The Importance and Challenges of Database Integration: MorphoBank, MorphoSource, and the Paleobiology Database

Julia M. Winchester,

Doug M. Boyer,

Maureen O'Leary,

Jocelyn A. Sessa

MORPHO SOURCE

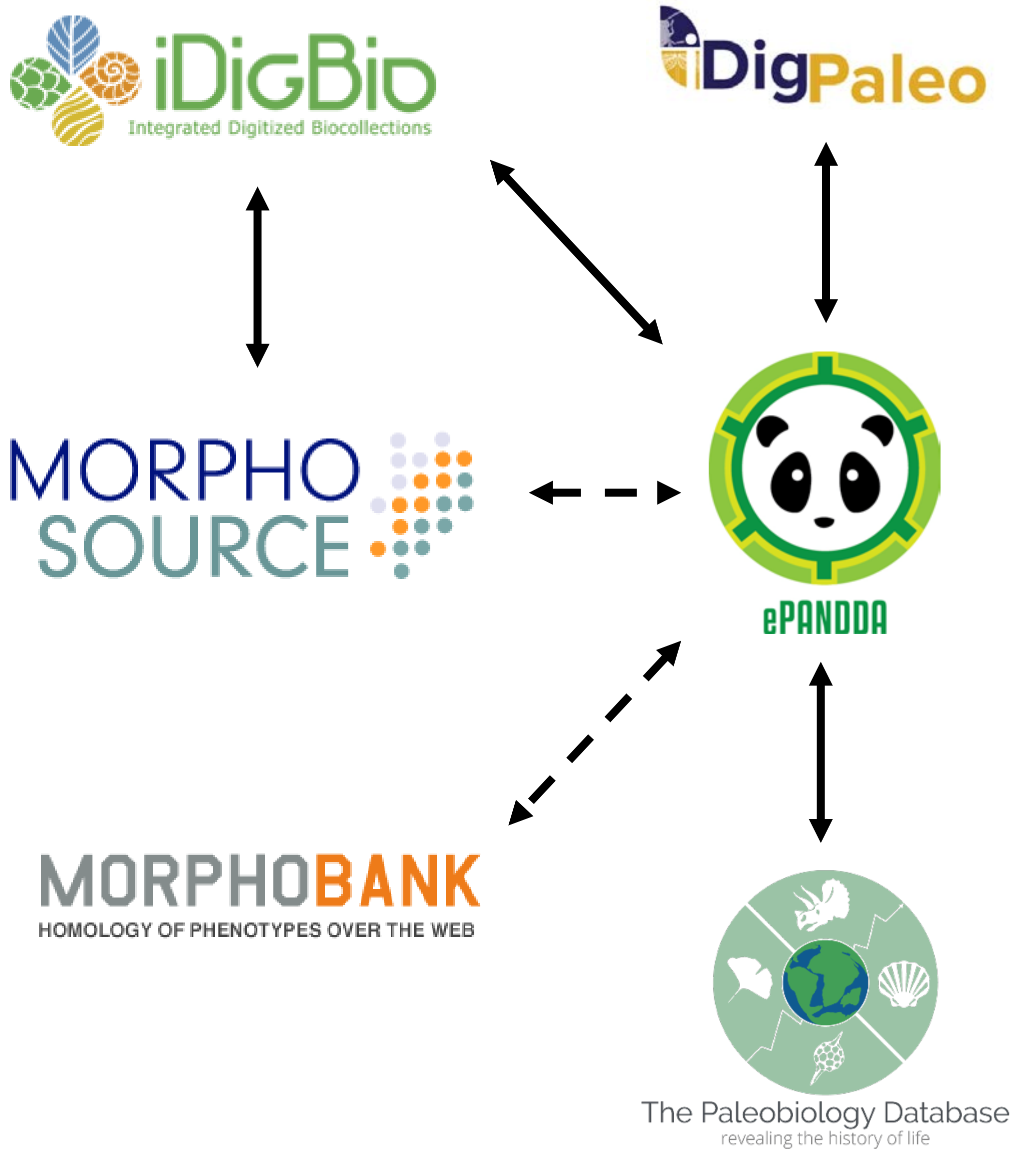
iDigBio Integrated Digitized Biocollections

MORPHOBANK HOMOLOGY OF PHENOTYPES OVER THE WEB

The Paleobiology Database revealing the history of life

Outline

- Introduction
  - Scientific data repositories
  - Niche specialization and data sustainability
  - Database integration
- Examples
  - iDigBio and MorphoSource
  - MorphoBank, PBDB, and MorphoSource
- Conclusions



Scientific data grows exponentially

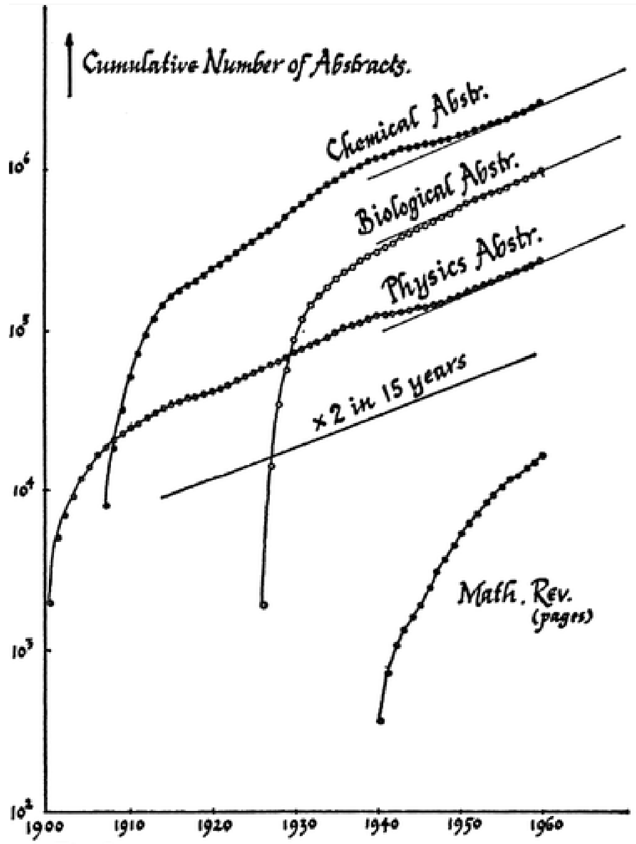
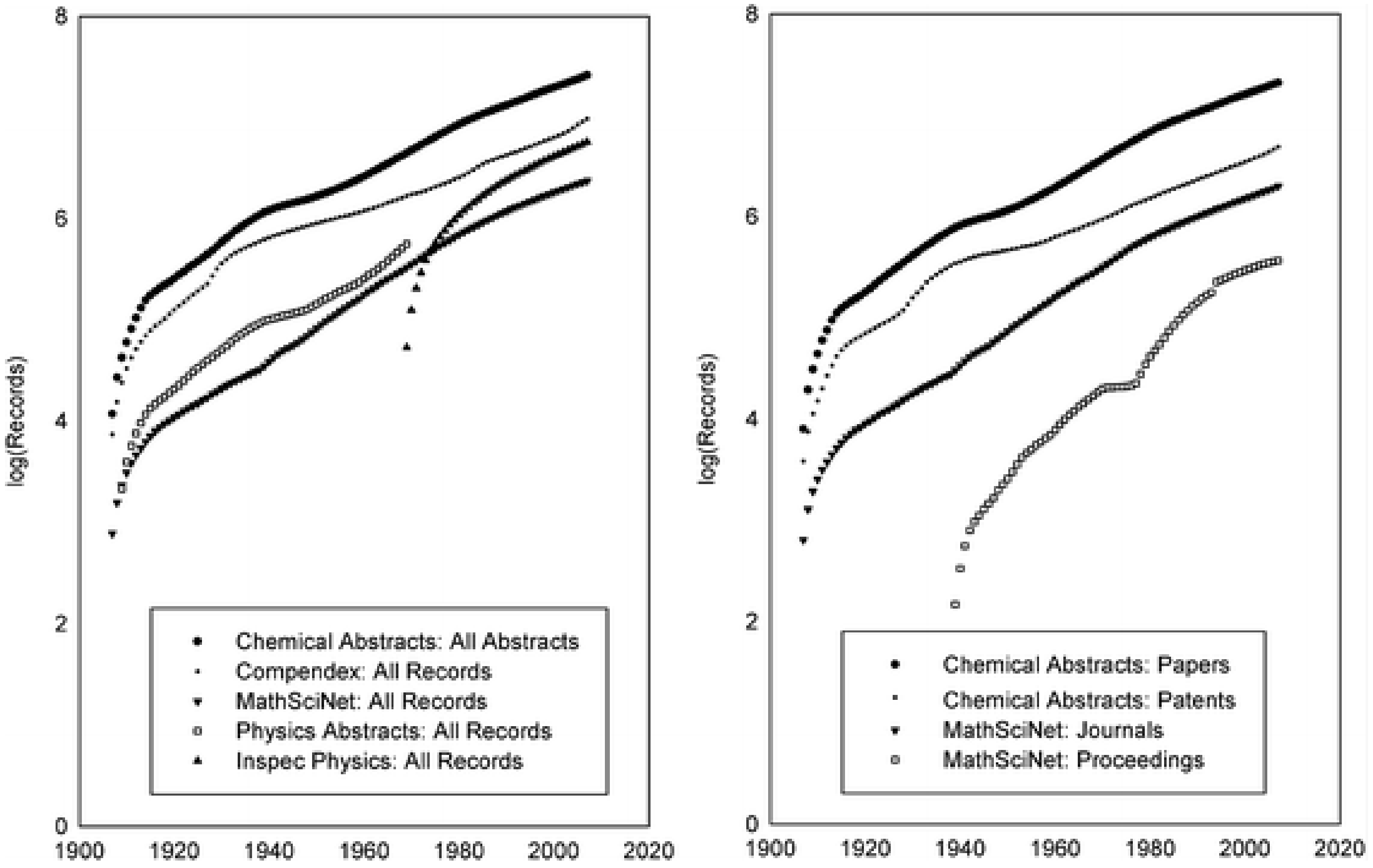


Fig. 2. CUMULATIVE NUMBER OF ABSTRACTS IN VARIOUS SCIENTIFIC FIELDS, FROM THE BEGINNING OF THE ABSTRACT SERVICE TO GIVEN DATE

Price (1960)



Larsen and von Ins (2010)

Scientific data grows exponentially

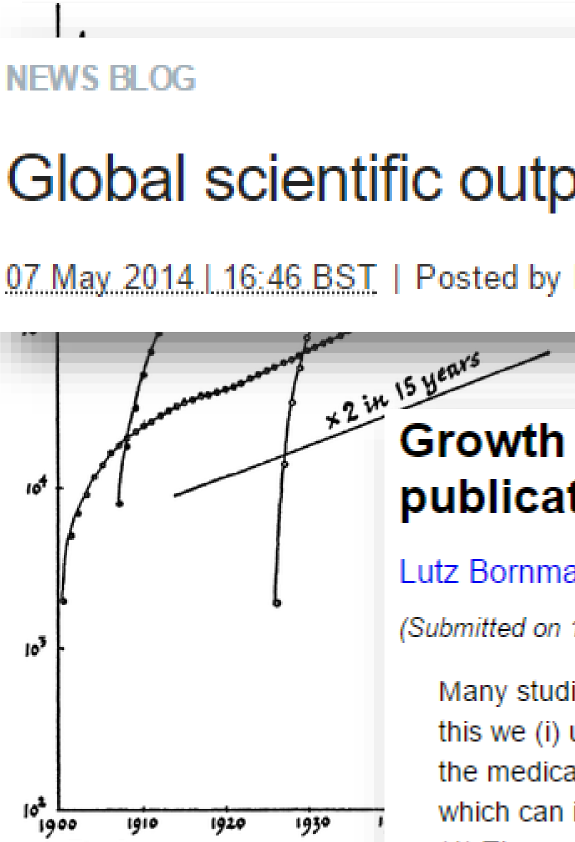
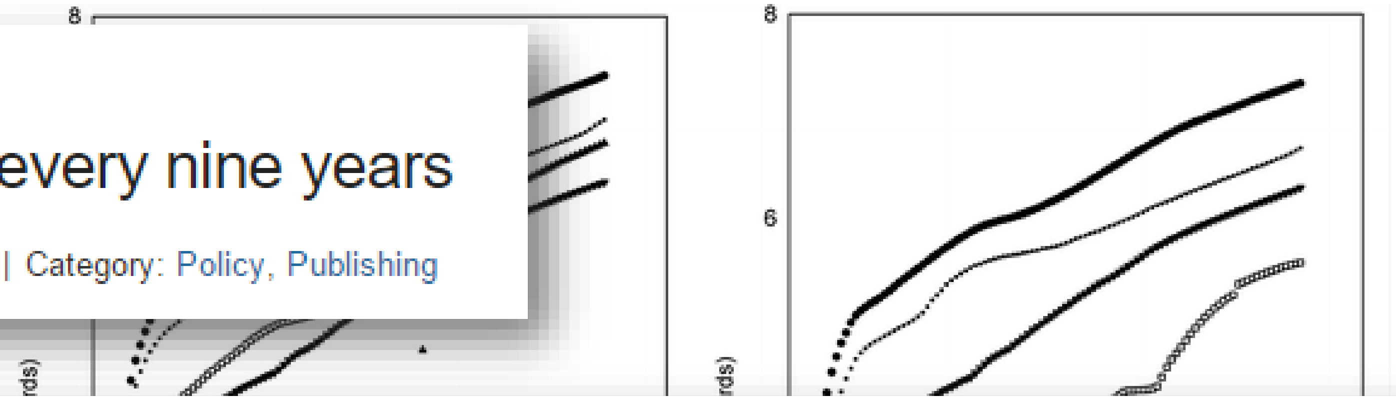


Fig. 2. CUMULATIVE NUMBER OF SCIENTIFIC FIELDS, FROM THE ABSTRACT SERVICE TO

Price (19



NEWS BLOG

Global scientific output doubles every nine years

07 May 2014 | 16:46 BST | Posted by Richard Van Noorden | Category: Policy, Publishing

Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references

Lutz Bornmann, Ruediger Mutz

(Submitted on 19 Feb 2014 (v1), last revised 8 May 2014 (this version, v3))

Many studies in information science have looked at the growth of science. In this study, we re-examine the question of the growth of science. To do this we (i) use current data up to publication year 2012 and (ii) analyse it across all disciplines and also separately for the natural sciences and for the medical and health sciences. Furthermore, the data are analysed with an advanced statistical technique - segmented regression analysis - which can identify specific segments with similar growth rates in the history of science. The study is based on two different sets of bibliometric data: (1) The number of publications held as source items in the Web of Science (WoS, Thomson Reuters) per publication year and (2) the number of cited references in the publications of the source items per cited reference year. We have looked at the rate at which science has grown since the mid-1600s. In our analysis of cited references we identified three growth phases in the development of science, which each led to growth rates tripling in comparison with the previous phase: from less than 1% up to the middle of the 18th century, to 2 to 3% up to the period between the two world wars and 8 to 9% to 2012.

Data repositories

Collections databases

VertNet

iDigBio
Integrated Digitized Biocollections

Lists of specimens and associated information

Research databases

iDigPaleo

Morphbank
http://morphbank.net

PANGAEA.

GBIF Global Biodiversity Information Facility

MORPHO SOURCE

MORPHOBANK
HOMOLOGY OF PHENOTYPES OVER THE WEB

The Paleobiology Database
revealing the history of life

EarthChem

NOAA Paleoclimatology

LinkedEarth

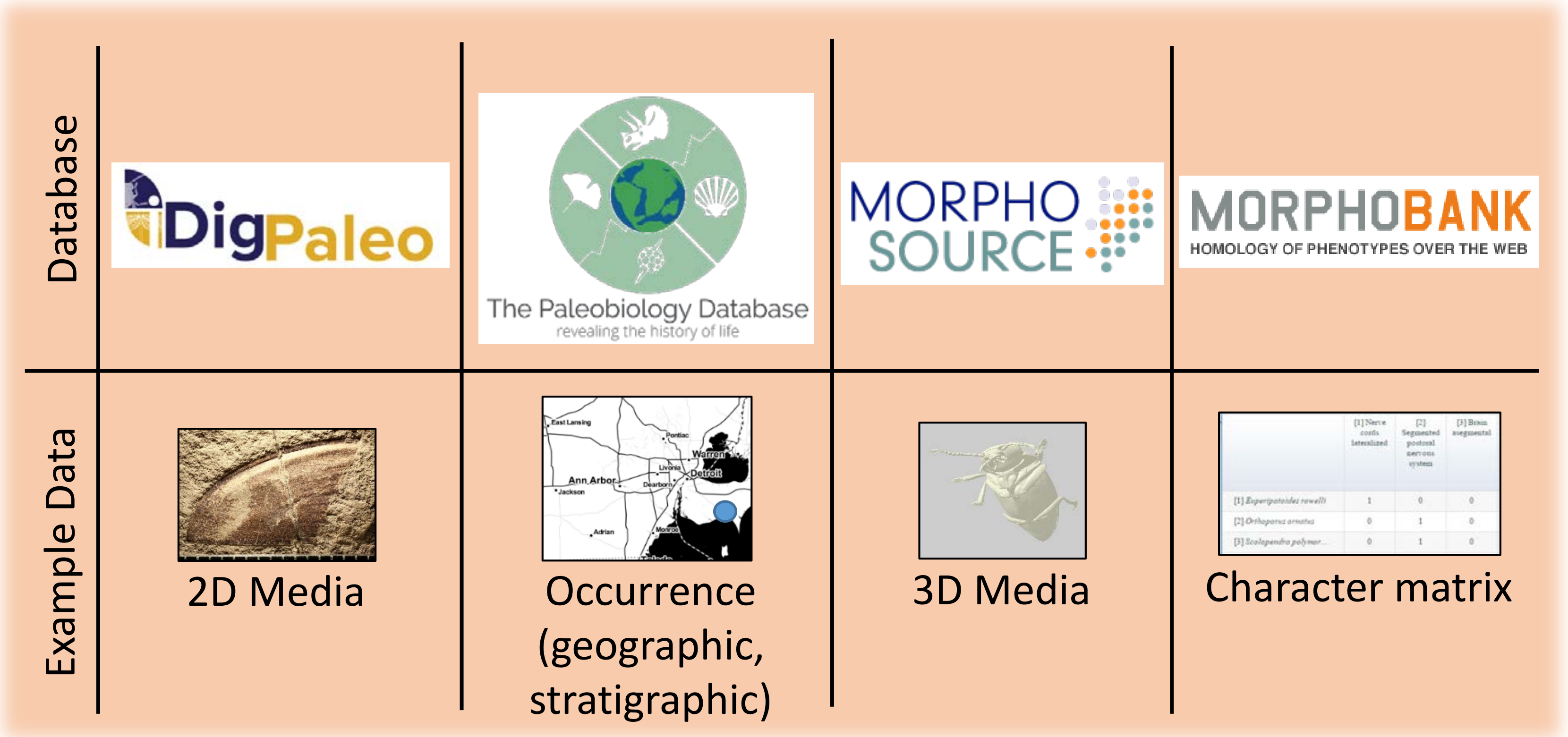
NEOTOMA PALEOECOLOGY DATABASE

Derived data, related to research question

(Uhen et al, 2013)

Research databases

- Distinct but overlapping spheres of data
- Some data more derived than others

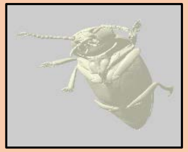


Some (but not all) data from a search for Dytiscidae water beetles

(Uhen et al, 2013)

Research databases

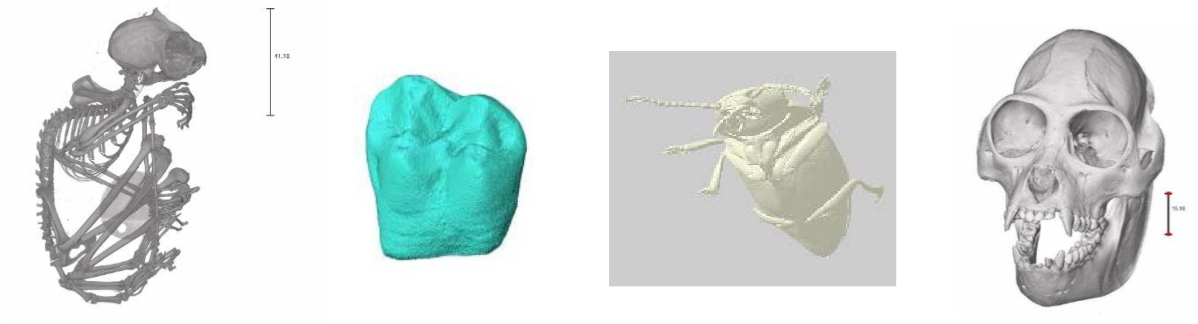
- Some databases specialize more than others
- Benefits of niche specialization
  - Reduced competition
  - Data sustainability

Database	DigPaleo	The Paleobiology Database revealing the history of life	MORPHO SOURCE	MORPHOBANK HOMOLOGY OF PHENOTYPES OVER THE WEB
Example Data	 2D Media	 Occurrence (geographic, stratigraphic)	 3D Media	 Character matrix

Some (but not all) data from a search for Dytiscidae water beetles

(Uhen et al, 2013)

MORPHO SOURCE



The Paleobiology Database
revealing the history of life

iDigBio
Integrated Digitized Biocollections

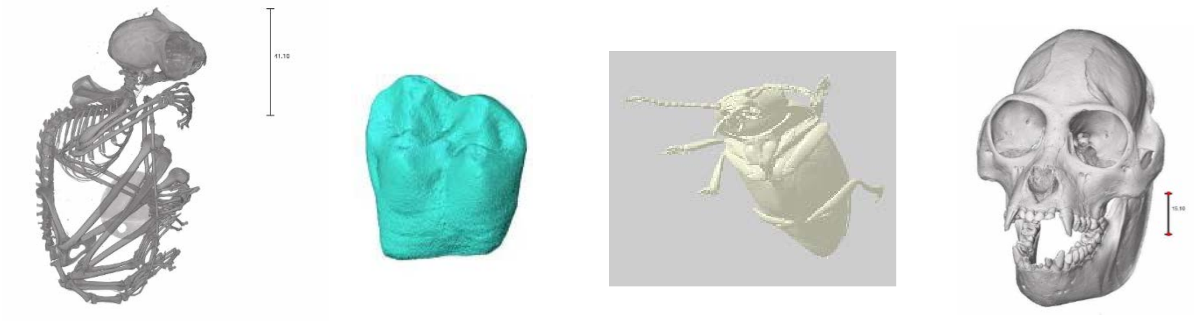
Integrative links allow data to be stored in most suitable repository, and made available to other repositories and the public

The Paleobiology Database
revealing the history of life

iDigBio
Integrated Digitized Biocollections

Concern: What if the specialized repository disappears?

MORPHO SOURCE



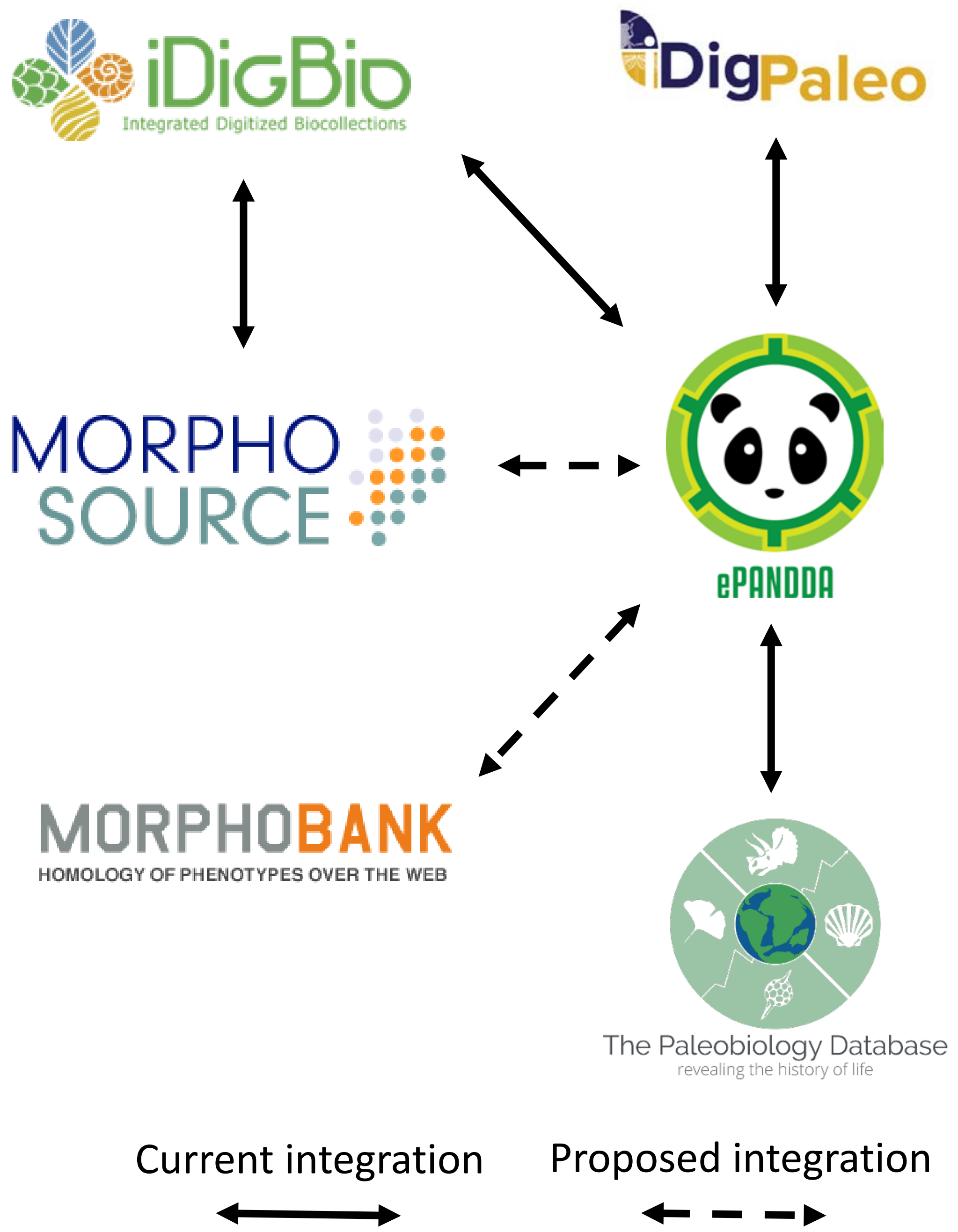
The Paleobiology Database
revealing the history of life

iDigBio
Integrated Digitized Biocollections

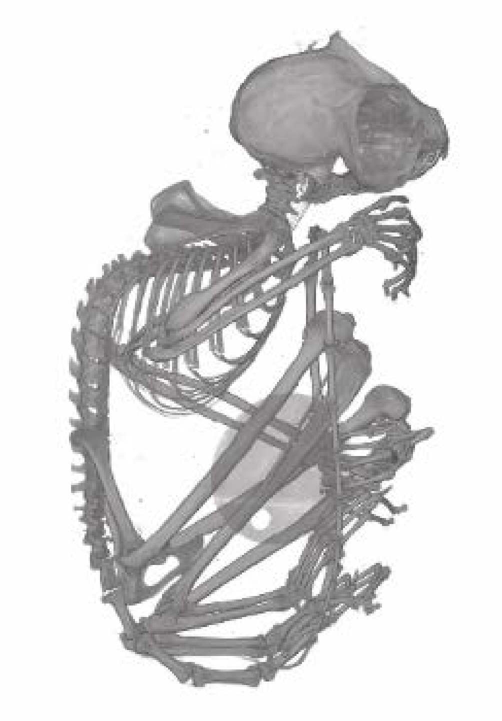
Solution: modular design conforming to community standards, easy to transport if necessary

Database integration

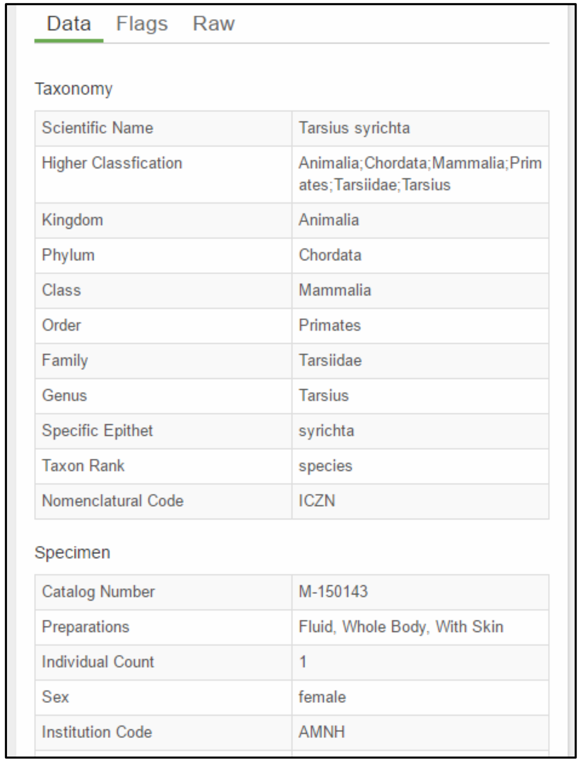
- Benefits
  - More sustainable data model
  - Easier for end users
  - New kinds of automated research
  - Force multiplier
- Challenges:
  - Collaborative development of integration models
  - Record matching



>25,000 3D media files of vouchered specimens



>104 million specimens, many from museum collections


Data Flags Raw

Taxonomy

Scientific Name	Tarsius syrichta
Higher Classfication	Animalia;Chordata;Mammalia;Primates;Tarsiidae;Tarsius
Kingdom	Animalia
Phylum	Chordata
Class	Mammalia
Order	Primates
Family	Tarsiidae
Genus	Tarsius
Specific Epithet	syrichta
Taxon Rank	species
Nomenclatural Code	ICZN

Specimen

Catalog Number	M-150143
Preparations	Fluid, Whole Body, With Skin
Individual Count	1
Sex	female
Institution Code	AMNH

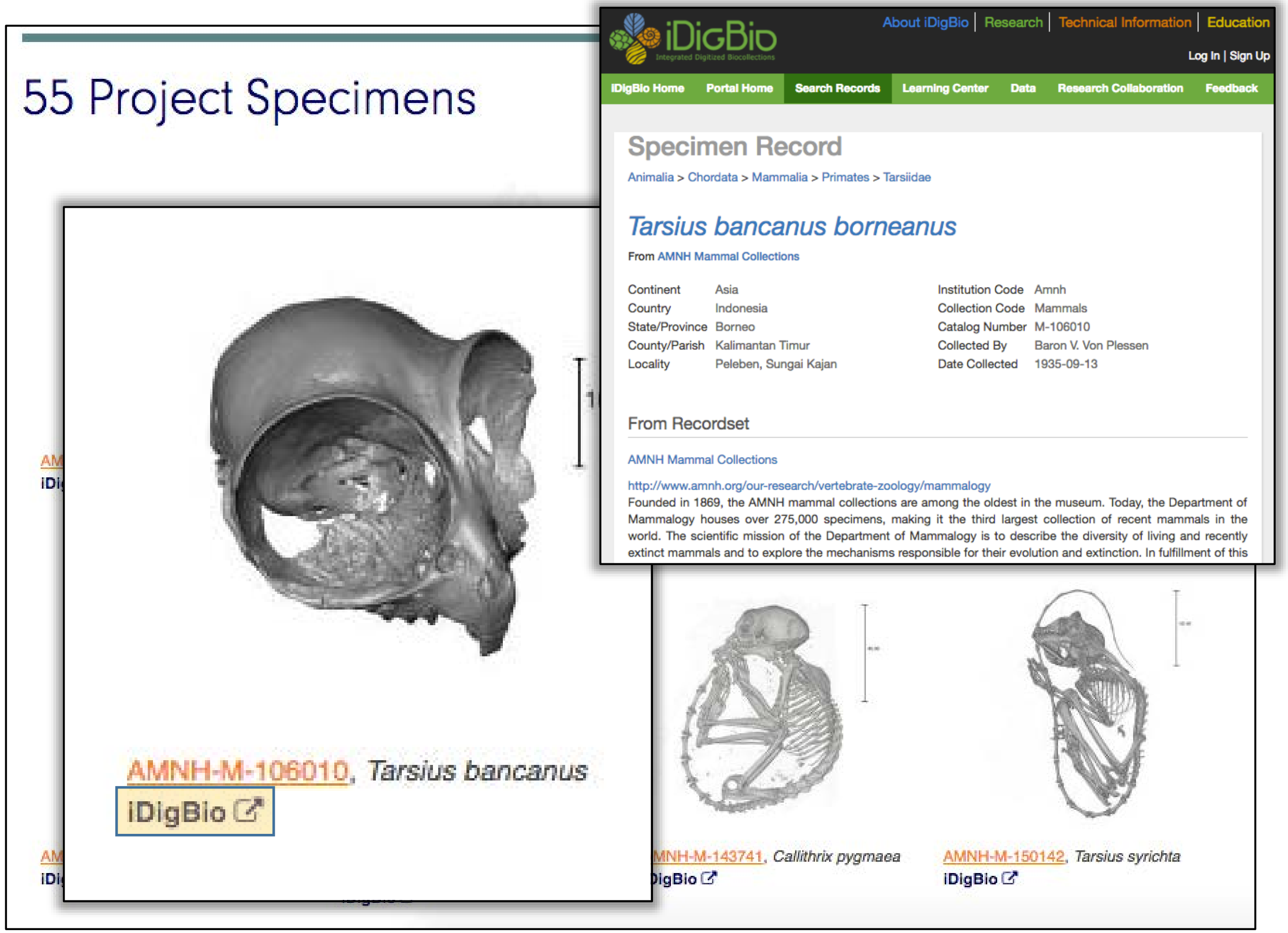
Integration model: associate specimen records in MorphoSource and iDigBio

Integrating pre-existing data

- Identifying iDigBio records matched to currently existing MorphoSource records

- Fuzzy specimen number matching

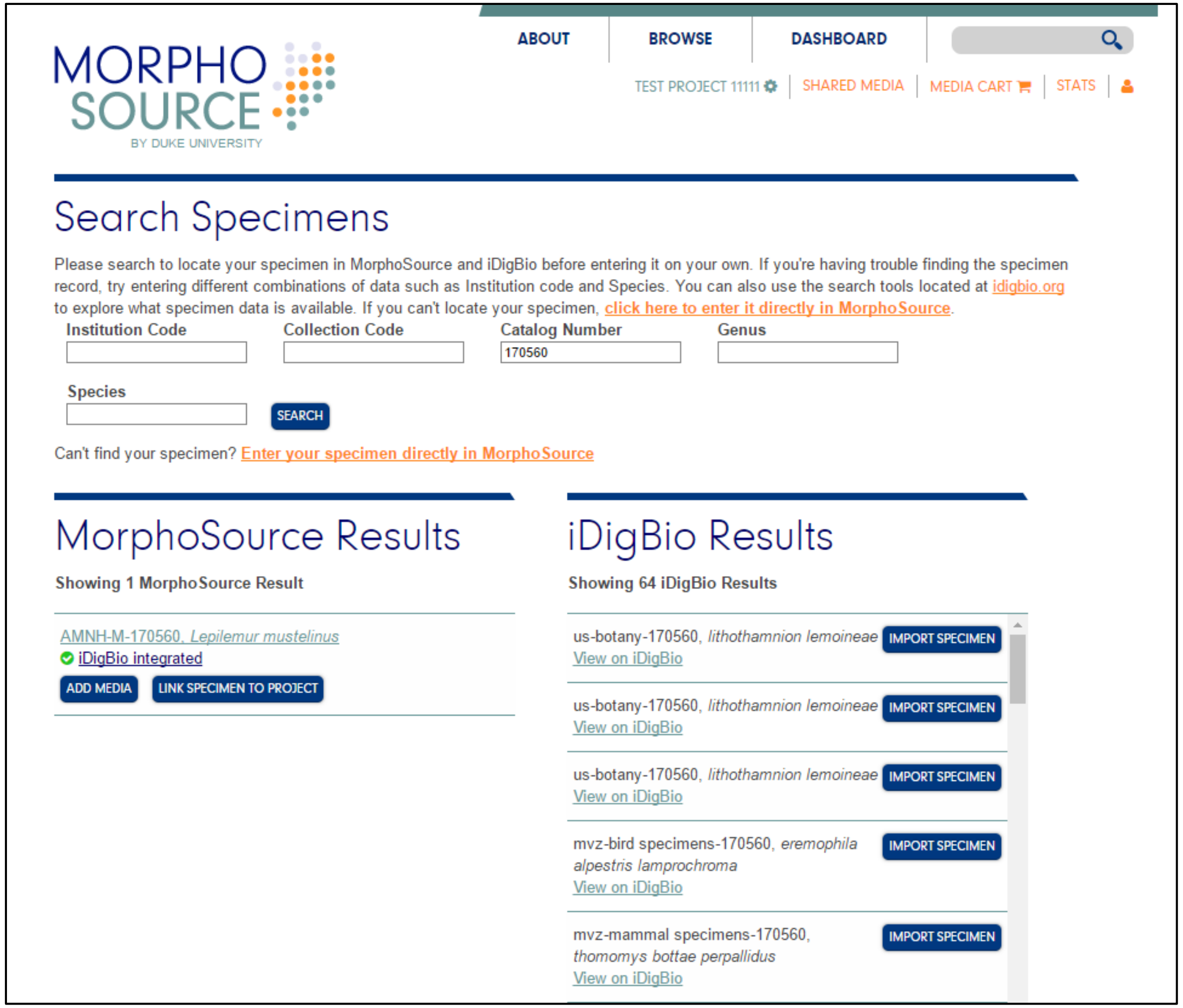
- For matches, occurrence ID gathered from iDigBio to create association



Integrating new data

- Users adding new specimen records search for pre-existing MorphoSource records

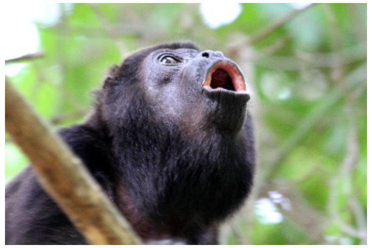
- Specimen import tool now searches iDigBio automatically
  - Associate occurrence ID



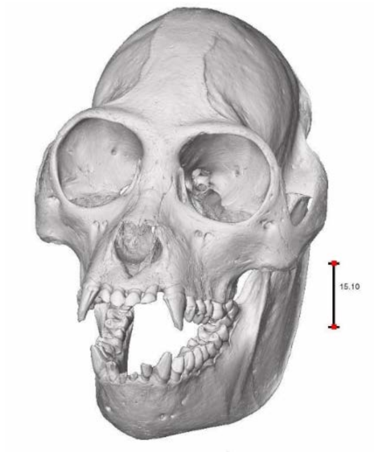
Good progress so far, but…

- Will never have perfect 100% matching between MorphoSource and iDigBio
- Reason: MorphoSource does not require pre-existing occurrence IDs (GUIDs) for specimen records when uploading data
  - Too many possible records without museum-provided occurrence IDs

TDWG GUID Applicability Statement, Recommendation 5: Providers should only assign GUIDs to objects for which they are the authority.



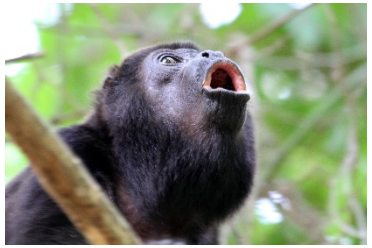
Specimen record MorphoSource can't authoritatively assign GUID



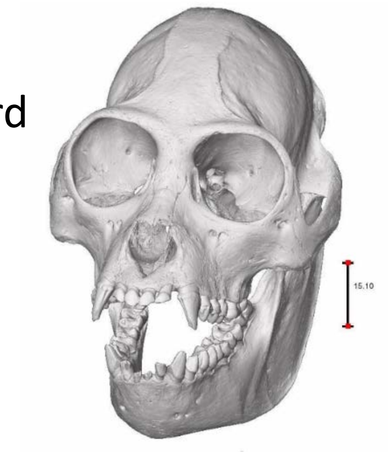
3D Media MorphoSource can authoritatively assign GUID

(Richards, 2010)

Ingestion of MorphoSource media records into iDigBio

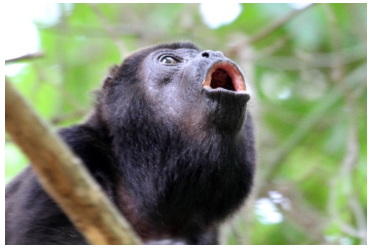


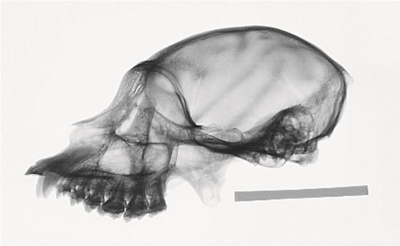
Alouatta palliata specimen record
No occurrence ID
MorphoSource-assigned GUID



3D Media
MorphoSource-assigned GUID

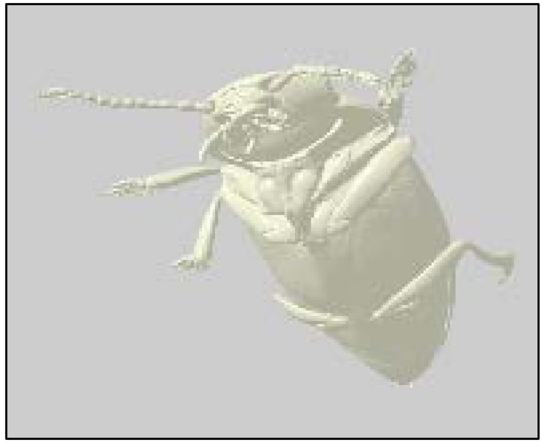
iDigBio
Integrated Digitized Biocollections





Already existing media, but no museum-provided occurrence ID for specimen

MORPHO SOURCE



3D Media

The Paleobiology Database

revealing the history of life

Occurrences, stratigraphy, references

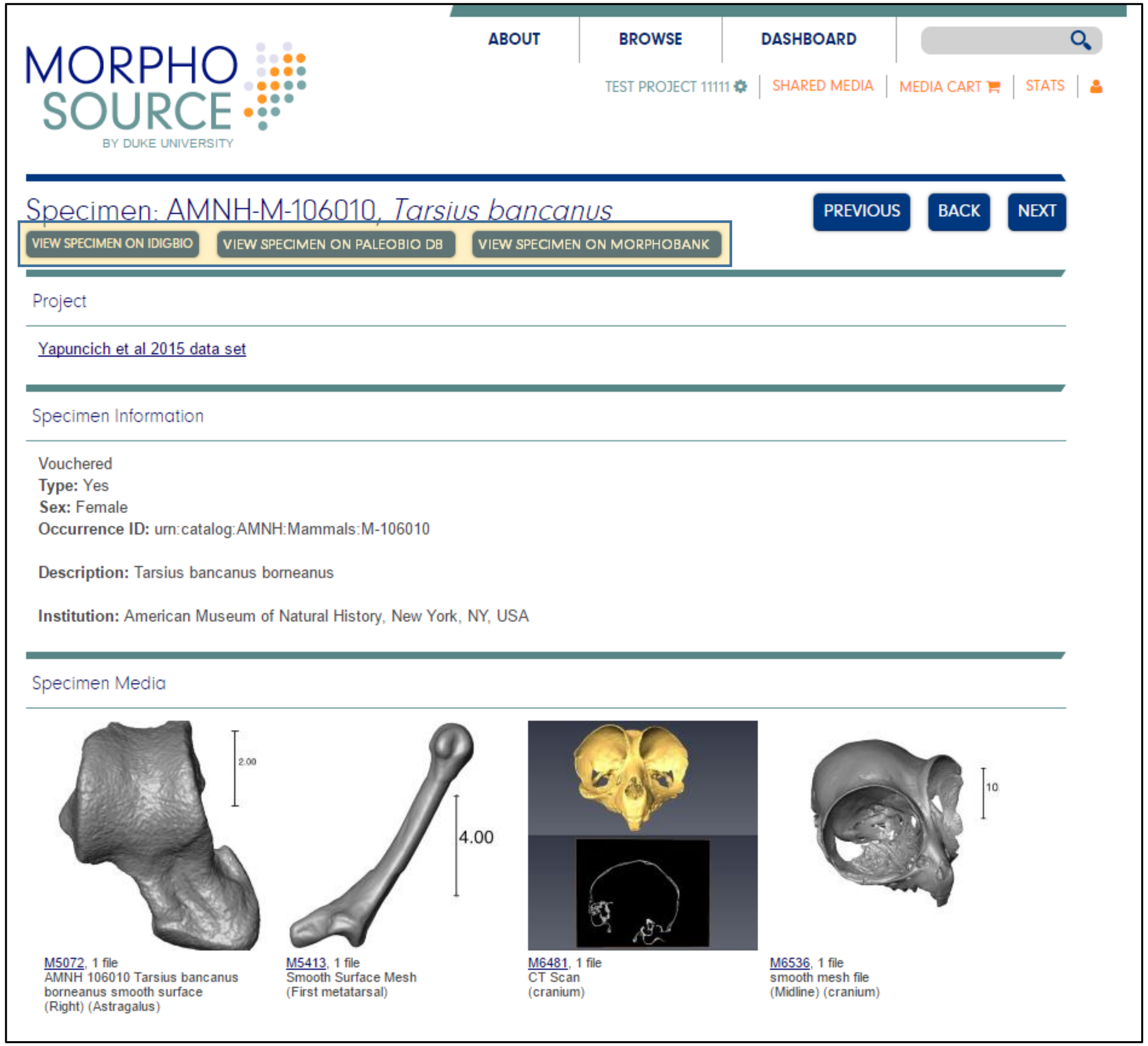
MORPHOBANK

HOMOLOGY OF PHENOTYPES OVER THE WEB

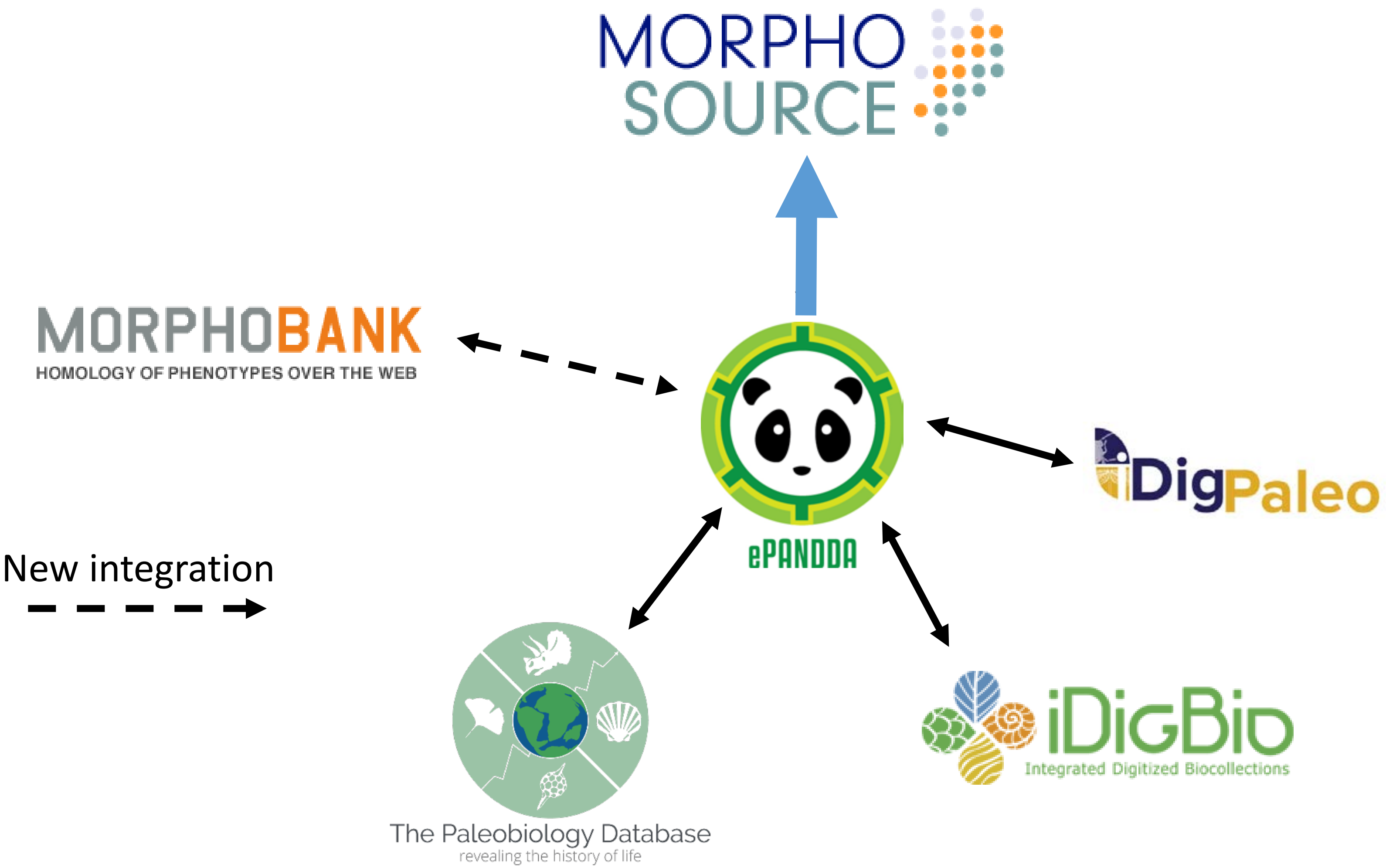
	[1] Nerve cords lateralized	[2] Segmented postoral nervous system	[3] Brain asegmental
[1] *Euperipatoides rowelli*	1	0	0
[2] *Orthoporus ornatus*	0	1	0
[3] *Scolopendra polymor…*	0	1	0

- Phenomic and phenomic/genomic character matrices
- Annotated phenomic data
  - 2D/3D Media files
- Collaborative matrix building tools
- >1,500 projects

- Integration model: Enhance individual resources through combined data access, deposition, and workflow tools



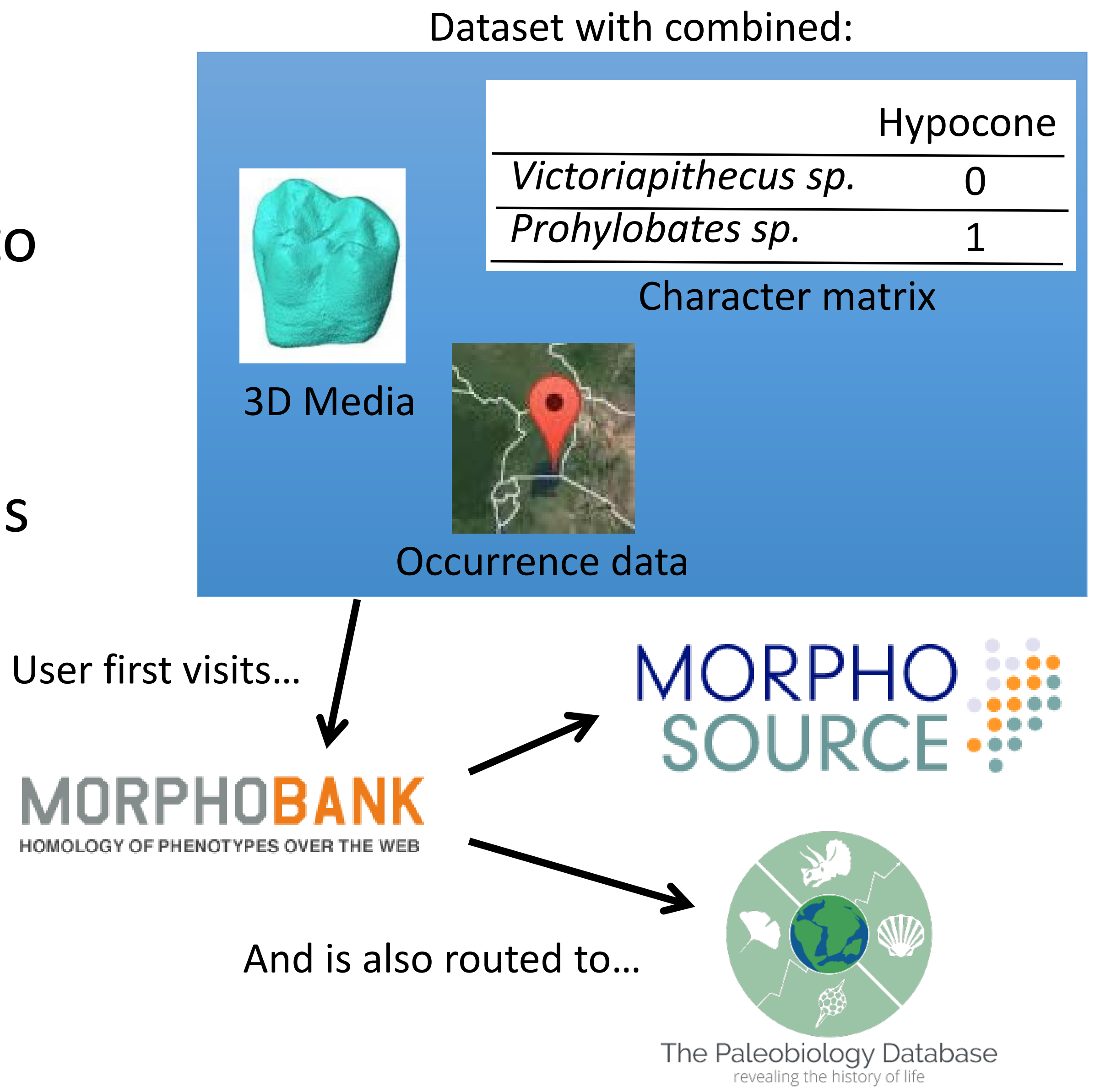
Links to data from other sites within each site



Implemented via ePANDDA integration

Data deposition

- Modular tools for depositing data to multiple sites
- Links/widgets within individual sites
- Benefits
  - Minimize duplicate data entry
  - Maximize metadata consistency



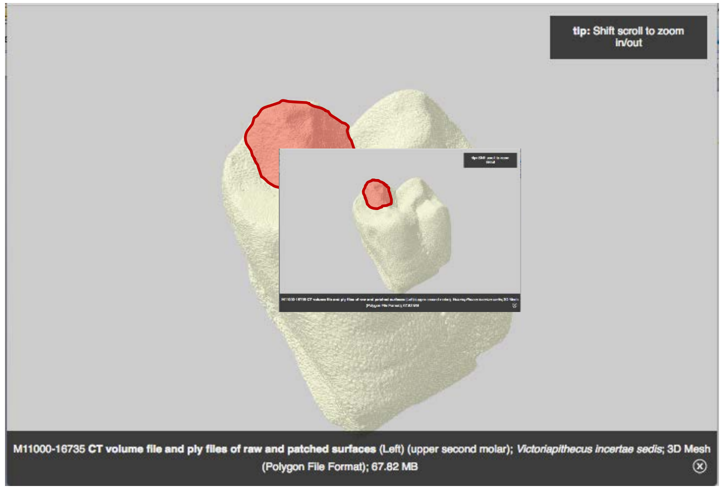
Data analysis tools

MorphoBank Collaborative Web Matrix Builder

no highlight Show names + images Characters Ontology Search Preferences

	[171] M 11: Upper M1 postmetaconule crista	[172] M 12: Upper M1 preprotocrista height relative t...	[173] M 13: Upper M1 waisted distal edge	[174] M 14: Upper M2 hypocone
[1] *Echinosorex gymnura*	absent	higher	absent	large
[2] *Erinaceus europaeus*	absent	higher	absent	large
[3] †*Aegyptopithecus zeu*…	absent	similar	absent	large

1. Search MorphoSource media, load into 3D media viewer with annotation tools (modular web applet)



2. Search PBDB for specimen, associate stratigraphy as character

Summary

- Two-forked approach to database integration
  - Front-end tools for ease of use
  - Back-end architecture (APIs, etc.) for future work
- Benefits
  - Improves data sustainability
  - Builds on previous infrastructure
  - Enables new automated data gathering methods
- Challenges
  - Collaborative development of integration model

ePANDDA

The Paleobiology Database
revealing the history of life

MORPHOBANK
HOMOLOGY OF PHENOTYPES OVER THE WEB

MORPHO SOURCE

Conclusion

- Database integration is beneficial and necessary for managing continually increasing amounts of scientific output

- Enables database niche specialization, ensuring data preservation and increasing data quality in terms of meeting best standards

- Requires robust community-approved standards and careful thought concerning integration models

Acknowledgements

Duke University

Gabe Yapuncich

University of Washington
Friday Harbor Laboratories

Adam Summers

NSF

CAREER Grant BCS 155284
EarthCube IA 540902

iDigBio
Integrated Digitized Biocollections

Gil Nelson
Alex Thompson
Kevin Love
Dan Stoner

UF
University of Florida

David Blackburn
Ed Stanley

Whirl-i-gig

Seth Kaufman
Maria Passarotti