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.
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)

- **InvertNet**: 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

• Associated media

• Global

• All taxa (neo- & paleo)
Why collections publish/share data with iDigBio

• Discovery and use
• Data quality, improvement

- Attribution, credit, collection value metrics

---

[iDigBio Integrated Digital Bioresources]

**Specimen Record**

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>scientificname_added</td>
<td></td>
</tr>
<tr>
<td>geopoint_low_precision</td>
<td>Geographic Coordinate has Low Precision.</td>
</tr>
<tr>
<td>dwc_phylum_added</td>
<td>Darwin Core Phylum Added.</td>
</tr>
<tr>
<td>dwc_country_replaced</td>
<td>Darwin Core Country Corrected.</td>
</tr>
<tr>
<td>idigbio_isocountrycode_added</td>
<td>iDigBio ISO 3166-1 alpha-3 Country Code Added.</td>
</tr>
<tr>
<td>dwc_kingdom_added</td>
<td>Darwin Core Kingdom Added.</td>
</tr>
</tbody>
</table>

**Recordset**

The table below represents monthly iDigBio portal use statistics for this recordset. **Search** indicates how many instances a record from this recordset matched a search query. **Download** indicates how many instances a record from this recordset was downloaded. **Seen** indicates how many instances a record from this recordset appeared (visually) in the search results as 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.

<table>
<thead>
<tr>
<th>Month of</th>
<th>Search</th>
<th>Download</th>
<th>Seen</th>
<th>Records Viewed</th>
<th>Media Viewed</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 / 2015</td>
<td>4,644,439</td>
<td>39,693</td>
<td>249</td>
<td>185</td>
<td>36</td>
</tr>
<tr>
<td>02 / 2015</td>
<td>14,478,438</td>
<td>18,174</td>
<td>872</td>
<td>93</td>
<td>6</td>
</tr>
<tr>
<td>03 / 2015</td>
<td>11,134,952</td>
<td>92,788</td>
<td>604</td>
<td>136</td>
<td>8</td>
</tr>
<tr>
<td>----------------------</td>
<td>------</td>
<td>------</td>
<td>------------------</td>
<td>------------------</td>
<td></td>
</tr>
<tr>
<td>PRI</td>
<td>112</td>
<td>28,474</td>
<td>505</td>
<td>149</td>
<td></td>
</tr>
<tr>
<td>SDSM-VP</td>
<td>767</td>
<td>31,749</td>
<td>23</td>
<td>149</td>
<td></td>
</tr>
<tr>
<td>NCSM-Bot</td>
<td>1,848</td>
<td>10,031</td>
<td>4,615</td>
<td>5,007</td>
<td></td>
</tr>
<tr>
<td>VMNH-IP</td>
<td>1,953</td>
<td>820,941</td>
<td>168</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>OU-IP</td>
<td>2,412</td>
<td>3,725</td>
<td>234</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>AMNH-IP</td>
<td>7,218</td>
<td>20,316</td>
<td>24</td>
<td>782</td>
<td></td>
</tr>
<tr>
<td>NCSM-IP</td>
<td>11,877</td>
<td>375,128</td>
<td>32,422</td>
<td>4,347</td>
<td></td>
</tr>
<tr>
<td>ROM_nonMam</td>
<td>27,312</td>
<td>881,618</td>
<td>266</td>
<td>269,540</td>
<td></td>
</tr>
<tr>
<td>NCSM-VP</td>
<td>31,394</td>
<td>212,267</td>
<td>63,018</td>
<td>47,976</td>
<td></td>
</tr>
<tr>
<td>UAM</td>
<td>36,657</td>
<td>951,925</td>
<td>124,371</td>
<td>154,354</td>
<td></td>
</tr>
<tr>
<td>ROM-Mam</td>
<td>49,672</td>
<td>672,794</td>
<td>706</td>
<td>173,092</td>
<td></td>
</tr>
<tr>
<td>Yale-VP</td>
<td>71,805</td>
<td>332,854</td>
<td>11,876</td>
<td>35,482</td>
<td></td>
</tr>
<tr>
<td>NRM (Sweden)</td>
<td>73,974</td>
<td>943,447</td>
<td>149</td>
<td>17,732</td>
<td></td>
</tr>
<tr>
<td>CU-IP</td>
<td>74,862</td>
<td>304,508</td>
<td>290</td>
<td>8,410</td>
<td></td>
</tr>
<tr>
<td>CM</td>
<td>79,254</td>
<td>1,451,295</td>
<td>224,338</td>
<td>66,530</td>
<td></td>
</tr>
<tr>
<td>KU-VP</td>
<td>79,979</td>
<td>219,905</td>
<td>75</td>
<td>165,043</td>
<td></td>
</tr>
<tr>
<td>Yale-Bot</td>
<td>91,261</td>
<td>4,711,819</td>
<td>2,957</td>
<td>253,575</td>
<td></td>
</tr>
<tr>
<td>UT-NPL</td>
<td>99,463</td>
<td>6,337,018</td>
<td>358</td>
<td>49,224</td>
<td></td>
</tr>
<tr>
<td>OMNH-IP</td>
<td>210,019</td>
<td>197,664</td>
<td>167</td>
<td>25,302</td>
<td></td>
</tr>
<tr>
<td>KU-IP</td>
<td>236,371</td>
<td>634,381</td>
<td>318,213</td>
<td>46,499</td>
<td></td>
</tr>
<tr>
<td>UF-IP</td>
<td>252,944</td>
<td>1,714,693</td>
<td>364</td>
<td>373,671</td>
<td></td>
</tr>
<tr>
<td>MNHN (France)</td>
<td>270,867</td>
<td>39,548</td>
<td>48</td>
<td>68,710</td>
<td></td>
</tr>
<tr>
<td>Yale-IP</td>
<td>361,213</td>
<td>7,192,437</td>
<td>7,082</td>
<td>2,234,510</td>
<td></td>
</tr>
<tr>
<td>UF-VP</td>
<td>377,038</td>
<td>2,049,541</td>
<td>1,162</td>
<td>3,059,768</td>
<td></td>
</tr>
<tr>
<td>UCMP</td>
<td>432,668</td>
<td>3,175,747</td>
<td>515,810</td>
<td>1,423,913</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2,882,940</strong></td>
<td><strong>33,313,825</strong></td>
<td><strong>1,309,241</strong></td>
<td><strong>8,483,699</strong></td>
<td></td>
</tr>
<tr>
<td>per record</td>
<td>11.6</td>
<td>0.5</td>
<td>2.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

YTD, 2016
Paleo data fields (GeologicalContext)

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

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

- 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


• 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

Enhancing Paleontological and Neontological Data API

Open Source cluster computer framework

Leaflet

Fresh data
Thanks!

• To all the data providers and publishers
• iDigBio team

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