



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

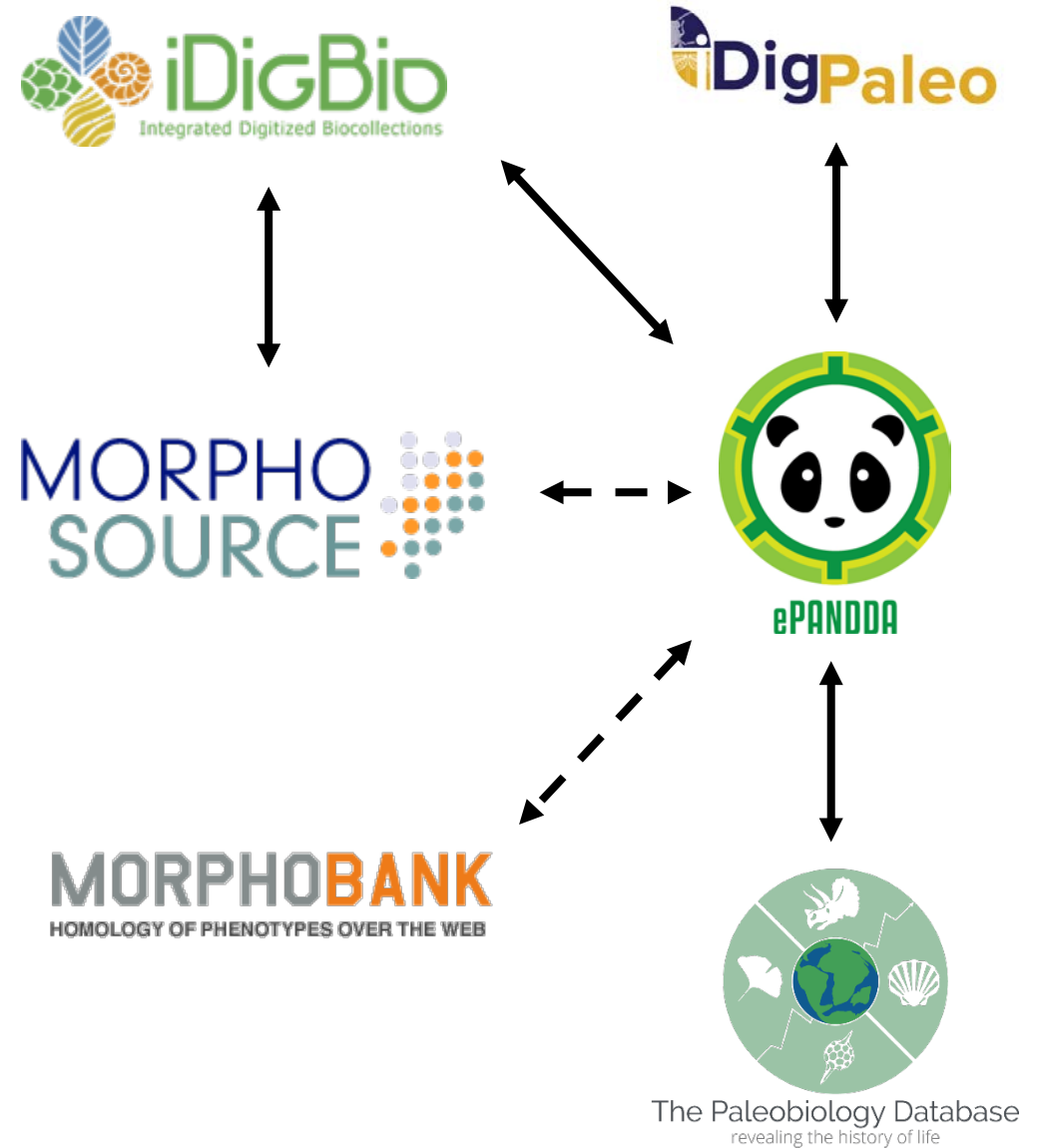
Julia M. Winchester,
Doug M. Boyer,
Maureen O'Leary,
Jocelyn A. Sessa



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

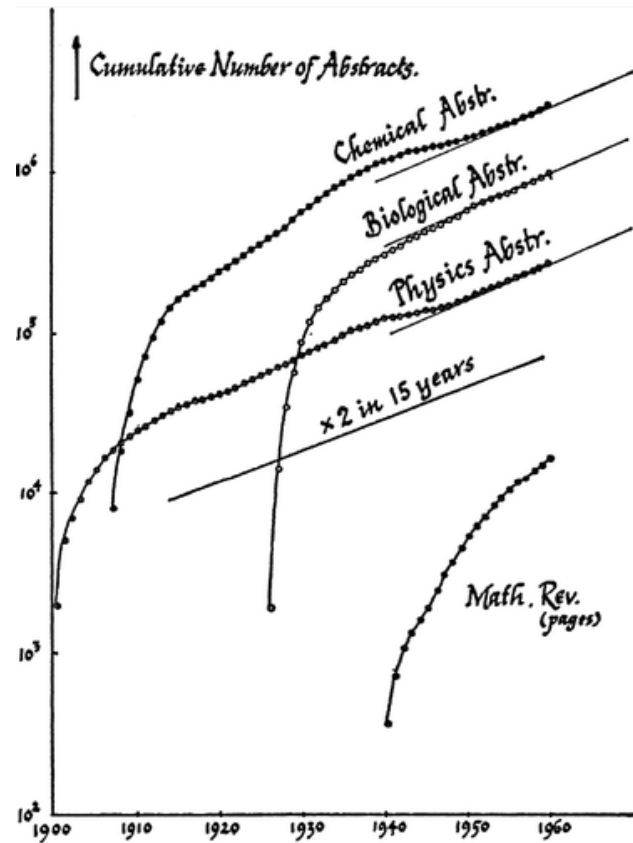
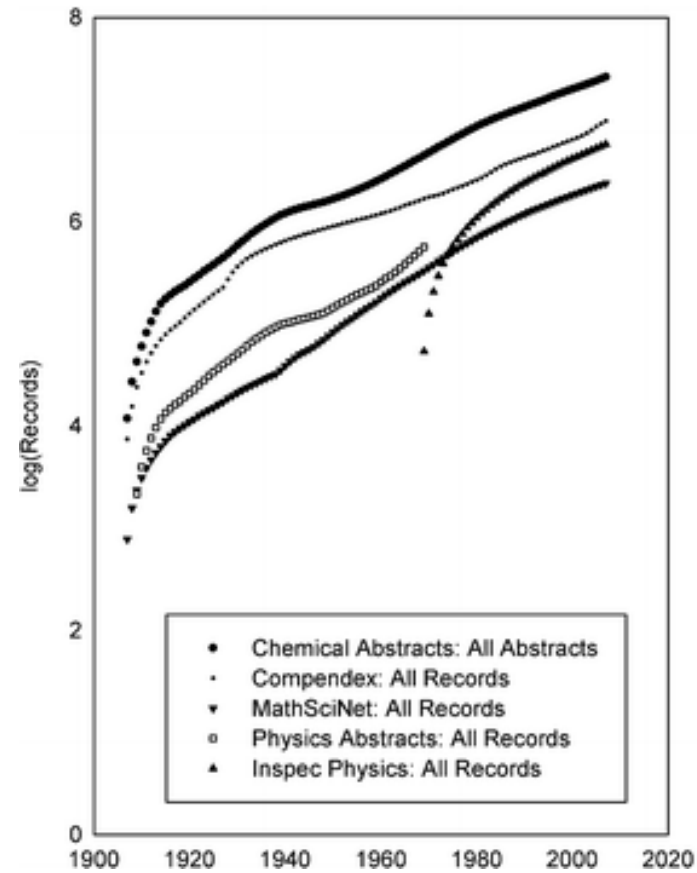
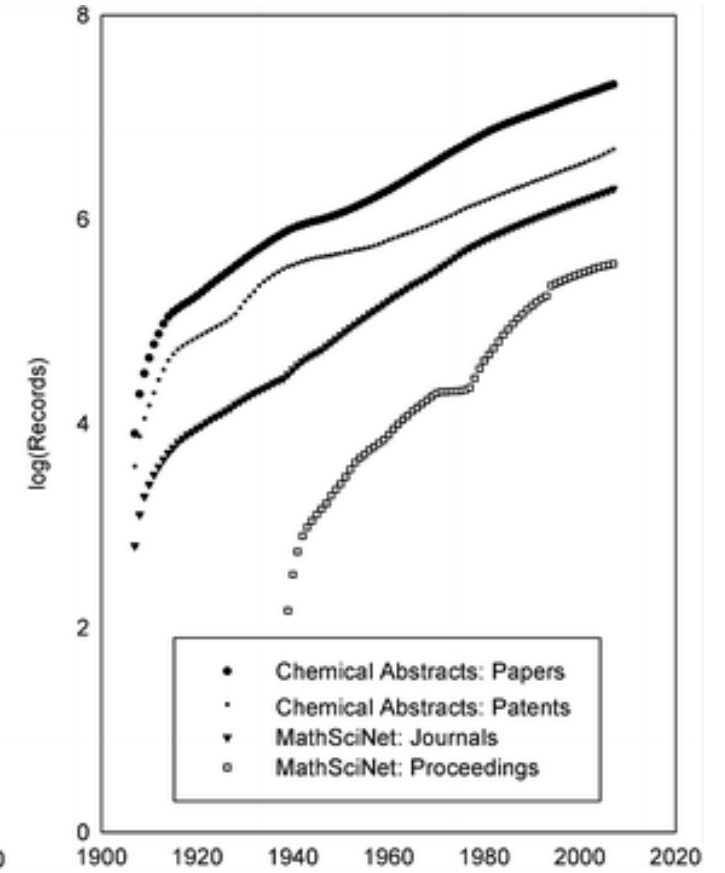


Fig. 2. CUMULATIVE NUMBER OF ABSTRACTS IN VARIOUS SCIENTIFIC FIELDS, FROM THE BEGINNING OF THE ABSTRACT SERVICE TO GIVEN DATE

Price (1960)



Larsen and von Ins (2010)



Scientific data grows exponentially

NEWS BLOG

Global scientific output doubles every nine years

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

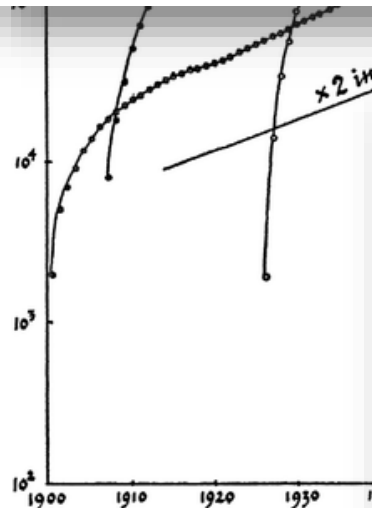


Fig. 2. CUMULATIVE NUMBER OF SCIENTIFIC PUBLICATIONS, FROM THE ABSTRACT SERVICE TO

Price (19

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

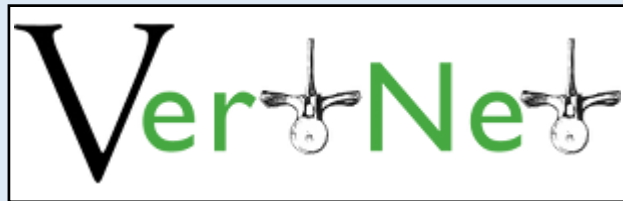
Lutz Bornmann, Ruediger Mutz

(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.

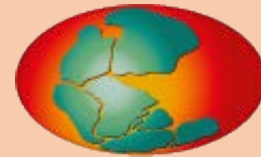
Data repositories

Collections databases



Lists of specimens and associated information

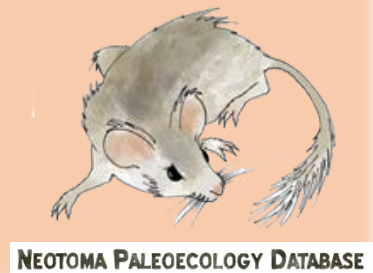
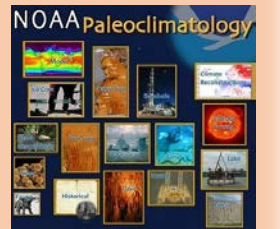
Research databases



PANGAEA.








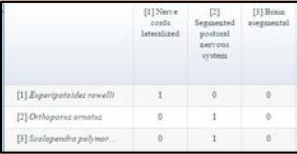


Derived data, related to research question



Research databases




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

Database				
Example Data	 2D Media	 Occurrence (geographic, stratigraphic)	 3D Media	 Character matrix

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

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	<table border="1"><thead><tr><th></th><th>[1] Nerve cords lateralized</th><th>[2] Segmented postoral nervous system</th><th>[3] Brain segmented</th></tr></thead><tbody><tr><td>[1] <i>Eisneripetolides rowelli</i></td><td>1</td><td>0</td><td>0</td></tr><tr><td>[2] <i>Orthoporus ornatus</i></td><td>0</td><td>1</td><td>0</td></tr><tr><td>[3] <i>Scalopandra polymor...</i></td><td>0</td><td>1</td><td>0</td></tr></tbody></table> Character matrix		[1] Nerve cords lateralized	[2] Segmented postoral nervous system	[3] Brain segmented	[1] <i>Eisneripetolides rowelli</i>	1	0	0	[2] <i>Orthoporus ornatus</i>	0	1	0	[3] <i>Scalopandra polymor...</i>	0	1	0
	[1] Nerve cords lateralized	[2] Segmented postoral nervous system	[3] Brain segmented																	
[1] <i>Eisneripetolides rowelli</i>	1	0	0																	
[2] <i>Orthoporus ornatus</i>	0	1	0																	
[3] <i>Scalopandra polymor...</i>	0	1	0																	

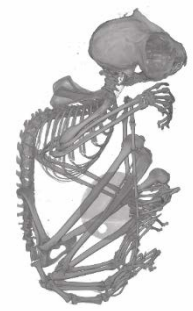
Some (but not all) data from a search for Dytiscidae water beetles



MORPHO
SOURCE



The Paleobiology Database
revealing the history of life



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



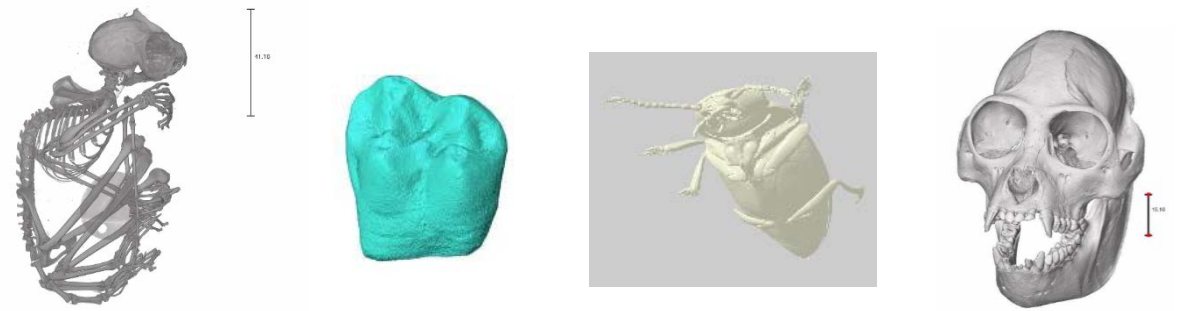
Concern: What if the specialized repository disappears?



MORPHO
SOURCE



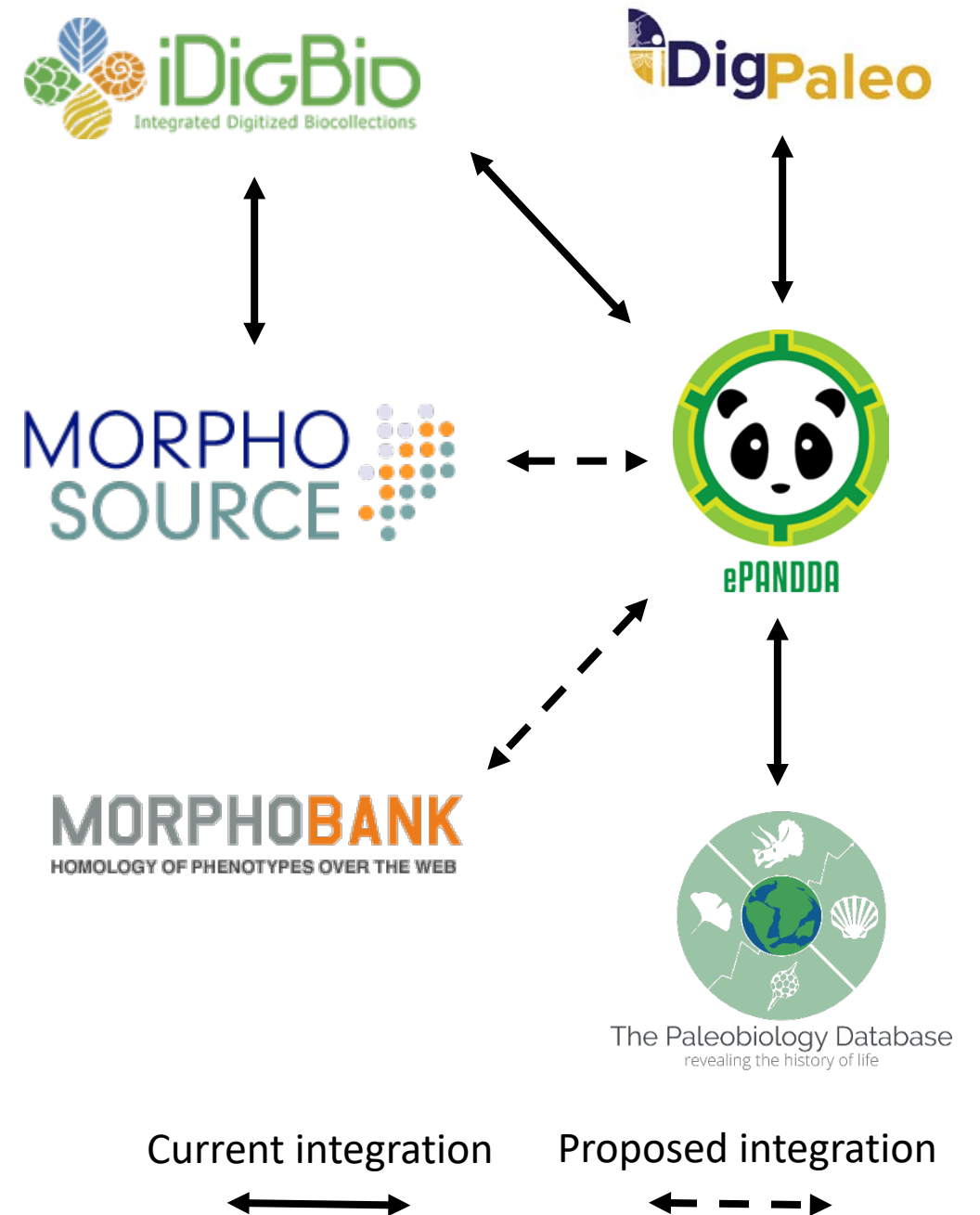
The Paleobiology Database
revealing the history of life



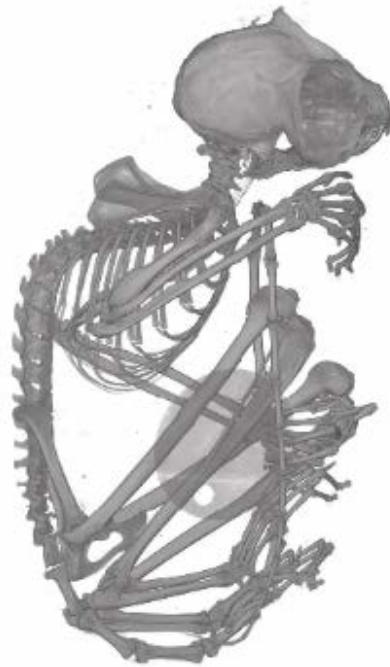
Solution: modular design conforming to community standards, easy to transport if necessary

Database integration

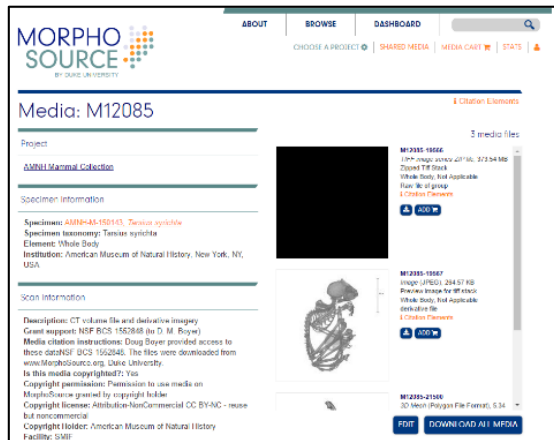
- Benefits
 - More sustainable data model
 - Easier for end users
 - New kinds of automated research
 - Force multiplier
- Challenges:
 - Collaborative development of integration models
 - Record matching



>25,000 3D media files of vouchered specimens



>104 million specimens, many from museum collections



The screenshot shows the MorphoSource interface for specimen M12085. It includes a navigation bar with 'ABOUT', 'BROWSE', and 'DASHBOARD'. The main content area displays 'Media: M12085' and lists three media files with their respective download links. The specimen information section identifies the species as *Tarsius syrichta* and provides details about the specimen's location and collection date.

Data		Flags	Raw
Taxonomy			
Scientific Name	Tarsius syrichta		
Higher Classification	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

The image shows a screenshot of the iDigBio website interface. At the top, there is a navigation bar with links for 'About iDigBio', 'Research', 'Technical Information', 'Education', 'Log In', and 'Sign Up'. Below this is a secondary navigation bar with 'iDigBio Home', 'Portal Home', 'Search Records', 'Learning Center', 'Data', 'Research Collaboration', and 'Feedback'. The main content area displays a 'Specimen Record' for *Tarsius bancanus borneanus*. The record includes taxonomic classification (Animalia > Chordata > Mammalia > Primates > Tarsiidae), source information (From AMNH Mammal Collections), and detailed collection data: Continent (Asia), Country (Indonesia), State/Province (Borneo), County/Parish (Kalimantan Timur), Locality (Peleben, Sungai Kajau), Institution Code (Amnh), Collection Code (Mammals), Catalog Number (M-106010), Collected By (Baron V. Von Plessen), and Date Collected (1935-09-13). Below the record, there is a link to the recordset and a paragraph of text about the AMNH mammal collections. On the left side of the screenshot, a large 3D model of a tarsus is shown, labeled '55 Project Specimens' and 'AMNH-M-106010, Tarsius bancanus'. Below the model is an 'iDigBio' logo with a link icon. At the bottom of the screenshot, two smaller 3D models of other specimens are shown, labeled 'AMNH-M-143741, Callithrix pygmaea' and 'AMNH-M-150142, Tarsius syrichta', each with its own 'iDigBio' link.

Integrating new data

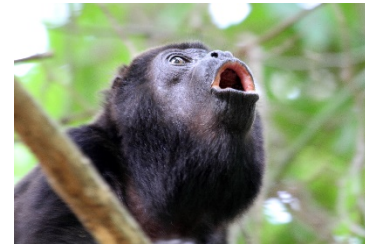
- Users adding new specimen records search for pre-existing MorphoSource records
- Specimen import tool now searches iDigBio automatically
 - Associate occurrence ID

The screenshot displays the MorphoSource website interface. At the top, there is a navigation bar with links for 'ABOUT', 'BROWSE', and 'DASHBOARD'. A search bar is located on the right side of the navigation bar. Below the navigation bar, the MorphoSource logo is visible, along with the text 'BY DUKE UNIVERSITY'. The main content area is titled 'Search Specimens' and includes a search form with fields for 'Institution Code', 'Collection Code', 'Catalog Number' (containing '170560'), and 'Genus'. A 'Species' field and a 'SEARCH' button are also present. Below the search form, there is a link: 'Can't find your specimen? [Enter your specimen directly in MorphoSource](#)'. The results section is divided into two columns: 'MorphoSource Results' and 'iDigBio Results'. The MorphoSource Results section shows one result: 'AMNH-M-170560, *Lepilemur mustelinus*' with a green checkmark and the text 'iDigBio integrated'. Below this result are two buttons: 'ADD MEDIA' and 'LINK SPECIMEN TO PROJECT'. The iDigBio Results section shows 64 results, with the first four visible: 'us-botany-170560, *lithothamnion lemoineae*', 'us-botany-170560, *lithothamnion lemoineae*', 'us-botany-170560, *lithothamnion lemoineae*', and 'mvz-bird specimens-170560, *eremophila alpestris lamprochroma*'. Each result includes a 'View on iDigBio' link and an 'IMPORT SPECIMEN' button.

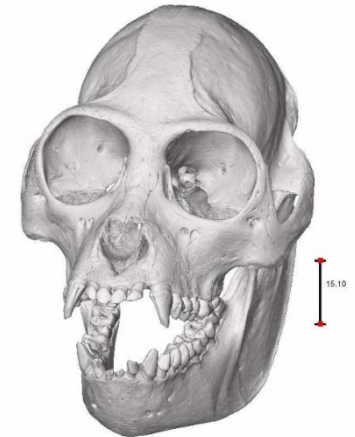
Good progress so far, but...

- 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.

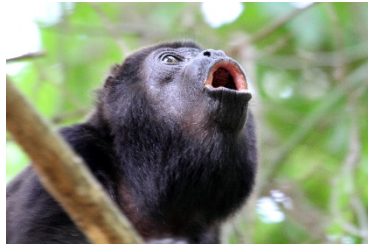


Specimen record
MorphoSource can't
authoritatively assign
GUID

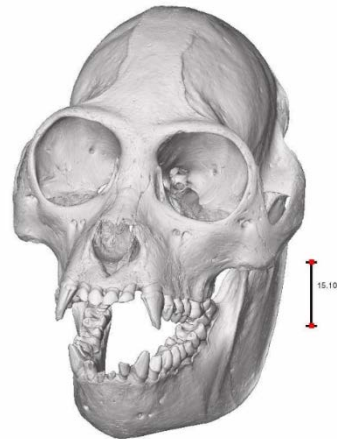


3D Media
MorphoSource can
authoritatively assign GUID

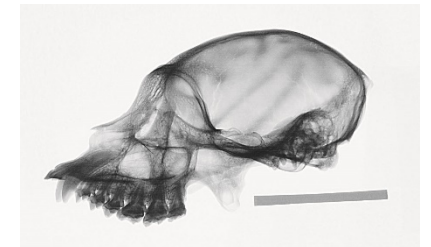
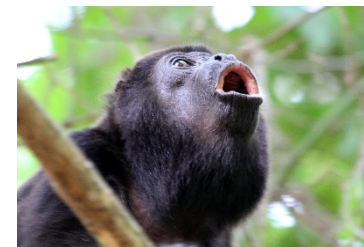
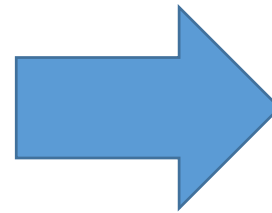
Ingestion of MorphoSource media records into iDigBio



Alouatta palliata specimen record
No occurrence ID
MorphoSource-assigned GUID



3D Media
MorphoSource-assigned GUID



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] <i>Euperipatoides rowelli</i>	1	0	0
[2] <i>Orthoporus ornatus</i>	0	1	0
[3] <i>Scolopendra polymor...</i>	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

MORPHO SOURCE BY DUKE UNIVERSITY

ABOUT | BROWSE | DASHBOARD

TEST PROJECT 11111 | SHARED MEDIA | MEDIA CART | STATS

Specimen: *AMNH-M-106010, Tarsius bancanus* [PREVIOUS] [BACK] [NEXT]

[VIEW SPECIMEN ON IDIGBIO](#) [VIEW SPECIMEN ON PALEOBIO DB](#) [VIEW SPECIMEN ON MORPHOBANK](#)

Project

[Yapuncich et al 2015 data set](#)

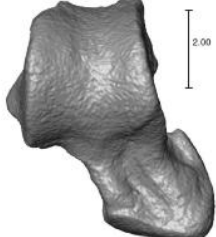

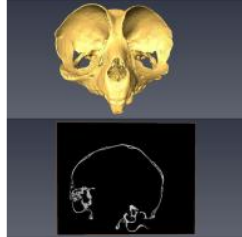
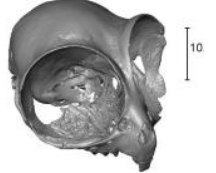
Specimen Information

Vouchered
Type: Yes
Sex: Female
Occurrence ID: urn:catalog:AMNH:Mammals:M-106010

Description: *Tarsius bancanus borneanus*

Institution: American Museum of Natural History, New York, NY, USA

Specimen Media

			
M5072 , 1 file AMNH 106010 <i>Tarsius bancanus borneanus</i> smooth surface (Right) (Astragalus)	M5413 , 1 file Smooth Surface Mesh (First metatarsal)	M6481 , 1 file CT Scan (cranium)	M6536 , 1 file smooth mesh file (Midline) (cranium)

Links to data from other sites within each site



MORPHO SOURCE

MORPHOBANK
HOMOLOGY OF PHENOTYPES OVER THE WEB

New integration
----->



iDigPaleo

iDigBio
Integrated Digitized Biocollections



The Paleobiology Database
revealing the history of life

MORPHO SOURCE
BY DUKE UNIVERSITY

ABOUT | BROWSE | DASHBOARD | TEST PROJECT 11111 | SHARED MEDIA | MEDIA CART | STATS

Specimen: AMNH-M-106010, *Tarsius bancanus* [PREVIOUS] [BACK] [NEXT]

VIEW SPECIMEN ON EDRO | VIEW SPECIMEN ON PALEOBIO DB | VIEW SPECIMEN ON MORPHOBANK

Project
[Yapuncich et al. 2015 data set](#)

Specimen Information

Vouchered
Type: Yes
Sex: Female
Occurrence ID: um.catalog.AMNH Mammals M-106010

Description: *Tarsius bancanus borneanus*

Institution: American Museum of Natural History, New York, NY, USA

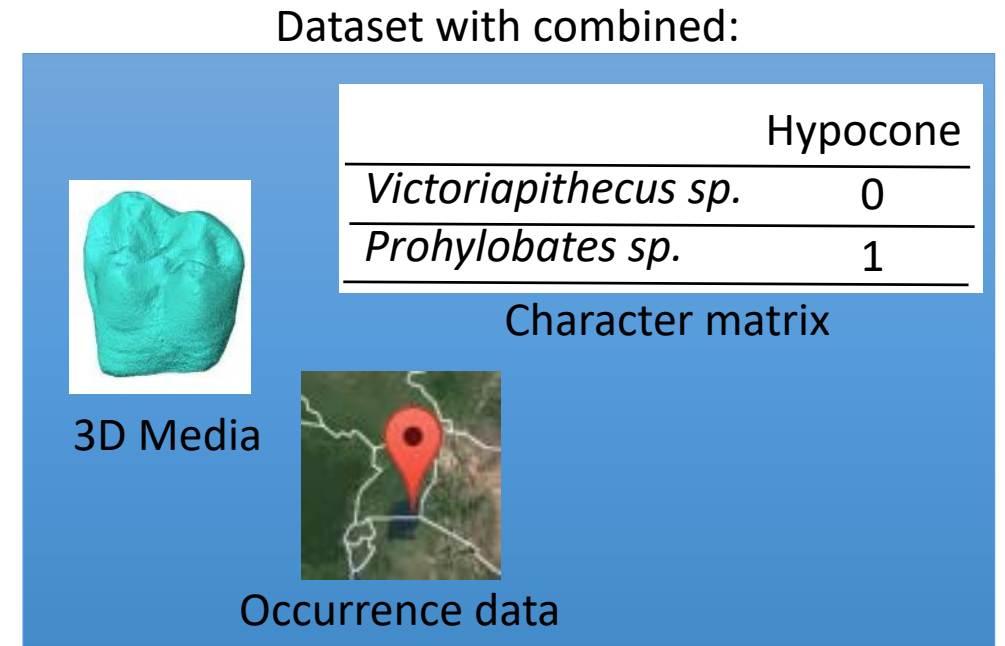
Specimen Media

MS022 1 file AMNH 106010 Tarsius bancanus borneanus smooth surface (Right) (Isobagalis)	MS413 1 file Smooth Surface Mesh (First metatarsal)	MS483 1 file CT Scan (cranium)	MS026 1 file smooth mesh file (skelene) (cranium)

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



User first visits...

MORPHOBANK
HOMOLOGY OF PHENOTYPES OVER THE WEB

MORPHO
SOURCE

And is also routed to...



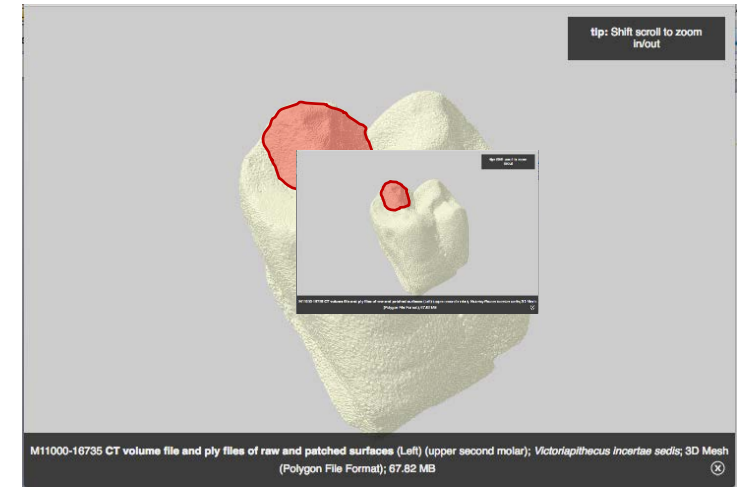
The Paleobiology Database
revealing the history of life

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] <i>Echinosorex gymmura</i>	absent	higher	absent	large		
[2] <i>Erinaceus europaeus</i>	absent	higher	absent	large		
[3] † <i>Aegyptopithecus zeu...</i>	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



Gabe Yapuncich



Gil Nelson

Alex Thompson

Kevin Love

Dan Stoner

Whirl-i-gig

Seth Kaufman

Maria Passarotti



Adam Summers



David Blackburn

Ed Stanley



CAREER Grant BCS 155284

EarthCube IA 540902