Horses in the cloud: Big data exploration and mining of fossil and extant *Equus* (Mammalia, Equidae)

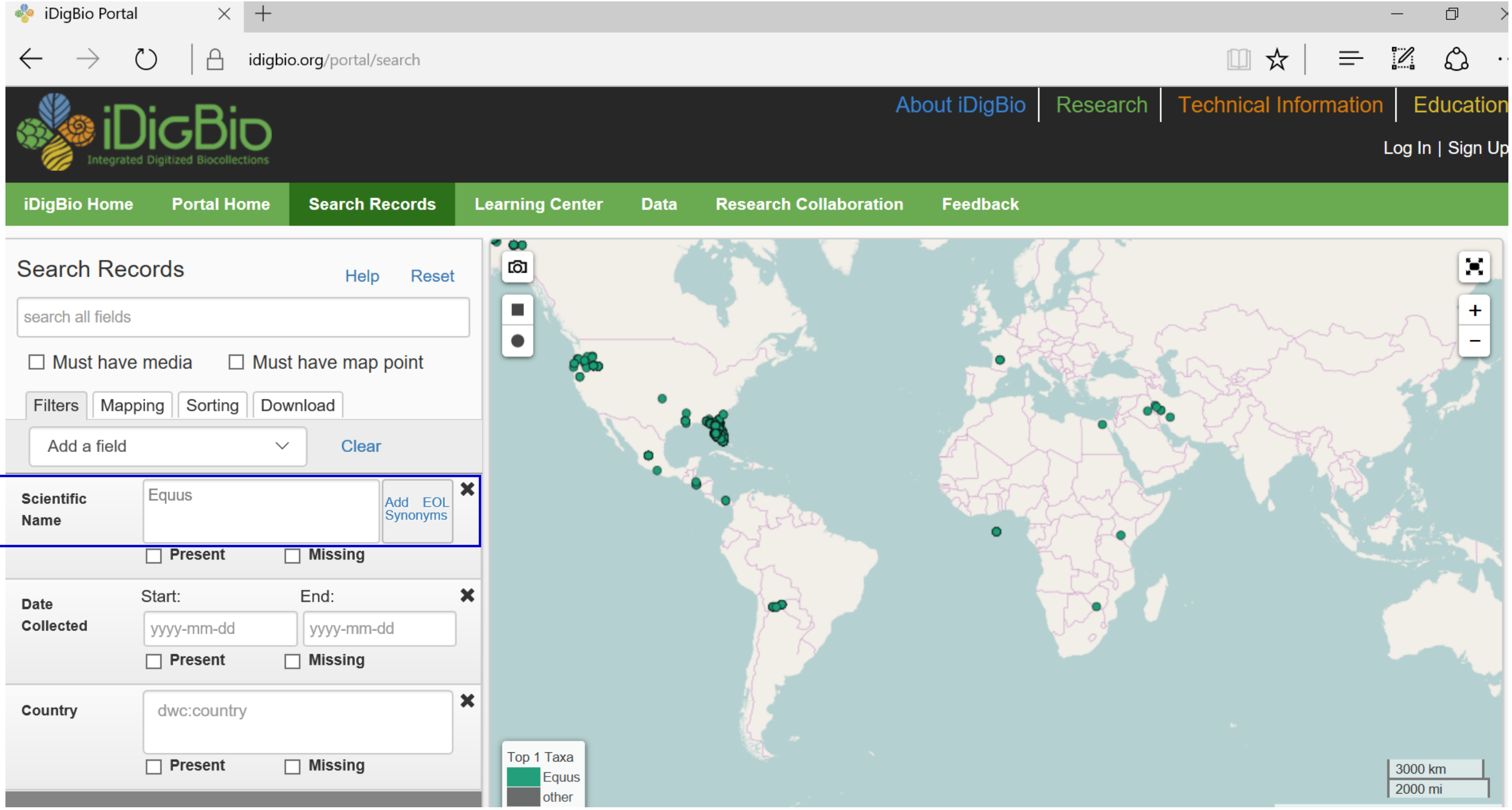
Bruce J. MacFadden & Robert P. Guralnick

Florida Museum of Natural History

University of Florida

Gainesville FL 32611

Initial query November 2015



Scientific Name: Equus

iDigBio
Integrated Digitized Biocollections

Talk outline—*Equus* use case

- Data exploration in 2016
- Big biodiversity databases and mining results
  - iDigBio, PBDB, GBIF
- Analysis
  - Integration
  - Geographic Bias
  - Holy Grail—integrated chronological data
- Future
  - *Equus* extinction geography
  - Ancillary data attached to vouchered specimens

MacFadden & Guralnick 2016 *Paleobiology* DOI: http://dx.doi.org/10.1017/pab.2016.42

iDigBio Integrated Digitized Biocollections

Equus: Initial exploration and metaresearch

- Question I wanted to answer:

What was the extinction geography of *Equus* since the Last Glacial Maximum?

Available databases did not have sufficient age data

- Then became a "metaresearch" analysis:

The scientific examination of how research is designed, carried out, and communicated (Kousta et al. 2016)

MacFadden & Guralnick 2016 *Paleobiology* DOI: http://dx.doi.org/10.1017/pab.2016.42

iDigBio
Integrated Digitized Biocollections

Which big database is optimal?

- Depends upon
  - Taxon or taxa studied
  - Question to be asked
  - Chronological precision required

- Use case example
  - *Equus*, fossil and extant
  - Late Pleistocene extinction geography

- Perhaps best to integrate multiple databases?

MacFadden & Guralnick 2016 *Paleobiology* DOI: http://dx.doi.org/10.1017/pab.2016.42

Big biodiversity databases

- Over past decade number has grown
- Goal: aggregate big data to ask novel questions
- Six were investigated here--

iDigBio
Integrated Digitized Biocollections

The Paleobiology Database
revealing the history of life

GBIF
Global Biodiversity
Information Facility

VertNet

Neotoma Paleoecology Database

USGS
science for a changing world
Biodiversity Information Serving Our Nation (BISON)

MacFadden & Guralnick 2016 *Paleobiology* DOI: http://dx.doi.org/10.1017/pab.2016.42

Big biodiversity databases

- Over the past decade number has grown
- Goal: aggregate big data to ask novel questions
- Six were investigated here
- Three were most useful for this study

iDigBio
Integrated Digitized Biocollections

The Paleobiology Database
revealing the history of life

GBIF
Global Biodiversity
Information Facility

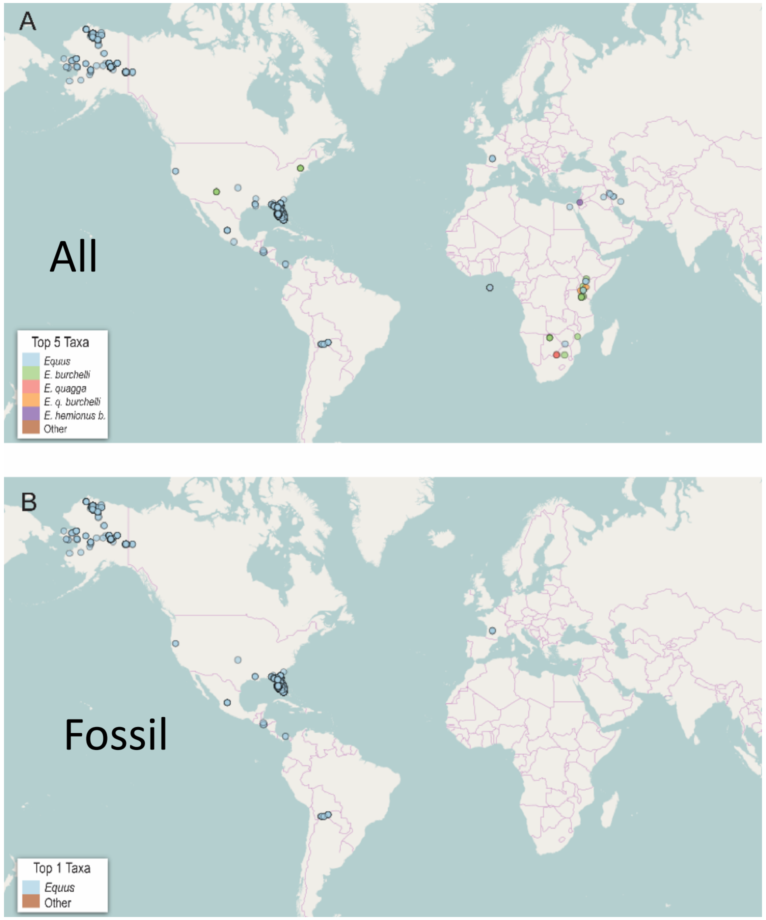
VertNet

Neotoma Paleoecology Database

USGS
science for a changing world
Biodiversity Information Serving Our Nation (BISON)

MacFadden & Guralnick 2016 *Paleobiology* DOI: http://dx.doi.org/10.1017/pab.2016.42

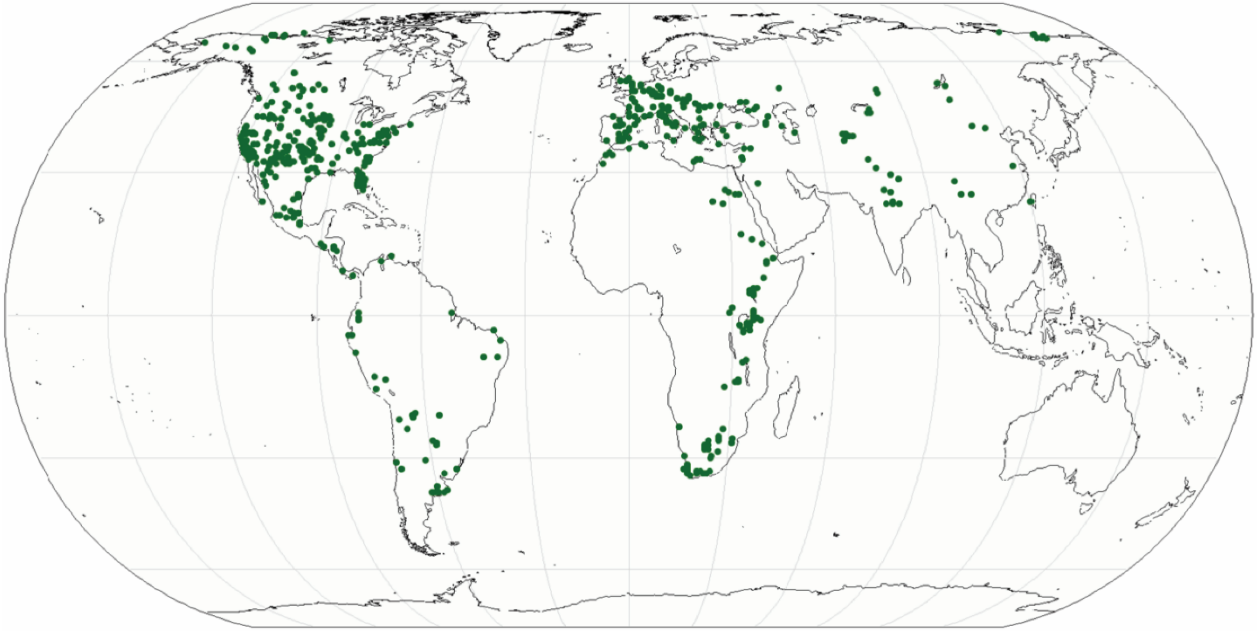
iDigBio (Integrated Digitized Biocollections)



- 64.6 million records
- vouchered specimens
- 22.4 K *Equus* records; 21.9 K fossil
- Concentrated (e.g., Alaska, Florida)
- Primary coverage North America

MacFadden & Guralnick 2016 *Paleobiology* DOI: http://dx.doi.org/10.1017/pab.2016.42

Paleobiology Database (PBDB)

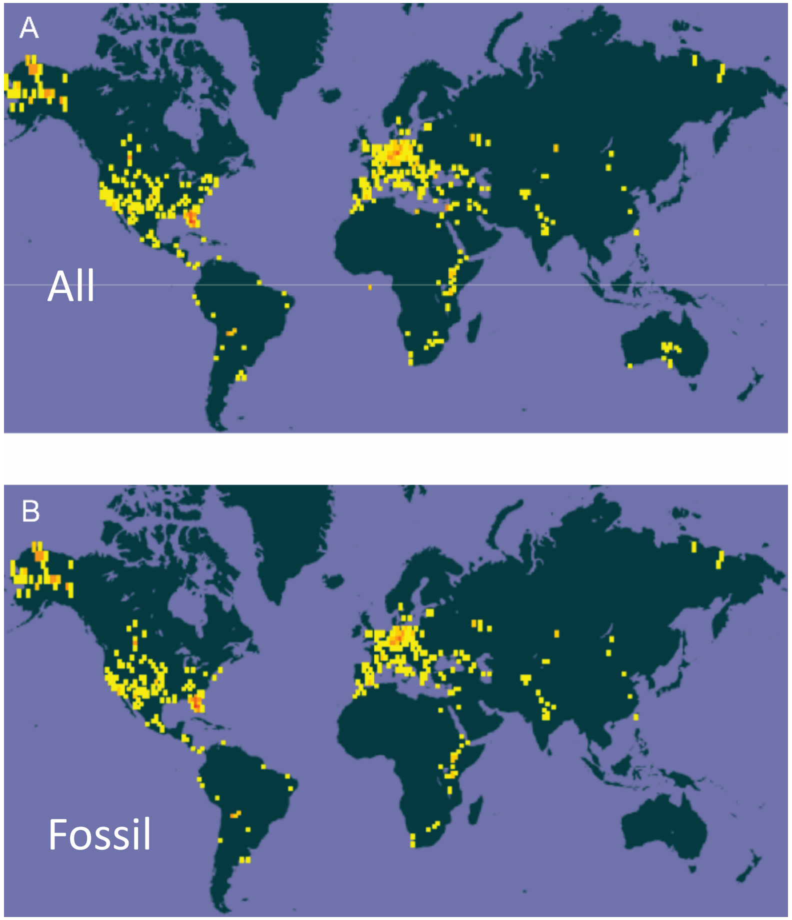


- 1.3 million occurrence records; not directly vouchered specimens
- 1.6 K fossil records for *Equus*
- More global coverage
- Age data not sufficiently binned for late Pleistocene

MacFadden & Guralnick 2016 *Paleobiology* DOI: http://dx.doi.org/10.1017/pab.2016.42

iDigBio Integrated Digitized Biocollections

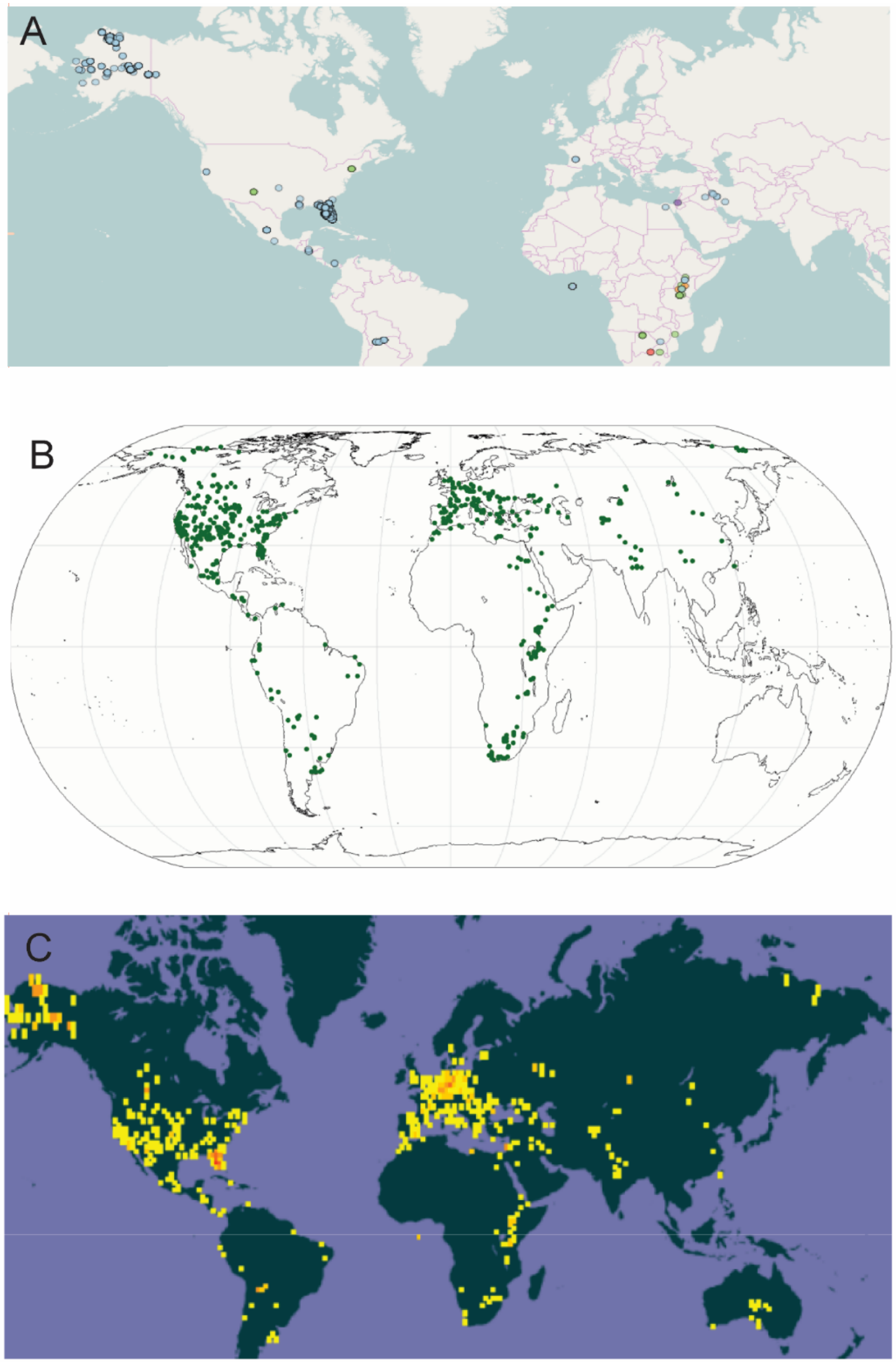
GBIF (Global Biodiversity Information Facility)



- 642 million total location data from > 400 data providers
- Vouchered and non-vouchered observations
- 44.5 K *Equus* records, including 42.4 fossil
- Broader coverage than iDigBio
- Age data still problematical

MacFadden & Guralnick 2016 *Paleobiology* DOI: http://dx.doi.org/10.1017/pab.2016.42

iDigBio
Integrated Digitized Biocollections



Summary Comparison

- All databases yielded 124 K *Equus* records; 116 K fossil
- Massive amounts of data
- iDigBio—vouchered specimen records, DarwinCore standards
- PBDB—relatively good (fossil) coverage despite only 1.6 K *Equus* records
- GBIF—Most complete fossil and extant coverage for *Equus;* mixed records perhaps problematic.

iDigBio
Integrated Digitized Biocollections

Database integration

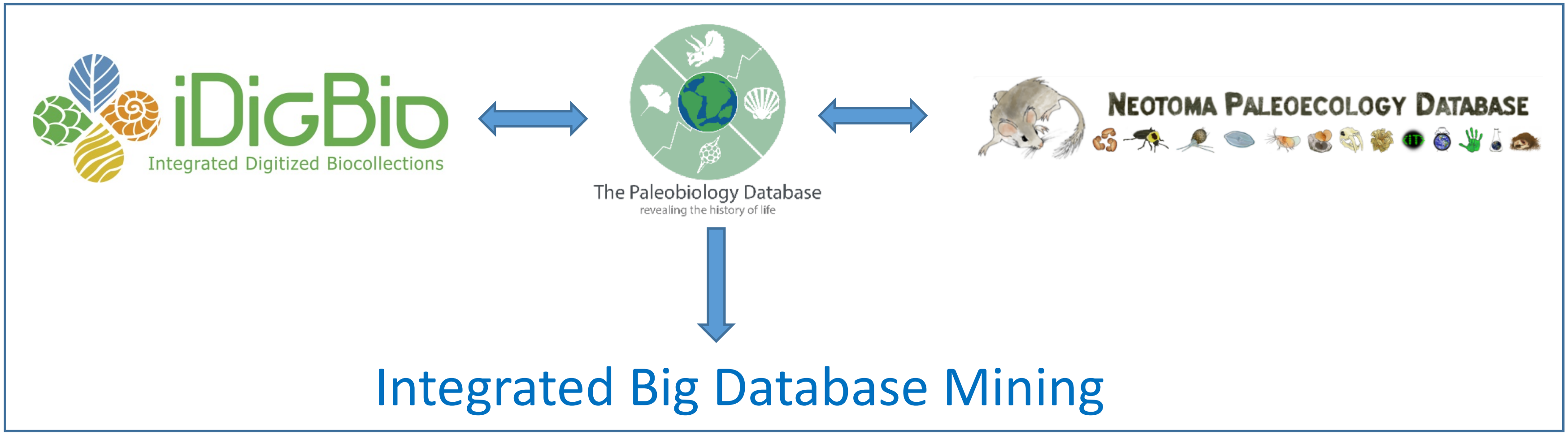
- Optimal scenario would be to simultaneous mine data from all relevant databases.
- But, current problem is that data semantics and standards are not universal across platforms.
- For example, ‘occurrence’ in PBDB equals DarwinCore ‘location’ in iDigBio and GBIF.
- These need to be made equivalent.

MacFadden & Guralnick 2016 *Paleobiology* DOI: http://dx.doi.org/10.1017/pab.2016.42

iDigBio
Integrated Digitized Biocollections

Which database is optimal?

- Depends upon the question being asked
- Perhaps a better approach would be to integrate multiple relevant databases.
- ePANDDA is currently doing this.


iDigBio
Integrated Digitized Biocollections
The Paleobiology Database
revealing the history of life
Neotoma Paleoecology Database

Integrated Big Database Mining

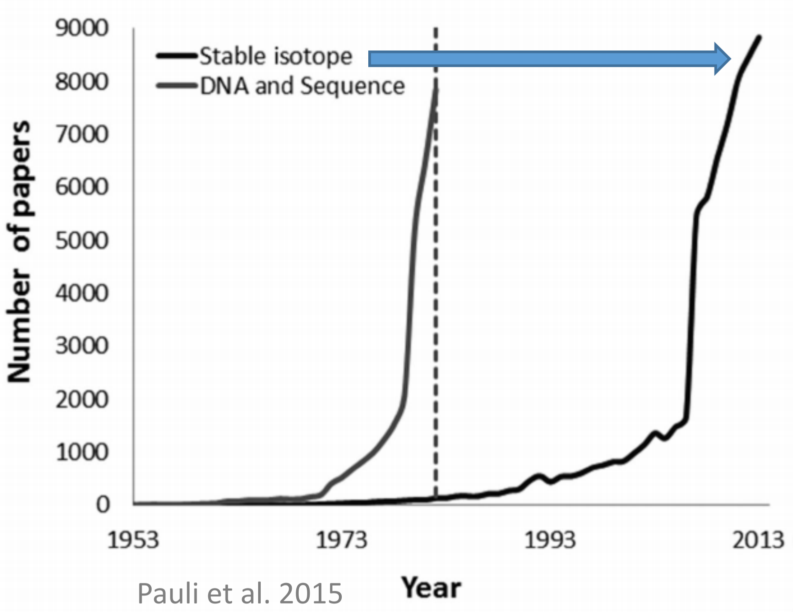
Holy grail—integrated age data

To study extinction geography of *Equus*—

- Big biodiversity databases need to integrate precise and binned chronological data.
- Neotoma currently has the lead in this regard, although with only a few hundred relevant records.
- The big advances in paleo will come once this is done; or other research is envisioned that does not require precise chronology (e.g., distributions).

MacFadden & Guralnick 2016 *Paleobiology* DOI: http://dx.doi.org/10.1017/pab.2016.42

Leveraging big data: ancillary fields

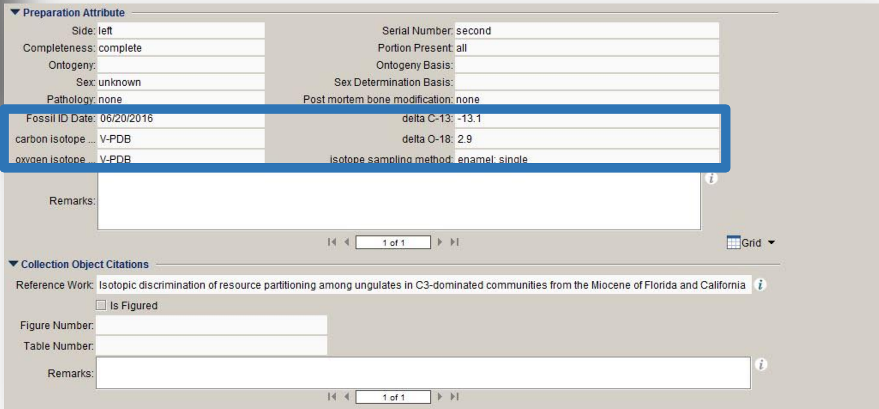


Pauli et al. 2015

Specify

ISOBANK

Our insertion of isotope data fields



Preparation Attribute

Side:	left	Serial Number:	second
Completeness:	complete	Portion Present:	all
Ontogeny:		Ontogeny Basis:	
Sex:	unknown	Sex Determination Basis:	
Pathology:	none	Post mortem bone modification:	none
Fossil ID Date:	06/20/2016	delta C-13:	-13.1
carbon isotope ...	V-PDB	delta O-18:	2.9
oxygen isotope ...	V-PDB	isotope sampling method:	enamel; single
Remarks:			

Collection Object Citations

Reference Work: Isotopic discrimination of resource partitioning among ungulates in C3-dominated communities from the Miocene of Florida and California

□ Is Figured

Figure Number:

Table Number:

Remarks:

Moran et al. 2016. GSA Annual Meeting talk, Denver

Concluding comments

- Big biodiversity databases in paleontology
    - Massive amounts of data (*Equus* use case 124 K records)
    - **Potential** to answer new questions
- *Equus* paleo(geographic) data are dense, but biased towards N America.
- Ancillary data fields will greatly increase utility
- “Big data” Museum bioinformatics will advance with
    - More precise age data
    - Standards integration (Darwin Core), ePANDDA, etc.