iDigBio Cyberinfrastructure, Portal and Data

José Fortes
Advanced Computing and Information Systems Laboratory (ACIS)
University of Florida

fortes@acis.ufl.edu

iDigBio is funded by a grant from the National Science Foundation’s Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). 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.
Outline

• Cyberinfrastructure
  – Web portal
  – Portal
  – Appliances
  – Research applications

• Data
  – Ingestion
  – Use
  – Integration
Evolution of iDigBio capabilities

Data ingestion → Data access, provision and visualization → Provide and enable data feedback → Data linking and federation → Process and visualize integrated data → Time

Increasing storage and server hosting in support of the above
Increasing number of appliances in support of the above
Web site for interaction with public, community, education and above
iDigBio Website

The iDigBio website is dedicated to making data and images of millions of biological specimens available online. The site offers resources for digitization, sharing collections, working groups, and proposals, as well as opportunities for citizen scientists. It also includes information for researchers and collections staff, highlighting how their collections can benefit from the work. Upcoming events are also featured, including workshops and symposiums focused on digitization and biodiversity.
Search across all data, all/individual fields, customize, use autocompletion, synonyms, ...
## View search results as table, labels, images...

<table>
<thead>
<tr>
<th>Family</th>
<th>Scientific Name</th>
<th>Genus</th>
<th>Country</th>
<th>State/Province</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endodontidae</td>
<td>Aaadonta</td>
<td>Aaadonta</td>
<td>Palau</td>
<td>Ulebechel Island</td>
</tr>
<tr>
<td>Endodontidae</td>
<td>Aaadonta</td>
<td>Aaadonta</td>
<td>Palau</td>
<td></td>
</tr>
<tr>
<td>Endodontidae</td>
<td>Aaadonta</td>
<td>Aaadonta</td>
<td>Palau</td>
<td></td>
</tr>
</tbody>
</table>

---

**Table view**

---

**Label view**

---

**Images**

---

## Search matched 65 records

---

## Search matched 3,512,348 records

---

## Search matched 749,373 records

---

**Images:**

2. *Adlumia fungosa Greene ex Britton, Stern & Poggenb.* Dicorylidae, Papaveraceae, USA, Vermont, David’s Den, Unknown, UConn, CONN.
7. *Adlumia fungosa Greene ex Britton, Stern & Poggenb.* Dicorylidae, Papaveraceae, USA, Connecticut, Ashford, Boston Hollow Road, Leslie J. Meirhoffs, UConn, CONN.

---

**Images:**

1. *Acanthonyx petiveri,* H. Milne-Edwards
2. *Acantholepis granulata,* Grelin
3. *Acar domingensis,* Lamarck
4. *Acar domingensis,* Lamarck
5. *Aedes sebae,* Milne-Edwards
Results mapped/rendered and downloadable
Specimen record page with details, info on associated media, georeference and provider
Media records with metadata, other media, provider, links to specimen record, data set ...

Media Record: Agrilus planipennis

Download Media File

Other Media

Record Provided By
C.A. Triplehorn Insect Collection (OSUC), Ohio State University
http://osuc.osu.edu
Vouchered occurrence records for insects from the C.A. Triplehorn Insect Collection at the Ohio State University.

Contacts
Publishers page with record counts, links to provider details

### iDigBio Data Publishers

This page shows all iDigBio data contributors. If you are interested in providing data, consult the data ingestion guide for more information.

<table>
<thead>
<tr>
<th>Record Count</th>
<th>Media Record Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total from Providers</td>
<td>22,556,733</td>
</tr>
<tr>
<td>Total in API</td>
<td>22,606,602</td>
</tr>
<tr>
<td>Total Published (all data incorporated in new workflow)</td>
<td>22,605,241</td>
</tr>
<tr>
<td>Total Indexed (all data) *</td>
<td>22,605,241</td>
</tr>
</tbody>
</table>

* Data that is marked deleted in iDigBio remains indexed until a cleanup is run.

### Publisher Summary

<table>
<thead>
<tr>
<th>Publisher Name</th>
<th>Digest</th>
<th>API</th>
<th>Index</th>
<th>Digest</th>
<th>API</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berkeley Natural History Museums IPT</td>
<td>1,860,584</td>
<td>1,859,985</td>
<td>1,859,985</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Florida Museum of Natural History IPT Service</td>
<td>1,047,587</td>
<td>1,047,587</td>
<td>1,047,584</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Northern Great Plains Herbaria Darwin Core Archive rss feed</td>
<td>43,012</td>
<td>43,012</td>
<td>43,012</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MyCoPortal Darwin Core Archive rss feed</td>
<td>1,679,459</td>
<td>1,679,458</td>
<td>1,679,458</td>
<td>371,346</td>
<td>371,346</td>
<td>371,346</td>
</tr>
<tr>
<td>KU Biodiversity Institute IPT</td>
<td>2,010,071</td>
<td>2,011,170</td>
<td>2,011,170</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>The University of Connecticut Biological Collections</td>
<td>172,098</td>
<td>171,936</td>
<td>171,198</td>
<td>166,689</td>
<td>166,519</td>
<td>166,022</td>
</tr>
<tr>
<td>xBioD IPT in the Museum of Biological Diversity at the Ohio State University</td>
<td>521,710</td>
<td>521,782</td>
<td>521,782</td>
<td>2,593</td>
<td>2,593</td>
<td>2,593</td>
</tr>
<tr>
<td>CMC Specify</td>
<td>9,131</td>
<td>9,131</td>
<td>9,131</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Consortium of North American Bryophyte Herbaria Darwin Core Archive rss feed</td>
<td>1,690,014</td>
<td>1,690,014</td>
<td>1,690,014</td>
<td>816,932</td>
<td>816,932</td>
<td>816,932</td>
</tr>
<tr>
<td>Museum of Comparative Zoology, Harvard University</td>
<td>1,736,357</td>
<td>1,736,471</td>
<td>1,736,471</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CNALH Darwin Core Archive rss feed</td>
<td>1,232,891</td>
<td>1,232,891</td>
<td>1,232,891</td>
<td>649,241</td>
<td>649,241</td>
<td>649,241</td>
</tr>
<tr>
<td>SCAN Darwin Core Archive rss feed</td>
<td>873,024</td>
<td>873,160</td>
<td>873,160</td>
<td>68,718</td>
<td>68,718</td>
<td>68,718</td>
</tr>
<tr>
<td>iDigBio Feeder RSS Feed</td>
<td>1,316,574</td>
<td>1,316,574</td>
<td>1,316,574</td>
<td>19,024</td>
<td>19,024</td>
<td>19,024</td>
</tr>
</tbody>
</table>
Recordset page with provider info, record counts, links to search and raw data

Recordset: University of Washington Herbarium

The University of Washington Herbarium (also known as WTU) is an international resource for research into the diversity, distribution and ecology of Pacific Northwest vascular plants, non-vascular plants, fungi, lichen, and algae. With over 600,000 specimens currently in the collections and between 5,000-10,000 specimens added annually, WTU is one of the largest herbaria in the region. The University of Washington Herbarium (also known as WTU) is an international resource for research into the diversity, distribution and ecology of Pacific Northwest vascular plants, non-vascular plants, fungi, lichen, and algae. With over 600,000 specimens currently in the collections and between 5,000-10,000 specimens added annually, WTU is one of the largest herbaria in the region.

Specimen Fields Used for Search

This table represents the fields in specimen records that are used for iDigBio search. The first column represents the field name and equivalent DWC term. The last two columns represent the number and percentage of records that provide the field.

<table>
<thead>
<tr>
<th>Field</th>
<th>(% Percent Used)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kingdom (dwc:kingdom)</td>
<td>0</td>
</tr>
<tr>
<td>Phylum (dwc:phylum)</td>
<td>0</td>
</tr>
<tr>
<td>Class (dwc:class)</td>
<td>0</td>
</tr>
<tr>
<td>Order (dwc:order)</td>
<td>0</td>
</tr>
<tr>
<td>Family (dwc:family)</td>
<td>66.675</td>
</tr>
<tr>
<td>Scientific Name (dwc:scientificName)</td>
<td>66.011</td>
</tr>
<tr>
<td>Genus (dwc:genus)</td>
<td>66.004</td>
</tr>
<tr>
<td>Specific Epithet (dwc:specificEpithet)</td>
<td>64.589</td>
</tr>
<tr>
<td>Infraspecific Epithet (dwc:infraspecificEpithet)</td>
<td>3.769</td>
</tr>
<tr>
<td>Higher Taxon (dwc:higherClassification)</td>
<td>0</td>
</tr>
<tr>
<td>Common Name (dwc:vernacularName)</td>
<td>0</td>
</tr>
<tr>
<td>Lat (dwc:decimalLatitude)</td>
<td>13.301</td>
</tr>
<tr>
<td>Lon (dwc:decimalLongitude)</td>
<td>13.110</td>
</tr>
</tbody>
</table>

Last Update: 2014-06-24
Total Specimen Records: 74,141
Total Media Records: 23,069

Website

Contacts
egbot@asu.edu
David Giblin, Collections Manager
wtu@u.washington.edu
Over 300 publishers, 27M specimen records, 4.3M media records

Publishing technologies: IPT, Symbiota, RSS (DwC-a, CSV)

Media data using Audubon core terms

... and many more.
The what and how of data ingestion

iDigBio Data Flow Diagram

- **Collections**
  - Specify, EMu, Symbiota, ...

- **Publishers**
  - IPT, Symbiota, iDigBio Feeder

- **Data Ingestion**
  - Python, PostgreSQL, JSON, Redis
    - iDigBio API
      - PostgreSQL, Riak
    - Searchable Indexed Data
      - Elasticsearch

- iDigBio Portal Web Site
  - HTML5, jQuery, Backbone, Node.js, Express
  - https://www.idigbio.org/portal

- Scientific Community
  - (Researchers, Scientists, Developers, and downstream consumers)

- IPT – RSS of DwC-A
- Specify, EMu, Arctos, VertNet Migrator, etc.
- Symbiota portals – RSS of DwC-A
- iDigBio Feeder – DwC-A, CSV, ...

If you can export specimen data from your database/spreadsheet into DwC-A (or even CSV), then you can share data with iDigBio.
Architecture Components

iDigBio Specimen Portal
HTML5, jQuery, React, Node.js, Express

iDigBio Search API
Elasticsearch

Text Indexes
Elasticsearch

SQL Indexes
PostgreSQL

Bulk Text Storage
Riak

iDigBio Metadata API
Node.js, Restify, REST, JSON

iDigBio Object API
Python, CherryPy, REST, JSON

Third Party API Consumers
Python, jQuery

iDigBio Authentication and Management
Apache, Ruby, Sinatra

iDigBio Worker Services
Apache, CherryPy, Python, Celery

Cloud Node

Cloud Node

Cloud Node

Cloud Node

Cloud Node

Cloud Node

Cloud Node
iDigBio infrastructure (54 servers): Proxy/load balance (2); Portal (5); API (5); Media API (10) Celery task (5) Ceph Object Storage (3) CSV generators (3), Redis cache (3), Application and database (18)
Appliances, e.g. upload of images and Specify package

- Reliable approach to upload batches of images with metadata
- Upload starts with CSV file with image paths, identifier, and metadata
- Successfully helped users to upload 290,000+ images.
Upload Images with CSV File

• Upload starts with CSV file with image paths, identifier, and additional metadata
• 10 threads used to speed up transmission.
CSV File Generation

- Appliance helps users generate a CSV file (with GUID and path) for all images within a directory hierarchy.
- Optionally, users can manually edit or define new metadata fields in the CSV file.
Viewing the History

The local upload history can be viewed/saved to CSV files.
Current upload results are also shown after each upload.

### Batch Information Table

(Click on each row to see the details)

<table>
<thead>
<tr>
<th>ID</th>
<th>CSV File Path</th>
<th>Start Time</th>
<th>Total Records</th>
<th>Failed Records</th>
<th>Skipped Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C:\Users\winday0215\Desktop\idigbio\images2\media_records.csv</td>
<td>2014-08-17 22:15:34</td>
<td>336</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>C:\Users\winday0215\Desktop\idigbio\images2\media_records.csv</td>
<td>2014-08-17 22:15:53</td>
<td>336</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>C:\Users\winday0215\Desktop\idigbio\images2\media_records.csv</td>
<td>2014-08-17 22:16:21</td>
<td>336</td>
<td>0</td>
<td>220</td>
</tr>
<tr>
<td>4</td>
<td>C:\Users\winday0215\Desktop\idigbio\idigbio-ingestion-tool-</td>
<td>2014-09-07 22:16:21</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Specify Appliance

- Appliance packages Ubuntu 12.04 LTS, MySQL, Java 7, Specify 6.5, Demo database
- User installs free software and appliance from iDigBio
iDigBio Research Section

Links to TCN research
List of iDigBio publications

• Expanding: https://www.idigbio.org/research
iDigBio Research Tools

Welcome your contributions!

Community Research Tools

To facilitate the study of biodiversity, a number of research tools are being developed to take advantage of the data being digitized at US institutions and made available by iDigBio through web services. You can find below some of these online tools developed by the community. If you would like your tool to be included in this list, please use the feedback form to tell us about your work.

List of Tools Integrating iDigBio Web Services

- https://www.idigbio.org/content/community-research-tools
- Welcome your contributions!
PhyloJIVE instance in iDigBio (biodiversity data + phylogeny)

- Developed by Garry Jolley-Rogers, Joe Miller, and Temi Varghese
- Displays phylogenetic trees in Newick format
- Displays up to 10 characters (traits); color scale indicates numerical intensity/categories
- Tree branches colored per predicted first character, calculated via reverse parsimony
- Integrated w/iDigBio search and mapping; linked to other sites (ALA, EOL, DiscoverLife)
- User-created trees/characters, sample trees, canned searches,...
Research tools integrated with iDigBio

- PhyloJIVE + OpenTree + iDigBio
- OpenRefine + OpenTree + iDigBio
- Arbor + OpenTree + iDigBio

- Others – contact Andrea Matsunaga (ammatsun@ufl.edu) if you are interested in integration of your research tool(s)

- See Demos and attend Discussion Sessions
Acknowledgements

• National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210)
• Dr. Anne Maglia and Dr. Judith Skog @NSF
• iDigBio faculty, students and staff at UF and FSU
  – in particular, the iDigBio IT team
  • in particular, the iDigBio IT team members at ACIS
Acknowledgements