diffenbaugh, N.S. & Webb, T. III (2010) Rapid, time-transgressive, and variable responses to early-Holocene midcontinental drying in North America. *Geology*, 38,135-138.
- Willis, K.J. & Birks, H.J.B. (2006) What is natural? The need for a long-term perspective in biodiversity conservation. *Science*, 314, 1261-1265.
- Willis, K.J. & MacDonald, G.M. (2011) Long-term ecological records and their relevance to climate change predictions for a warmer world. *Annual Review of Ecology, Evolution, and Systematics*, 42, 267-287.
- Wright, N., Zahirovic, S., Müller, R.D. & Seton, M. (2013) Towards community-driven paleogeographic reconstructions: integrating open-access paleogeographic and paleobiology data with plate tectonics. *Biogeosciences*, 10, 1529-1541.
- Submitted: 5 November 2020
First decision: 7 December 2020
Accepted: 5 January 2021
- Edited by Robert J. Whittaker