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

David Lazarus *[1] Jeremy Young[2], Shanan Peters[3] and Johan Renaudie[1]

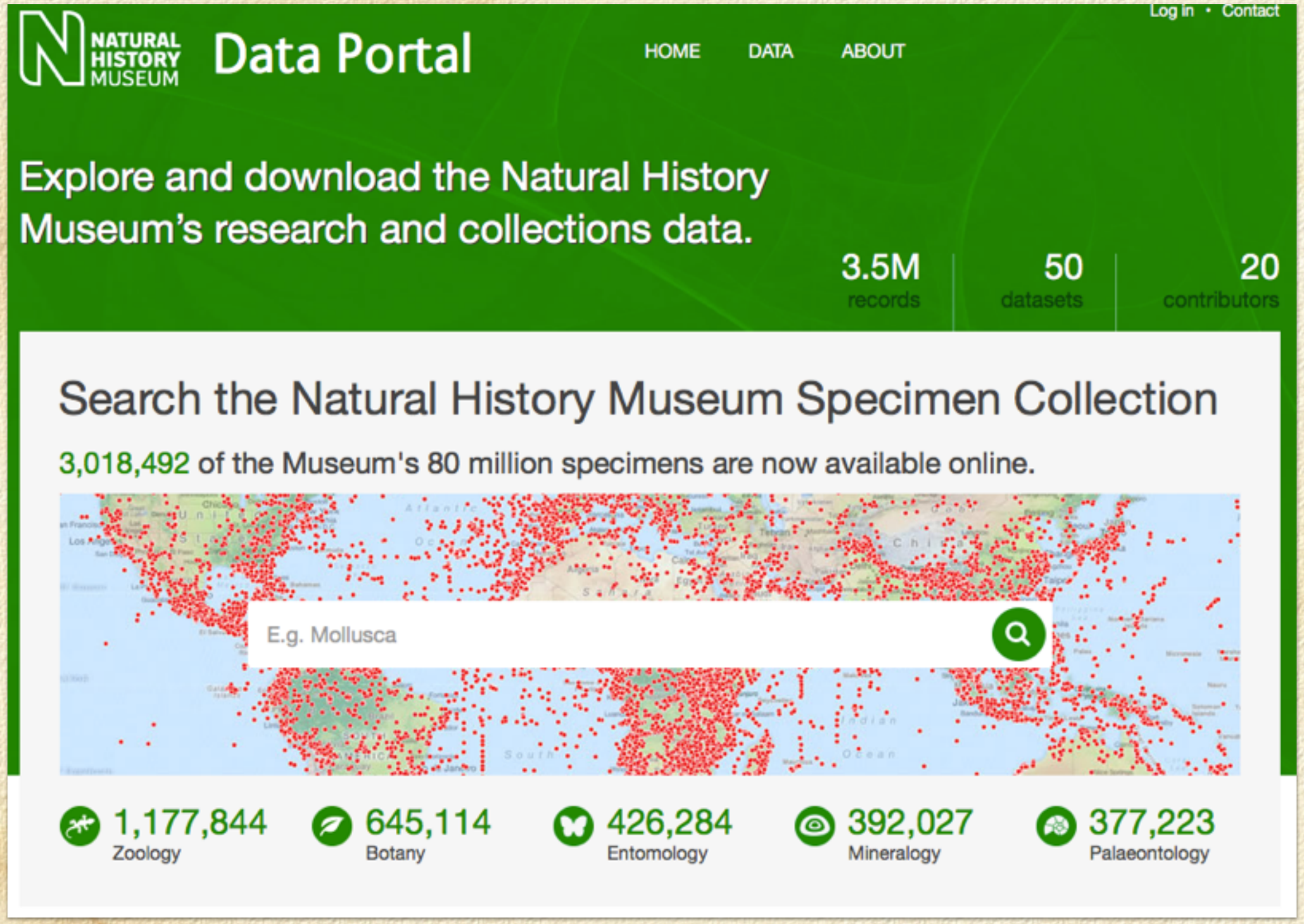
[1] Museum für Naturkunde, Invalidenstrasse 43, 10115 Berlin, Germany.
david.lazarus@mfn-berlin.de

[2] University College London, Earth Sciences, Gower Street, WC1E 6BT, UK
[3] University Wisconsin, Department of Geoscience, Madison, 1215 W Dayton St, WI 53706, USA

Outline of talk

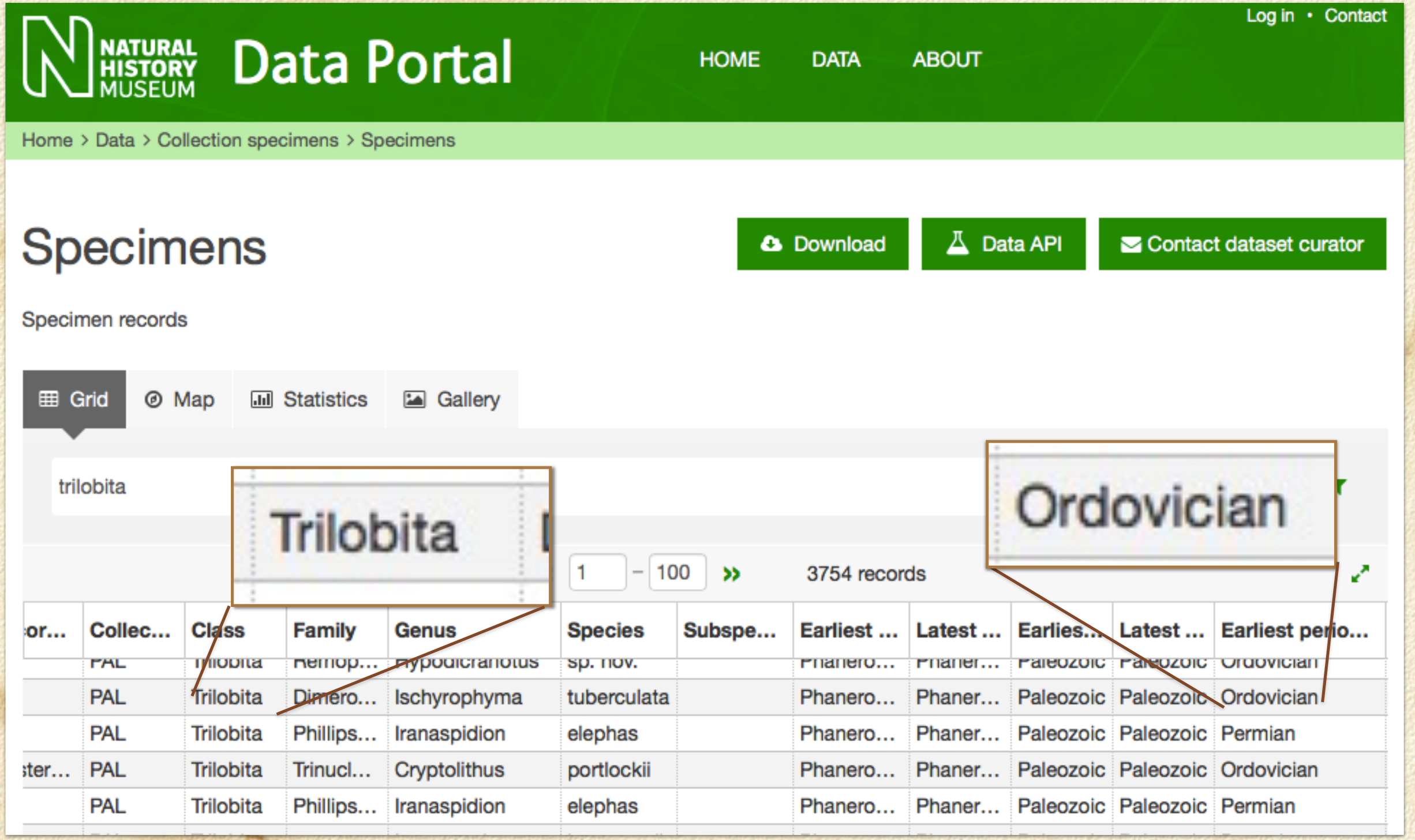
- Collection databases, portals and limitations
- Paleontologic research databases
- Future directions and summary

Collection Databases



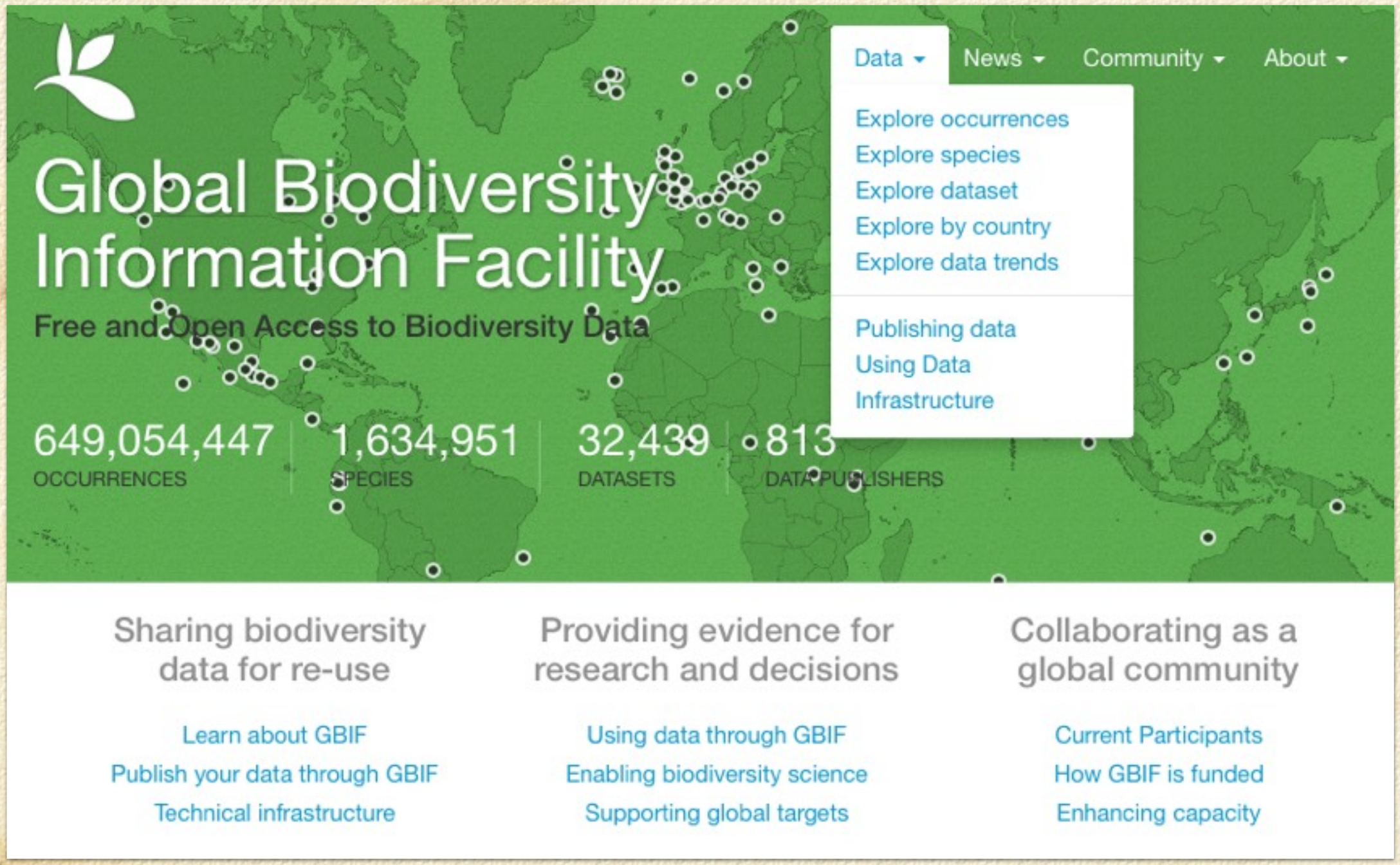
- Most digital occurrence data for biologic objects is in collection databases
- Most can be accessed online via individual museum websites

Collection DB Access at Host Museum



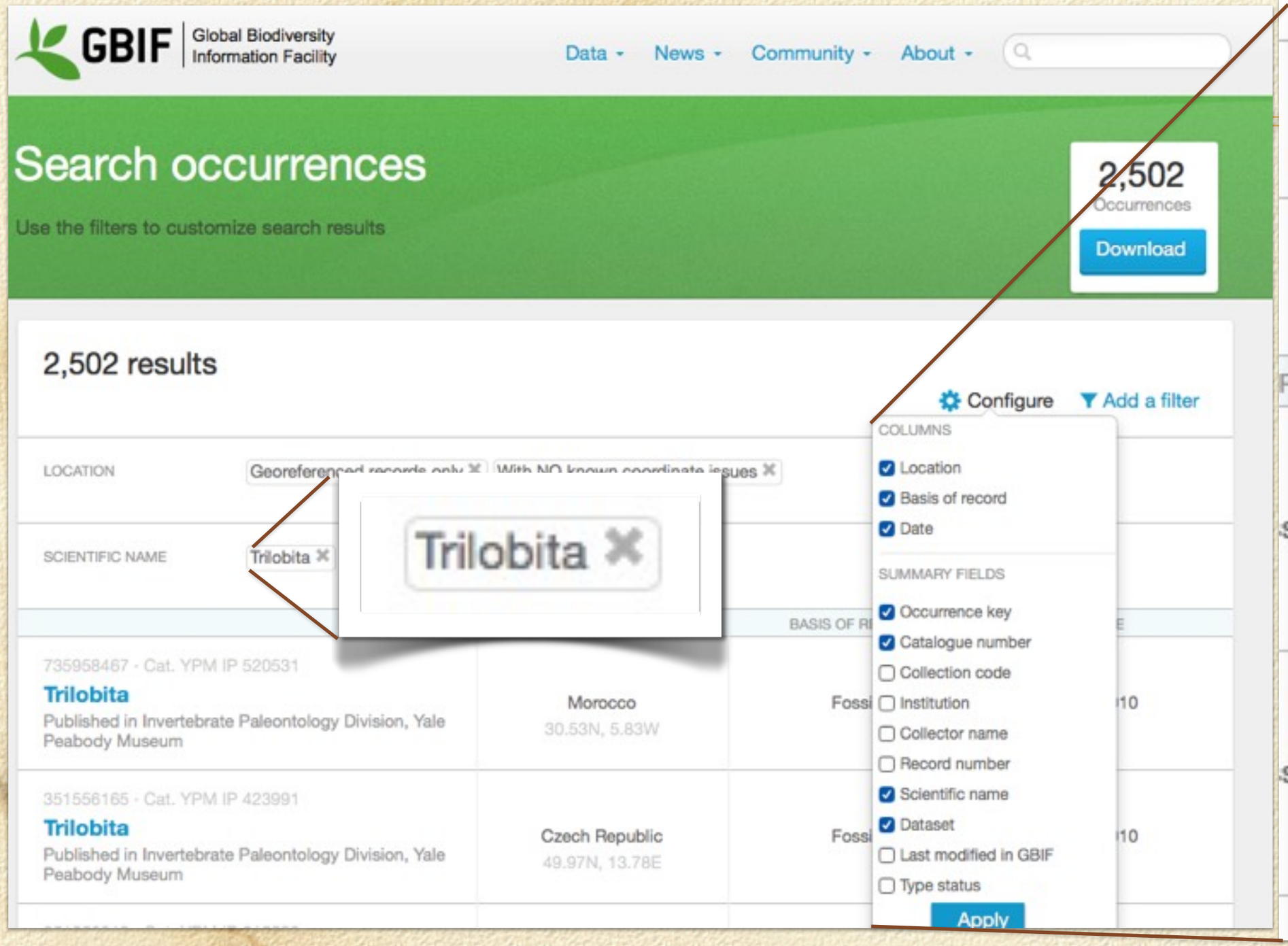
- All 3 parts of basic paleo record usually provided: taxa, locality, age
- Usually little else given: content is rather skeletal from research view

Portals



- 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



- 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

- 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):
Taxon name:
Country/continent:
State/province:
County/parish:
Reference #:
Time interval (or age in Ma): to
Group, formation, or member:
Paleoenvironment:
Lithology:
Data authorizer:
Group/project:

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: collection number | ascending
Number of records per page: 30

- Occurrences have (ideally) detailed lithology & paleoenvironment data

Neptune Database (NSB)

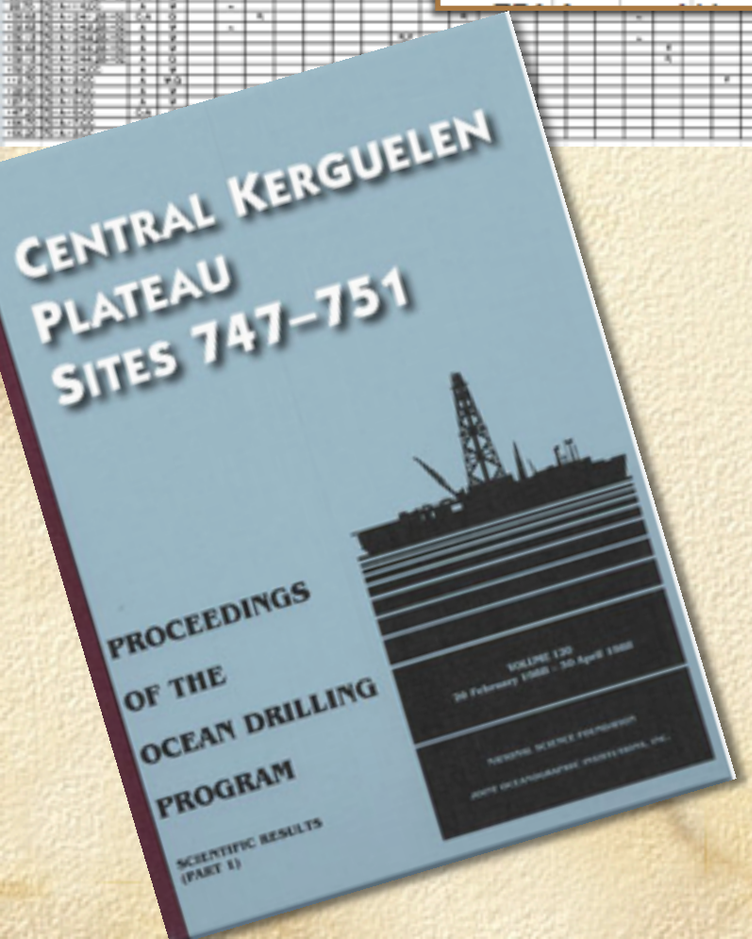
www.nsb-mfn-berlin.de

- 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

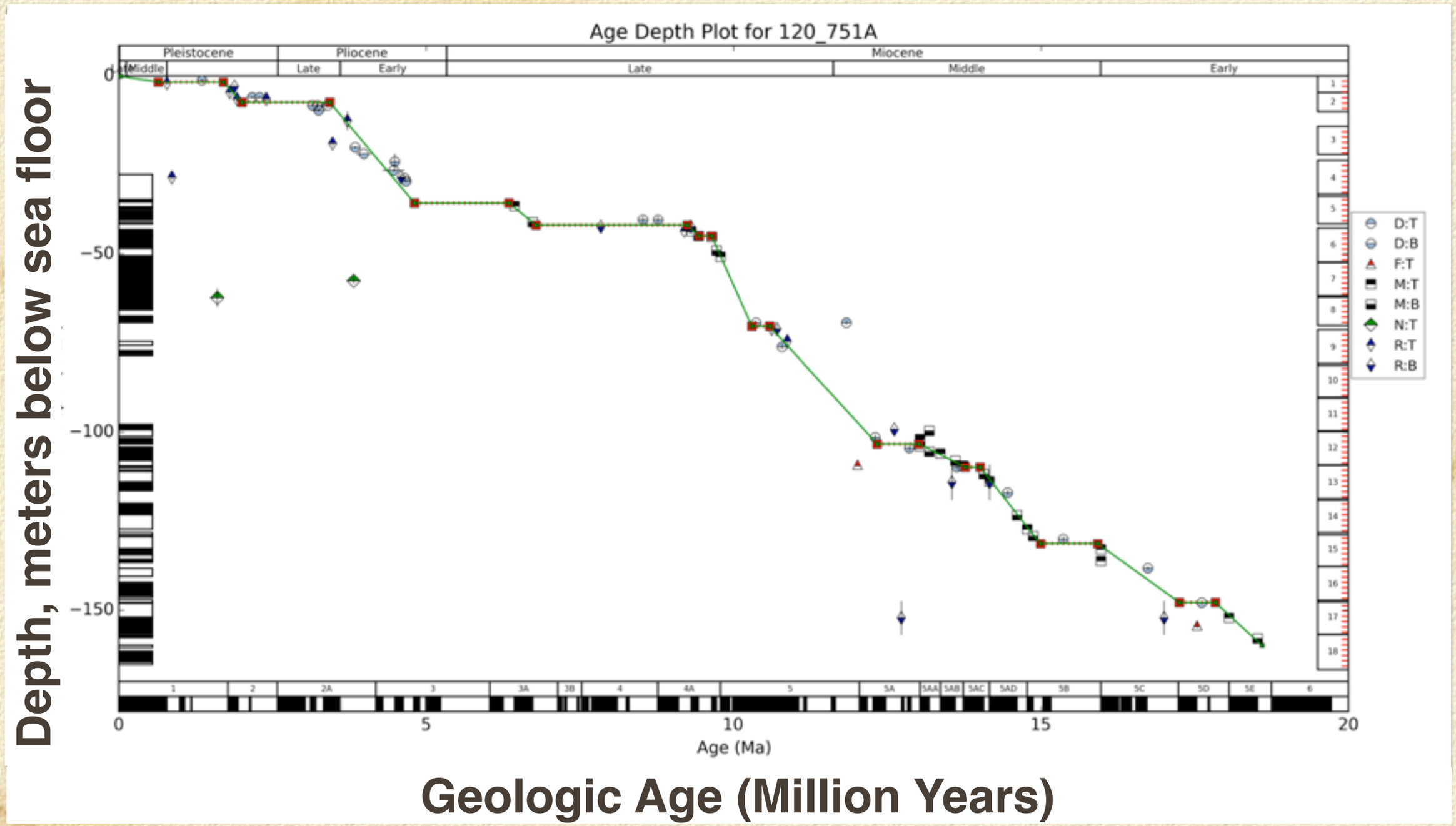
NSB Content (1)

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

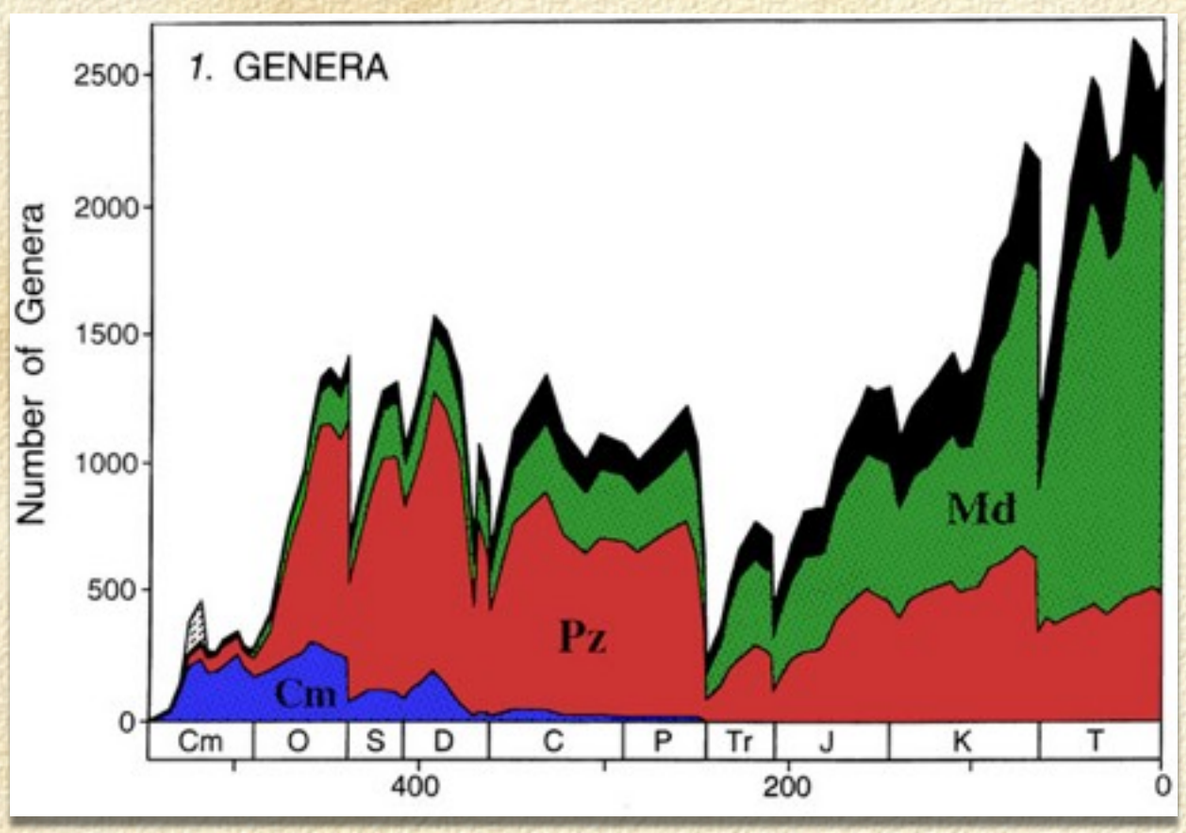
NSB Content (2)



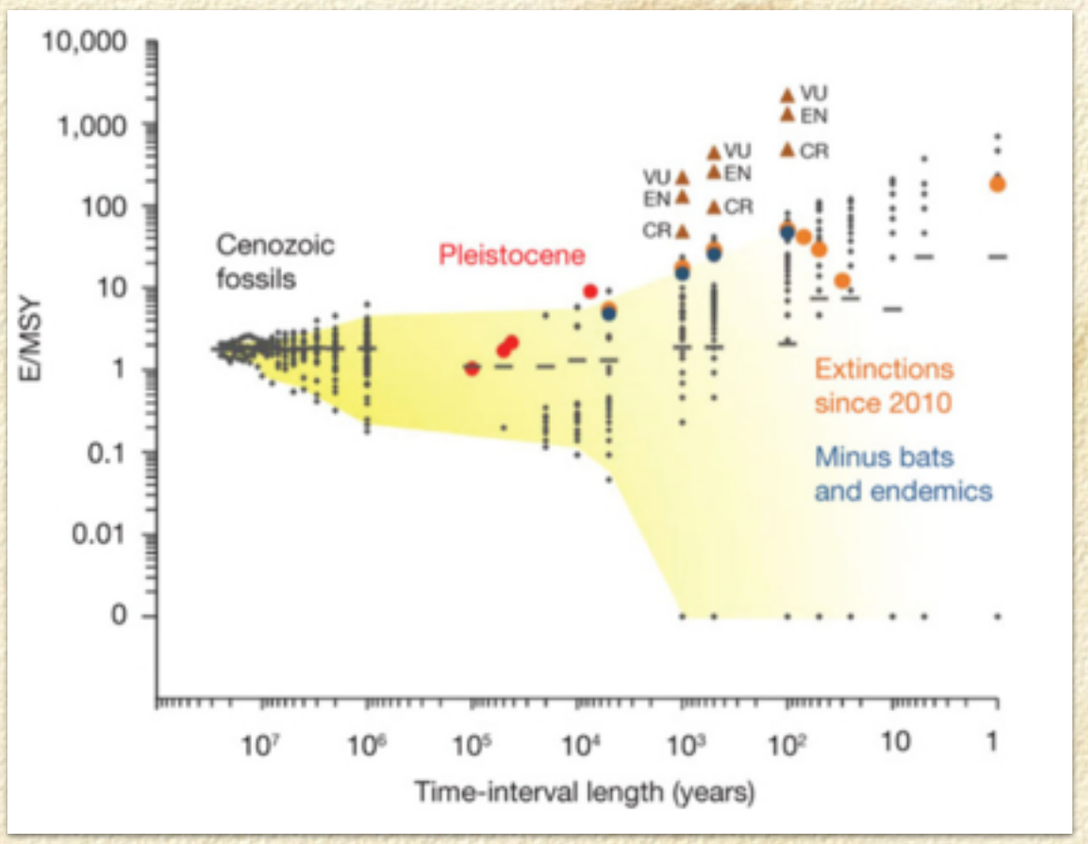
- Geologic age of occurrences are internally *calculated* based on age model function (green line)
- NSB also contains all data & calibrations used to construct age model

Research Use

Sepkoski 3-Fauna Model (1975-1995)



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

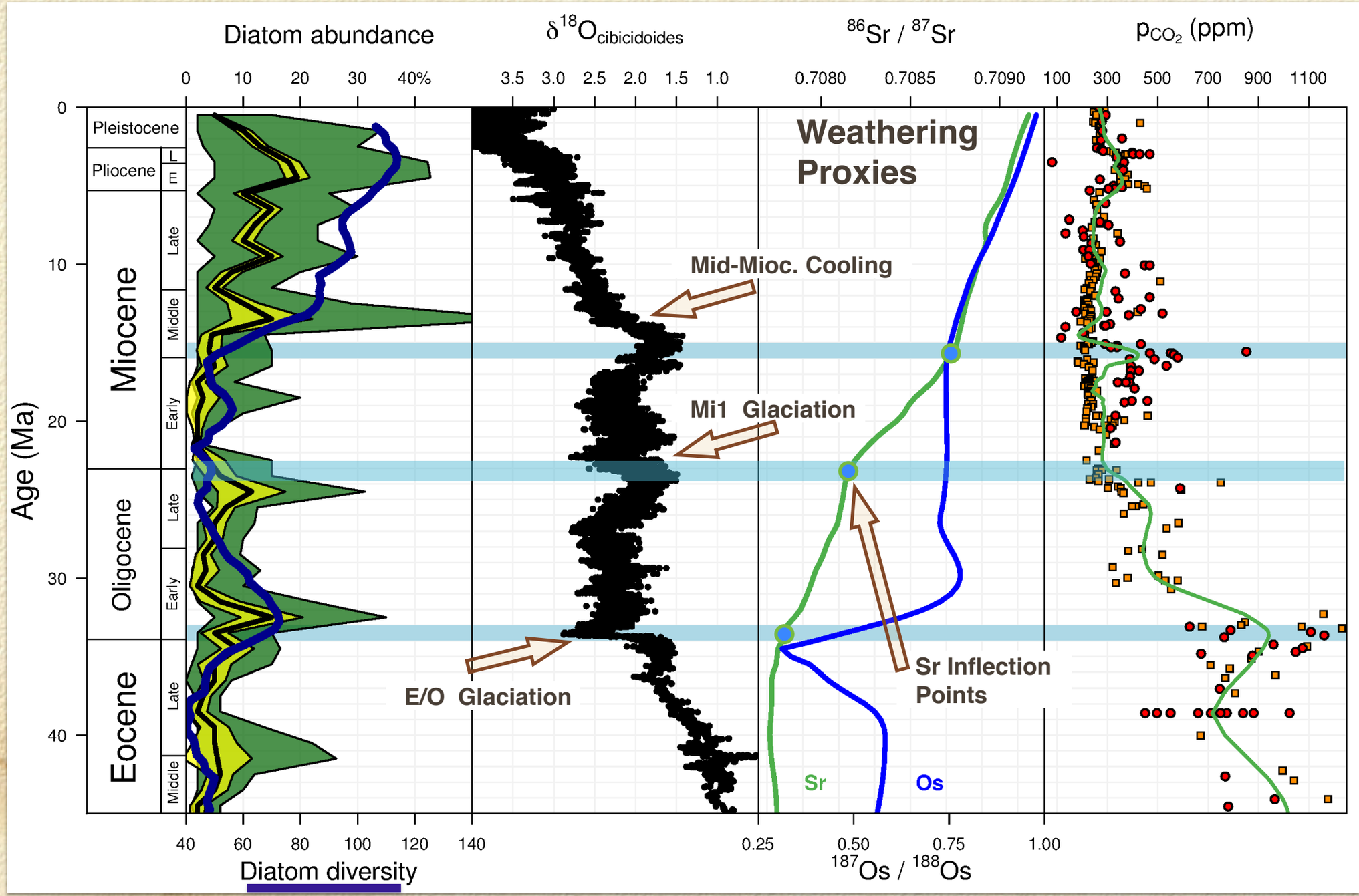


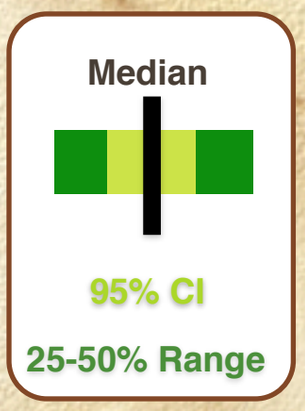
- Classic Biodiversity Dynamics - testing models such as the Red Queen, Evolutionary carrying capacity of environments
- Mass Extinctions and Recoveries
- *Extinction Risk from Global Warming*
- *Bio-Geo Interactions on evolutionary timescales*

Integrated Bio-Geosphere Studies

- NSB geochronology + Deep-sea sediment data = **Biogenic silica flux history**
- Silica Flux → Global Weathering, **pCO2**, (**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)





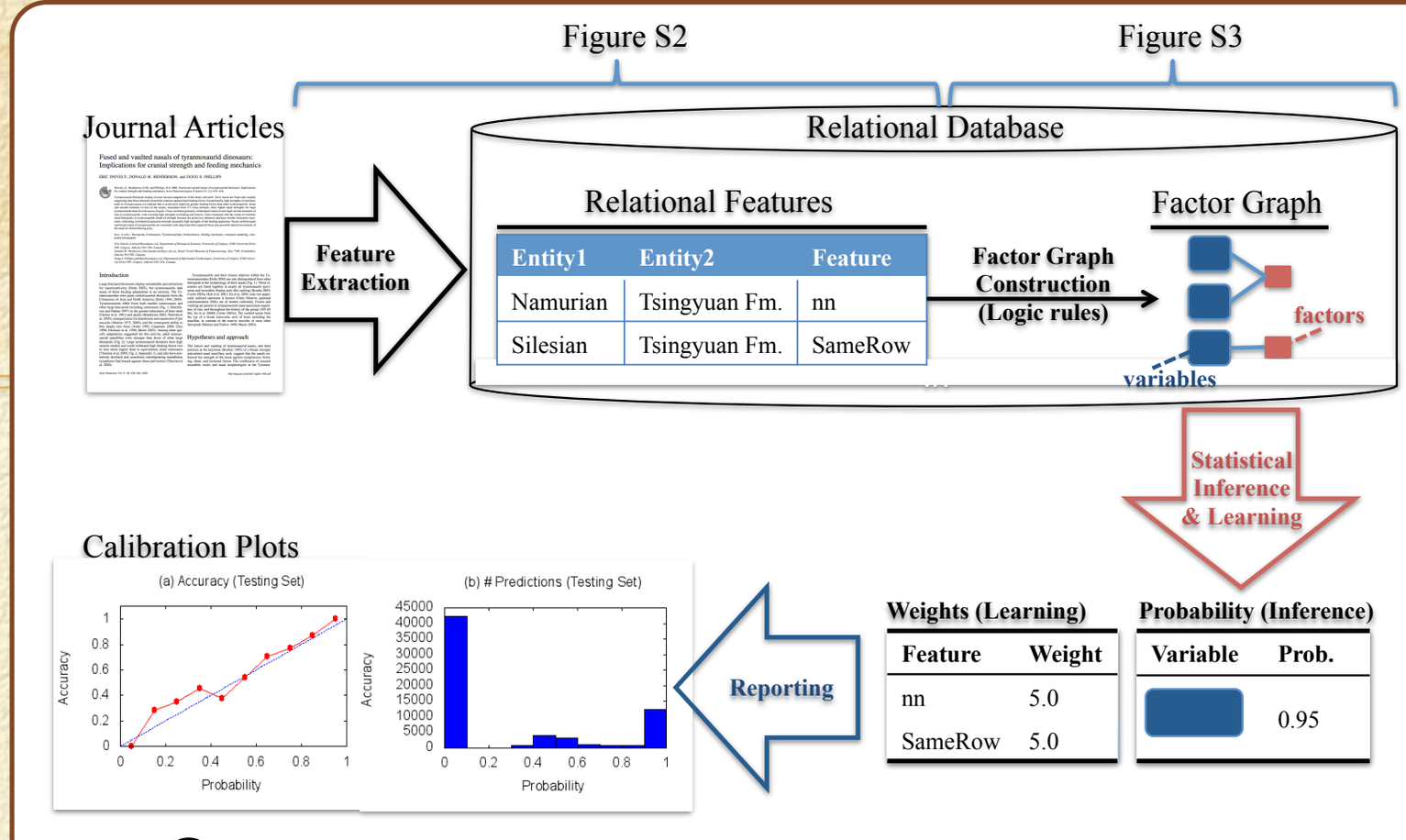
Future Development - Linking Research Database Systems

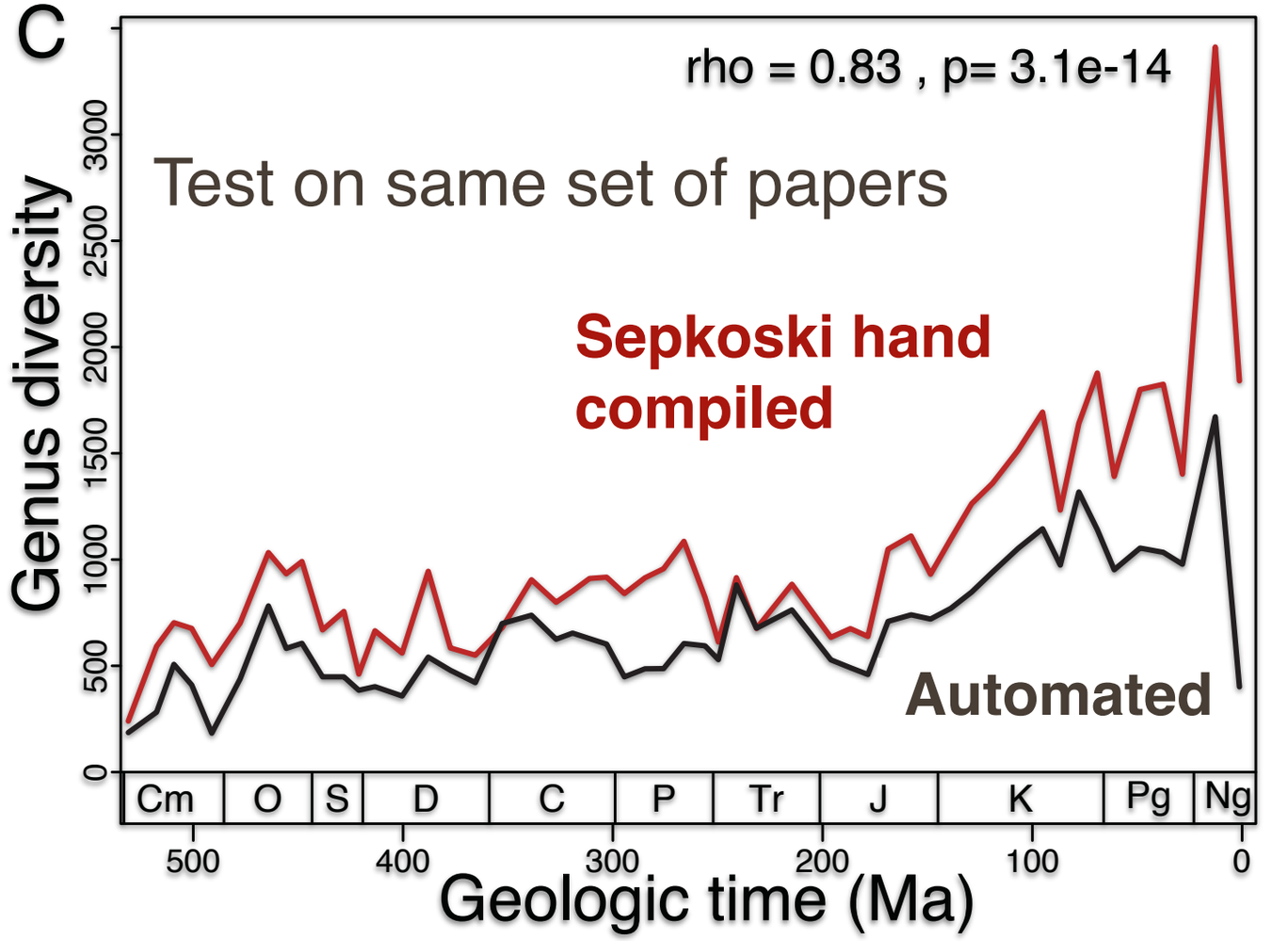
- Monolithic database for paleontologic occurrence data neither needed or desired: federation of databases better
- Fossil occurrence data have very different quality by source
  - Deep-sea marine microfossil data : standardized species, ‘matrix’ datasets, age-modeled sections, earth science measurements 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
- lrst-order patterns similar
- ? 2nd-order patterns
- Massively hi-tech requirements





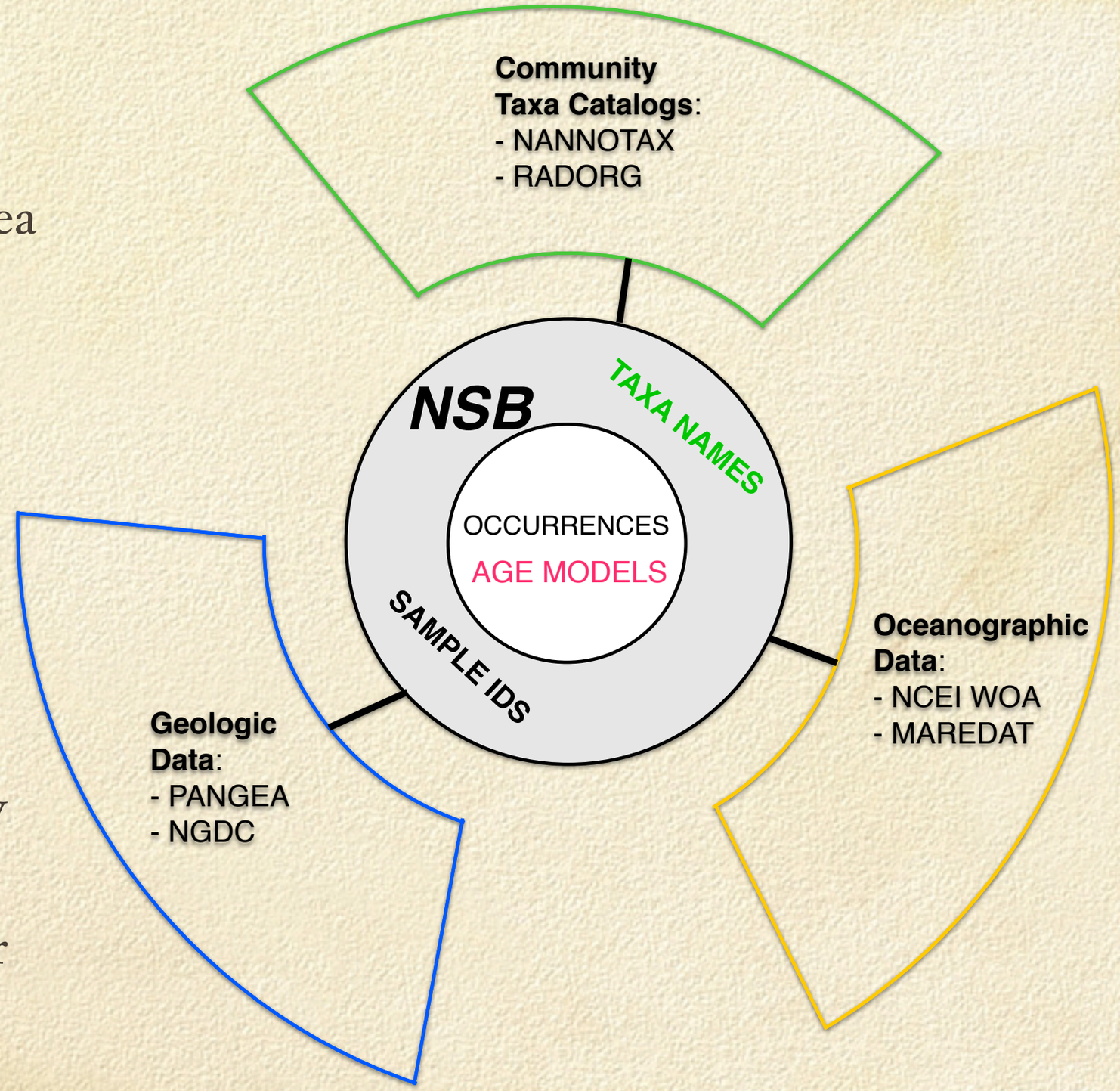
Future Development - Integration with other Earth System Data

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

- 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