palaeoverse

Data acquisition for paleobiological analysis

Erin Dillon

(updated from content created by Bethany Allen & Harriet B. Drage)

Objectives

1. Learn about different types of databases
2. Know how to access raw data
3. Understand why it is important to keep raw data raw
4. Acquire and load an example fossil dataset into R

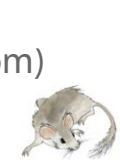
Many different data types!

- Occurrences
- Relative abundances
- Taxonomy
- Traits
- Morphology
- Stratigraphic information
- Geological data
- Paleoenvironmental information
- Phylogenies
- ...

Data acquisition: online databases

Many databases for different data types/origins/groups, e.g.:

- Geobiodiversity Database (stratigraphic approach; geobiodiversity.com)
- Neotoma (Pliocene–Quaternary; neotomadb.org)
- Global Biodiversity Information Facility, GBIF (modern and fossil occurrences, gbif.org)
- iDigBio (mostly US museums; idigbio.org)
- Neptune/Triton (planktonic microfossils; nsb.mfn-berlin.de)
- BioDeepTime (time series; doi.org/10.1111/geb.13735)
- Phylacine (Quarternary mammals; megapast2future.github.io/)
- PARED (paleo reefs; paleo-reefs.pal.uni-erlangen.de)



Data acquisition: online databases

3D scans or morphological data:

- MorphoBank: morphobank.org
- MorphoSource: morphosource.org
- Phenome10k: phenome10k.org

Trait data:

- Open Traits Network: opentraits.org
- Coral Trait Database: coraltraits.org

Phylogenetic data:

- TreeBASE: treebase.org
- Fossil Calibration Database: fossilcalibrations.org

Conservation status data:

- IUCN: iucnredlist.org

Stratigraphic and geological data:

- Macrostrat: macrostrat.org
- USGS geological maps: mrdata.usgs.gov/geology/state

Paleoenvironmental and paleoclimatic data:

- EarthByte: earthbyte.org/category/resources/data-models/
- BRIDGE palaeoclimate models: bristol.ac.uk/geography/research/bridge
- CHELSA: chelsa-climate.org

General repositories:

- Zenodo: zenodo.org
- Open Science Framework: osf.io

Data acquisition: online databases

Name	Data Description	Temporal Scope	Spatial Scope	Taxonomic Scope	Website
AFORO (Shape Analysis of Fish Otoliths)	Sagitta otolith images and shape analysis	N/A	Global	Fish	http://aforo.cmima.csic.es/index.jsp
American Society of Mammalogists (ASM) Mammal Diversity Database	Taxonomic checklist of extant and recently extinct (since ~1500 CE) mammals	Modern	Global	Mammals	https://www.mammaldiversity.org/
Anthromes	Land cover data and maps showing global ecological patterns created by humans	Holocene – Modern	Global	N/A	https://anthroecology.org/datasets/
Arctic Data Center	Data and software repository for the Arctic section of National Science Foundation's Office of Polar Programs	Modern	Arctic	Multiple taxa	https://arcticdata.io/
Arctos	Biological specimen data	Modern	Global	Multiple taxa	https://arctosdb.org/
Biodiversity Atlas – India	Biodiversity data, including butterflies, moths, cicadas, mammals, birds, reptiles, amphibians, and odonates	Modern	India	Multiple taxa	https://www.bioatlasindia.org/bai-websites
Biodiversity Information Serving Our Nation (BISON)	Modern, historical, and fossil species occurrence data (subset of GBIF)	Precambrian – Phanerozoic	United States	Multiple taxa	https://bison.usgs.gov/
Biological and Chemical Oceanography Data Management Office (BCO-DMO)	Biological and chemical oceanography data	Modern	Global	N/A	https://www.bco-dmo.org/
BioTime	Species abundance time series data	Modern	Global	Multiple taxa	https://biotime.st-andrews.ac.uk
BugsCEP	Coleopteran occurrence and (paleo)ecological data	Quaternary	Global	Beetles	http://bugscep.com/
Catalogue of Life (COL)	Taxonomic classifications and catalogue of species	N/A	Global	Multiple taxa	https://www.catalogueoflife.org/
Consortium of European Taxonomic Facilities (CETAF)	Biological and geological specimen data	N/A	Global	Multiple taxa	https://cetaf.org/
COPEPOD Project (Global Plankton Database)	Plankton abundance, biomass, and composition data	Modern	Global	Plankton	https://www.st.nmfs.noaa.gov/copepod/
Coral Trait Database	Scleractinian coral life history trait, phylogenetic, and biogeographic data	N/A	Global	Scleractinian corals	https://coraltraits.org
Data Observation Network for Earth (DataONE)	Searches across databases for ecological, environmental, and earth science data	N/A	Global	Multiple taxa	https://www.dataone.org/

Dillon & Dunne *et al.* 2023: zenodo.org/records/7340036

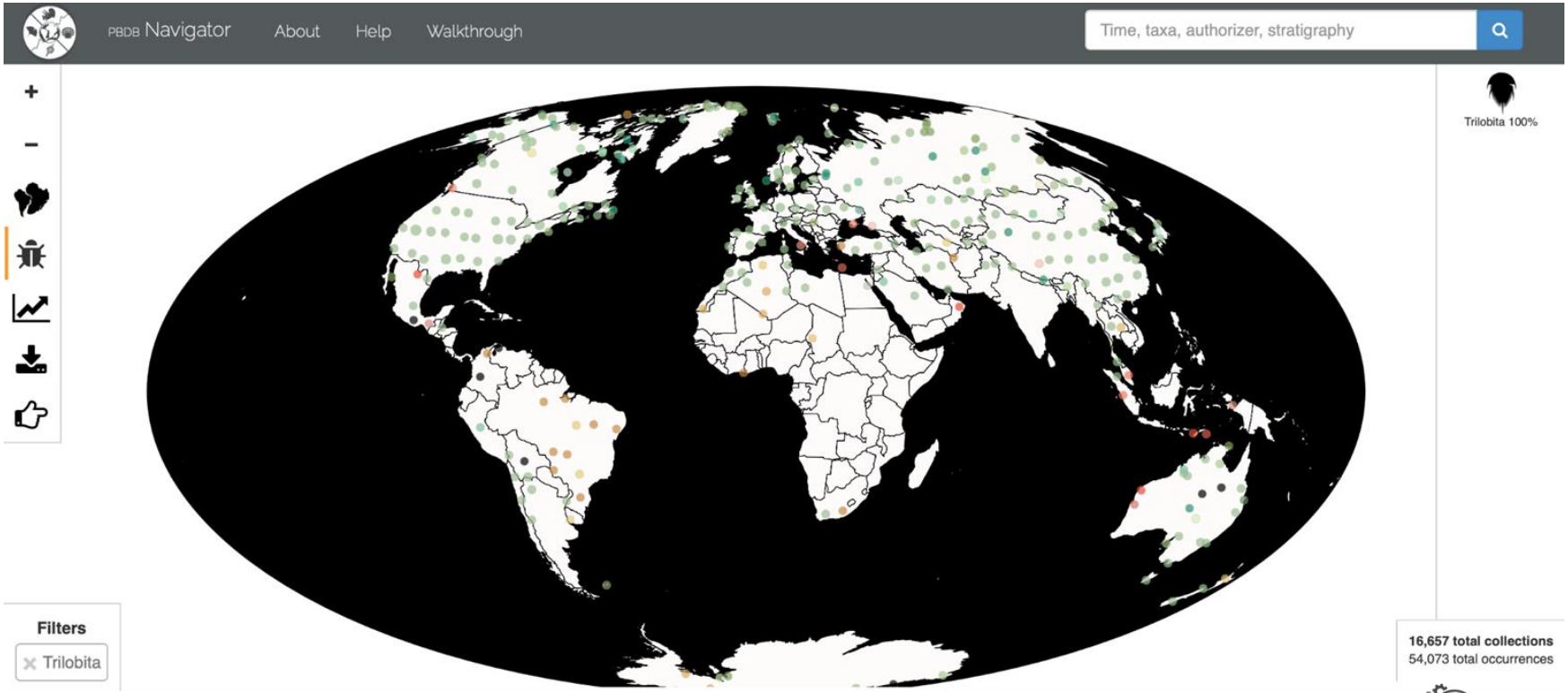
PAGES PAST GLOBAL CHANGES

PAGES databases: pastglobalchanges.org/science/data/databases

Paleobiology Database (PBDB)

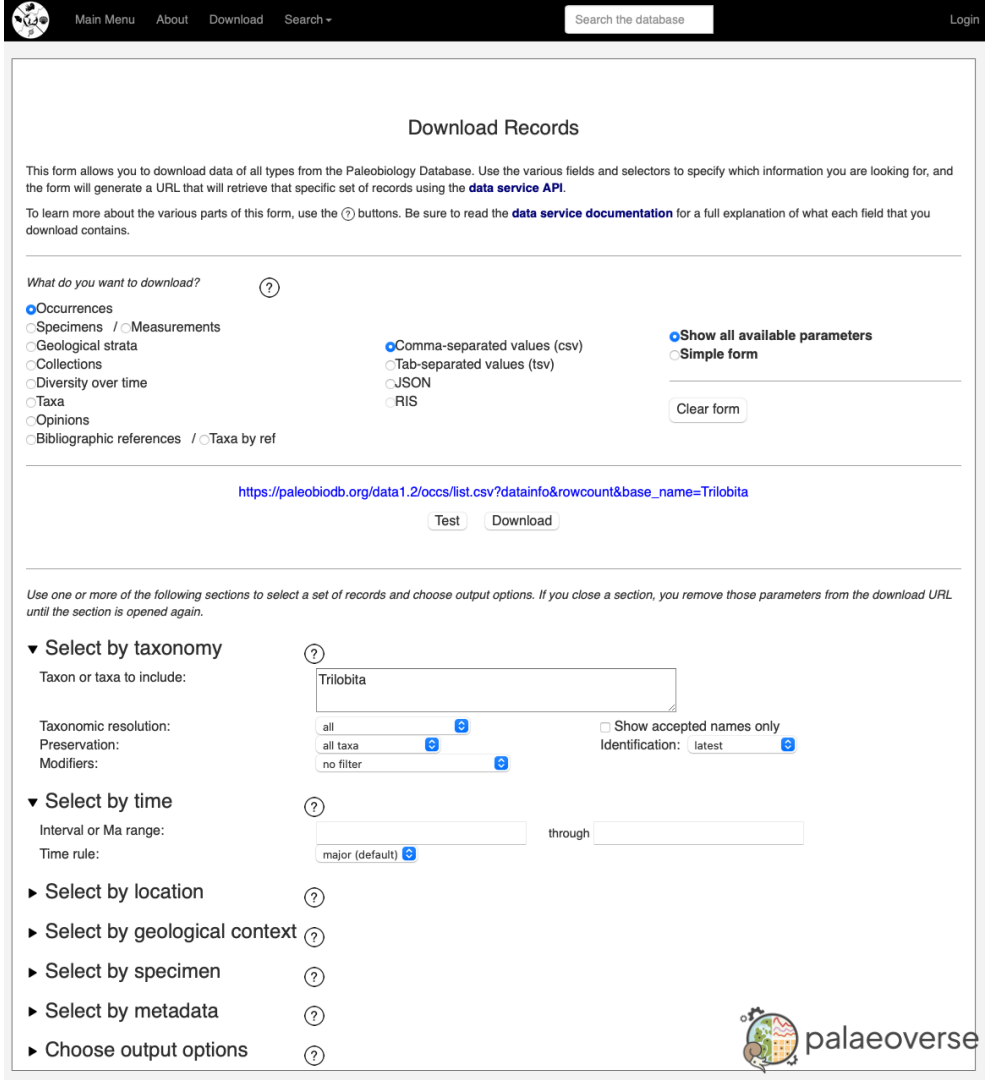
- paleobiodb.org
- Over 20 years old, mostly funded by NSF
- Coverage is global, good for macrofossils throughout geological time
- Can be explored using the Navigator, but data can also be downloaded
- All fossils entered by paleontologists from the published literature

PBDB user interface



palaeoverse

PBDB user interface



PBDB user interface

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Data Provide	The Paleobiology Database															
2	Data Source	The Paleobiology Database															
3	Data License	Creative Commons CC0															
4	License URL	https://creativecommons.org/publicdomain/zero/1.0/															
5	Documentat	http://paleobiodb.org/data1.2/occs/list_doc.html															
6	Data URL	http://paleobiodb.org/data1.2/occs/list.csv?datainfo&rowcount&base_name=Trilobita															
7	Access Time	Thu 2025-07-03 13:54:09 GMT															
8	Title	PBDB Data Service															
9	Parameters:																
10		base_name	Trilobita														
11		timerule	major														
12		taxon_status	all														
13	Elapsed Time	3.84															
14	Records Fou	54073															
15	Records Retu	54073															
16	Records:																
17	occurrence_	record_type	reid_no	flags	collection_n	identified_na	identified_ra	identified_no	difference	accepted_na	accepted_ra	accepted_no	early_interva	late_interval	max_ma	min_ma	reference_no
18	1	occ			1	Australosutu	species	349412		Australosutu	species	349412	Ivorian		353.7	346.7	1
19	2	occ			1	Carbonocory	species	349526	recombined	Phillibole pla	species	349526	Ivorian		353.7	346.7	1
20	3	occ			1	Thigriffides rc	species	349420		Thigriffides rc	species	349418	Ivorian		353.7	346.7	1

palaeoverse

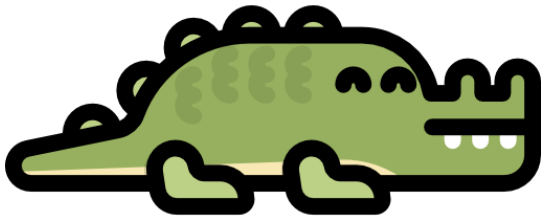
What is raw data, and why keep it raw?

- **The data you originally downloaded — no changes!**
- Why?
  - identification of later errors
  - reproducibility
  - online databases are not static
- How?
  - store files locally and clearly
  - check file encoding
  - read-only (copy file to make changes)
  - clean using R/other language
- Archive along with all other materials at project end

Today’s research question

How do paleolatitudinal range and paleodiversity vary for Crocodylia across the Cenozoic?

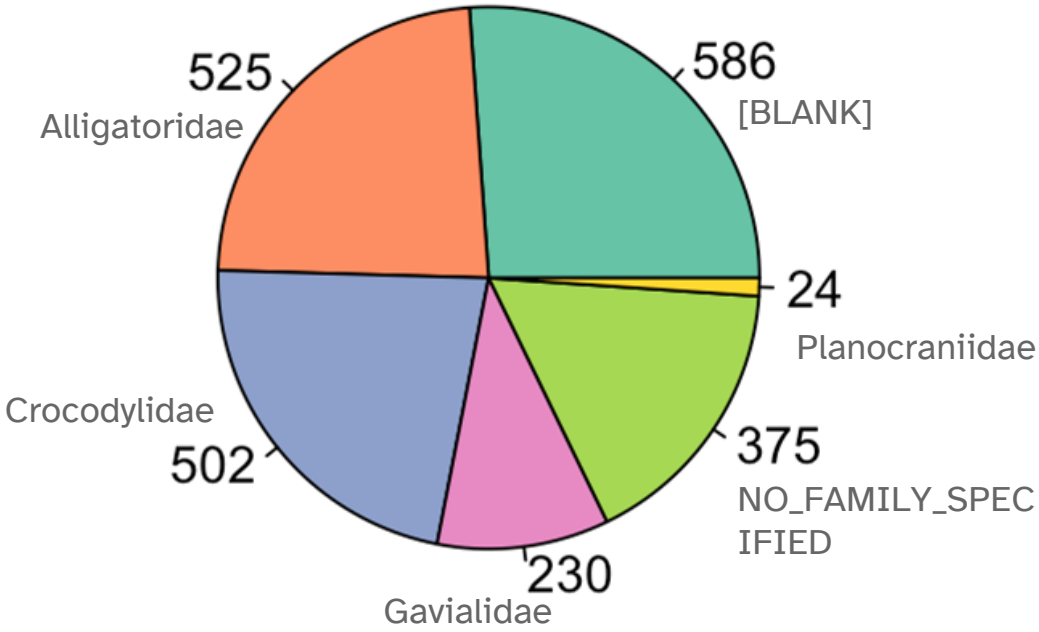
Crocs are a good group to look at for paleo/ecology - fossil record, modern data, responsive to temperature, global record

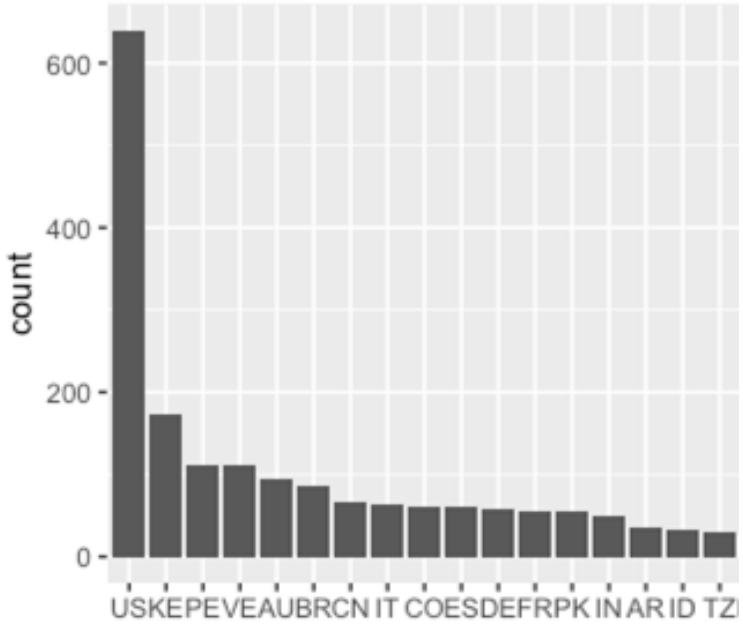


palaeoverse

Let’s load our data!

Our paleo dataset





palaeoverse

On to data processing!