

## Cyberinfrastructure and Data (What a difference a year makes ...)

José Fortes

Advanced Computing and Information Systems Laboratory (ACIS)

University of Florida

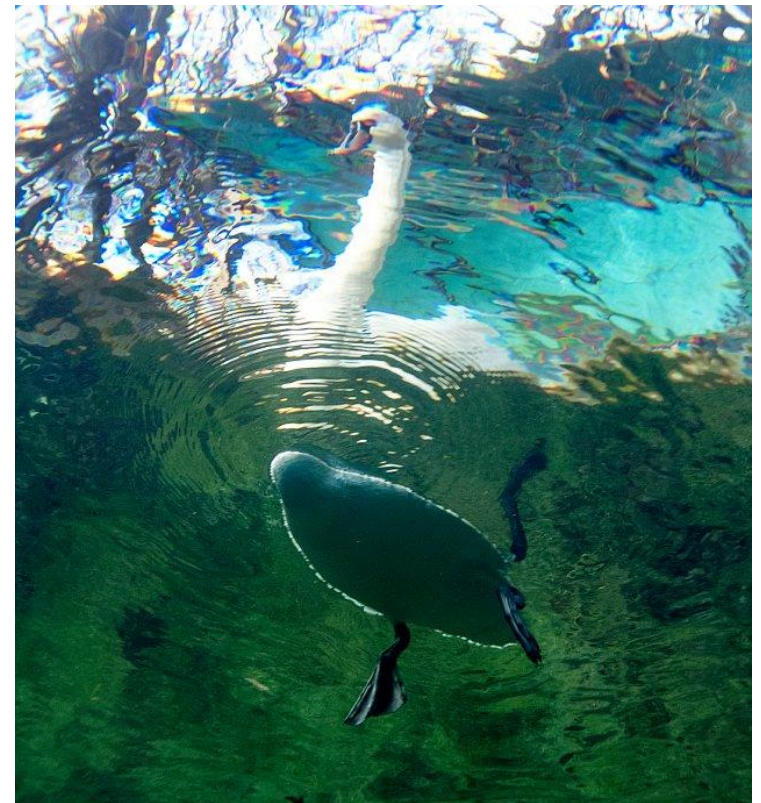
 [fortes@acis.ufl.edu](mailto:fortes@acis.ufl.edu)



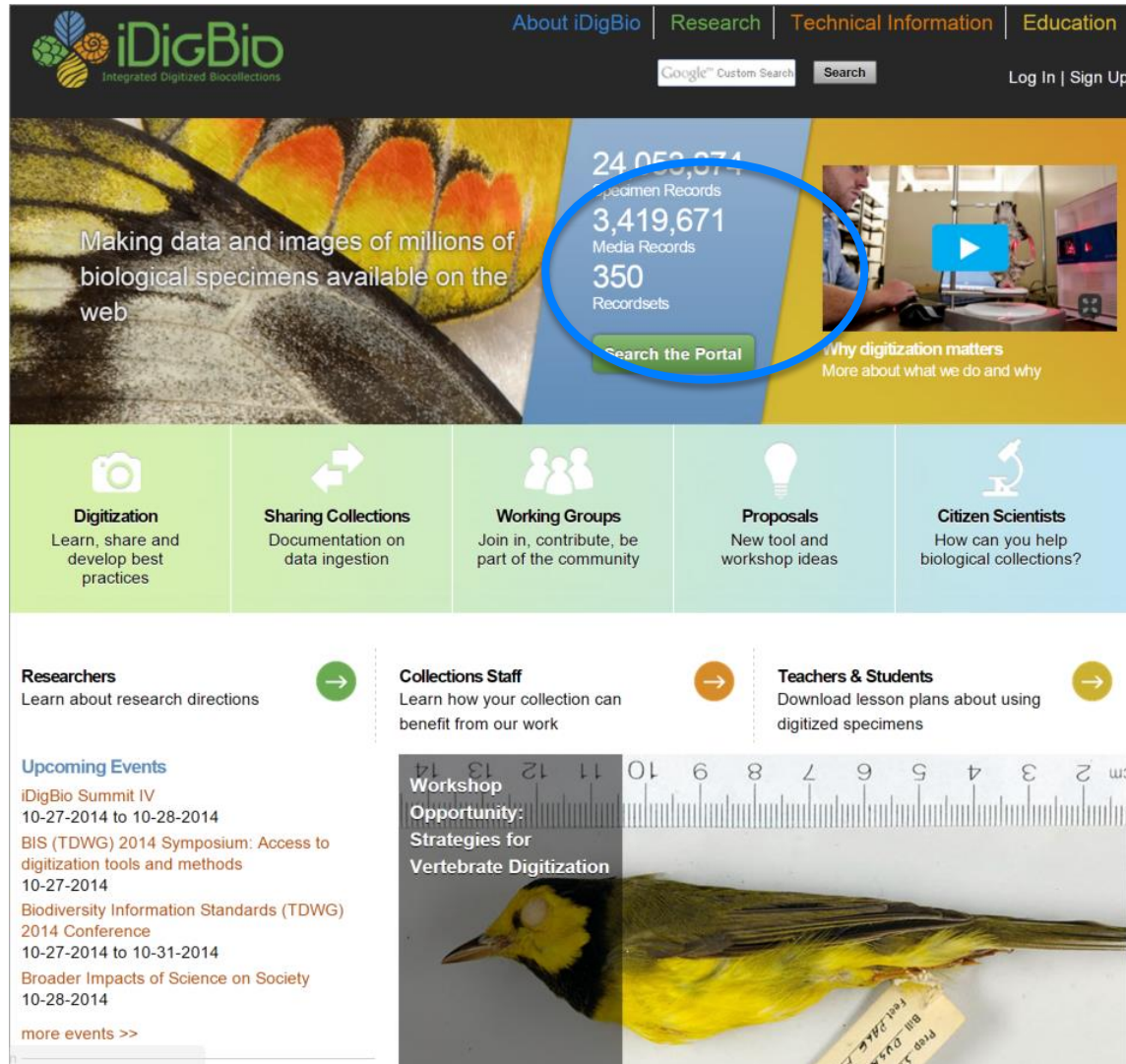
*iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.*

# Outline

- Cyberinfrastructure
  - Web site
  - Portal
  - Appliances
  - Research applications
- Data = Specimen records and media
  - Ingestion
  - Use
  - Integration



# iDigBio Website



The screenshot shows the iDigBio website homepage. At the top, there is a navigation bar with links for "About iDigBio", "Research", "Technical Information", and "Education". A search bar with "Google Custom Search" and a "Search" button is also present, along with "Log In" and "Sign Up" links. The main banner features a butterfly image on the left with the text "Making data and images of millions of biological specimens available on the web". On the right, a statistics box lists: "24,050,074 Specimen Records", "3,419,671 Media Records", and "350 Recordsets". A blue circle highlights this statistics box. Below the banner is a row of five icons representing different user groups: Digitization, Sharing Collections, Working Groups, Proposals, and Citizen Scientists. The bottom section includes "Upcoming Events" with dates for iDigBio Summit IV, BIS (TDWG) 2014 Symposium, Biodiversity Information Standards (TDWG) 2014 Conference, and Broader Impacts of Science on Society. There are also sections for "Researchers", "Collections Staff", and "Teachers & Students" with right-pointing arrows. A video player for "Why digitization matters" is visible on the right side of the banner. At the bottom right, there is a photo of a yellow bird specimen next to a ruler, with the text "Workshop Opportunity: Strategies for Vertebrate Digitization" overlaid.

# Search data, all/individual fields, w/ autocomplete, synonyms, customize; map and download results

**Search Records**

Full Text Search

only records with images [Hide Advanced Search](#)

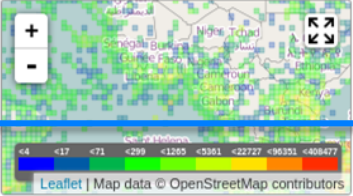
**Current Results**

Query: Match all. Sort by Genus c

Records: 22,556,636

Approx. Download Time: 4hrs 23mins 9secs

Email:



**Advanced Search**

Family   [What is EOL?](#)

Present  Missing

Scientific Name   [What is EOL?](#)

Present  Missing

Genus   [What is EOL?](#)

Present  Missing

Country

Present  Missing

State/Province

Present  Missing

Add a field  Sort by  Direction     [Tips & Hints](#)

Table view [Label view](#) [Images](#) Search Matched 22,556,636 Records

Family	Scientific Name	Genus ▼	Country	State/Province	Lat	Lon
Poaceae	×Elymordeum littorale	×Elymordeum	United States	Alaska	61.5	-149.53
Poaceae	×Elymordeum littorale	×Elymordeum	United States	Alaska	61.5	-149.53
Poaceae	×Elymordeum littorale	×Elymordeum	United States	Alaska	61.5	-149.53

# View search results as table, labels, images...

Table view **Label view** Images Search Matched 65 Records

Family	Scientific Name ^	Genus	Country	State/Province
Endodontidae	Aadonta	Aadonta	Palau	Ulebsechel Island
Endodontidae	Aadonta	Aadonta	Palau	
Endodontidae	Aadonta	Aadonta	Palau	

Table view **Label view** Images Search Matched 3,512,348 Records






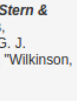


<p><b>"<i>Adlumia fungosa</i> Greene ex Britton, Stern &amp; Poggenb."</b>, Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Lower slope of north side of Bear Mountain, Salisbury, Mt. Riga, Leslie J. Mehrhoff, UConn, CONN</p> 	<p><b>"<i>Adlumia fungosa</i> Greene ex Britton, Stern &amp; Poggenb."</b>, Dicotyledonae, Papaverales, Papaveraceae, USA, Vermont, Devil's Den, Unknown, UConn, CONN</p> 	<p><b>"<i>Adlumia fungosa</i> Greene ex Britton, Stern &amp; Poggenb."</b>, Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Mount Tom Pond, E.H. Eames, UConn, CONN</p> 	<p><b>"<i>Adlumia fungosa</i> Greene ex Britton, Stern &amp; Poggenb."</b>, Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Pistapaug Mountain, Durham, Leslie J. Mehrhoff, UConn, CONN</p> 
<p><b>"<i>Adlumia fungosa</i> Greene ex Britton, Stern &amp; Poggenb."</b>, Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Wooster Mountain, Leslie J. Mehrhoff, UConn, CONN</p> 	<p><b>"<i>Adlumia fungosa</i> Greene ex Britton, Stern &amp; Poggenb."</b>, Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Mrs. G. J. Mendel 111 Highland Ave., So. Norwalk, "Wilkinson, A.E.", UConn, CONN</p> 	<p><b>"<i>Adlumia fungosa</i> Greene ex Britton, Stern &amp; Poggenb."</b>, Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Ashford, Boston Hollow Road, Leslie J. Mehrhoff, UConn, CONN</p> 	<p><b>"<i>Adlumia fungosa</i> Greene ex Britton, Stern &amp; Poggenb."</b>, Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Roxbury Iron Mine; area Mine Hill, Leslie J. Mehrhoff, UConn, CONN</p> 

Table view **Label view** Images Search Matched 749,373 Records

 <p>Acanthonyx petiverii, H. Milne-Edwards</p>	 <p>Acanthopleura granulata, Gmelin</p>	 <p>Acar domingensis, Lamarck</p>	 <p>Acar domingensis, Lamarck</p>	 <p>Achelous sebae, Milne-Edwards</p>
---	--	---	--	--

# Specimen record page with details, info on provider associated media and georeference

## *Astrophyton* FLMNH, Invertebrate Zoology, 11905-Echinodermata

Taxonomy **Specimen** Collection Event Locality Other

Class Ophiuroidea  
 Order Euryalida  
 Family Gorgonocephalidae  
 Genus *Astrophyton*

### Record Provided By

#### invertebratezoology

The UF Invertebrate collection holds ~510,000 databased lots of mollusks and marine invertebrates. It began as a Malacology collection almost 100 years ago and ~85% of the holdings are still mollusks. Since 2000 the collection was expanded to cover all invertebrate phyla, focusing on marine taxa. Today it holds >40,000 species from 28 phyla.

#### Contacts

Gustav Paulay Curator of Invertebrate Zoology <a href="mailto:paulay@flmnh.ufl.edu">paulay@flmnh.ufl.edu</a>	William Paine IT Director <a href="mailto:netadmin@flmnh.ufl.edu">netadmin@flmnh.ufl.edu</a>
---	--

Go To Recordset

View Raw Data

### Associated Media



### Specimen Georeference



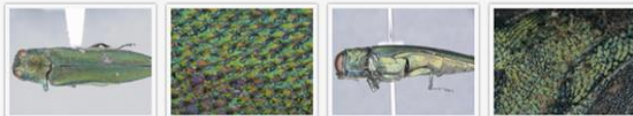
# Media records with metadata, other media, provider, links to specimen record, data set ...

## 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](#)

[Go To Specimen Record](#)

[Go To Recordset](#)

[View Raw Data](#)

### Media Metadata

<b>Resource Creation Technique</b>	AutoMontage extended-focus imaging
<b>Creator</b>	ELIJAH
<b>Identifier</b>	<a href="http://bioguid.osu.edu/osuc_occurrence_images/2828">http://bioguid.osu.edu/osuc_occurrence_images/2828</a>
<b>Modified</b>	2014-08-07T11:14:45Z
<b>Rights</b>	Copyright Norman Johnson 2014
<b>Title</b>	Image for occurrence record OSUC 211414
<b>Type of Resource</b>	StillImage
<b>Subtype</b>	Photograph
<b>Credit</b>	Norman Johnson
<b>License Terms</b>	CC-BY-NC-SA
<b>License URL</b>	<a href="http://creativecommons.org/licenses/by-nc-sa/3.0/">http://creativecommons.org/licenses/by-nc-sa/3.0/</a>
<b>Published Source</b>	<a href="http://specimage.osu.edu/getImageInfo.html?image_id=2828">http://specimage.osu.edu/getImageInfo.html?image_id=2828</a>
<b>Caption</b>	head, lateral view
<b>Access URI</b>	<a href="http://osuc.osu.edu/spm_images/OSUC211414/2828_compress.jpg">http://osuc.osu.edu/spm_images/OSUC211414/2828_compress.jpg</a>
<b>Format</b>	image/jpeg

# 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.

	Record Count	Media Record Count
Total from Providers	22,556,733	3,167,330
Total in API	22,606,002	3,232,444
Total Published (all data incorporated in new workflow)	22,605,241	3,231,947
Total Indexed (all data) *	22,605,241	3,231,947

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

## Publisher Summary

Publisher Name	Record Count			Media Record Count		
	Digest	API	Index	Digest	API	Index
<a href="#">Berkeley Natural History Museums IPT</a>	1,860,584	1,859,985	1,859,985	0	0	0
<a href="#">Florida Museum of Natural History IPT Service</a>	1,047,587	1,047,587	1,047,564	0	0	0
<a href="#">Northern Great Plains Herbaria Darwin Core Archive rss feed</a>	43,012	43,012	43,012	0	0	0
<a href="#">MyCoPortal Darwin Core Archive rss feed</a>	1,679,459	1,679,458	1,679,458	371,346	371,346	371,346
<a href="#">KU Biodiversity Institute IPT</a>	2,010,071	2,011,170	2,011,170	0	0	0
<a href="#">The University of Connecticut Biological Collections</a>	172,098	171,936	171,198	166,689	166,519	166,022
<a href="#">xBioD IPT in the Museum of Biological Diversity at the Ohio State University</a>	521,710	521,782	521,782	2,593	2,593	2,593
<a href="#">CMC_specify</a>	9,131	9,131	9,131	0	0	0
<a href="#">Consortium of North American Bryophyte Herbaria Darwin Core Archive rss feed</a>	1,690,014	1,690,014	1,690,014	816,932	816,932	816,932
<a href="#">Museum of Comparative Zoology, Harvard University</a>	1,736,357	1,736,471	1,736,471	0	0	0
<a href="#">CNALH Darwin Core Archive rss feed</a>	1,232,891	1,232,891	1,232,891	649,241	649,241	649,241
<a href="#">SCAN Darwin Core Archive rss feed</a>	873,024	873,160	873,160	68,696	68,718	68,718
<a href="#">iDigBio Feeder RSS Feed</a>	1,316,574	1,316,574	1,316,574	19,024	19,024	19,024



# Recordset page with provider info, record counts, links to search and raw data

## Recordset: CUMV Reptile Collection (Arctos)



The CUMV Amphibian & Reptile Collection became one of the leading university based herp collections in North America during the first half of this century, largely because of the efforts of Professor Albert Hazen Wright and his wife, Anna Allen Wright. The major strengths of the collection, amphibians from the southeastern United States and both reptiles and amphibians from the Northeast, reflects the intensive collection by the Wrights. Much of the material collected by the Wrights in New York and Georgia is not duplicated elsewhere. The last 15 years have been seen important acquisitions for the collection. To complement our traditional strength in North American taxa, we have made a concerted effort to obtain foreign material, especially synoptic series representing geographic areas. Through collecting, exchanges and acquisition of other various collections we now have good representation of Costa Rican viperids, lizards from Western and South Australia, amphibians and reptiles from Puerto Rico, snakes and lizards from Mexico, and a more representative collection of African and European species.

Last Update: 2014-07-14

Total Specimen Records: 13,037

Total Media Records: 4

[Search This Recordset](#)

[View Raw Data](#)

### Contacts

John Friel  
Curator  
[john.friel@cornell.edu](mailto:john.friel@cornell.edu)

### Specimen Fields Used for Search

Total Records: 13,037

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.

Field	Records With This Field	(%) Percent Used
Kingdom (dwc:kingdom)	11,403	87.466
Phylum (dwc:phylum)	11,431	87.681
Class (dwc:class)	11,470	87.98
Order (dwc:order)	11,409	87.512
Family (dwc:family)	11,472	87.996
Scientific Name (dwc:scientificName)	13,037	100
Genus (dwc:genus)	11,612	89.07
Specific Epithet (dwc:specificEpithet)	11,267	86.423
Infraspecific Epithet (dwc:infraspecificEpithet)	3,160	24.239
Higher Taxon (dwc:higherClassification)	11,569	88.74
Common Name (dwc:vernacularName)	0	0
Lat (dwc:decimalLatitude)	5,936	45.532
Lon (dwc:decimalLongitude)	5,936	45.532
Country (dwc:country)	11,703	89.768

Over 300 providers, 24M specimen records, 3.4M media records

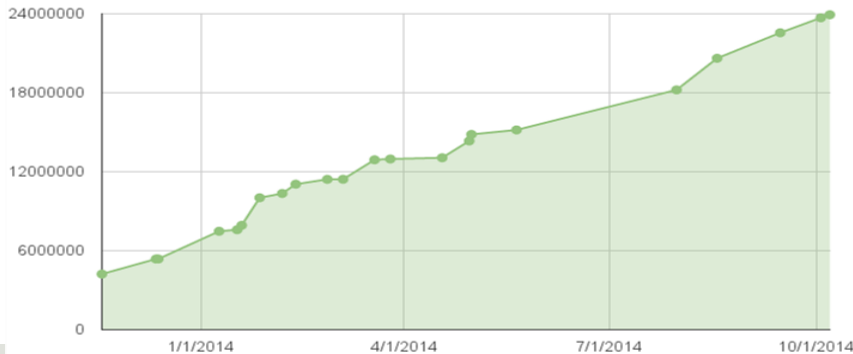
Publishing technologies: IPT, Symbiota, RSS (DwC-a, CSV)

Media data using Audubon core terms

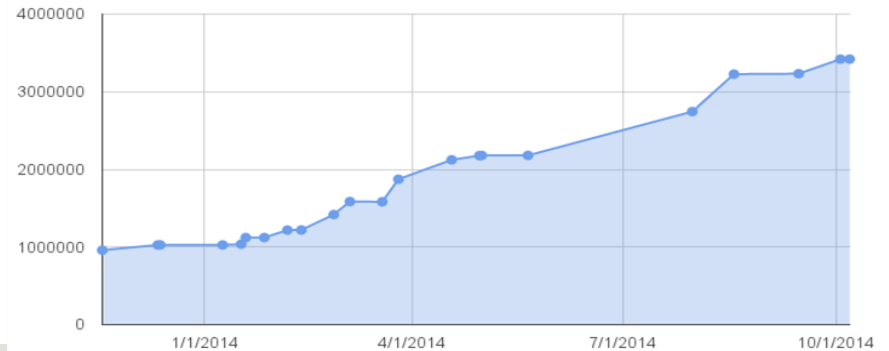


... and many more.

iDigBio Data Ingestion - Specimen Records

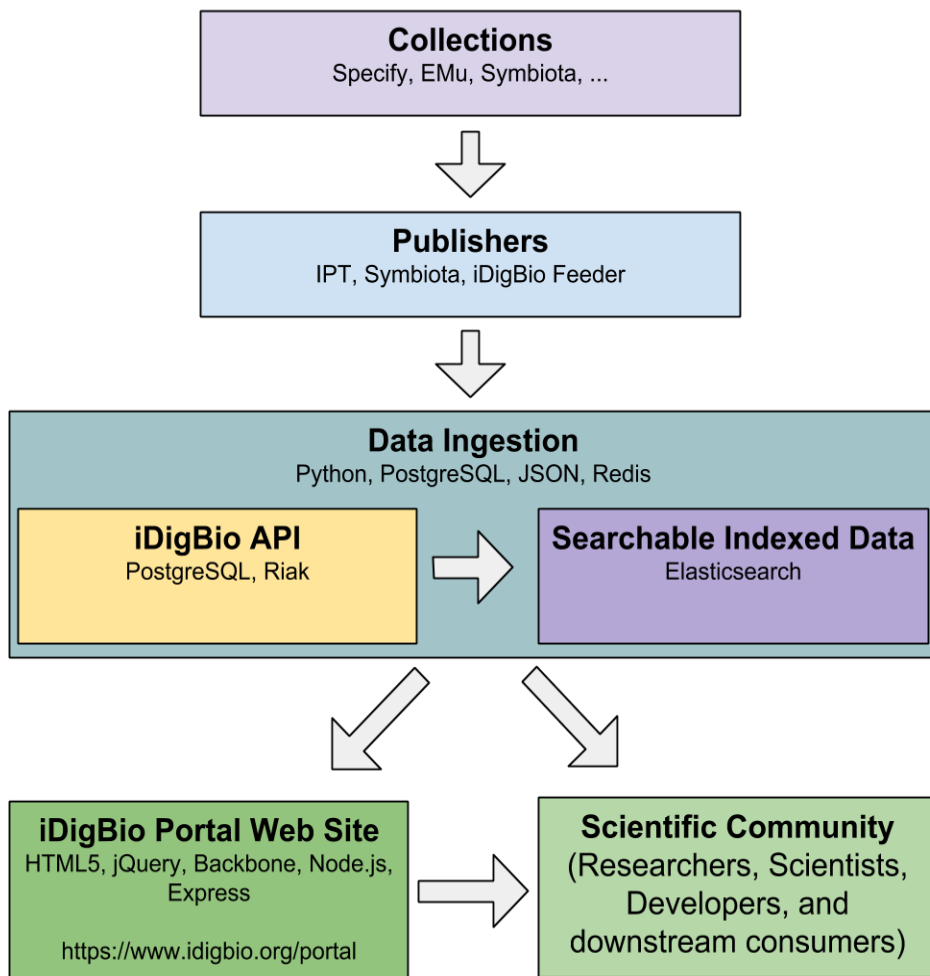


iDigBio Data Ingestion - Media Records



# The what and how of data ingestion

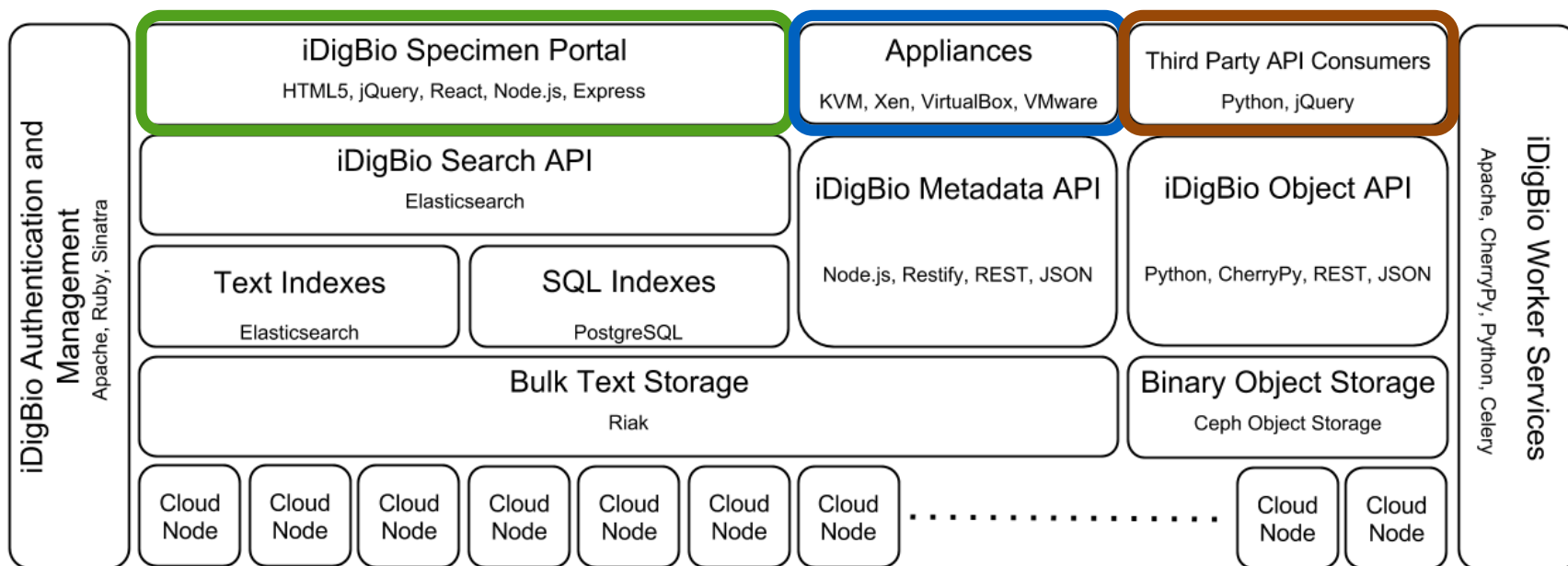
iDigBio Data Flow Diagram



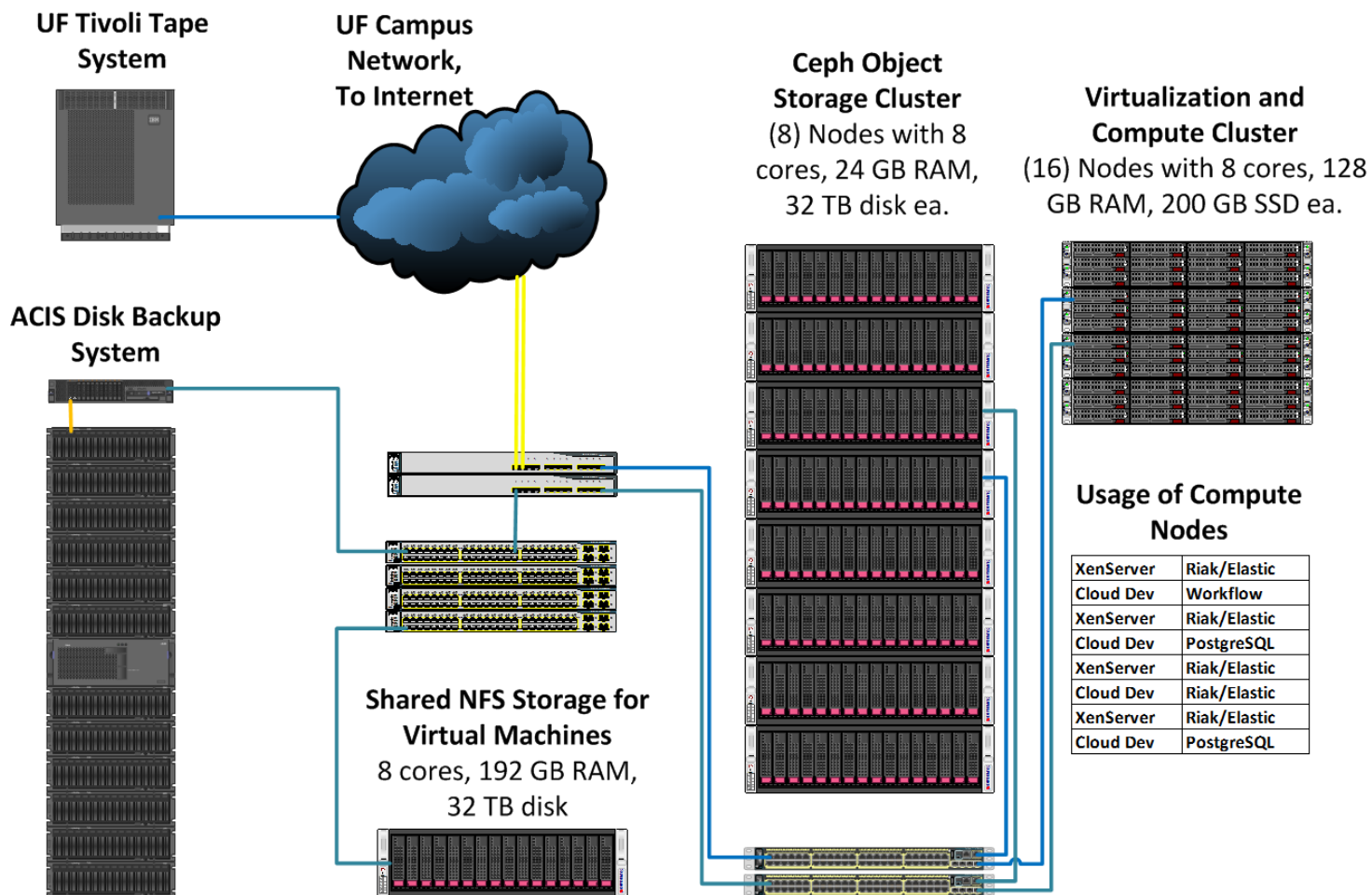
- IPT – RSS of DwC-A
  - Specify, EMu, Arctos, VertNet Migrator, etc.
- Symbiota portals – RSS of DwC-A
- iDigBio Feeder – DwC-A, CSV, ...

**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 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)

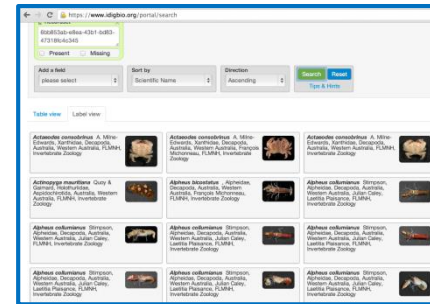


# Appliances, e.g. upload of images and Specify package

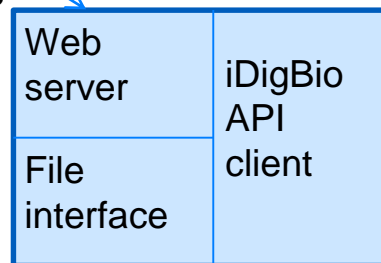
- Reliable approach to upload batches of images with metadata
- Upload starts with CSV file with image paths, identifier, and metadata
- Successfully helped users to upload 290,000+ images.



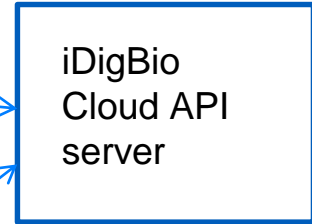
Web UI  
browser



HTTP



HTTP



Appliance

Images stored  
on local media



## Upload Images with CSV File

[Generate CSV](#)[Upload Via CSV](#)[Upload History](#)

Image License

CC0 (Public Domain) ▼

CSV File Full Path

C:\Users\winday0215\Desktop\idigbio\images\media\_records.

 Upload CSV file format

Progress: (Successful:350, Skipped: 810, Failed: 0, Total to upload: 2821. )



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

# CSV File Generation

Generate CSV   Upload Via CSV   Upload History

Upload Path \*

Also Search Files in the Sub-directories.

GUID Syntax \*

CSV Save Path

Note: Fields with \* are mandatory.

- 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

[Generate CSV](#)
[Upload Via CSV](#)
[Upload History](#)

## Batch Information Table

(Click on each row to see the details)

 Show  entries

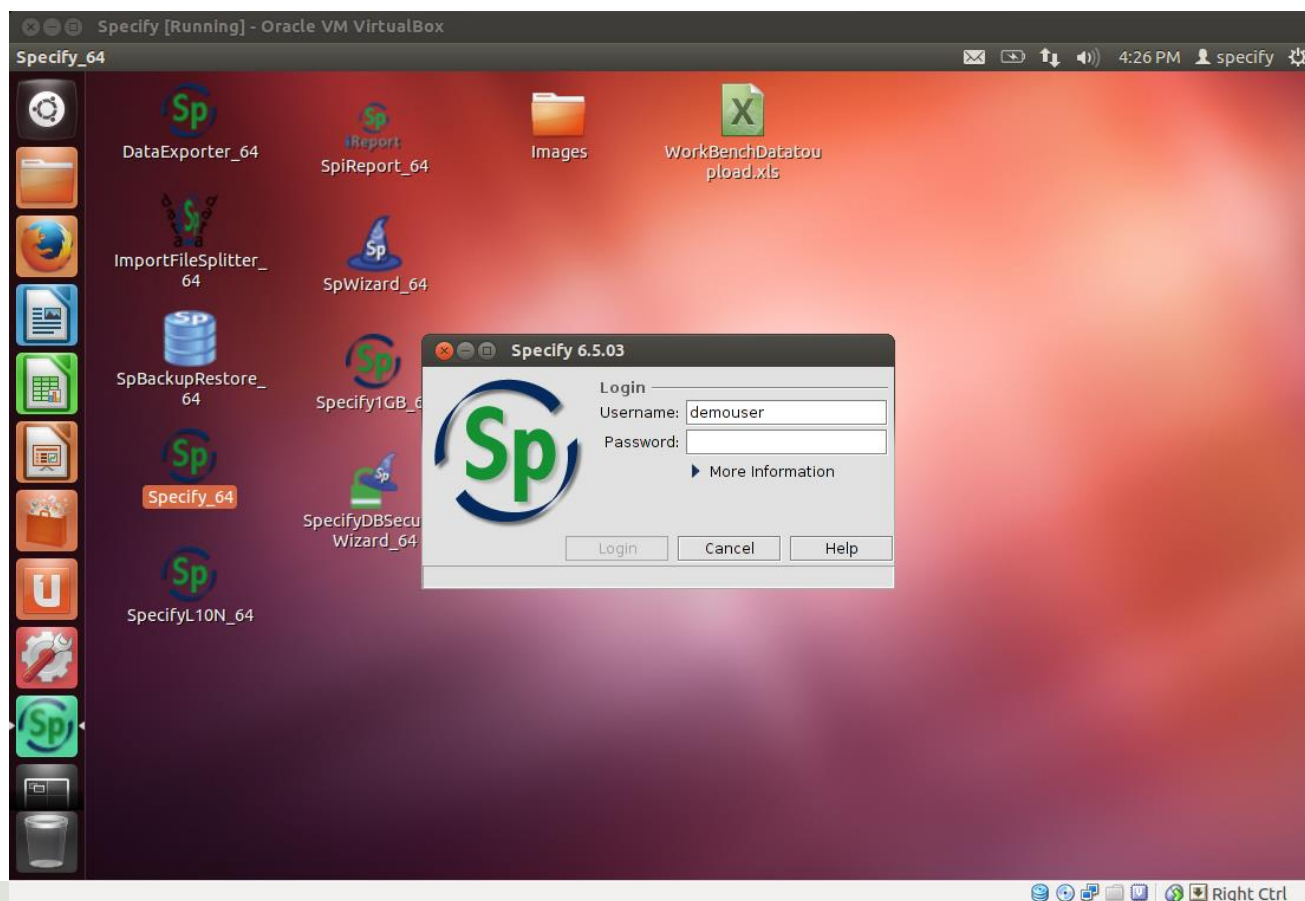
← Previous    1    Next →

ID	CSV File Path	Start Time	Total Records	Failed Records	Skipped Records
1	C:\Users\winday0215\Desktop\idigbio\images2\media_records.csv	2014-08-17 22:15:34	336	0	0
2	C:\Users\winday0215\Desktop\idigbio\images2\media_records.csv	2014-08-17 22:15:53	336	0	0
3	C:\Users\winday0215\Desktop\idigbio\images2\media_records.csv	2014-08-17 22:16:21	336	0	220
4	C:\Users\winday0215\Desktop\idigbio\idigbio-ingestion-tool-	2014-09-07	3	0	0

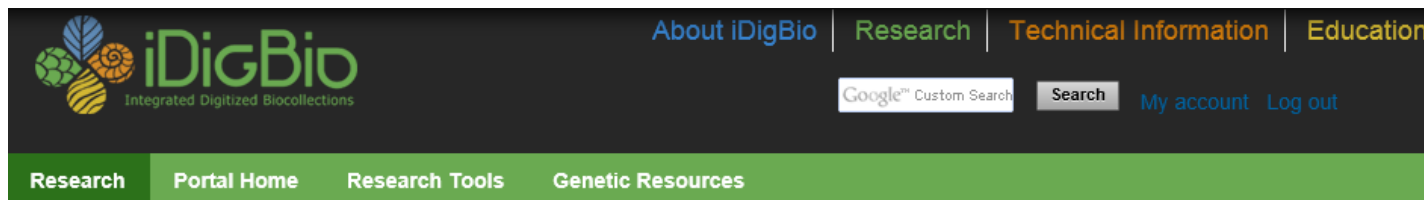
- The local upload history can be viewed/saved to CSV files
- Current upload results are also shown after each upload

# 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



## Research

### Researchers

Browse our specimen portal



### Collections Staff

Learn how your collection can benefit from our work



### Teachers & Students

Learning resources & opportunities to engage



### Looking for research ideas?

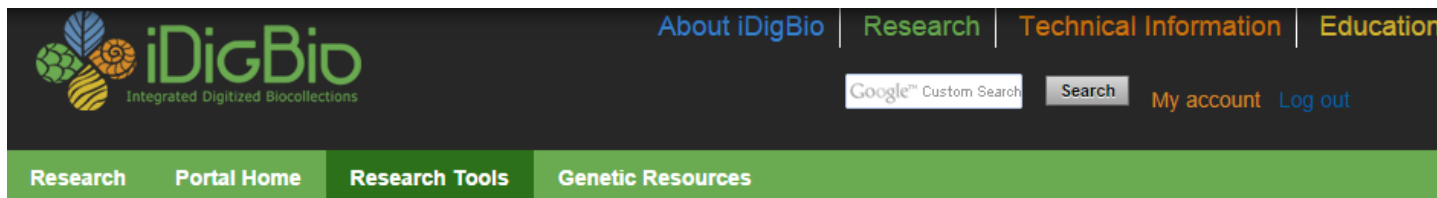
Read about the research questions proposed by each TCN:

2011	<ul style="list-style-type: none"> <li>InvertNet</li> <li>Tri-trophic</li> <li>Lichens &amp; Bryophytes (LBCC)</li> </ul>
2012	<ul style="list-style-type: none"> <li>New England Vascular Plants (NEVP)</li> <li>PaleoNICHES</li> <li>Macrofungi Collection Consortium (MaCC)</li> <li>Southwest Collections of Arthropods Network (SCAN)</li> </ul>
2013	<ul style="list-style-type: none"> <li>Fossil Insect Collaborative (FIC)</li> <li>Vouchered Animal Communication Signals (VACS)</li> <li>Macroalgal Herbarium Consortium (MHC)</li> </ul>
2014	<ul style="list-style-type: none"> <li>Great Lakes Invasives</li> <li>InvertEBase</li> <li>SouthEast Regional Network of Expertise and Collections (SERNEC)</li> </ul>

Links to TCN research  
List of iDigBio publications

- Expanding: <https://www.idigbio.org/research>

# iDigBio Research Tools



## 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.

### Researchers

Browse our specimen portal



### Collections Staff

Learn how your collection can benefit from our work

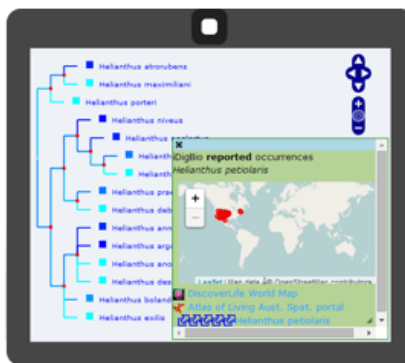


### Teachers & Students

Learning resources & opportunities to engage



## List of Tools Integrating iDigBio Web Services



Solutions to fundamental questions about biodiversity require a new approach that integrates across phylogeny, biogeography, geology, and paleobiology. **PhyloJIVE**, developed by Garry Jolley-Rogers, Joe Miller, and Temi Varghese, integrates biodiversity data with phylogeny. Through **PhyloJIVE**, occurrence records can be viewed in a phylogenetic context, and user-supplied character data can be visualized on the phylogeny. Exploration of the linkages between phylogeny, distributions, and character states can lead to new

- <https://www.idigbio.org/content/community-research-tools>
- Welcome your contributions!



## Research tools integrated with iDigBio

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



- Others – contact Andrea Matsunaga ([ammatsun@ufl.edu](mailto:ammatsun@ufl.edu)) if you are interested in working jointly to integrate of your research tool(s)
- See Demos and attend Discussion Sessions related to research applications, tools and APIs

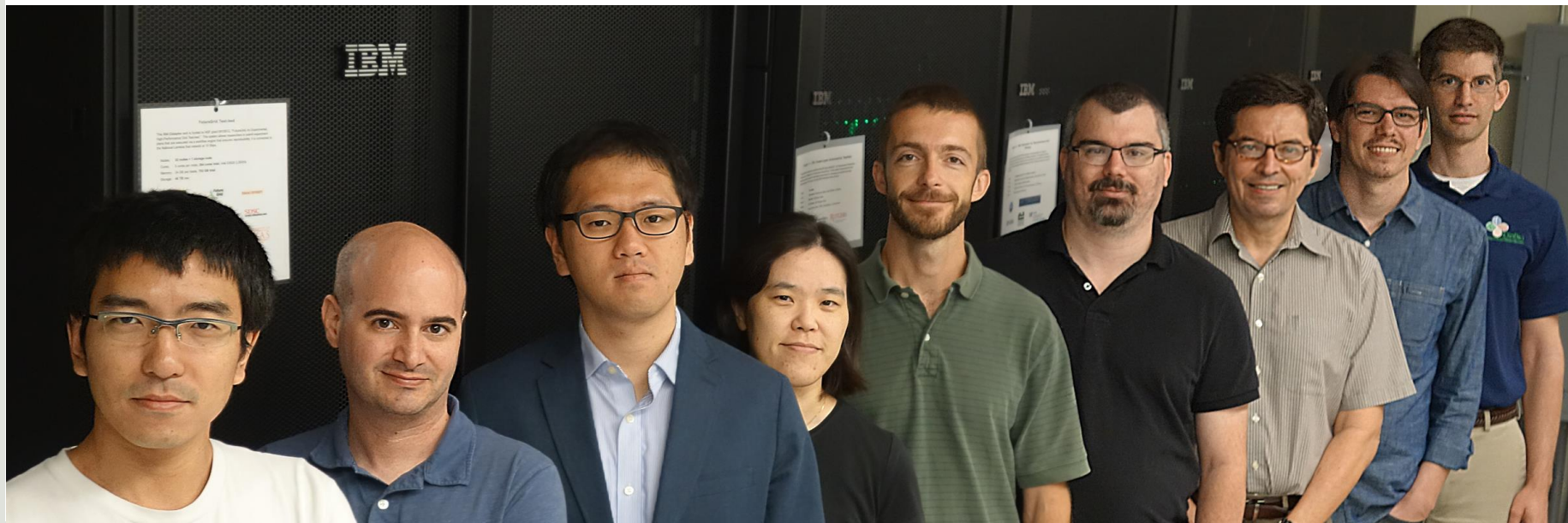
# Conclusions

- Stable cyberinfrastructure and data ingestion
- Growing number of data/media sets
- Emerging capabilities for collections-based research leveraging 3<sup>rd</sup> party community tools
- Linking and integrating with other cyberinfrastructure (GBIF, BISON, EOL,...)
- It is an ADBC village effort!!
- Much more left to do
  - Parallelize more parts of Ingestion process (such as media processing), Support for additional publisher types (beyond IPT, Symbiota, iDigBio RSS Feeder), Improved ingestion logging and error detection, Support for additional media types (audio, 3D scans, ...), Data quality, Provider attribution, Feedback, Links to publications, APIs to broaden support for research tools, ....



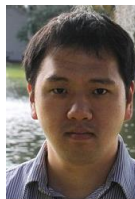
# Acknowledgements

- 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





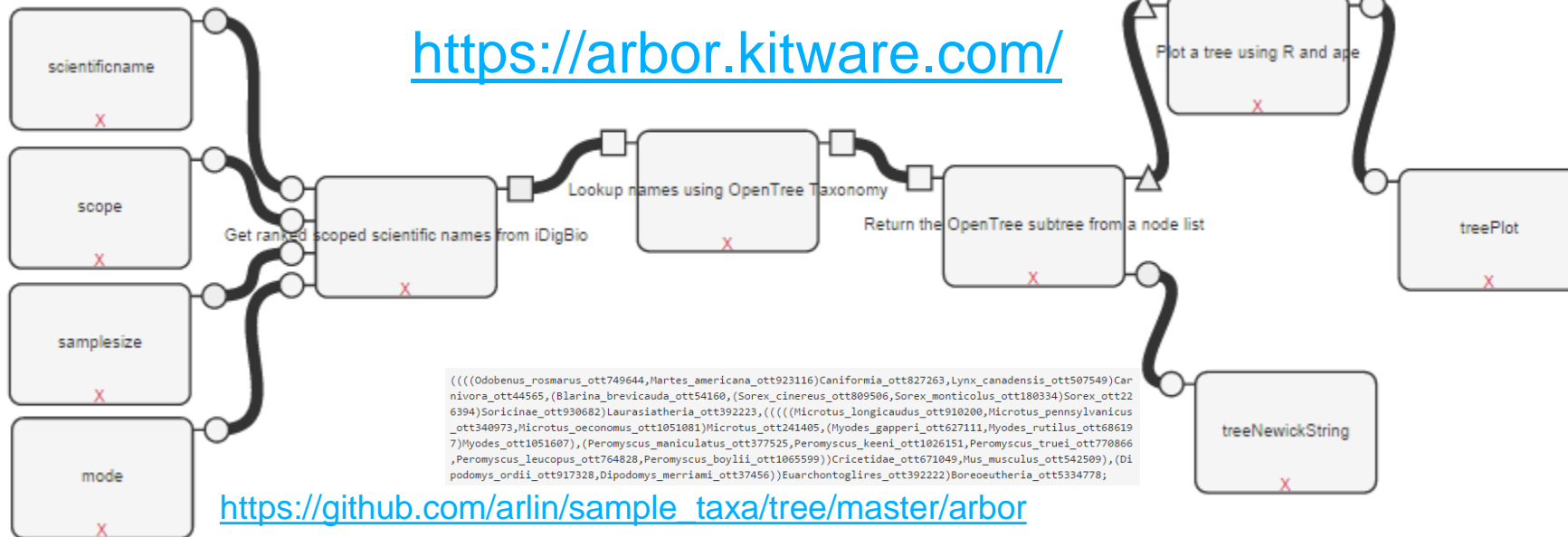
# Acknowledgements





# Arbor, OpenTree, and iDigBio

<https://arbor.kitware.com/>



[https://github.com/arlin/sample\\_taxa/tree/master/arbor](https://github.com/arlin/sample_taxa/tree/master/arbor)

Workflow to get an induced tree from a configurable iDigBio query

scientificname  
Mammalia

scope  
\_all

samplesize  
20

mode  
top

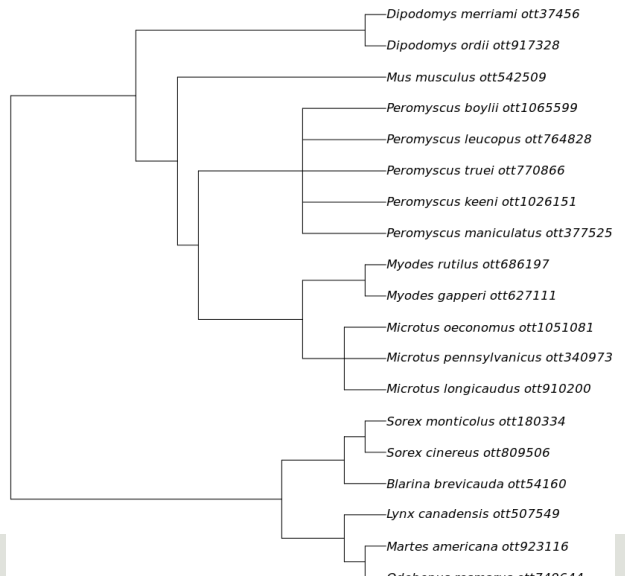
Run Close

```

1 {
2   "query": {
3     "query_string" : {
4       "default_field" : "order",
5       "query" : "rodentia"
6     }
7   },
8   "aggregations": {
9     "my_agg": {
10      "terms": {
11        "field": "scientificname",
12        "size": 100
13      }
14    }
15  }
16 }
  
```

Success! Produced the following outputs:

- Workflow to get an induced tree from a configurable iDigBio query treeNewickString [string]
- Workflow to get an induced tree from a configurable iDigBio query treePlot [image]



# PhyloJIVE, OpenTree, and iDigBio

OpenTree ▾ Sample Trees

OpenTree Studies

iDigBio Frequency

Query

```

1 {
2   "query": {
3     "query_string": {
4       "query": "mammalia"
5     }
6   },
7   "size": 0,
8   "aggregations": {
9     "my_agg": {
10      "terms": {
11        "field": "scientificname",
12        "size": 100
13      }
14    }
15  },
16  "filter": {
17    "exists": { "field": "geopoint" }
18  }
19 }

```

Response

```

"aggregations": {
  "my_agg": {
    "buckets": [
      {
        "key": "peromyscus maniculatus",
        "doc_count": 36597
      },
      {
        "key": "peromyscus leucopus",
        "doc_count": 20699
      },
      {
        "key": "myodes rutilus",
        "doc_count": 19017
      }
    ]
  }
}

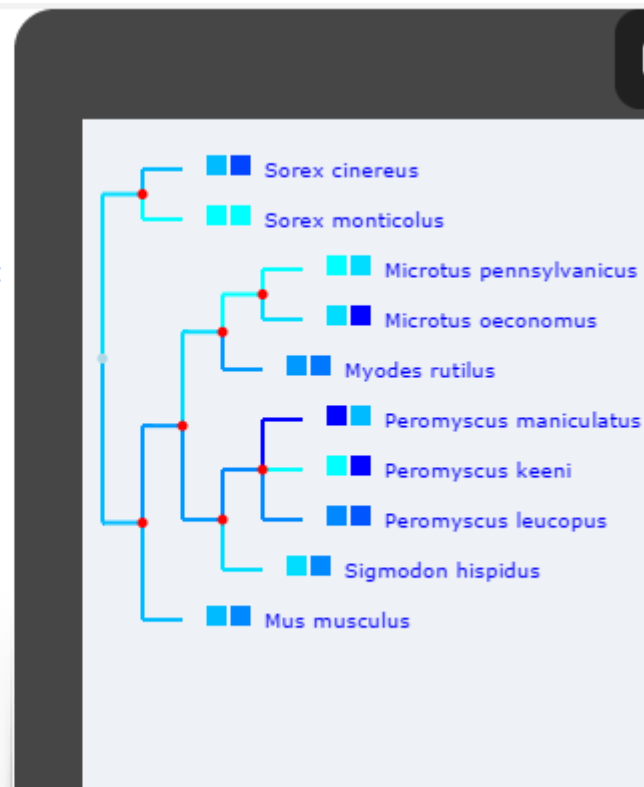
```

Criteria:  
top\_10\_Mammalia

Select another criteria:

top\_10\_Mammalia ▾

- top\_10\_Diptera
- top\_50\_Diptera
- top\_10\_Felidae
- top\_50\_Felidae
- top\_10\_Insecta
- top\_50\_Insecta
- top\_10\_Mammalia
- top\_50\_Mammalia
- top\_10\_Nematoda
- top\_50\_Nematoda
- top\_10\_Rodentia
- top\_50\_Rodentia
- top\_10\_Rupchand
- rand\_50\_Diptera
- rand\_50\_Felidae
- rand\_50\_Insecta
- rand\_50\_Mammalia
- rand\_50\_Nematoda
- rand\_50\_Rupchand
- rand\_50\_Rodentia



- Searching for any occurrence of the word “Mammalia” with lat/long
- Selects the top or random scientific names occurring in iDigBio
- Keep only binomials, ignore names with certain special characters
- [http://search.idigbio.org/idigbio/records/\\_search](http://search.idigbio.org/idigbio/records/_search)

# OpenRefine, OpenTree and iDigBio

Google refine *A power tool for working with messy data.*

Create Project  
Open Project  
Import Project

Create a project by importing data. What kinds of data files can I import?  
TSV, CSV, \*SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with Google Refine extensions.

Get data from

This Computer

Web Addresses (URLs)

Clipboard

Google Data

Enter one or more web addresses (URLs) pointing to data to download:

`http://search.idigbio.org/idigbio/records/_search?source={%22size%22:0,%22query%22:{%22query_string%22:{%22query%22:mammalia%22}},%22aggregations%22:{%22my_agg%22:{%22terms%22:{%22field%22:scientificname%22,%22size%22:10}}}}`

Add Another URL Next »

1. Create a new project from "Web addresses"

[http://search.idigbio.org/idigbio/records/\\_search?source={%22size%22:0,%22query%22:{%22query\\_string%22:{%22query%22:mammalia%22}},%22aggregations%22:{%22my\\_agg%22:{%22terms%22:{%22field%22:scientificname%22,%22size%22:10}}}}](http://search.idigbio.org/idigbio/records/_search?source={%22size%22:0,%22query%22:{%22query_string%22:{%22query%22:mammalia%22}},%22aggregations%22:{%22my_agg%22:{%22terms%22:{%22field%22:scientificname%22,%22size%22:10}}}})

2. Parse as JSON and select the my\_agg node

# OpenRefine, OpenTree and iDigBio

Google refine *A power tool for working with messy data.*

Create Project   « Start Over   Configure Parsing Options

Open Project

Import Project

	my_agg - buckets - __anonymous__ - key	my_agg - buckets - __anonymous__ - doc_count
1.	peromyscus maniculatus	36597
2.	peromyscus leucopus	20699
3.	myodes rutilus	19017
4.	sorex cinereus	16331
5.	mus musculus	14354
6.	peromyscus maniculatus sonoriensis	13185
7.	mammalia	12675
8.	microtus oeconomus	12230
9.	sigmodon hispidus	12047
10.	peromyscus leucopus noveboracensis	11909
11.	peromyscus maniculatus nebrascensis	10753
12.	dipodomys merriami merriami	10192
13.	peromyscus maniculatus rufinus	10185
14.	peromyscus maniculatus gambelii	10130
15.	microtus pennsylvanicus	9855
16.	peromyscus keeni	8623

Parse data as

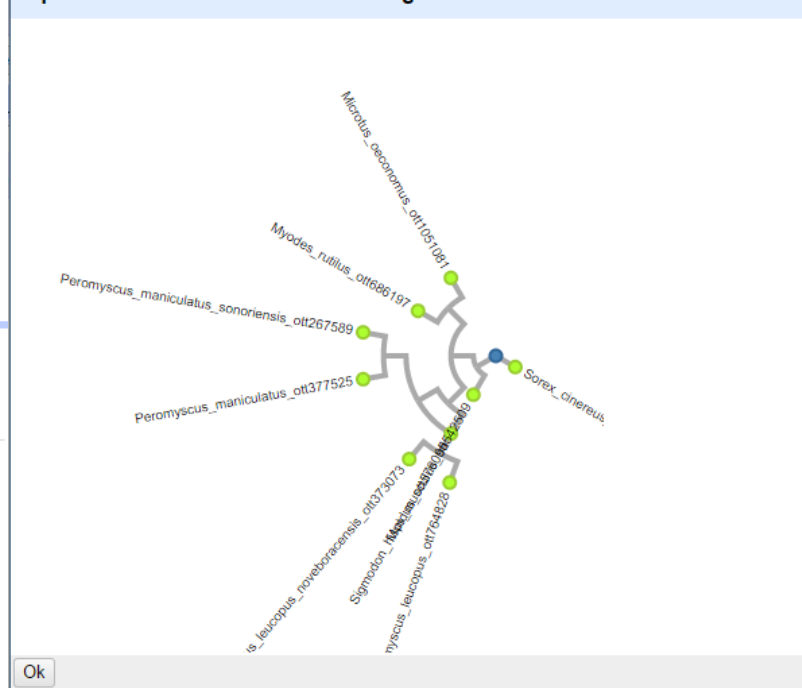
- CSV / TSV / separator-based files
- Line-based text files
- Fixed-width field text files
- PC-Axis text files
- JSON files

Load at most 0

Store file source (file names, URLs) in each row

4. Filter out monomials
5. Generate a tree visualization

Opentree induced sub-tree from Google Refine data



3. Create new column to hold OTT IDs

```
import tnrs
return tnrs.getOttId(value)
```

## Darwin Core Archive / DwC-A

<http://rs.tdwg.org/dwc/terms/guides/text/>

A Darwin Core Archive is a zip file that includes metadata about the dataset, the data itself, and any optional extension data.

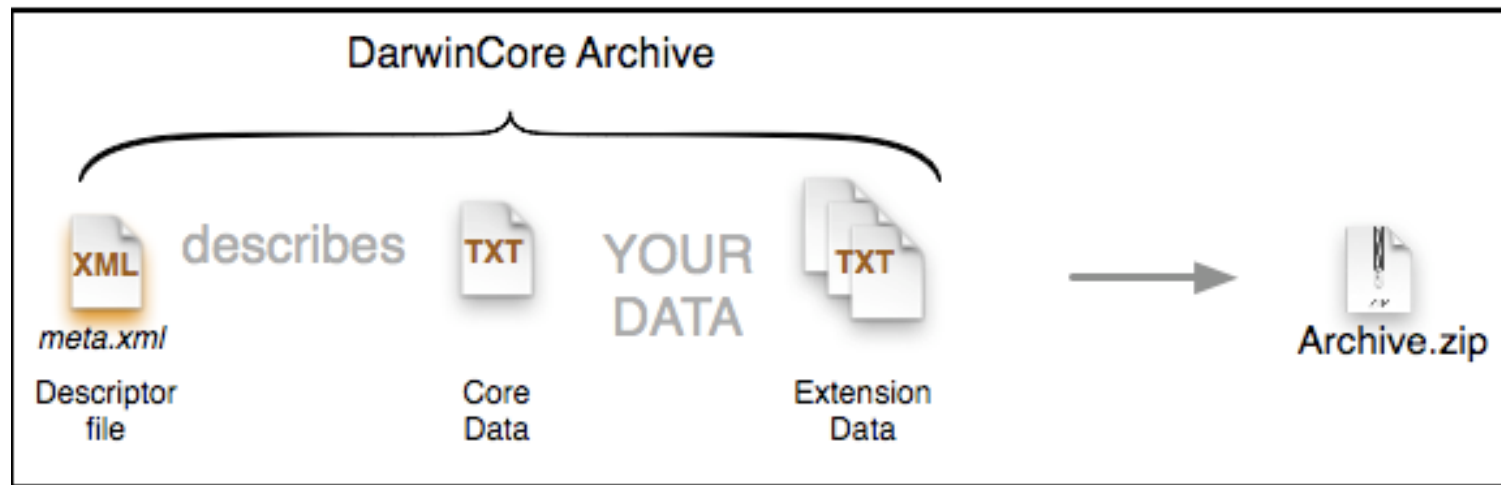


Image source: <http://tools.gbif.org/dwca-assistant/>

# Specimen Data – Darwin Core Standard

<http://rs.tdwg.org/dwc/terms/>

Field	Records With This Field	(%) Percent Used
<b>Institution Code</b> (dwc:institutionCode)	41,262	100
<b>Catalog Number</b> (dwc:catalogNumber)	41,262	100
<b>Collection Code</b> (dwc:collectionCode)	41,262	100
<b>Occurrence ID</b> (dwc:occurrenceID)	41,262	100
<b>Basis of Record</b> (dwc:basisOfRecord)	41,262	100
<b>Kingdom</b> (dwc:kingdom)	41,261	99.998
<b>Phylum</b> (dwc:phylum)	41,261	99.998
<b>Class</b> (dwc:class)	41,261	99.998
<b>Order</b> (dwc:order)	41,261	99.998
<b>Family</b> (dwc:family)	41,261	99.998
<b>Scientific Name</b> (dwc:scientificName)	41,261	99.998
<b>Locality</b> (dwc:locality)	41,248	99.966
<b>Specific Epithet</b> (dwc:specificEpithet)	41,157	99.746
<b>Genus</b> (dwc:genus)	41,124	99.666
<b>Continent</b> (dwc:continent)	40,963	99.275



## Recommended minimum Darwin Core fields for iDigBio Ingestion:

Record ID

Scientific Name

Occurrence ID

Event Date

Collector Name

Locality Data

Barcode, catalog number, accession id or collection number

Paleo specimens should also include geological context

## Media Data – Audubon Core / AC

[http://terms.tdwg.org/wiki/Audubon\\_Core\\_Term\\_List](http://terms.tdwg.org/wiki/Audubon_Core_Term_List)



*Images Source: Arizona State University Lichen Herbarium (Accessed through iDigBio Specimen Data Portal, <https://www.idigbio.org/portal>, 2014-09-18)*

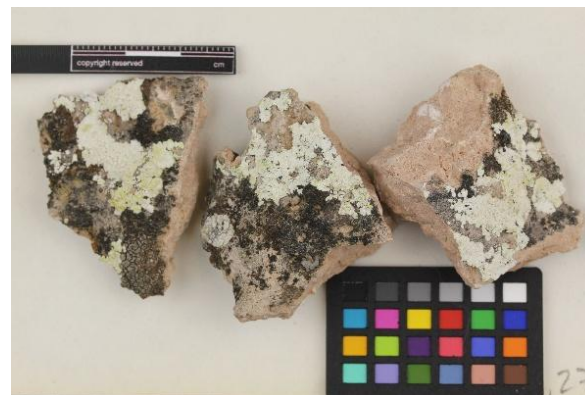
GBIF has a nice write-up on the benefits of AC over dwc:associatedMedia:

<http://gbif.blogspot.com/2014/05/multimedia-in-gbif.html>

## Audubon Core vocabularies

### address such concerns as:

- the management of the media and collections
- descriptions of their content
- their taxonomic, geographic, and temporal coverage
- appropriate ways to retrieve, attribute and reproduce them



### Media Metadata

<b>Associated Specimen Reference</b>	<a href="http://lichenportal.org/portal/collections/individual/index.php?occid=1374628">http://lichenportal.org/portal/collections/individual/index.php?occid=1374628</a>
<b>Type of Resource</b>	StillImage
<b>Subtype</b>	Photograph
<b>Metadata Date</b>	2013-04-24 02:00:19
<b>Provider-managed ID</b>	urn:uuid:9a77ed32-7fa4-4831-938e-a499078058a8
<b>Credit</b>	Arizona State University Lichen Herbarium (ASU)
<b>License Terms</b>	CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
<b>License URL</b>	<a href="http://creativecommons.org/licenses/by-nc-sa/3.0/">http://creativecommons.org/licenses/by-nc-sa/3.0/</a>
<b>Access URI</b>	<a href="http://storage.idigbio.org/asu/lichens/ASU0068/ASU0068021a_lg.jpg">http://storage.idigbio.org/asu/lichens/ASU0068/ASU0068021a_lg.jpg</a>
<b>Format</b>	image/jpeg

## Audubon Core can support “new” media types

### Specimen Data

dwc:catalogNumber: UF 105199  
dwc:scientificName:  
    Carcharocles megalodon  
dwc:stateProvince: Florida  
dwc:county: Duval  
dwc:latestPeriodOrHighestSystem:  
    Late Miocene  
dwc:decimalLatitude: 30.39211

### Media Data

dwc:scientificName:  
    Carcharocles megalodon  
dc:type: image  
ac:subtype:  
    <http://www.fabbers.com/StL.asp>  
ac:subtypeLiteral: 3dModel  
ac:tag: tooth



Image source: Aaron Wood, Florida Museum of Natural History

## Recommended minimum Audubon Core fields for iDigBio Data Ingestion:

Access URI

Rights

Provider

Scientific name

Title

Description

Tags

# Practical Details

## Data Formats

- ISO 8601 Dates

- WGS84 Decimal Lat/Long

## Controlled Vocabularies

- ISO Country Names and Codes

- State/Province names

## Identifier Formats (UUID, ARK, URN, DOI, URI, URL, LSID, ...)

## Copyright and Standard Licenses

## Apple Core guidelines for herbaria

- <http://code.google.com/p/applecore/wiki/Introduction>

# Ingestion Reporting

<https://www.idigbio.org/portal/publishers>

## Publisher Summary

Publisher Name	Record Count			Media Record Count		
	Digest	API	Index	Digest	API	Index
<a href="#">Berkeley Natural History Museums IPT</a>	1,860,584	1,859,985	1,859,985	0	0	0
<a href="#">Florida Museum of Natural History IPT Service</a>	1,047,587	1,047,587	1,047,587	0	0	0
<a href="#">MyCoPortal Darwin Core Archive rss feed</a>	1,679,459	1,679,458	1,679,458	371,346	371,346	371,346
<a href="#">Northern Great Plains Herbaria Darwin Core Archive rss feed</a>	43,012	43,012	43,012	0	0	0
<a href="#">KU Biodiversity Institute IPT</a>	2,010,071	2,011,170	2,011,170	0	0	0
<a href="#">The University of Connecticut Biological Collections</a>	172,098	172,102	172,102	166,689	166,707	166,707
<a href="#">xBioD IPT in the Museum of Biological Diversity at the Ohio State University</a>	521,710	521,782	521,782	2,593	2,593	2,593
<a href="#">CMC_specify</a>	9,131	9,131	9,131	0	0	0
<a href="#">Consortium of North American Bryophyte Herbaria Darwin Core Archive rss feed</a>	1,690,014	1,690,014	1,690,014	816,932	816,932	816,932
<a href="#">Museum of Comparative Zoology, Harvard University</a>	1,736,357	1,736,471	1,736,471	0	0	0
<a href="#">CNALH Darwin Core Archive rss feed</a>	1,232,891	1,232,891	1,232,891	649,241	649,241	649,241
<a href="#">SCAN Darwin Core Archive rss feed</a>	873,024	873,160	873,160	68,696	68,718	68,718
<a href="#">iDigBio Feeder RSS Feed</a>	1,316,574	1,316,574	1,316,574	19,024	19,024	19,024
<a href="#">Consortium of Intermountain Herbaria Darwin Core Archive rss feed</a>	204,129	204,131	204,131	74,014	74,015	74,015
<a href="#">CAS-IPT</a>	1,875,928	1,875,979	1,875,979	0	0	0
<a href="#">Macroalgal Herbarium Portal Darwin Core Archive rss feed</a>	2,145	2,145	2,145	1,937	1,937	1,937
<a href="#">CNH portal Darwin Core Archive rss feed</a>	89,199	89,199	89,199	56,557	56,557	56,557
<a href="#">IPT - Hosted by VertNet</a>	5,070,222	5,070,222	5,070,222	479,440	479,440	479,440
<a href="#">North American Network of Small Herbaria Darwin Core Archive rss feed</a>	4,162	4,162	4,162	4,273	4,273	4,273
<a href="#">Harvard University Herbaria IPT installation</a>	412,331	412,331	412,331	295,055	295,055	295,055
<a href="#">SNOMNH IPT</a>	310,328	310,328	310,328	0	0	0
<a href="#">Morphbank IPT Feed</a>	48,567	97,127	97,127	0	65,167	65,167
<a href="#">SEINet Darwin Core Archive rss feed</a>	347,210	347,216	347,216	161,533	161,627	161,627