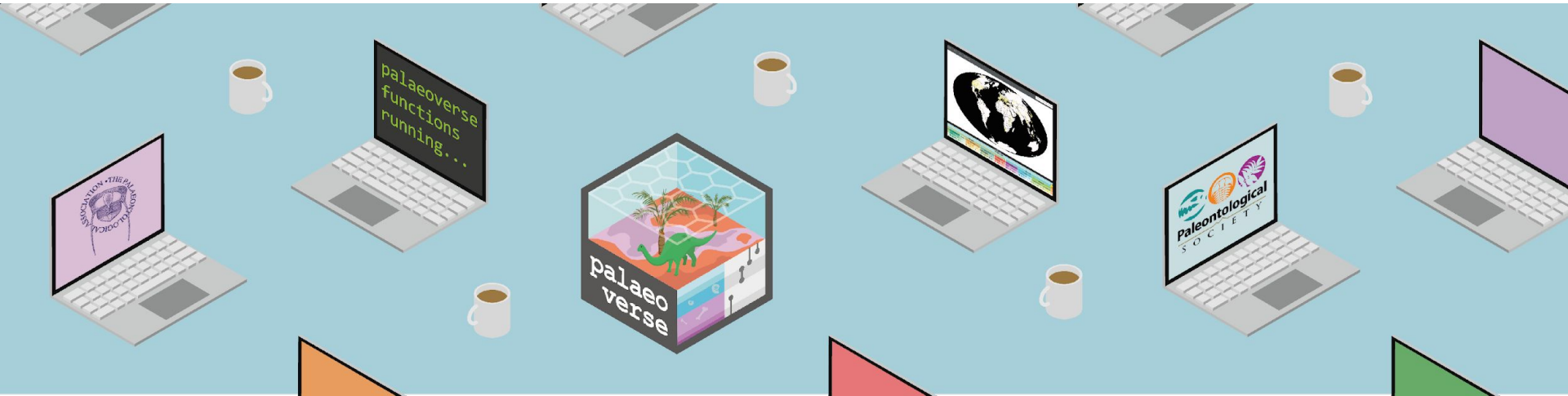


# Data acquisition



Bethany Allen & Harriet B. Drage

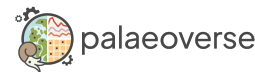
# Objectives

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

# Data acquisition: online databases

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

- Geobiodiversity Database (stratigraphic approach; [geobiodiversity.com](http://geobiodiversity.com))
- Neotoma (recent; [neotomadb.org](http://neotomadb.org))
- iDigBio (mostly US museums; [idigbio.org](http://idigbio.org))
- Neptune/Triton (planktonic microfossils; [nsb.mfn-berlin.de](http://nsb.mfn-berlin.de))
- BioDeepTime (time series; [doi.org/10.1111/geb.13735](https://doi.org/10.1111/geb.13735))
- Phylacine (Quarternary mammals; [megapast2future.github.io/](https://megapast2future.github.io/))
- PARED (palaeo reefs; [paleo-reefs.pal.uni-erlangen.de](http://paleo-reefs.pal.uni-erlangen.de))



# Data acquisition: online databases

- IUCN: [iucnredlist.org](http://iucnredlist.org)
- TreeBASE: [treebase.org](http://treebase.org)
- MorphoBank: [morphobank.org](http://morphobank.org)
- Fossil Calibration Database: [fossilcalibrations.org](http://fossilcalibrations.org)
- MorphoSource: [morphosource.org](http://morphosource.org)
- Phenome10k: [phenome10k.org](http://phenome10k.org)
- Macrostrat: [macrostrat.org](http://macrostrat.org)
- EarthByte: [earthbyte.org/category/resources/data-models/](http://earthbyte.org/category/resources/data-models/)
- BRIDGE palaeoclimate models: [bristol.ac.uk/geography/research/bridge](http://bristol.ac.uk/geography/research/bridge)
- CHELSA: [chelsa-climate.org](http://chelsa-climate.org)

# Paleobiology Database (PBDB)

[paleobiodb.org](http://paleobiodb.org)

Over 20 years old, mostly funded by NSF

Coverage is global, good for macrofossils throughout geological time

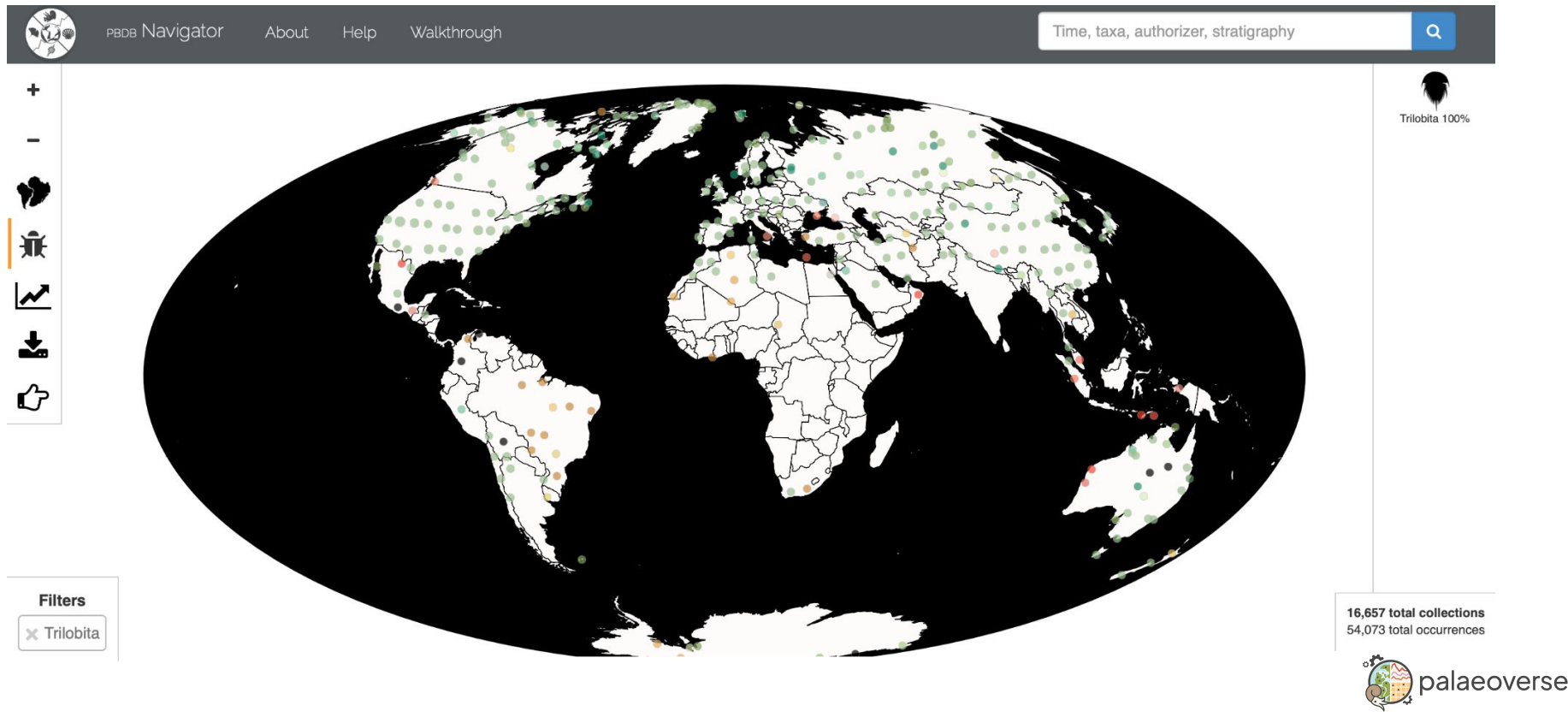
Can be explored using the Navigator, but data can also be downloaded

All fossils entered by palaeontologists from the published literature




The Paleobiology Database  
revealing the history of life

# PBDB user interface



# PBDB user interface



[Main Menu](#) [About](#) [Download](#) [Search](#)

[Login](#)

Search the database

## Download Records

This form allows you to download data of all types from the Paleobiology Database. Use the various fields and selectors to specify which information you are looking for, and the form will generate a URL that will retrieve that specific set of records using the [data service API](#).

To learn more about the various parts of this form, use the [?](#) buttons. Be sure to read the [data service documentation](#) for a full explanation of what each field that you download contains.

What do you want to download? [?](#)

☒ Occurrences

☐ Specimens / ☐ Measurements

☐ Geological strata

☐ Collections

☐ Diversity over time

☐ Taxa

☐ Opinions

☐ Bibliographic references / ☐ Taxa by ref

☒ Comma-separated values (csv)

☐ Tab-separated values (tsv)

☐ JSON

☐ RIS

☒ Show all available parameters

☐ Simple form

Clear form

[https://paleobiodb.org/data1.2/occs/list.csv?datainfo&rowcount&base\\_name=Trilobita](https://paleobiodb.org/data1.2/occs/list.csv?datainfo&rowcount&base_name=Trilobita)

Test

Download

Use one or more of the following sections to select a set of records and choose output options. If you close a section, you remove those parameters from the download URL until the section is opened again.

▼ Select by taxonomy [?](#)

Taxon or taxa to include:

Trilobita

Taxonomic resolution:

all

Preservation:

all taxa

Modifiers:

no filter

☐ Show accepted names only

Identification:

latest

▼ Select by time [?](#)

Interval or Ma range:

through

Time rule:

major (default)


► Select by location [?](#)

► Select by geological context [?](#)

► Select by specimen [?](#)

► Select by metadata [?](#)

► Choose output options [?](#)

 palaeoverse

# PBDB user interface

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q														
1	Data Provider	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 Four	54073																													
15	Records Retu	54073																													
16	Records:																														
17	occurrence_	record_type	reid_no	flags	collection_no	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														



# Global Biodiversity Information Facility (GBIF)



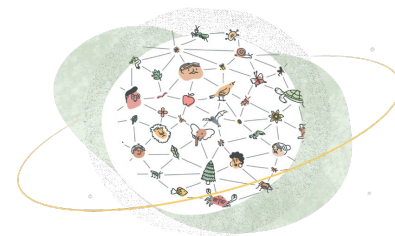
[gbif.org](http://gbif.org)


Compiles modern and fossil occurrence data from a wide variety of sources,  
including direct reporting from museum collections

(includes PBDB data)





Also has extensive taxonomic records

# GBIF user interface





Get dataHow-toToolsCommunityAbout



hdage

Occurrences2

Search all fields

Simple filtersAll filters

Occurrence status

Present

Licence

Scientific name

Trilobita

Basis of record

Year

Month

Location

Administrative areas (gadm.org)

Country or area

Continent

Dataset

Publisher

IUCN Global Red List Category

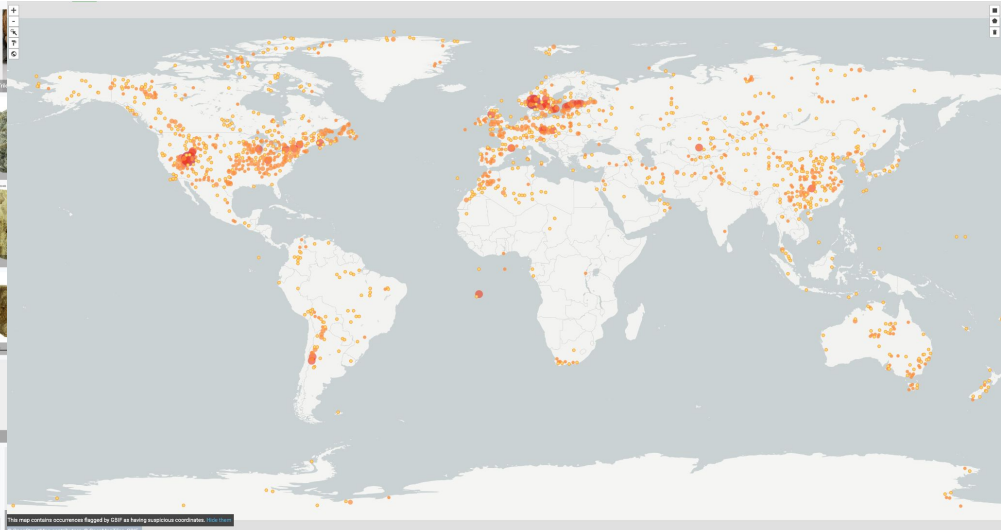
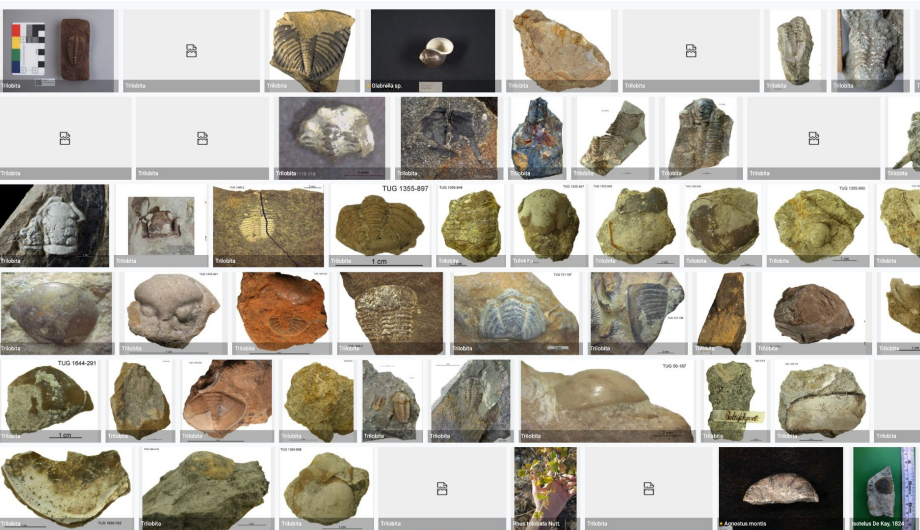
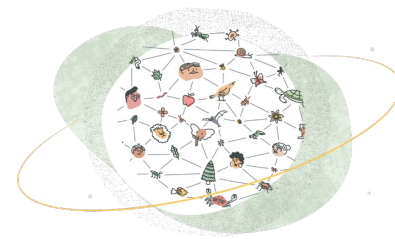
Issues and flags

TABLEGALLERYMAPTAXONOMYMETRICSDOWNLOAD

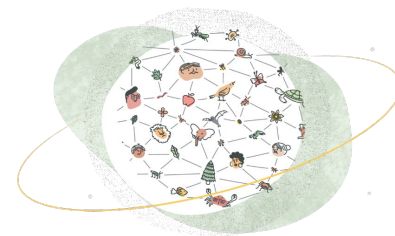
Scientific name	Country or area	Coordinates	Event date	Occurrence status	Basis of record	Dataset	Kingdom	Phylum	Class	Order	Family	Genus	Species
Conocephalus spec.	Austria	47.9N, 13.0E	2024 May 26	Present	Human observation	Biodiversitätsdatenbank Salzburg	Animalia	Arthropoda	Trilobita				Conocephalus
Modocia typicilis	United States of America		2024 May 01 - 2024 May ...	Present	Preserved specimen	Museum of Comparative Zoology, Harvard...	Animalia	Arthropoda	Trilobita	Agnostida	Marjumiidae	Modocia	
Bolaspidella Reser, 1937	United States of America		2024 May 01 - 2024 May ...	Present	Preserved specimen	Museum of Comparative Zoology, Harvard...	Animalia	Arthropoda	Trilobita	Ptychoparida	Menomonitidae	Bolaspidella	
Bolaspidella Reser, 1937	United States of America		2024 May 01 - 2024 May ...	Present	Preserved specimen	Museum of Comparative Zoology, Harvard...	Animalia	Arthropoda	Trilobita	Ptychoparida	Menomonitidae	Bolaspidella	
Bolaspidella Reser, 1937	United States of America		2024 May 01 - 2024 May ...	Present	Preserved specimen	Museum of Comparative Zoology, Harvard...	Animalia	Arthropoda	Trilobita	Ptychoparida	Menomonitidae	Bolaspidella	
Ekrathia kingi (Meek, 1870)	United States of America	33.4N, 113.3W	2024	Present	Preserved specimen	Museum of Comparative Zoology, Harvard...	Animalia	Arthropoda	Trilobita	Ptychoparida	Alokestocaridae	Ekrathia	Ekrathia kingi
Modocia Walcott, 1924	United States of America		2023 May 01 - 2023 May ...	Present	Preserved specimen	Museum of Comparative Zoology, Harvard...	Animalia	Arthropoda	Trilobita	Agnostida	Marjumiidae	Modocia	
Conocephalus spec.	Austria	47.4N, 9.7E	2023 Jun 29	Present	Human observation	Biodiversitätsdatenbank Salzburg	Animalia	Arthropoda	Trilobita				Conocephalus
Conocephalus spec.	Austria	48.2N, 16.2E	2023 Jul 10	Present	Human observation	Biodiversitätsdatenbank Salzburg	Animalia	Arthropoda	Trilobita				Conocephalus
Conocephalus spec.	Austria	47.8N, 13.0E	2023 Oct 12	Present	Human observation	Biodiversitätsdatenbank Salzburg	Animalia	Arthropoda	Trilobita				Conocephalus
Asaphus Brongniart, 1822	Finland		2022 May 01	Present	Fossil specimen	Fossil of Finland	Animalia	Arthropoda	Trilobita	Asaphida	Asaphidae	Asaphus	
Hypagnostus parvifrons (Linnaeus, 1869)	Norway	61.0N, 11.5E	2022 May 27	Present	Fossil specimen	Palaeontological collection (PMO) at UiO N...	Animalia	Arthropoda	Trilobita	Agnostida	Spinagnostidae	Hypagnostus	Hypagnostus
Hypagnostus parvifrons (Linnaeus, 1869)	Norway	61.0N, 11.5E	2022 May 27	Present	Fossil specimen	Palaeontological collection (PMO) at UiO N...	Animalia	Arthropoda	Trilobita	Agnostida	Spinagnostidae	Hypagnostus	Hypagnostus
Onicalymene Dean, 1962	Norway		2022 May 09	Present	Fossil specimen	Palaeontological collection (PMO) at UiO N...	Animalia	Arthropoda	Trilobita	Phacopida	Calymenidae	Onicalymene	
Pseudomergalaspis patagula (Törnquist, 18...	Norway		2022 May 21	Present	Fossil specimen	Palaeontological collection (PMO) at UiO N...	Animalia	Arthropoda	Trilobita	Asaphida	Asaphidae	Pseudomergalas...	Pseudomergal...
Ogmasaphus Jeanussou, 1953	Norway		2022 May 21	Present	Fossil specimen	Palaeontological collection (PMO) at UiO N...	Animalia	Arthropoda	Trilobita	Asaphida	Asaphidae	Ogmasaphus	
Ogmasaphus Jeanussou, 1953	Norway		2022 May 21	Present	Fossil specimen	Palaeontological collection (PMO) at UiO N...	Animalia	Arthropoda	Trilobita	Asaphida	Asaphidae	Ogmasaphus	
Ogmasaphus Jeanussou, 1953	Norway		2022 May 21	Present	Fossil specimen	Palaeontological collection (PMO) at UiO N...	Animalia	Arthropoda	Trilobita	Asaphida	Asaphidae	Ogmasaphus	
Ogmasaphus Jeanussou, 1953	Norway		2022 May 21	Present	Fossil specimen	Palaeontological collection (PMO) at UiO N...	Animalia	Arthropoda	Trilobita	Asaphida	Asaphidae	Ogmasaphus	
Amplexus natutus (Dalman, 1827)	Norway		2022 May 16	Present	Fossil specimen	Palaeontological collection (PMO) at UiO N...	Animalia	Arthropoda	Trilobita	Asaphida	Raphiophoridae	Amplexus	Amplexus nat...

Previous12345...Next

# GBIF user interface



# GBIF user interface



[Get data](#)
[How-to](#)
[Tools](#)
[Community](#)
[About](#)

Occurrences

Search all fields

Simple filters All filters

Occurrence status

Present

Licence

Scientific name

Triobita

Basis of record

Year

Month

Location

Administrative areas (padm.org)

Country or area

Continent

Dataset

Publisher

IUCN Global Red List Category

Issues and flags

TABLE

GALLERY

MAP

TAXONOMY

METRICS

DOWNLOAD

SEARCH OCCURRENCES | 294,449 RESULTS

DOWNLOAD OPTIONS

	Raw data	Interpreted data	Multimedia	Coordinates	Format	Estimated data size
Simple	X	✓	X	✓ (if available)	Tab-delimited CSV (for use in Excel, etc.)	159 MB (35 MB zipped for download)
Darwin Core Archive	✓	✓	✓ (links)	✓ (if available)	Tab-delimited CSV (for use in Excel, etc.)	485 MB (107 MB zipped for download)
Species List	X	✓	X	X	Tab-delimited CSV (for use in Excel, etc.)	
Cube	X	✓	X	✓ (if selected)	Tab-delimited CSV (for use in Excel, etc.)	

DOWNLOAD REPORT

**Total:** 294,449  
**Licence:** CC BY-NC 4.0  
**Year range:** 1818–2024  
**With year:** 23 %  
**With coordinates:** 62 %  
**With taxon match:** 100 %

**Known issues**

A part of the GBIF processing is to flag occurrences that have suspicious fields

87,042 Taxon match highmark 45,641 Occurrence status inferred from individual count 87,893 Continent derived from coordinates  
 39,776 Continent derived from country 49,243 Taxon match fuzzy 46,313 Reference URI invalid 4,201 Basis of record invalid  
 3,673 Occurrence status separable 5,664 Type status invalid 5,213 Country coordinate mismatch 4,584 Modified date invalid  
 4,001 Country derived from coordinates 2,035 Recorded date invalid 1,975 Recorded date mismatch 1,796 Continent coordinate mismatch  
 677 Continent country mismatch 343 Country invalid 253 Identified date invalid 238 Coordinate invalid 229 Modified date unlikely  
 142 Coordinate uncertainty metres invalid 114 Individual count conflicts with occurrence status 68 Recorded date unlikely  
 45 Presumed negated longitude 40 Geodetic datum invalid 27 occurrenceIssue.SUSPECTED\_TYPE 25 Continent invalid  
 14 Taxon match name and ID ambiguous 10 Occurrence status inferred from basis of record 6 Depth min/max swapped  
 5 Individual count invalid 4 Multimedia URI invalid 2 Depth unlikely 1 Coordinate out of range 1 Identified date unlikely  
 1 Multimedia data invalid

**Fossils**

There are fossils among your results. That can mean species occurrences at unexpected locations

# What is raw data and why keep it raw?

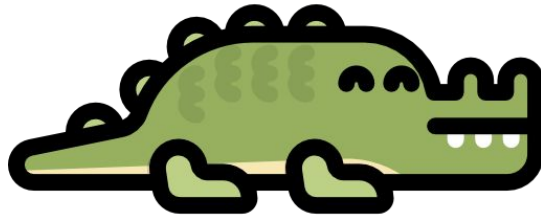
- **The data you originally downloaded - no changes!**
- Why?:
  - identification of later errors
  - reproducibility
  - online databases are not static
- How?
  - store files locally and clearly
  - check file encoding
  - read-only - copy file to make changes
  - clean using R/other language



# Today's research question

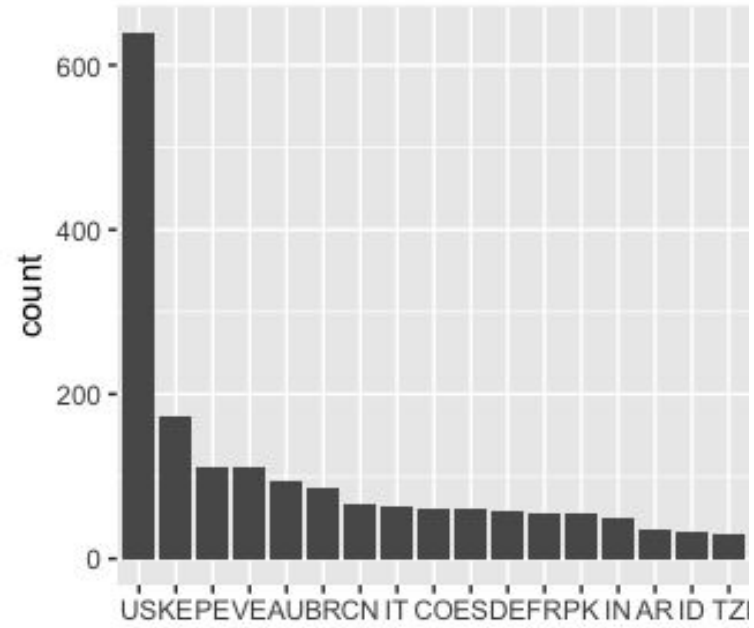
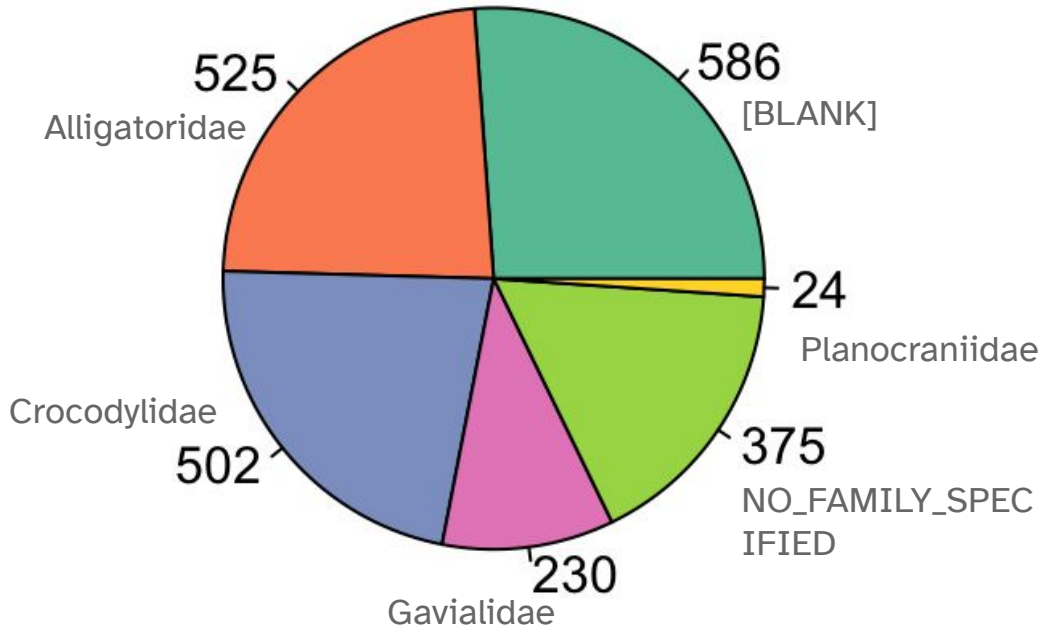
**How does range of Crocodylia change across the Cenozoic with environmental conditions?**

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



# Let's load our data!

# Our palaeo dataset





# **On to data processing!**