Paleobiology Specimen Data and the Role of Data Aggregators:

The iDigBio Perspective

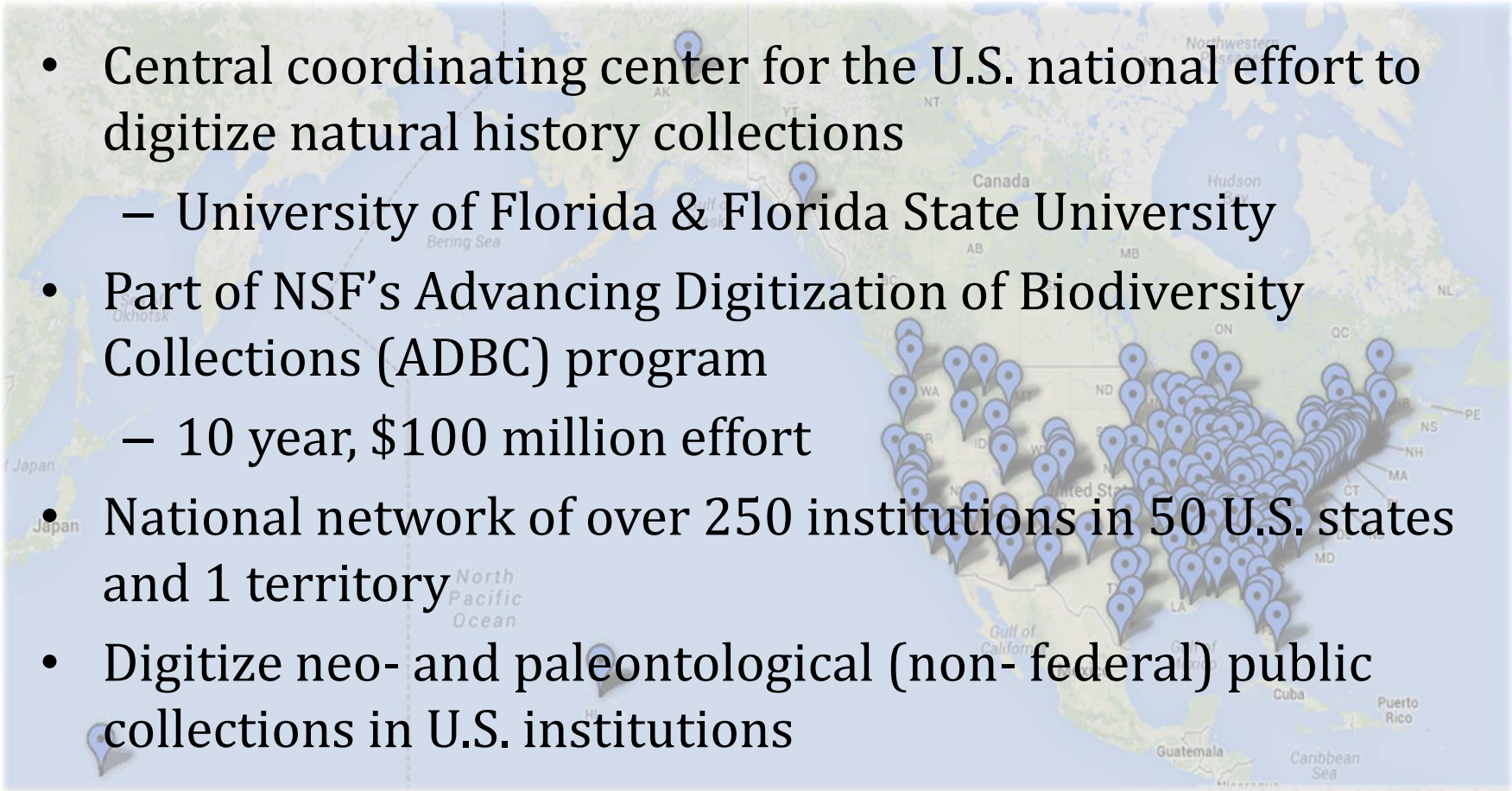
Shelley James & Gil Nelson

iDigBio

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

GSA 2016

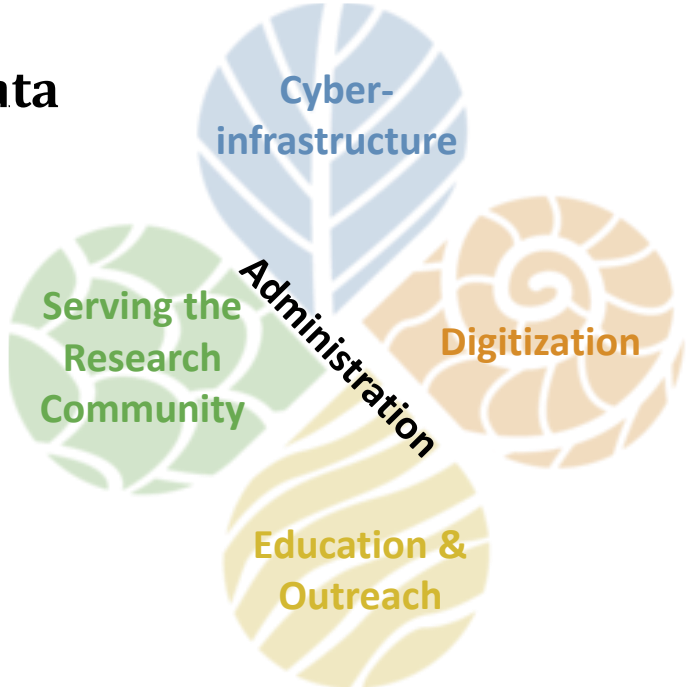
Integrated Digitized Biocollections: who we are



- Central coordinating center for the U.S. national effort to digitize natural history collections
  - University of Florida & Florida State University
- Part of NSF's Advancing Digitization of Biodiversity Collections (ADBC) program
  - 10 year, $100 million effort
- National network of over 250 institutions in 50 U.S. states and 1 territory
- Digitize neo- and paleontological (non- federal) public collections in U.S. institutions

What do we do?

- **Enable digitization of biodiversity collections data**
  - Develop efficient & effective standards & workflows
  - Workforce education & training
- **Provide portal access to biodiversity data in a cloud computing environment**
  - Respond to cyberinfrastructure needs
  - Enable access & discoverability
- **Facilitate use of biodiversity data to address key environmental and economic challenges**
  - Researchers, educators, general public, policy-makers, …
- **Plan for long-term sustainability of the national digitization network & effort**
  - Expand participation: partners, data sources, public, …
  - Proliferate and broaden uses of biodiversity data



18 Thematic Collections Networks (TCNs)

- **Invert**Net: An Integrative Platform for Research on Environmental Change, Species Discovery and Identification
- **Plants, Herbivores, and Parasitoids**: A Model System for the Study of Tri-Trophic Associations
- North American **Lichens and Bryophytes**: Sensitive Indicators of Environmental Quality and Change

Digitizing Fossils to Enable New Syntheses in Biogeography - Creating a PALEONICHES-TCN

- The **Macrofungi** Collection Consortium: Unlocking a Biodiversity Resource for Understanding Biotic Interactions, Nutrient Cycling and Human Affairs
- Mobilizing New England **Vascular Plant** Specimen Data to Track Environmental Change
- Southwest Collections of **Arthropods** Network (SCAN): A Model for Collections Digitization to Promote Taxonomic and Ecological Research

iDigPaleo: Fossil Insect Collaborative: A Deep-Time Approach to Studying Diversification and Response to Environmental Change

- Developing a Centralized Digital Archive of Vouchered **Animal Communication Signals**
- The **Macroalgal** Herbarium Consortium: Accessing 150 Years of Specimen Data to Understand Changes in the Marine/Aquatic Environment
- Documenting the Occurrence through Space & Time of Aquatic **Non-indigenous Fish, Mollusks, Algae, & Plants** Threatening North America's Great Lakes
- The Key to the Cabinets: Building and Sustaining a Research Database for a Global Biodiversity Hotspot (**plants**)
- InvertEBase: reaching back to see the future: species-rich **invertebrate faunas** document causes and consequences of biodiversity shifts
- The **Microfungi** Collections Consortium: A Networked Approach to Digitizing Small Fungi with Large Impacts on the Function and Health of Ecosystems

Documenting Fossil Marine Invertebrate Communities of the Eastern Pacific - Faunal Responses to Environmental Change over the last 66 million years

- **Lepidoptera** of North America Network: Documenting Diversity in the Largest Clade of Herbivores

The Cretaceous World: Digitizing Fossils to Reconstruct Evolving Ecosystems in the Western Interior Seaway

- The Mid-Atlantic Megalopolis: Achieving a greater scientific understanding of our urban world (**plants**)

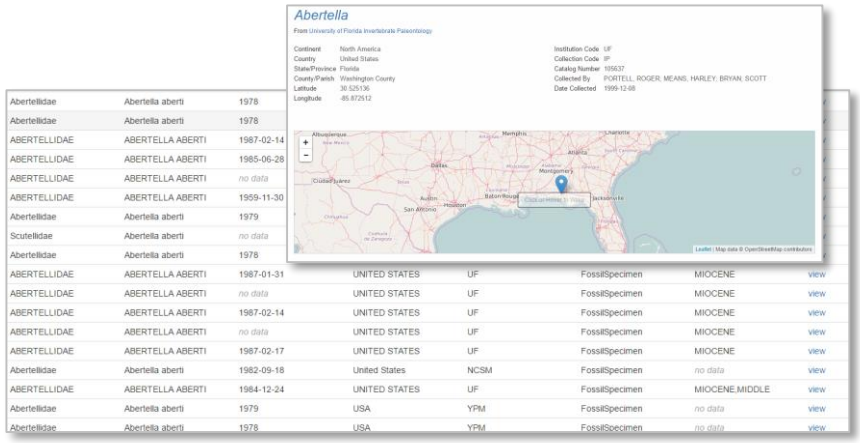
Increasing opportunities for collections research

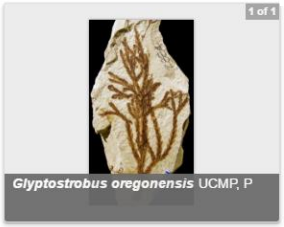
iDigBio: HUB and data aggregator

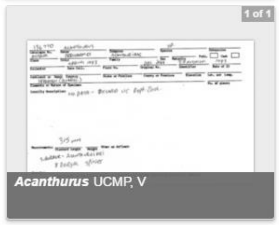
Our scope:

- Vouchered specimen data Darwin Core
- Associated media Audubon Core
- Global
- All taxa (neo- & paleo)

Biodiversity Information Standards TDWG







Why collections publish/share data with iDigBio

- Discovery and use
- Data quality, improvement

Specimen Record

Data | Flags | Raw

Type	Description
scientificname_added	
geopoint_low_precision	Geographic Coordinate has Low Precision.
dwc_phylum_added	Darwin Core Phylum Added.
dwc_country_replaced	Darwin Core Country Corrected.
idigbio_isocountrycode_added	iDigBio ISO 3166-1 alpha-3 Country Code Added.
dwc_kingdom_added	Darwin Core Kingdom Added.

- Attribution, credit, collection value metrics

Recordset

Data Corrected | Data Use | Raw

The table below represents monthly iDigBio portal use statistics for this recordset. ***Search*** indicates in how many instances a record from this recordset matched a search query. ***Download*** indicates in how many instances a record from this recordset was downloaded. ***Seen*** indicates in how many instances a record from this recordset appeared (visually) in the search results in a browser window. ***Records Viewed*** and ***Media Viewed*** indicate how many specimen and media records were opened and viewed in full detail. Note: Monthly statistics aggregation began on Jan 15th 2015; therefore, the month of (01 / 2015) represents approximately half a month of statistics reporting.

Month of	Search	Download	Seen	Records Viewed	Media Viewed
01 / 2015	4,644,439	39,693	249	185	36
02 / 2015	14,478,408	18,174	872	93	6
03 / 2015	11,134,952	92,788	604	136	8

Data use!

PRI	112	28,474	505	0
SDSM-VP	767	31,749	23	149
NCSM-Bot	1,848	10,031	4,615	5,007
VMNH-IP	1,953	820,941	168	83
OU-IP	2,412	3,725	234	0
AMNH-IP	7,218	20,316	24	782
NCSM-IP	11,877	375,128	32,422	4,347
ROM_nonMam	27,312	881,618	266	269,540
NCSM-VP	31,394	212,267	63,018	47,976
UAM	36,657	951,925	124,371	154,354
ROM-Mam	49,672	672,794	706	173,092
Yale-VP	71,805	332,854	11,876	35,482
NRM (Sweden)	73,974	943,447	149	17,732
CU-IP	74,862	304,508	290	8,410
CM	79,254	1,451,295	224,338	66,530
KU-VP	79,979	219,905	75	165,043
Yale-Bot	91,261	4,711,819	2,957	253,575
UT-NPL	99,463	6,337,018	358	49,224
OMNH-IP	210,019	197,664	167	25,302
KU-IP	236,371	634,381	318,213	46,499
UF-IP	252,944	1,714,693	364	373,671
MNHN (France)	270,867	39,548	48	68,710
Yale-IP	361,213	7,192,437	7,082	2,234,510
UF-VP	377,038	2,049,541	1,162	3,059,768
UCMP	432,668	3,175,747	515,810	1,423,913
Total	**2,882,940**	**33,313,825**	**1,309,241**	**8,483,699**
per record		11.6	0.5	2.9

YPM IP 037326

YTD, 2016

Paleo data fields (GeologicalContext)

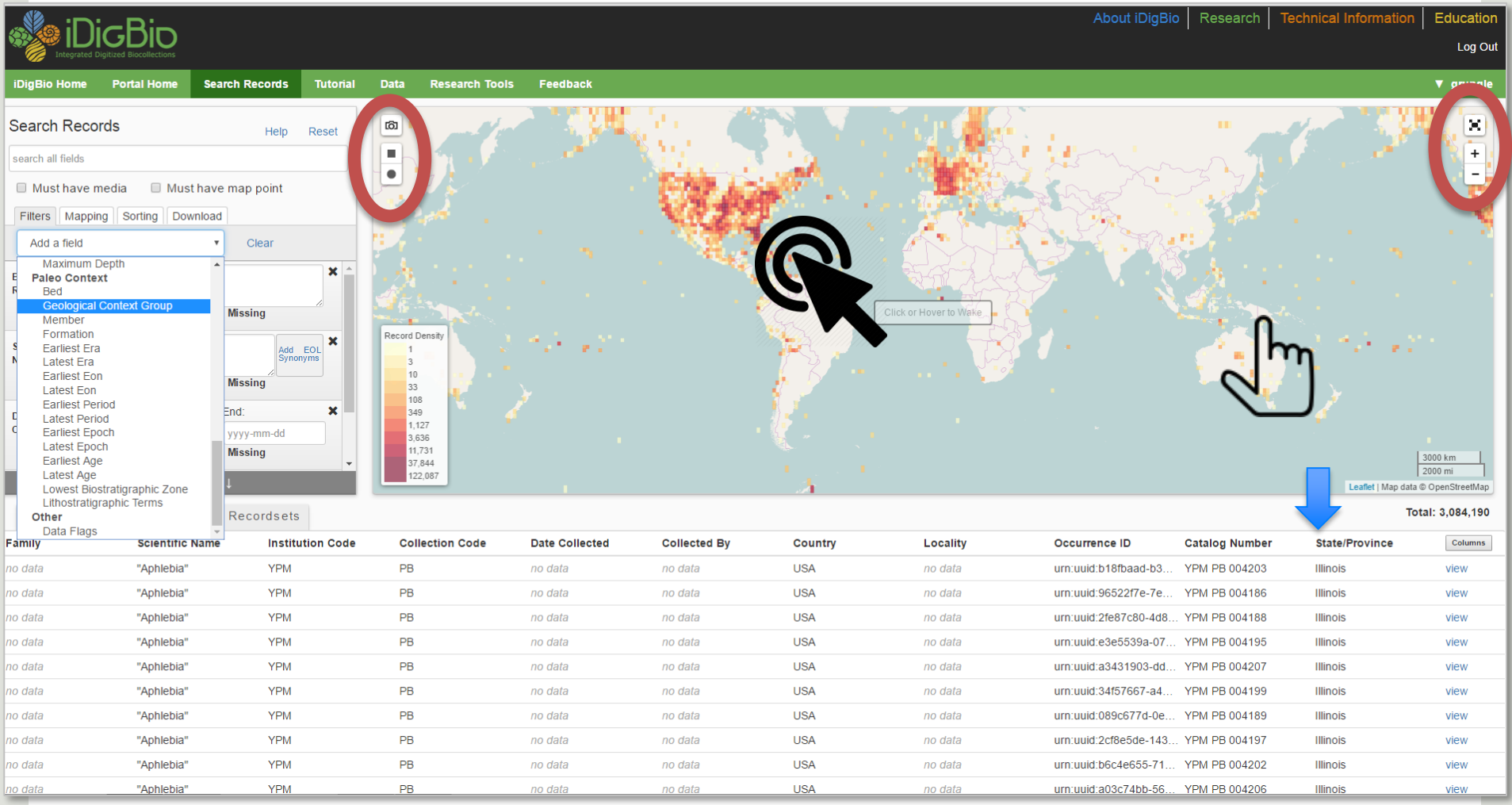
Biodiversity Information Standards TDWG

Darwin Core

- basisOfRecord = *fossilSpecimen*
- recordID
- **occurrenceID** (unique!)
- scientificName
- eventDate
- recordedBy
- Locality information
- catalogNumber
- **institutionID**
- **collectionID**
- [bibliographicCitation]

- bed
- group (Geological context group)
- member
- formation
- earliestEonOrLowestEonothem (Earliest Eon)
- latestEonOrHighestEonothem (Latest Eon)
- earliestEraOrLowestErathem (Earliest Era)
- latestEraOrHighestErathem (Latest Era)
- earliestPeriodOrLowestSystem (Earliest Period)
- latestPeriodOrHighestSystem (Latest Period)
- earliestEpochOrLowestSeries (Earliest Epoch)
- latestEpochOrHighestSeries (Latest Epoch)
- earliestAgeOrLowestStage (Earliest Age)
- latestAgeOrHighestStage (Latest Age)
- lowestBiostratigraphicZone
- lithostratigraphicTerms

The data within iDigBio: *portal.idigbio.org/portal/search*



Family	Scientific Name	Institution Code	Collection Code	Date Collected	Collected By	Country	Locality	Occurrence ID	Catalog Number	State/Province	
no data	"Aphlebia"	YPM	PB	no data	no data	USA	no data	urn:uuid:b18fbaad-b3...	YPM PB 004203	Illinois	view
no data	"Aphlebia"	YPM	PB	no data	no data	USA	no data	urn:uuid:96522f7e-7e...	YPM PB 004186	Illinois	view
no data	"Aphlebia"	YPM	PB	no data	no data	USA	no data	urn:uuid:2fe87c80-4d8...	YPM PB 004188	Illinois	view
no data	"Aphlebia"	YPM	PB	no data	no data	USA	no data	urn:uuid:e3e5539a-07...	YPM PB 004195	Illinois	view
no data	"Aphlebia"	YPM	PB	no data	no data	USA	no data	urn:uuid:a3431903-dd...	YPM PB 004207	Illinois	view
no data	"Aphlebia"	YPM	PB	no data	no data	USA	no data	urn:uuid:34f57667-a4...	YPM PB 004199	Illinois	view
no data	"Aphlebia"	YPM	PB	no data	no data	USA	no data	urn:uuid:089c677d-0e...	YPM PB 004189	Illinois	view
no data	"Aphlebia"	YPM	PB	no data	no data	USA	no data	urn:uuid:2cf8e5de-143...	YPM PB 004197	Illinois	view
no data	"Aphlebia"	YPM	PB	no data	no data	USA	no data	urn:uuid:b6c4e655-71...	YPM PB 004202	Illinois	view
no data	"Aphlebia"	YPM	PB	no data	no data	USA	no data	urn:uuid:a03c74bb-56...	YPM PB 004206	Illinois	view

Try the iDigBio API!

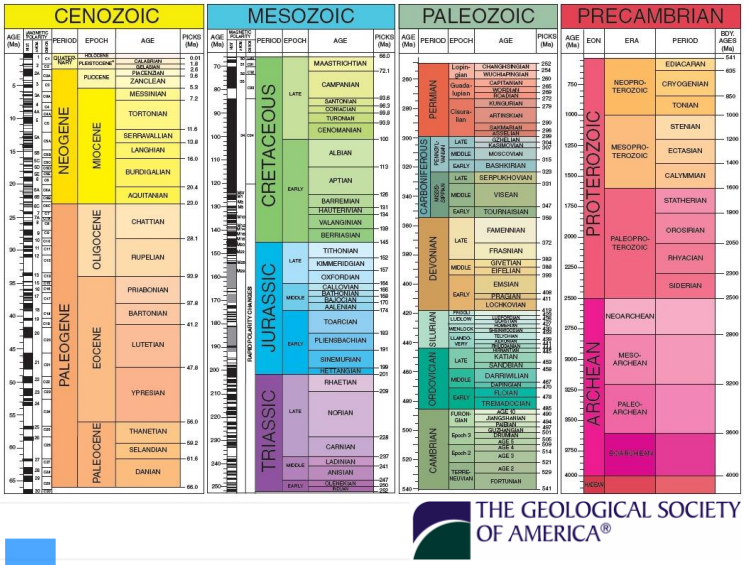
Paleo data shared by iDigBio

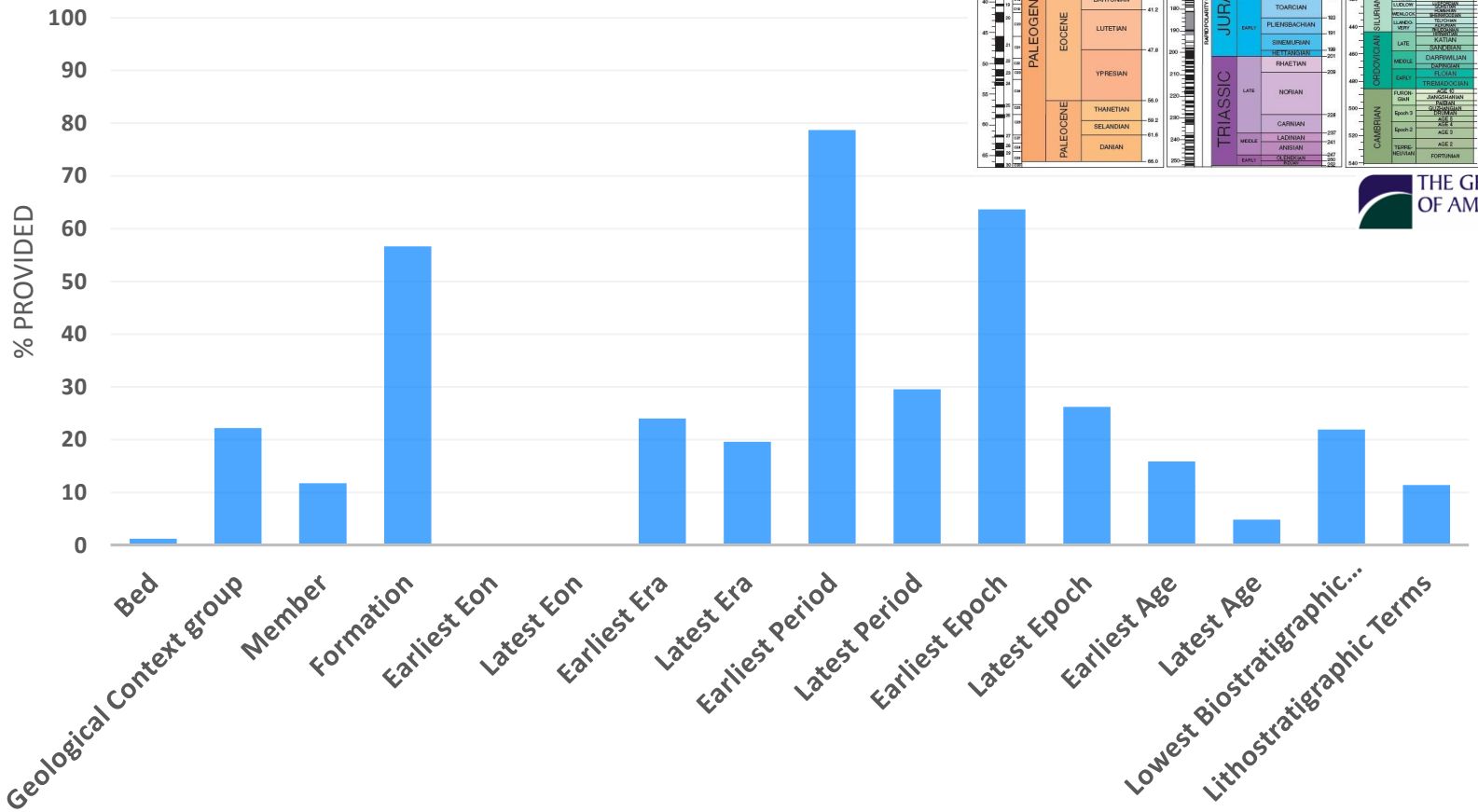
- 3,084,190 fossilSpecimen records (4%)
- 167,447 records have media
- 110,615 records have associated media
- 1,263,438 records have a geopoint

YPM PB 021063

- 50% of records from 3 institutions (University of Florida, Yale Peabody Museum, UC-Museum of Paleontology)
- 24 institutions from 4 countries (86% US)
- 35 collections

Paleo fields in iDigBio





GEOLOGICAL CONTEXT DWC FIELD

Data quality issues

- Random information included in DwC fields
- Non-standardized or consistent vocabulary; numbers instead of names
- Not using all available DwC terms
- Taxonomy!

miocene
miocène
miocene or pleistocene
miocene or pliocene
miocene, early
miocene, late
miocene, late or pliocene, early
miocene, late or pliocene, early or pleistocene, early
miocene, middle
miocene,early
miocene,late
miocene,middle
miocene; middle
miocene; upper
miocene?
miocene-pleistocene (mixed)

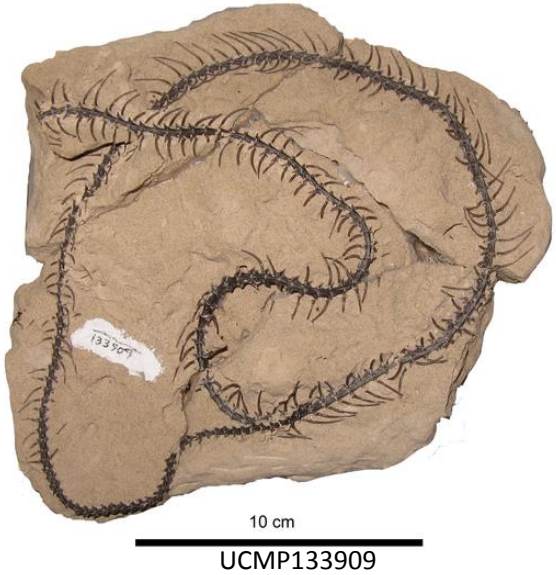
Paleo specific Data flags

https://github.com/iDigBio/idigbio-search-api/wiki/Data-Quality-Flags

- About 40 data flags, helping to correct higher taxonomy, geopoints

What other data flags would the community like?

- DwC fields to be indexed
  - geologicalContextID
  - highestBiostratigraphicZone



Research using specimen data

iDigBio Search API

`ridigbio: Interface to the iDigBio Data API`

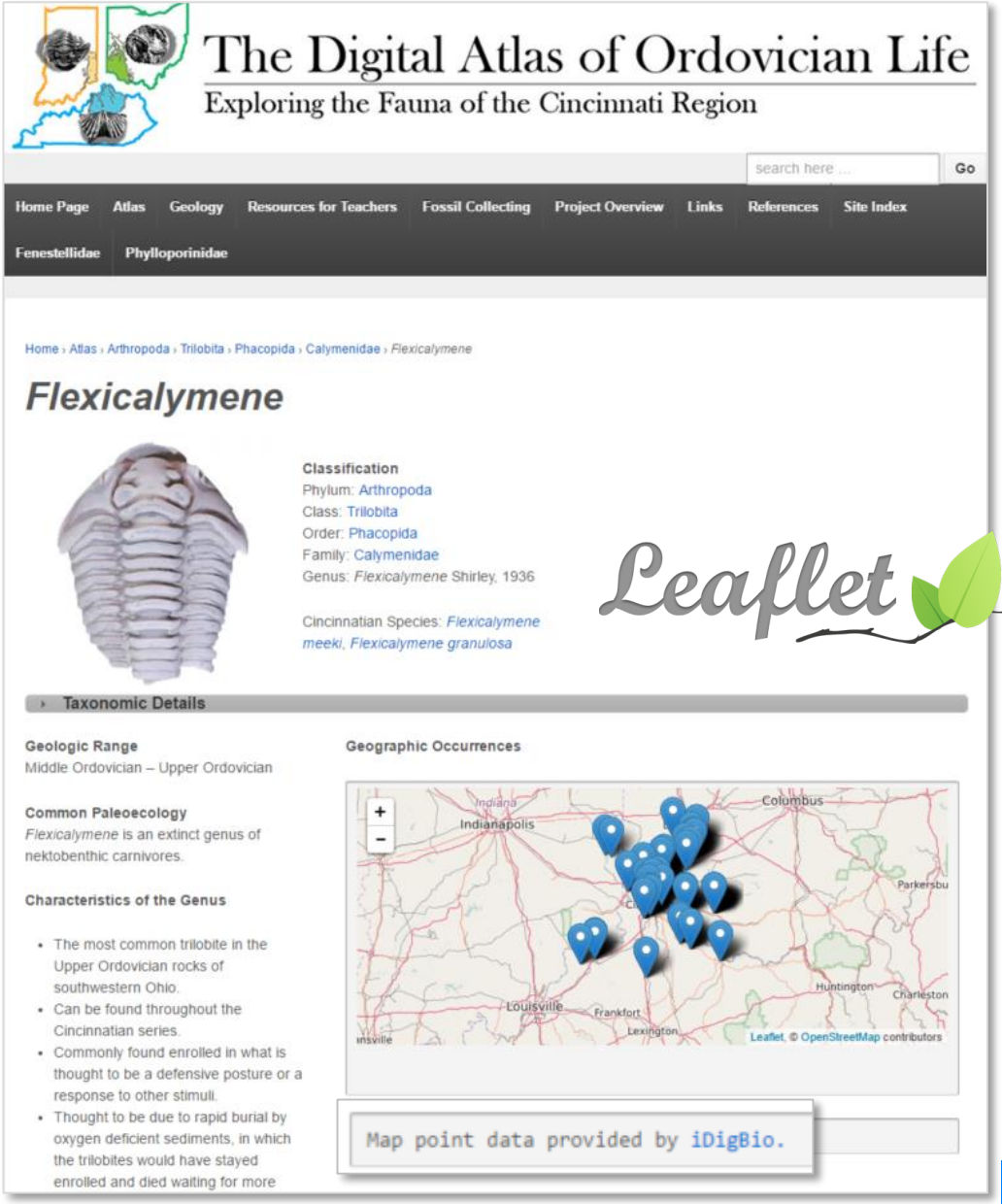
Enhancing Paleontological and Neontological Data API

ePANDDA

Spark

Effechecka
Gouda
Fresh data

Open Source cluster computer framework



www.iDigBio.org

I Dig Bio do you?

iDigBio
Integrated Digitized Biocollections

YPM VP 056914

Thanks!

- To all the data providers and publishers
- iDigBio team

Shelley James, sjames@flmnh.ufl.edu
Gil Nelson, gnelson@bio.fsu.edu

UF UNIVERSITY of FLORIDA
FLORIDA STATE UNIVERSITY 1851
FLORIDA MUSEUM OF NATURAL HISTORY
NSF