

# Paleontologic collection data in the broader context of paleontologic research data systems

---

**David Lazarus <sup>\*1</sup> Jeremy Young<sup>2</sup>, Shanan Peters<sup>3</sup> and Johan Renaudie<sup>1</sup>**

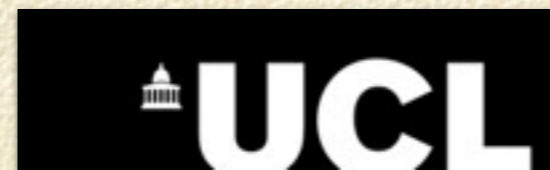
*<sup>1</sup> Museum für Naturkunde, Invalidenstrasse 43, 10115 Berlin, Germany.  
david.lazarus@mfn-berlin.de*

*<sup>2</sup> University College London, Earth Sciences, Gower Street, WC1E 6BT, UK*

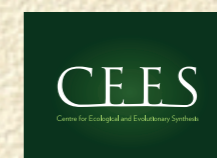
*<sup>3</sup> University Wisconsin, Department of Geoscience, Madison, 1215 W Dayton St, WI 53706, USA*



museum für  
naturkunde  
berlin



**DFG** Deutsche  
Forschungsgemeinschaft





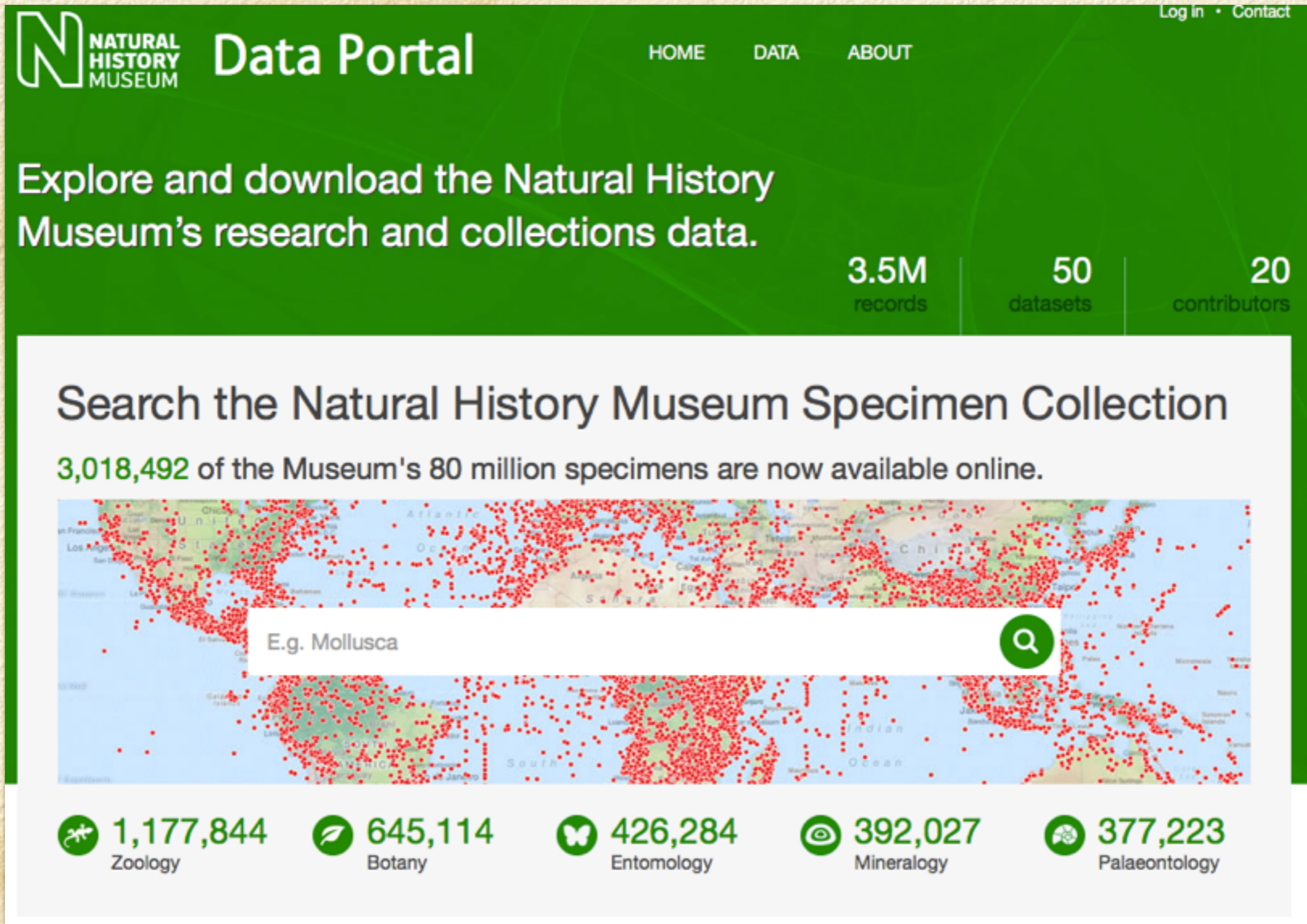
# Outline of talk

---

- Collection databases, portals and limitations
- Paleontologic research databases
- Future directions and summary



# Collection Databases



NATURAL HISTORY MUSEUM Data Portal

HOME DATA ABOUT Log In • Contact

Explore and download the Natural History Museum's research and collections data.

3.5M records 50 datasets 20 contributors

Search the Natural History Museum Specimen Collection

3,018,492 of the Museum's 80 million specimens are now available online.

E.g. Mollusca

1,177,844 Zoology 645,114 Botany 426,284 Entomology 392,027 Mineralogy 377,223 Palaeontology

- Most digital occurrence data for biologic objects is in collection databases
- Most can be accessed online via individual museum websites



# Collection DB Access at Host Museum

NATURAL HISTORY MUSEUM Data Portal

Log in • Contact

HOME DATA ABOUT

Home > Data > Collection specimens > Specimens

## Specimens

Download Data API Contact dataset curator

Specimen records

Grid Map Statistics Gallery

trilobita

Trilobita

Ordovician

1 - 100 >> 3754 records

or...	Collec...	Class	Family	Genus	Species	Subspe...	Earliest ...	Latest ...	Earlies...	Latest ...	Earliest perio...
	PAL	Trilobita	Remop...	Hypodiceranotus	sp. nov.		Phanero...	Phanero...	Paleozoic	Paleozoic	Ordovician
	PAL	Trilobita	Dimerop...	Ischyrophyma	tuberculata		Phanero...	Phaner...	Paleozoic	Paleozoic	Ordovician
	PAL	Trilobita	Phillips...	Iranaspidion	elephas		Phanero...	Phaner...	Paleozoic	Paleozoic	Permian
ster...	PAL	Trilobita	Trinucl...	Cryptolithus	portlockii		Phanero...	Phaner...	Paleozoic	Paleozoic	Ordovician
	PAL	Trilobita	Phillips...	Iranaspidion	elephas		Phanero...	Phaner...	Paleozoic	Paleozoic	Permian

- All 3 parts of basic paleo record usually provided: taxa, locality, age
- Usually little else given: content is rather skeletal from research view



# Portals

**Global Biodiversity Information Facility**  
Free and Open Access to Biodiversity Data

649,054,447 OCCURRENCES | 1,634,951 SPECIES | 32,439 DATASETS | 813 DATA PUBLISHERS

**Data** ▾ News ▾ Community ▾ About ▾

- Explore occurrences
- Explore species
- Explore dataset
- Explore by country
- Explore data trends

Publishing data  
Using Data  
Infrastructure

**Sharing biodiversity data for re-use**

- Learn about GBIF
- Publish your data through GBIF
- Technical infrastructure

**Providing evidence for research and decisions**

- Using data through GBIF
- Enabling biodiversity science
- Supporting global targets

**Collaborating as a global community**

- Current Participants
- How GBIF is funded
- Enhancing capacity

- Enormous amounts of data - mostly collection but significant other (surveys...)
- Major funding resources used to create
- 1790 GBIF pubs (n.b. - mix of major, minor use; or just mention of database)



# Portals: limitations

The screenshot shows the GBIF search interface. At the top, the GBIF logo and navigation menu are visible. The main search area displays "Search occurrences" with a result count of 2,502 and a "Download" button. Below this, there are filters for "LOCATION" (Georeferenced records only, With NO known coordinate issues) and "SCIENTIFIC NAME" (Trilobita). A table of results is shown, with two entries for Trilobita from Morocco and the Czech Republic. A configuration menu is open, showing options for "COLUMNS" and "SUMMARY FIELDS". The "COLUMNS" menu has "Location", "Basis of record", and "Date" checked. The "SUMMARY FIELDS" menu has "Occurrence key", "Catalogue number", "Scientific name", and "Dataset" checked. A "Trilobita" filter box is highlighted with a red arrow pointing to the "SCIENTIFIC NAME" filter.

LOCATION	SCIENTIFIC NAME	BASIS OF RECORD
735958467 - Cat. YPM IP 520531 <b>Trilobita</b> Published in Invertebrate Paleontology Division, Yale Peabody Museum	Morocco 30.53N, 5.83W	Fossil
351558165 - Cat. YPM IP 423991 <b>Trilobita</b> Published in Invertebrate Paleontology Division, Yale Peabody Museum	Czech Republic 49.97N, 13.78E	Fossil

- Portals tend to have lowest common denominator info
- Most do not support paleo data (tho iDigBio a bit better)

***No Geol.  
Info!***



# Paleontology Research Databases: A Different Road Taken

---

- Paleontologists began global occurrence data syntheses in early 1970s
  - long before portals etc developed in biology
  - generally occurrences of species or higher *taxa* in space and time
- Key decision made to target only published data
  - Accessibility: very little collection data available in db form (let alone 'online': internet was a Pentagon project back then)
  - Data Quality
    - Most collection objects are not well studied: determinations, geologic context imprecise, outdated or wrong [tho varies by taxon, collection]
    - Collection databases *must*, to manage collection, database all objects, good, bad or ugly - research goals are more selective
    - Metadata (paleoenvironment, etc) often minimal or absent in collection data records



# Major Paleontology Occurrences Databases

---

*not listing taxonomic catalogs, morphometric databases, etc*

- The Paleobiology Database (PBDB) - main community effort
- Neptune (NSB) - marine microfossils
- Geobiodiversity Database - Chinese, strong in geologic data
- Neotoma - continental data, last few million years
- New and Old World Mammals (NOW)
- Sepkoski Genus Database



# Paleobiology Database (PBDB)

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

*- major data types only, all numbers rough estimates!*

- N Occurrence Records (K): **1,300**
- N Taxa (valid, synonym...) (K): **342**
- N Records Total (K): **1,900**
- Geologic Age Range (MY): **600-0**
- Fossil Deposits: shallow marine, terrestrial
- N Publications (all uses/mentions): **260++**
- Dates originated / online: **2000**
- Comments:
  - Main paleontology community database effort, >>100 active participants, >40K papers entered (by same); multiple developers, funding agencies
  - Stratigraphic data/handling partially improved by complementary database Macrostrat (N. America rock formations)



The Paleobiology Database  
revealing the history of life



# PBDB Content



Paleobiology Database Classic

Quick search

Full search

Download

About

Log in

## Collection search form

### Search values

Collection name or number(s):	<input type="text"/>	Taxon name:	<input type="text"/>
Country/continent:	<input type="text"/>	State/province:	<input type="text"/>
County/parish:	<input type="text"/>	Reference #:	<input type="text"/>
Time interval (or age in Ma):	<input type="text"/>	to	<input type="text"/>
Group, formation, or member:	<input type="text"/>		
Paleoenvironment:	<input type="text"/>		
Lithology:	<input type="text"/>	<input type="text"/>	
Data <input type="text" value="authorizer"/> :	<input type="text"/>	Group/project:	<input type="text"/>
<input type="text"/>	<input type="text"/>		

*Hints: none of the fields are required, and you can search by filling out any combination of one or more fields. The search engine is not case sensitive and wildcards are allowed. You can use "\_" to match any single character or "%" for an open ended match. You may enter ranges and comma separated variables for collection numbers, i.e. 300, 400-405*

### Search and display options

Sort by:   Number of records per page:

- Occurrences have (ideally) detailed lithology & paleoenvironment data



# Neptune Database (NSB)

[www.nsb-mfn-berlin.de](http://www.nsb-mfn-berlin.de)

*- major data types only, all numbers rough estimates!*

- N Occurrence Records (K): **780**
- N Taxa (valid, synonym...) (K): **18**
- N Records Total (K): **1,000**
- Geologic Age Range (MY): **100-0**
- Fossil Deposits: deep-sea microfossils
- N Publications (all uses/mentions): **70**
- Dates originated / online: **1994 / 2003-8...; 2014+**
- Comments:
  - Both paleobiology and geochronology data
  - High geologic age resolution, high density species data but only marine plankton
  - Initiated and currently led by me in Berlin but complex history





# NSB Content (1)

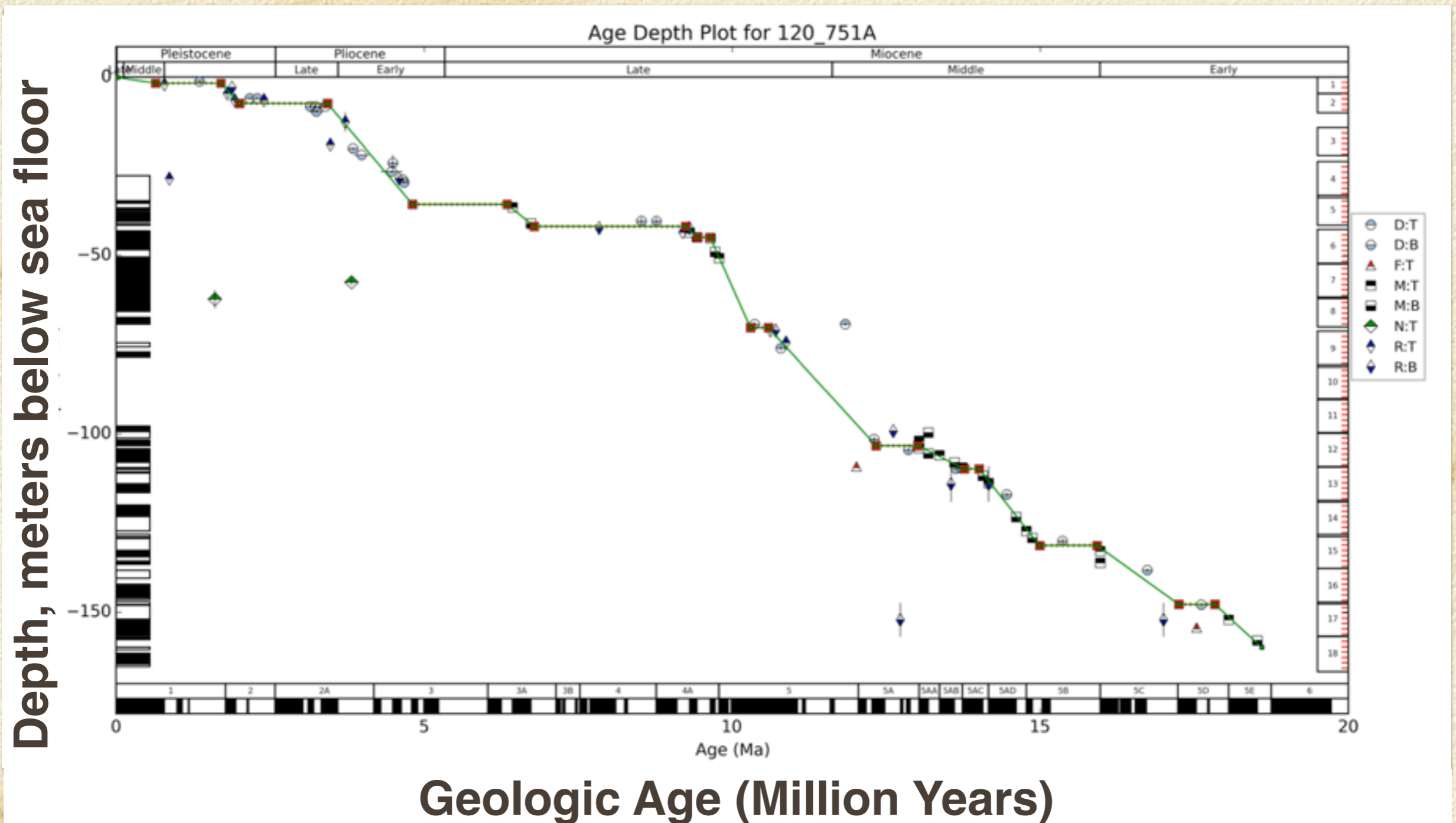
Site	H	Cor	T	Sc	Top(c)	Depth	Group	Group	Prunopyle ti-tan	Cycladophora pliocenica	Antarctissa strelkovi	Stichocorys peregrina	Antarctissa denticulata	Antarctissa cylindrica
751A	2H			1	98	5.68	A	G		+	C		C	C
751A	2H			2	98	7.18	A	G		C	F		F	A
751A	2H			3	98	8.68	C/A	G		C	A		C	C
751A	2H	CC			7	10.06	C/A	G	+	F	F		A	C
751A	3H			1	98	15.18	C/A	G		C	F		A	F
751A	3H			2	98	16.68	A	G	+	F	A		A	A
751A	3H			3	98	18.18	C	G		R	F		F	A
751A	3H			4	98	19.68	A	G	+	C	R		A	F
751A	3H			5	98	21.18	C	G	R	F	R		A	A
751A	3H	CC			9	22.01	A	G	R	F	F		A	C
751A	4H			2	98	26.18	C	G	F	R	A			A
751A	4H			3	98	27.68	A	M	F	C	A			C



- Occurrences in structured matrix (aka 'range chart') from deep-sea drilled geologic sections ('Holes') - usually digitally available, tho in many file formats
- Almost none of the fossil material held in Museums or other institutions (on PI's office shelf, or since lost...)
- Extensive paleoenvironmental data from same Holes are held in external dbs (IODP Janus; WDC Pangea)



# NSB Content (2)



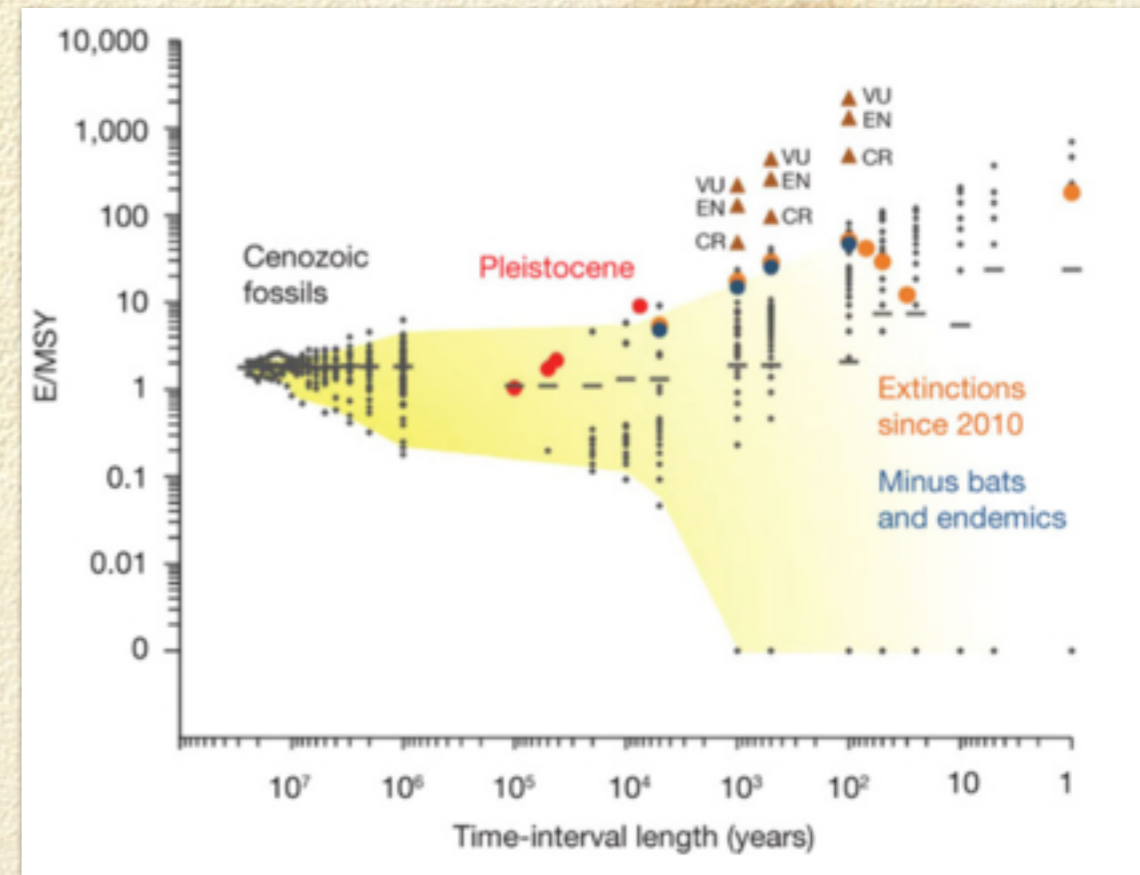
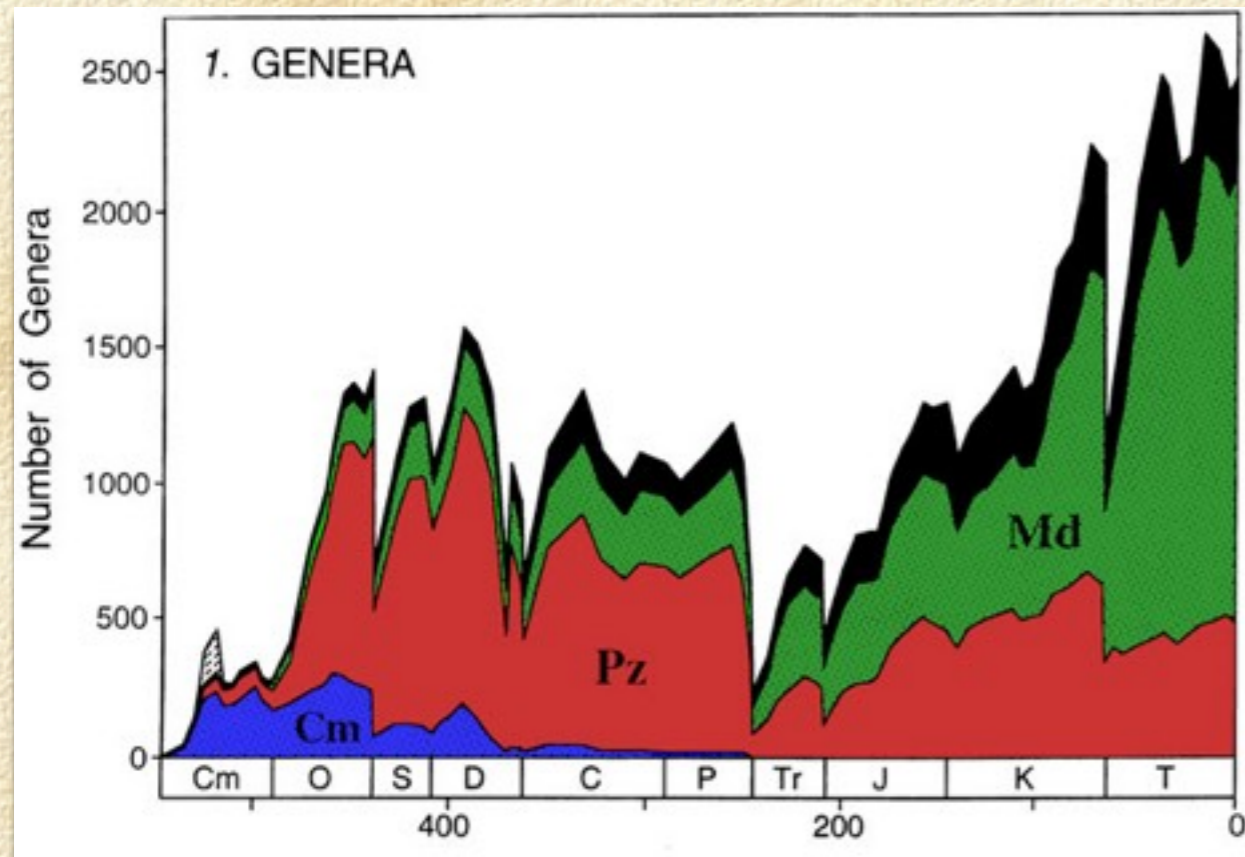
- Geologic age of occurrences are internally *calculated* based on age model function (green line)
- NSB also contains all data & calibrations used to construct age model



# Research Use

## Sepkoski 3-Fauna Model (1975-1995)

## Recent vs Geologic Extinction Rates (Barnosky et al. 2011)



- Classic Biodiversity Dynamics - testing models such as the Red Queen, Evolutionary carrying capacity of environments
- Mass Extinctions and Recoveries
- *Extinction Risk from Global Warming*
- *Bio-Geo Interactions on evolutionary timescales*

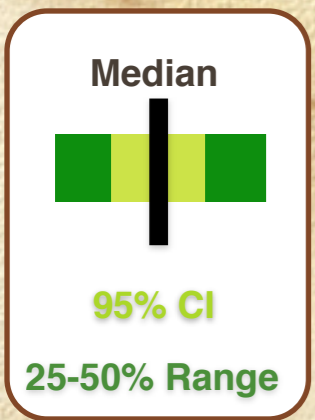
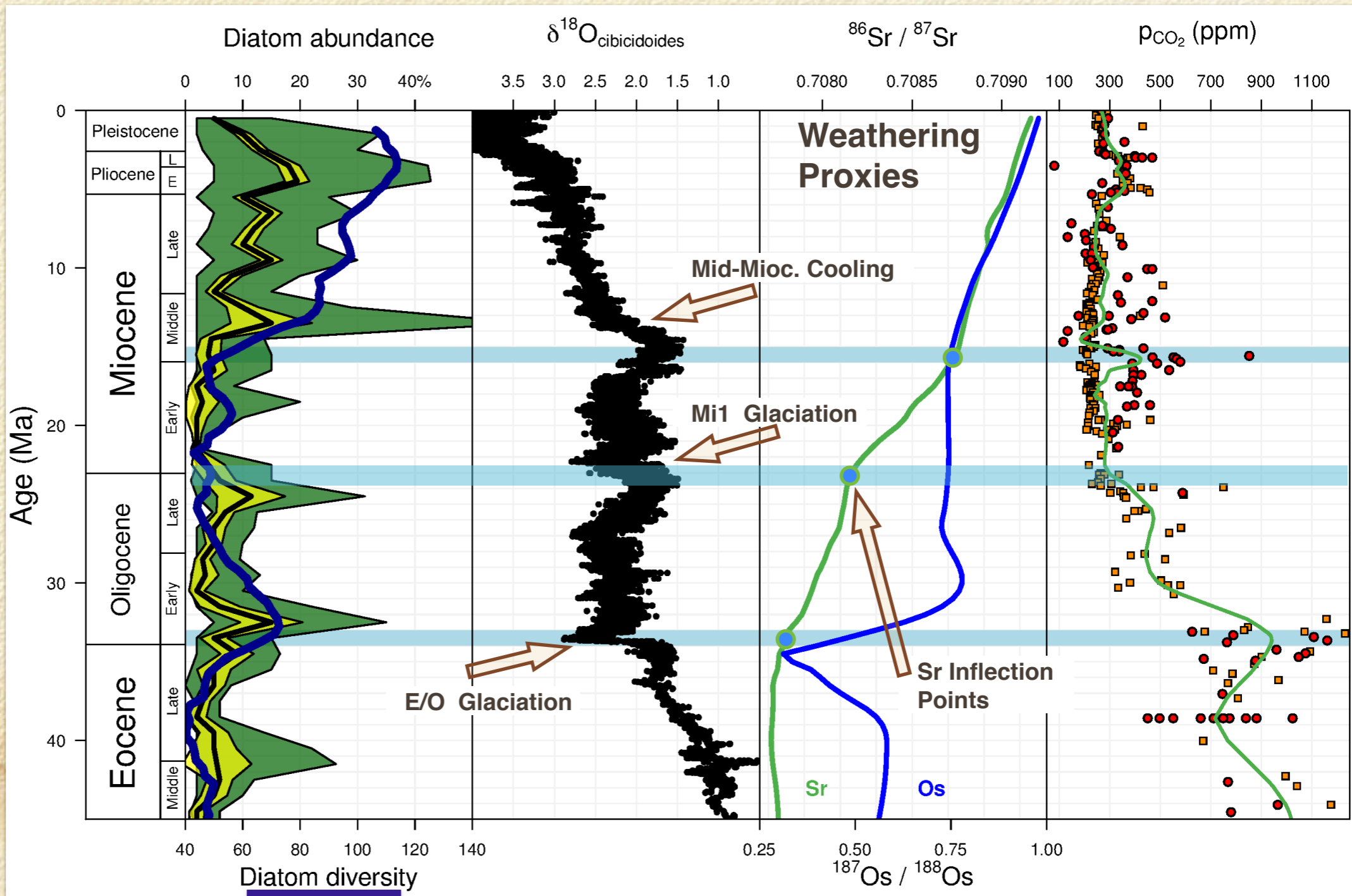


# Integrated Bio-Geosphere Studies

- NSB geochronology + Deep-sea sediment data = **Biogenic silica flux history**
- Silica Flux → Global Weathering, **pCO<sub>2</sub>**, (**biotic**) control of **global climate**
- Diatom diversity (blue line) also from NSB

• *Requires biodiversity-earth science data integration*

*from Lazarus et al (2014) PLoS One; Renaudie (in review)*





# Future Development - Linking Research Database Systems

---

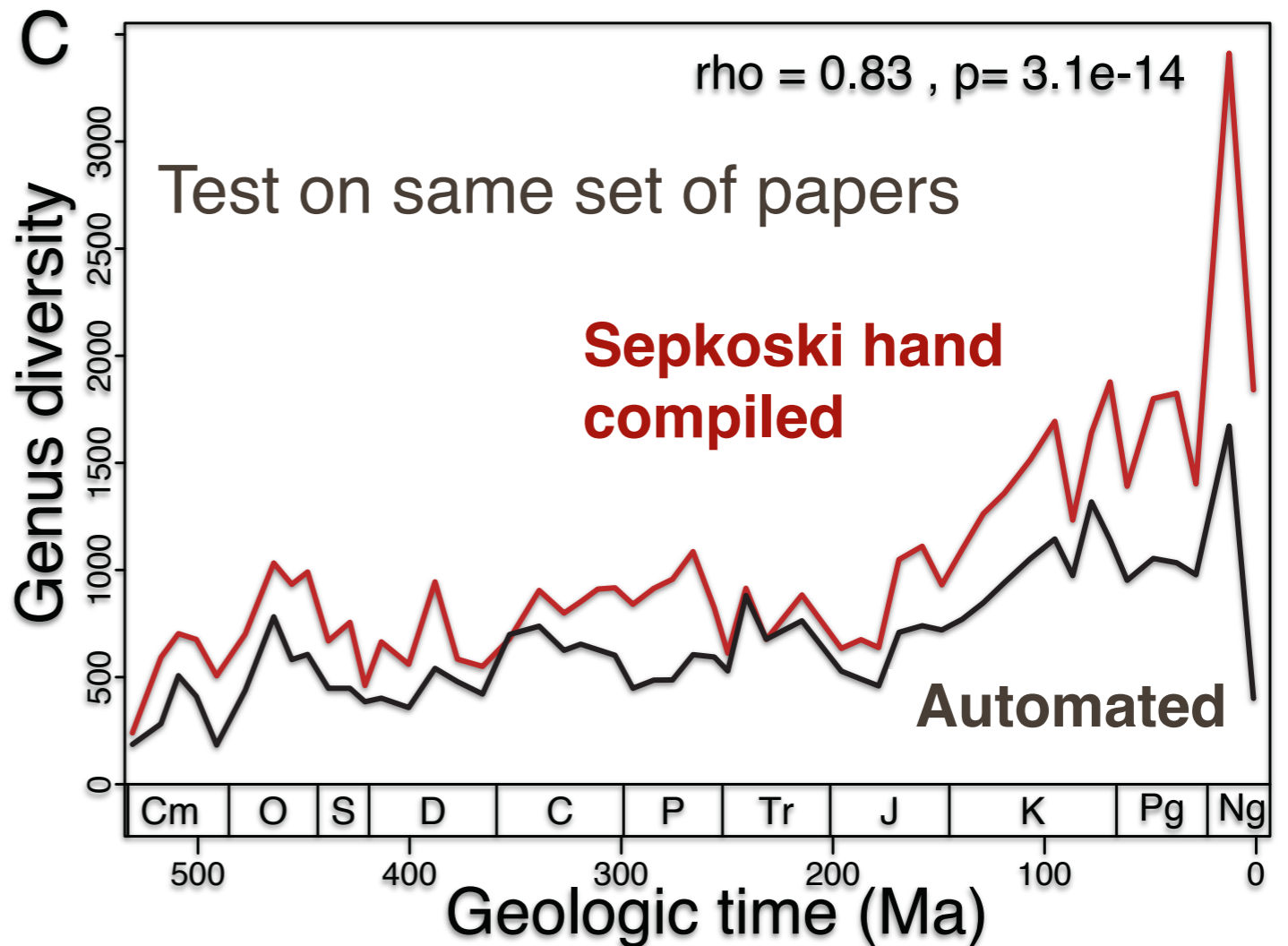
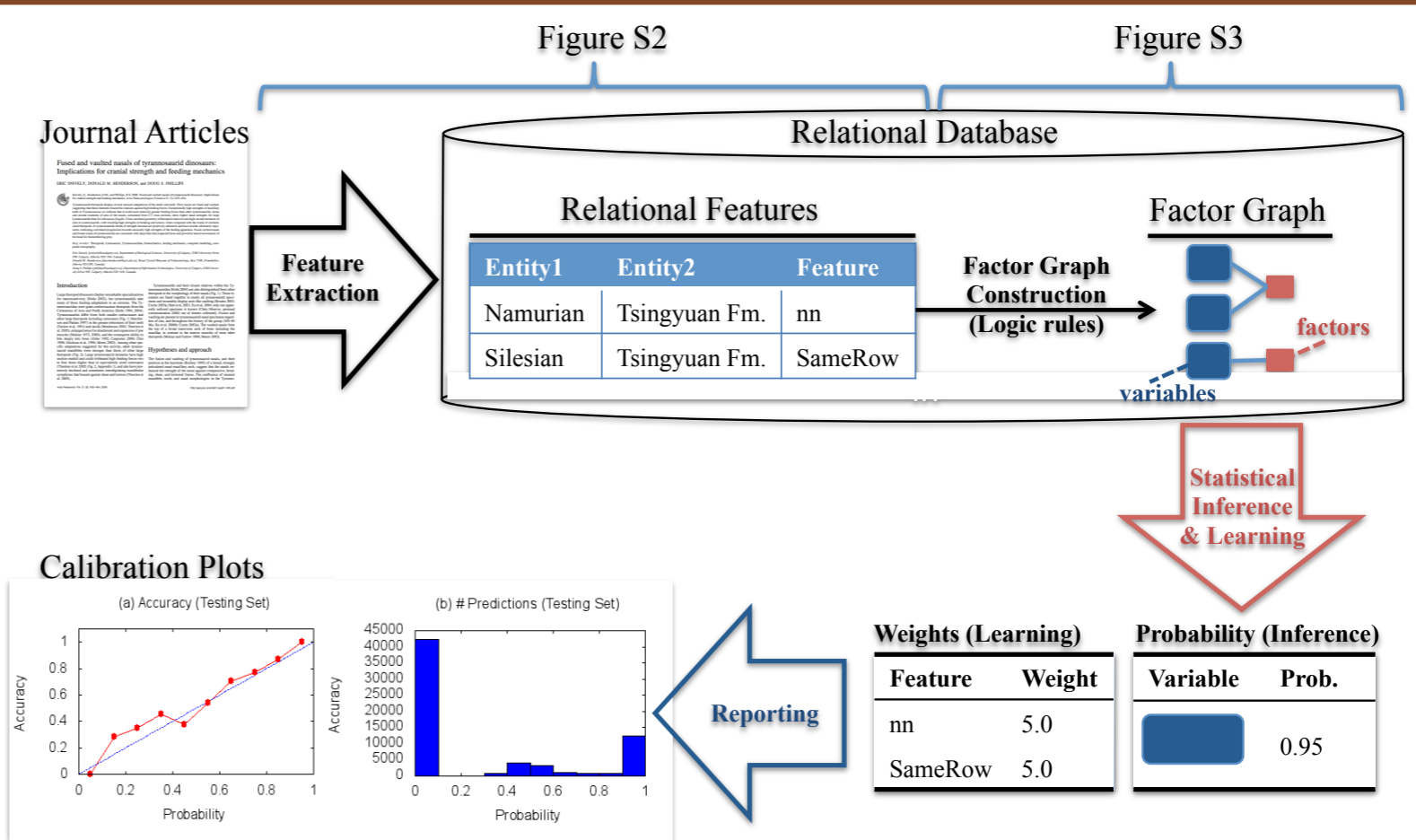
- Monolithic database for paleontologic occurrence data neither needed or desired: federation of databases better
- Fossil occurrence data have very different quality by source
  - Deep-sea marine microfossil data : standardized species, 'matrix' datasets, age-modeled sections, earth science measurements from same sections
- Multiple databases covering same data domain (PBDB/Macrostrat, GBDB, NOW) could be better integrated
  - No such initiatives at present but to be expected in future



# Future Development - Improving Literature-based Content

*Peters et al. 2014*

- PaleoDeepDive
- Automated literature data extraction
- 7X data volume vs PBDB
- 1st-order patterns similar
- ? 2nd-order patterns
- Massively hi-tech requirements





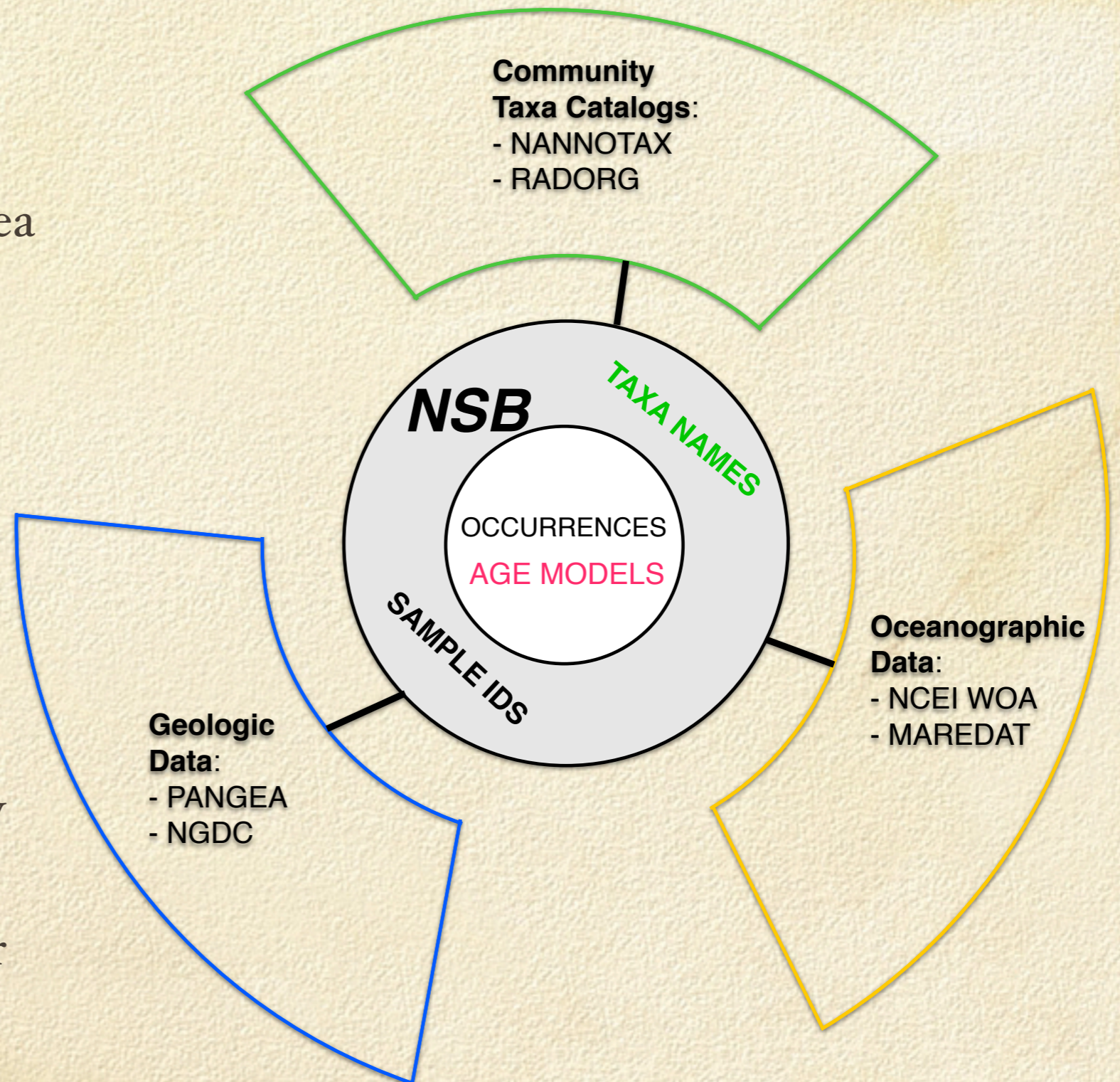
# Future Development - Integration with other Earth System Data

## Link:

- Modern ocean ecology & occurrence data
- Geologic data from deep-sea sections
- Fossil occurrences and geochronology from NSB
- Taxonomic information

## To support:

- Integrated study of taxa using modern and fossil occurrences, (paleo)ecology data
- Geochronology services for paleoceanography





# Summary: Paleontology Research Occurrence Databases

---

- Several large databases, some as offline systems in early 1970s
  - >2,000,000 occurrence records, ca 500,000 taxa names
- Published, vetted data on taxon occurrences: space & time; plus paleoenvironmental context
- ca 1,000 publications vs ca 1,900 for GBIF
  - but paleontology 1/100 size of neontology: very large footprint in paleo
- Biodiversity vs time; testing evolutionary ecologic models; integrated biodiversity and geologic environmental proxy studies: extinction risk, climate impacts
- Trend towards integration of paleontology and earth science data systems



# What is Research Role of Paleontologic Collection Databases?

---

- *Primary*: Manage and improve access to collection material: the role for which they were originally made
- *Secondary*: To complement existing infrastructure of paleontologic research databases
  - Primary role of data in research analyses no longer likely as better quality infrastructure in use - data quality, metadata, links to earth science data



# Benefits of Linking Paleontologic Research and Collection Databases

---

- Coverage: fill gaps in Research Databases
  - 10-100X more records in collection dbs vs research
  - Community-based input e.g. PBDB means personal preferences control what gets entered
- Abundance data from collection records
  - specimen based records vs taxon based in research dbs
  - are numbers of/in collection records a good proxy for original abundance of fossils in rocks?



# Benefits of Linking Paleontologic Collection Databases

---

- User: Efficient searches (one vs many searches)
- DB Manager: Best practice in database contents
  - Stratigraphic lists for regions (Litho- & Chronostratigraphy)
  - (Rare) expert taxonomic lists



# Future Development - Adding Collection Database Information



**The  
Micropalaeontological  
Society**



**The  
Geological  
Society**

*serving science & profession*

- Based on Lyell meeting “*Palaeoinformatics: Synthesizing data from the past to illuminate the future*” (March, Burlington House, London)
- Role of collection databases one (of many) themes discussed
- Most saw complementary value in research but need for further development:
  - Data exchange protocols & ontologies that fully support paleo data (e.g. geologic age, stratigraphy, etc):
    - Darwin extended and ABCDEFG useful but not enough
  - Advanced data filtering & cleaning abilities to extract ‘good’ data from messy collection database sources
- Workshop under discussion (late 2016 or early 2017)



# Summary

---

- Paleontologic research on occurrences of taxa primarily use well established literature-sourced databases
- This type of research is a central, highly successful theme in paleontology
- Collection databases seen as potentially useful to this research as secondary source of data
- Much work still needed to deal with data quality issues and limits to current database integration technologies