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
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

Price (1960)  
Larsen and von Ins (2010)
Scientific data grows exponentially.

**Global scientific output doubles every nine years**

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

Lutz Bornmann, Ruediger Mutz

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

- VerNe
- iDigBio

Lists of specimens and associated information

Research databases

- DigPaleo
- Morphbank
- MORPHO SOURCE
- EarthChem
- MorphoBANK
- PANGAEA
- GBIF
- LinkedEarth
- Neotoma Paleoeoecology 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

<table>
<thead>
<tr>
<th>Database</th>
<th>Example Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>DigPaleo</td>
<td>2D Media</td>
</tr>
<tr>
<td>The Paleobiology Database</td>
<td>Occurrence (geographic, stratigraphic)</td>
</tr>
<tr>
<td>MOPHOSOURCE</td>
<td>3D Media</td>
</tr>
<tr>
<td>MOPHOBANK</td>
<td>Character matrix</td>
</tr>
</tbody>
</table>

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

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

(Uhen et al, 2013)
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?
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

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

Already existing media, but no museum-provided occurrence ID for specimen
• Integration model: Enhance individual resources through combined data access, deposition, and workflow tools

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

Dataset with combined:

<table>
<thead>
<tr>
<th>Hypocone</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Victoriapithecus sp.</td>
<td>0</td>
</tr>
<tr>
<td>Prohylobates sp.</td>
<td>1</td>
</tr>
</tbody>
</table>

User first visits…

And is also routed to…
Data analysis tools

MorphoBank Collaborative Web Matrix Builder

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
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

iDigBio
Gil Nelson
Alex Thompson
Kevin Love
Dan Stoner

Whirl-i-gig
Seth Kaufman
Maria Passarotti

University of Washington
Adam Summers

Friday Harbor Laboratories

University of Florida
David Blackburn
Ed Stanley

NSF
CAREER Grant BCS 155284
EarthCube IA 540902