

Cyberinfrastructure Status Report (Got data...want juice!)

José Fortes

(with slides provided by Alex Thompson, Andrea Matsunaga, Dan Stoner,
Matthew Collins and Renato Figueiredo)

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

✉ fortes@acis.ufl.edu



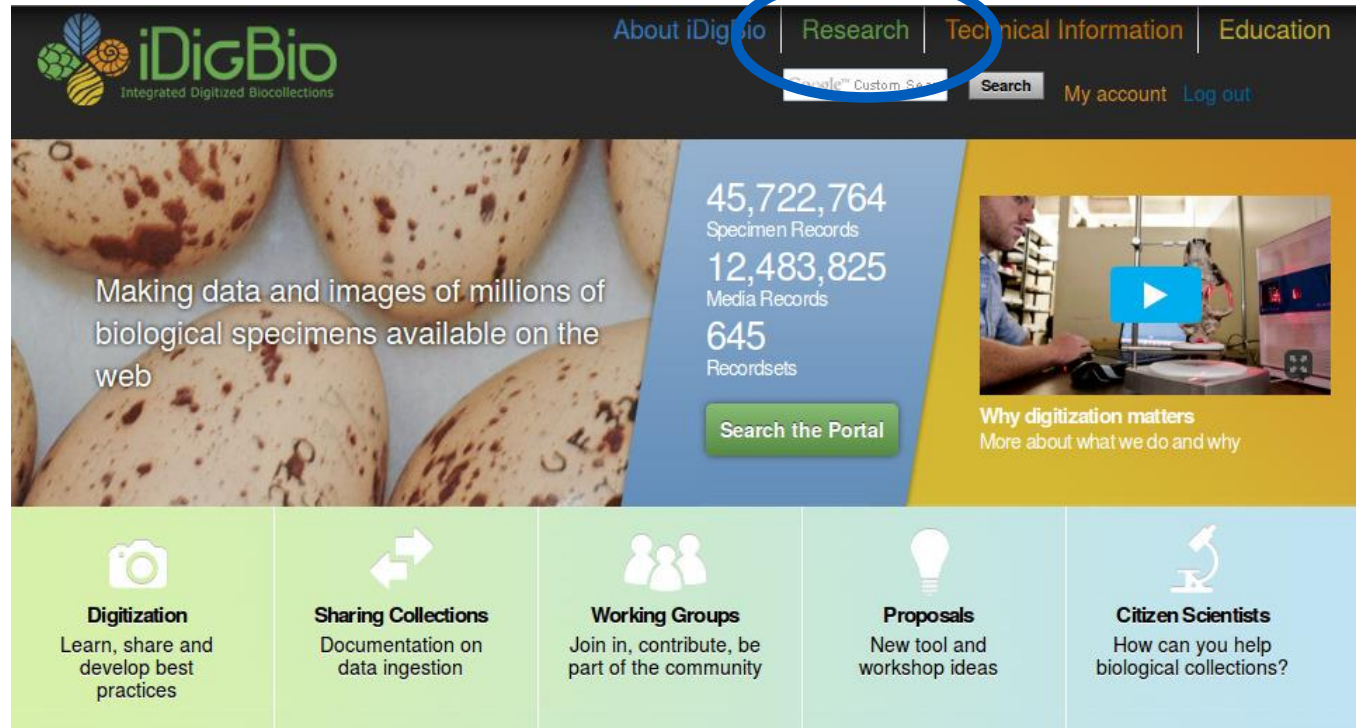
iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect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

Outline

- Cyberinfrastructure
 - Web site
 - Data portal
 - Data
 - Ingestion
 - Use
 - Integration
 - Appliances
 - Research applications
 - Parting messages



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'. The 'Technical Information' link is circled in blue. Below the navigation bar is a search bar with a 'Search' button and links for 'My account' and 'Log out'. The main content area features a large image of speckled bird eggs on the left. To the right of the eggs, there is a statistics box with the following data:

45,722,764	Specimen Records
12,483,825	Media Records
645	Recordsets

Below the statistics is a green button labeled 'Search the Portal'. To the right of the statistics is a video player with a play button and the text 'Why digitization matters' and 'More about what we do and why'. Below the main content area is a row of five green and blue boxes, each with an icon and a title:

- Digitization**: Learn, share and develop best practices
- Sharing Collections**: Documentation on data ingestion
- Working Groups**: Join in, contribute, be part of the community
- Proposals**: New tool and workshop ideas
- Citizen Scientists**: How can you help biological collections?

Researchers
Learn about research directions



Collections Staff
Learn how your collection can benefit from our work



Teachers & Students
Download lesson plans about using digitized specimens



Upcoming Events

Worldwide Engagement for Digitizing Biocollections (WeDigBio) Event
10-22-2015 to 10-25-2015

Improving Data Quality: iDigBio Recordset data cleaning method, tools, and data flags
10-23-2015

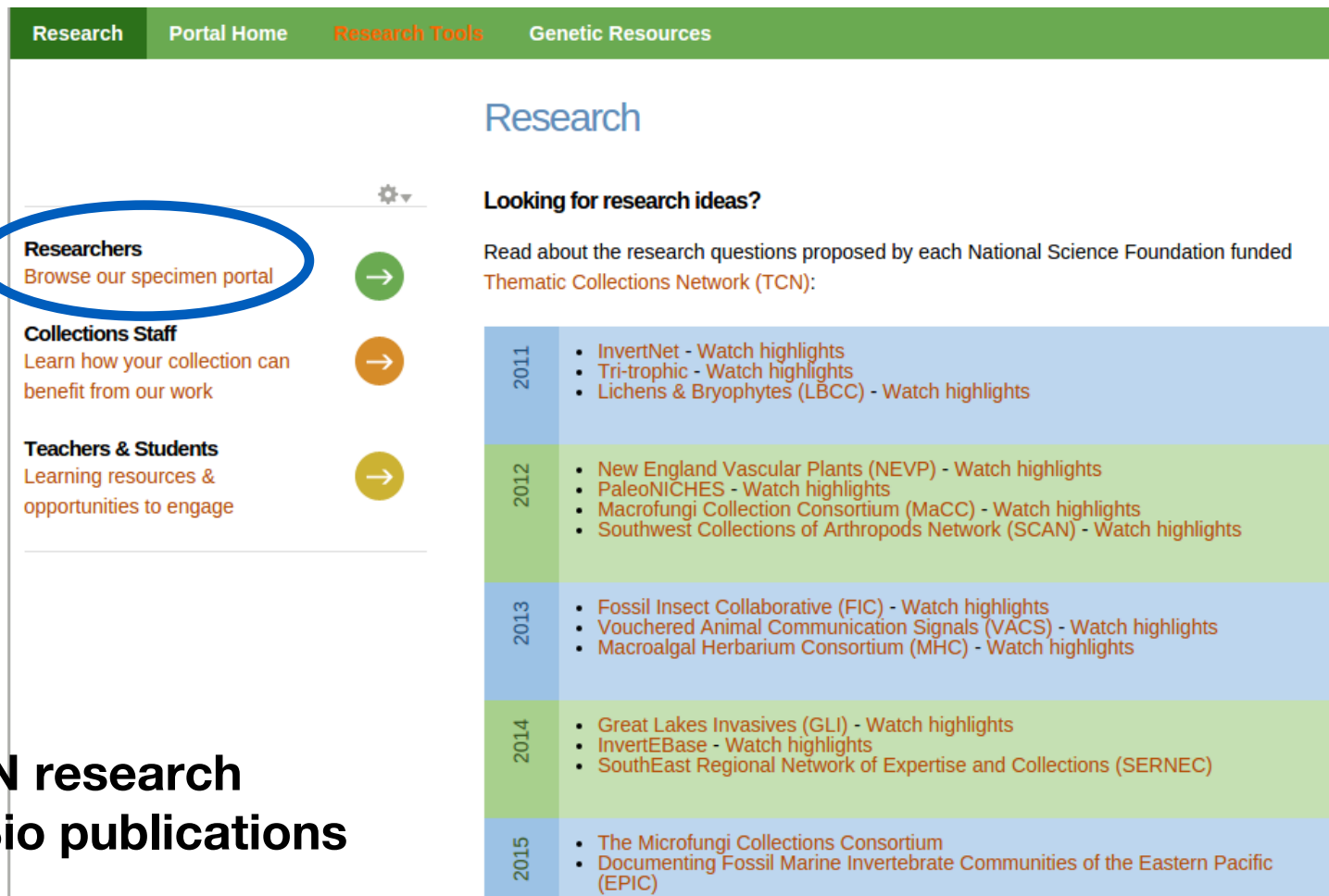
iDigBio Executive Committee Meetings 2015
10-28-2015

iDigBio Steering Committee Meetings 2015
10-28-2015

Undergraduates
Learn About Careers and Graduate Study in Biology at the North Carolina Museum of Natural Sciences



iDigBio Research Section



Research Portal Home Research Tools Genetic Resources

Research

Looking for research ideas?

Read about the research questions proposed by each National Science Foundation funded Thematic Collections Network (TCN):

2011	<ul style="list-style-type: none"> InvertNet - Watch highlights Tri-trophic - Watch highlights Lichens & Bryophytes (LBCC) - Watch highlights
2012	<ul style="list-style-type: none"> New England Vascular Plants (NEVP) - Watch highlights PaleoNICHES - Watch highlights Macrofungi Collection Consortium (MaCC) - Watch highlights Southwest Collections of Arthropods Network (SCAN) - Watch highlights
2013	<ul style="list-style-type: none"> Fossil Insect Collaborative (FIC) - Watch highlights Vouchered Animal Communication Signals (VACS) - Watch highlights Macroalgal Herbarium Consortium (MHC) - Watch highlights
2014	<ul style="list-style-type: none"> Great Lakes Invasives (GLI) - Watch highlights InvertEBase - Watch highlights SouthEast Regional Network of Expertise and Collections (SERNEC)
2015	<ul style="list-style-type: none"> The Microfungi Collections Consortium Documenting Fossil Marine Invertebrate Communities of the Eastern Pacific (EPIC)

Links to TCN research
List of iDigBio publications

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

Search across all data, all/individual fields, customize, use autocomplete, synonyms,...

iDigBio Home
Portal Home
Search Records
Tutorial
Our Data
Research Tools
Feedback

Search Records Help Reset

Must have image Must have map point

Filters Mapping Sorting Download

Clear

Kingdom ✕

 Present Missing

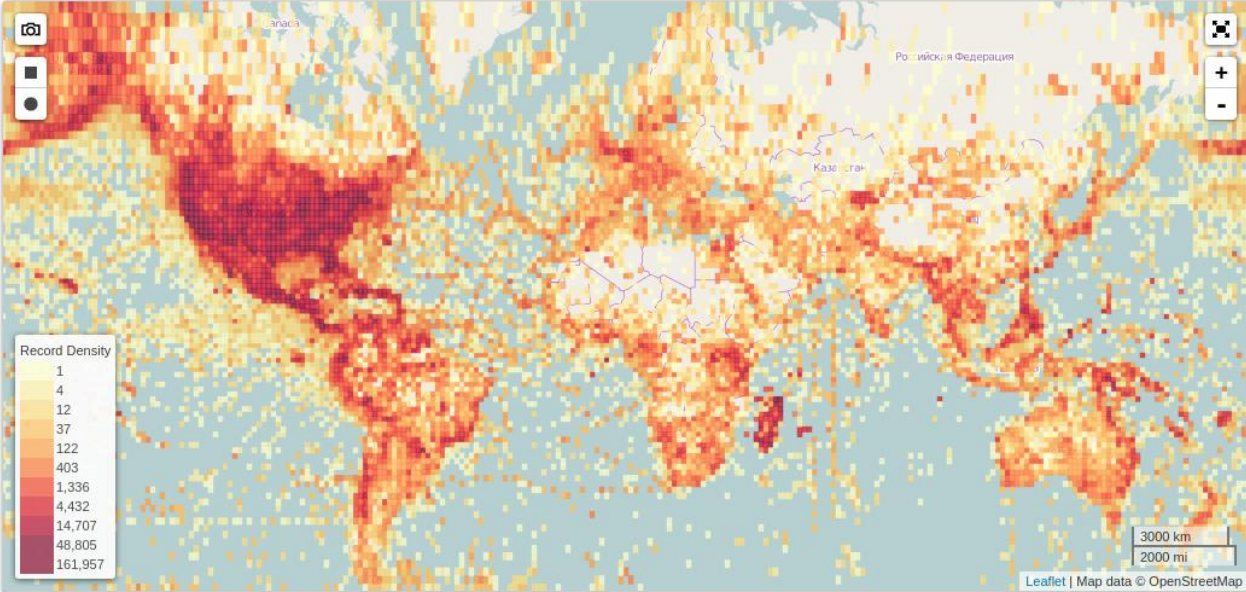
Scientific Name ✕

 Present Missing Add EOL Synonyms

Date Collected Start: End: ✕

 Present Missing

Scroll To Bottom



Total: 20,495,711






List
Labels
Images
Recordsets

Family	Scientific Name	Date Collected	Country	Institution Code	Basis of Record	Columns
<i>no data</i>	"Ambocoelia" sp.	<i>no data</i>	United States	MCZ	PreservedSpecimen	view
<i>no data</i>	"Ambocoelia" sp.	<i>no data</i>	United States	MCZ	PreservedSpecimen	view
Apterotonidae	"Apterotonus" apurensis	1981-01-08	Brasil	Instituto Nacional de Pesquisas da...	PreservedSpecimen	view
Apterotonidae	"Apterotonus" apurensis	1986-02-28	Brasil	Instituto Nacional de Pesquisas da...	PreservedSpecimen	view

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

List	Labels	Images	Recordsets	Total: 280,461		
Family	Scientific Name	Date Collected	Country	Institution Code	Basis of Record	Columns
Carabidae	<i>Abacetus rufitarsis</i>	1934-07-30	India	MCZ	PreservedSpecimen	view
Carabidae	<i>Abacetus straneoii</i>	1944-03-01/1944-07-31	Papua New Guinea	MCZ	PreservedSpecimen	view
Carabidae	<i>Abacetus permundus</i>	1878-05-01/1878-05-31	United States	MCZ	PreservedSpecimen	view
Abacionidae	<i>Abacion</i>	2014-09-01	United States	MCZ	PreservedSpecimen	view

List	Labels	Images	Recordsets	Total: 280,461		
<p>Abacetus rufitarsis Straneo</p> <p>India, Puducherry Territory, Karaikal District, Nedungadu MCZ, Ent, 29281, P. S. Nathan</p> <p><i>Abacetus rufitarsis</i> 3</p> <p><i>Animalia, Arthropoda, Insecta, Coleoptera</i></p> <p>1934-07-30</p>	<p>Abacetus straneoii Darlington (1962)</p> <p>Papua New Guinea, Oro Province, Dobodura MCZ, Ent, 30218, Philip Jackson Darlington, Jr.</p> <p><i>Abacetus straneoii</i> 3</p> <p><i>Animalia, Arthropoda, Insecta, Coleoptera</i></p> <p>1944-03-01/1944-07-31</p>	<p>Abacetus permundus</p> <p>United States, Illinois, Richland & Lawrence Co., Wabash Valley MCZ, Ent, 33042, [no agent data]</p> <p><i>Abacetus permundus</i> 3</p> <p><i>Animalia, Arthropoda, Insecta, Coleoptera</i></p> <p>1878-05-01/1878-05-31</p>				
<p>Abacion (Loomis, 1937)</p> <p>United States, Tennessee, Great Smoky Mountains National Park, Greenbrier picnic area Lat: 35°42' 46" Lon: -83°23' 3" MCZ, IZ, 46877, Gonzalo Giribet, Rosa Fernández García</p> <p><i>Abacion</i> 7</p> <p><i>Animalia, Arthropoda, Diplopoda, Callipodida</i></p> <p>2014-09-01</p>	<p>Abacetus nicippe (Cramer, 1779)</p> <p>Mexico, Tamaulipas, Victoria MCZ, Ent, 27382, Turner, Donald B. Stallings, Sr.</p> <p><i>Abacetus nicippe</i> 2</p> <p><i>Animalia, Arthropoda, Insecta, Lepidoptera</i></p> <p>1941-06-10</p>	<p>Abacetus nicippe (Cramer, 1779)</p> <p>USA, Arizona, Maricopa County, Phoenix Lat: 33°26' 53" Lon: -112°4' 23" ASU, ASUHC, ASUHC0079342, Roman S. Wielgus</p> <p><i>Abacetus nicippe</i> 2</p> <p><i>Animalia, Arthropoda, Insecta, Lepidoptera</i></p> <p>1963-09-01</p>				

List	Labels	Images	Recordsets	Total: 280,461				
				 <p>1 of 3</p> <p><i>Abacetus rufitarsis</i> MCZ, Ent</p>	 <p>2 of 3</p> <p><i>Abacetus rufitarsis</i> MCZ, Ent</p>	 <p>3 of 3</p> <p><i>Abacetus rufitarsis</i> MCZ, Ent</p>	 <p>1 of 3</p> <p><i>Abacetus straneoii</i> MCZ, Ent</p>	 <p>2 of 3</p> <p><i>Abacetus straneoii</i> MCZ, Ent</p>

Results mapped/rendered and downloadable

Search Records [Help](#) [Reset](#)

search all fields

Must have image
 Must have map point

Filters Mapping Sorting Download

Current Search

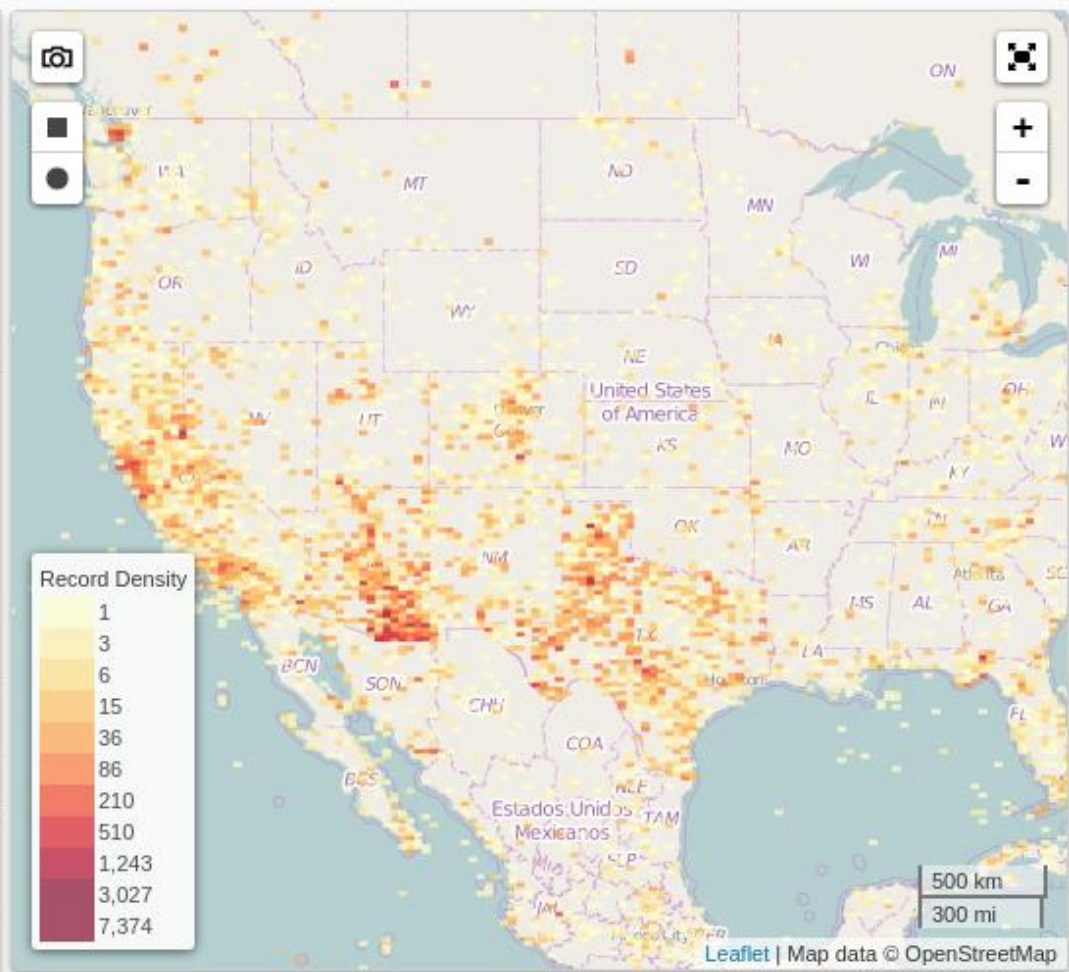
Image is present. Kingdom = Animalia. S [↻](#)

Download CSV - Build time: 0 hrs 3 mins 16 secs

Email [ⓧ](#)

Downloads

Search	Status
is present. Kingdom = Animalia...	Click To Download



Specimen record page with summary, details, flags, associated media, georeference and provider

Specimen Record

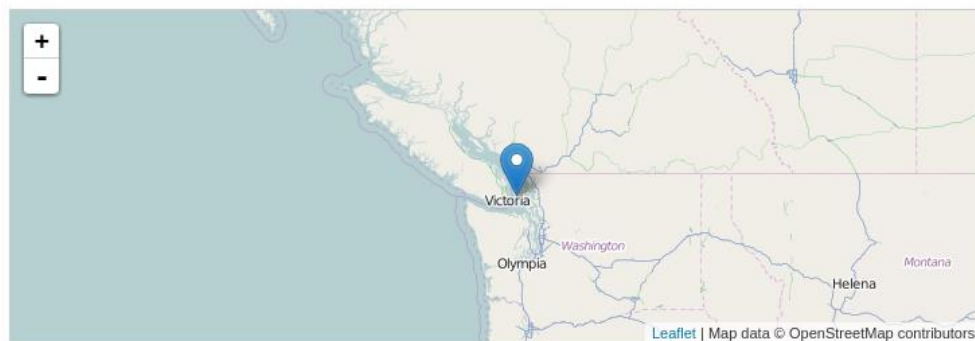
Animalia > Annelida > Polychaeta > Arenicolidae

Abarenicola pacifica Healy & Wells

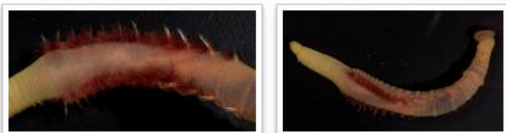
From invertebratezoology

Continent North America
 Country United States
 State/Province Washington
 County/Parish San Juan County
 Latitude 48.483333333333334
 Longitude -123.06944444444444

Institution Code FLMNH
 Collection Code Invertebrate Zoology
 Catalog Number 952 Annelida
 Collected By G Paulay



Media



Data Flags Raw

Contents

Summary
 Map
 Media
 Attribution
 All Data

Taxonomy

Scientific Name	Abarenicola pacifica
Kingdom	animalia
Phylum	annelida
Class	Polychaeta
Family	Arenicolidae
Genus	Abarenicola
Specific Epithet	pacifica
Scientific Name Authors	Healy & Wells

Data Flags Raw

Type	Description
geopoint_datum_missing	Geographic Coordinate Missing
dwc_phylum_added	Darwin Core Phylum Added.
dwc_continent_added	Darwin Core Continent Added.
dwc_country_replaced	Darwin Core Country Corrected.
idigbio_isocountrycode_added	iDigBio ISO 3166-1 alpha-3 Cou
dwc_kingdom_added	Darwin Core Kingdom Added.

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

Media Record

[Animalia](#) > [Annelida](#) > [Polychaeta](#) > [Arenicolidae](#)

Abarenicola pacifica Healy & Wells [view specimen record](#)

From Image Appliance Recordset for FLMNH Invertebrate Zoology



[Download Media File](#)

Other Media



Contents

- [Media](#)
- [Other](#)
- [Media](#)
- [Attribution](#)
- [All Data](#)

Publishers page with record counts, links to provider details

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	45,563,237	12,492,054
Total in API	45,722,764	12,483,825
Total Indexed (all data) *	45,722,764	12,483,825

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

Publisher Summary

Publisher Name	Records			Media		
	Digest	API	Index	Digest	API	Index
IPT - Hosted by VertNet	7,893,325	7,893,325	7,893,325	552,691	552,691	552,691
MNHN - Collections	6,743,820	6,743,820	6,743,820	5,581,957	5,581,957	5,581,957
speciesLink Network / INCT-HVFF IPT	3,331,822	3,331,822	3,331,822	0	0	0
KU Biodiversity Institute IPT	2,340,379	2,340,379	2,340,379	0	0	0
Berkeley Natural History Museums IPT	2,303,497	2,303,386	2,303,386	0	0	0
Consortium of North American Bryophyte Herbaria Darwin Core Archive rss feed	2,092,032	2,092,032	2,092,032	1,169,079	1,169,079	1,169,079
CAS-IPT	1,928,116	1,928,116	1,928,116	0	0	0

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

Recordset

Search Recordset

UF FLMNH Ichthyology

Specimen Records: 220,878 Media Records: 0 Last Update: 2015-10-14

The UF Fish Collection, dating to 1917, contains 214,205 lots and 2,300,803 specimens. Included are representatives of 8,250 species from 400 families. The collection includes 93 primary types and approximately 1,600 lots of secondary types representing 563 species. Also in the collection are 5,825 specimens of disarticulated and articulated skeletons representing 875 species. Especially notable are historic collections of large and important marine fishes as well as rapidly growing collections of freshwater fishes from Southeast Asia. In 2006, the museum expanded its program to archive frozen tissue samples with a newly established UF Genetic Resources Collection. Tissues of fishes are stored in -20°C freezers and number 4,150 samples of 900 species. All specimens and tissues are databased online and available for loan.

Contacts

Name Rob Robins
Role Ichthyology Collection Manager
Email rrobins@flmnh.ufl.edu
Phone none

Data Corrected Data Use Raw

This table shows any data corrections that were performed on this recordset to improve the capabilities of iDigBio Search. The first column represents the correction performed. The last two columns represent the number and percentage of records that were corrected. A complete list of the data quality flags and their descriptions can be found [here](#). Clicking on a data flag name will take you to a search for all records with this flag in this recordset.

Flag	Records With This Flag	(%) Percent With This Flag
dwc_kingdom_added ⓘ	219527	99.388
dwc_phylum_added ⓘ	219527	99.388
geopoint_datum_missing ⓘ	215241	97.448

Data Corrected Data Use Raw

Month of	Search	Download	Seen	Records Viewed	Media Viewed
01 / 2015	52,169,787	224,080	733	1,877	0
02 / 2015	160,296,710	223,895	2,372	1,035	0
03 / 2015	84,959,293	440,641	455	1,173	0
05 / 2015	257,737,259	213,932	2,486	7,539	0
06 / 2015	886,140,576	986,892	5,521	16,490	0
07 / 2015	2,446,086,020	221,546	21,805	10,133	0
08 / 2015	775,568,750	988,710	2,388	18,901	0
09 / 2015	3,259,157,586	225,407	1,580	32,974	0
10 / 2015	405,700,449	157,307	1,358	18,734	0

Over 700 providers, 46M specimen records, 12M media records

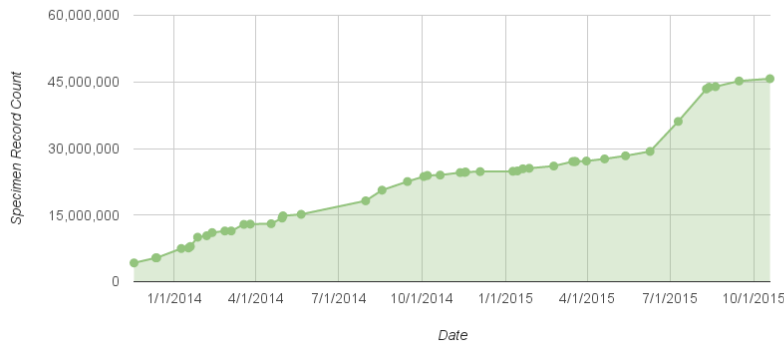
Publishing technologies: IPT, Symbiota, RSS (DwC-a, CSV)

Media data using Audubon Core terms

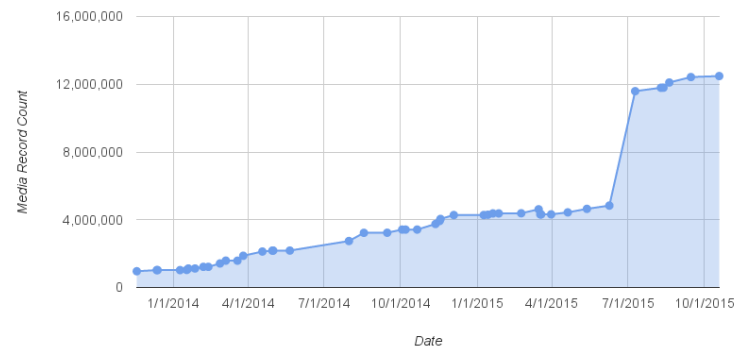


... and many more.

iDigBio Data Ingestion - Specimen Records

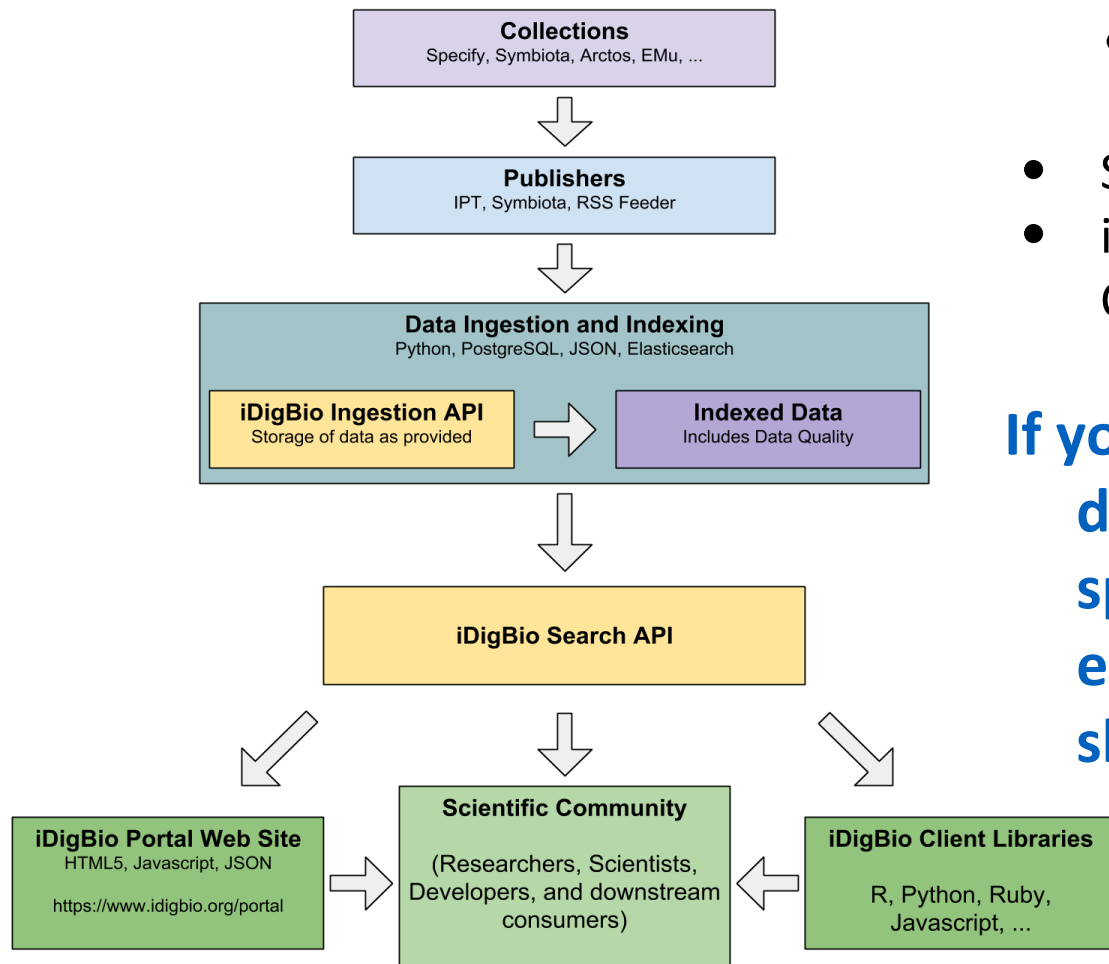


iDigBio Data Ingestion - Media Records



The what and how of data ingestion

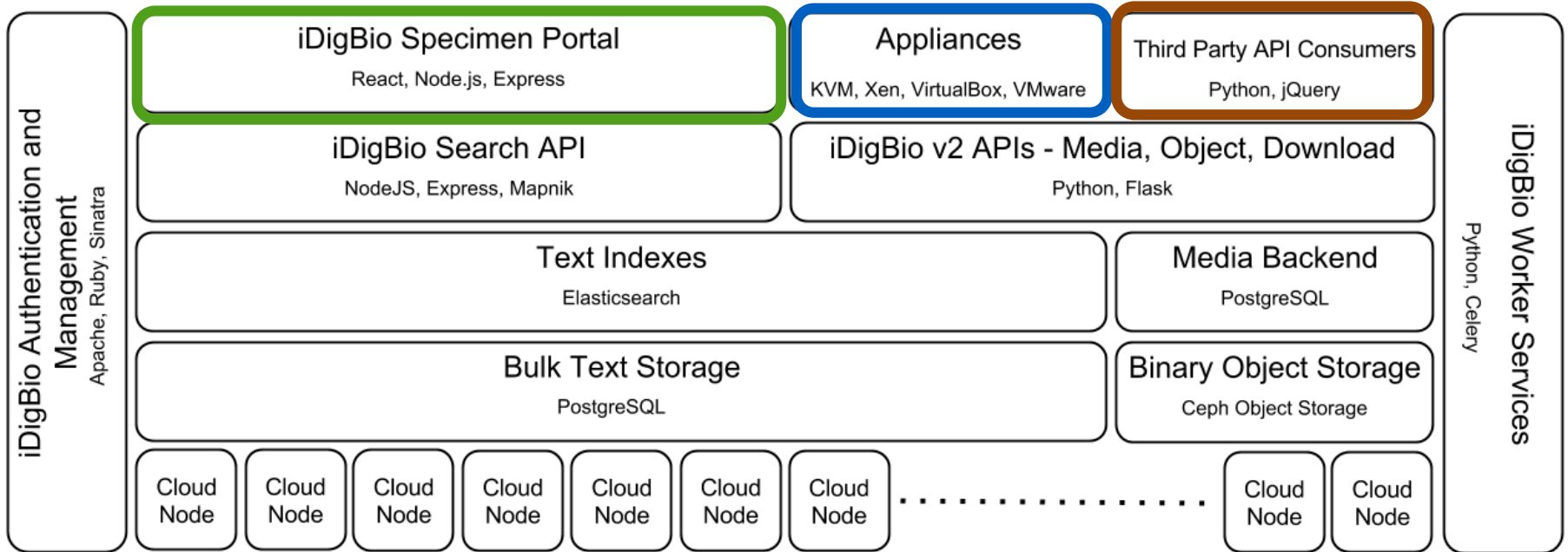
iDigBio Data Flow Diagram



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

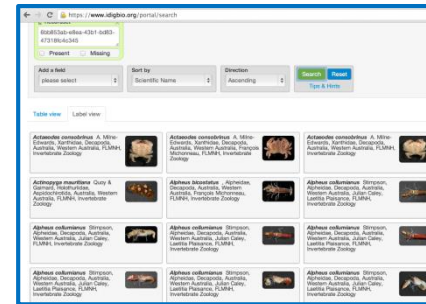


Appliances

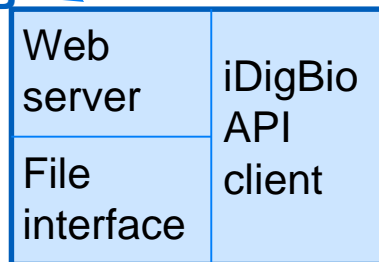
- Image upload appliance - reliable upload of images+metadata batches
 - users include NYBG and SCAN (Symbiota) - NAU, UHIM, UC Boulder
- Specify appliance
 - used in 10 training workshops, 4 countries; ~200 people/~25 institutions



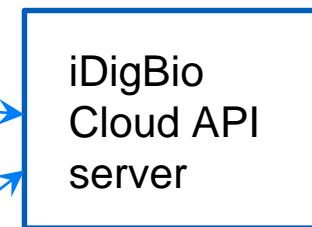
Web UI browser



HTTP



HTTP



Appliance

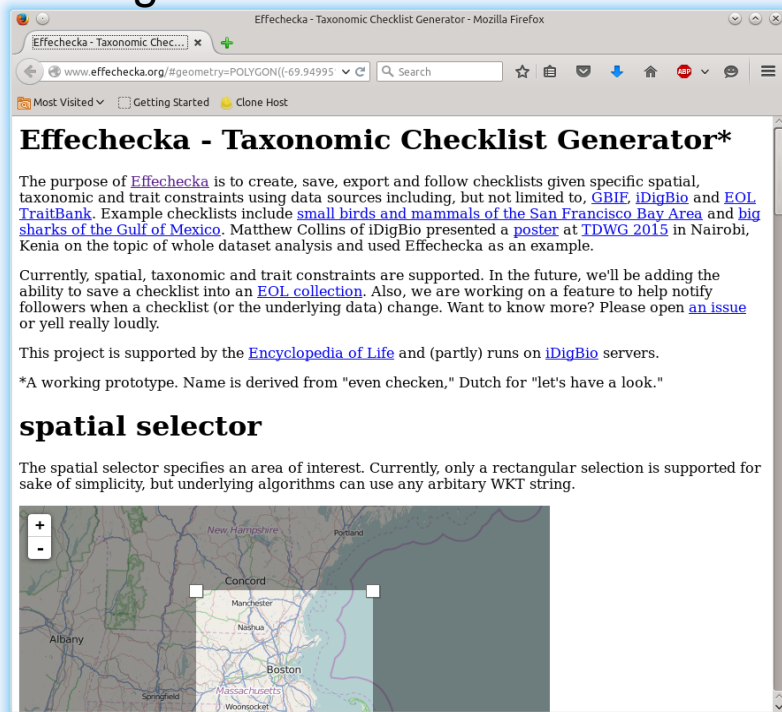
Images stored on local media



Emerging Research Tools from iDigBio and the Community

Checklist generation & “Spark” data processing

Jorrit Poelen (GLOBI) &
Jen Hammock
working with EOL

Effechecka - Taxonomic Checklist Generator*

The purpose of [Effechecka](#) is to create, save, export and follow checklists given specific spatial, taxonomic and trait constraints using data sources including, but not limited to, [GBIE](#), [iDigBio](#) and [EOL TraitBank](#). Example checklists include [small birds and mammals of the San Francisco Bay Area](#) and [sharks of the Gulf of Mexico](#). Matthew Collins of iDigBio presented a [poster](#) at [TDWG 2015](#) in Nairobi, Kenya on the topic of whole dataset analysis and used Effechecka as an example.

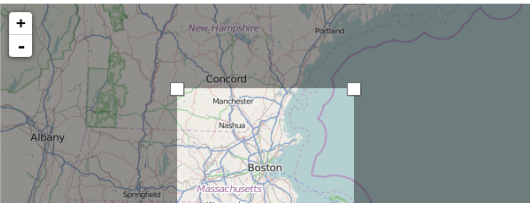
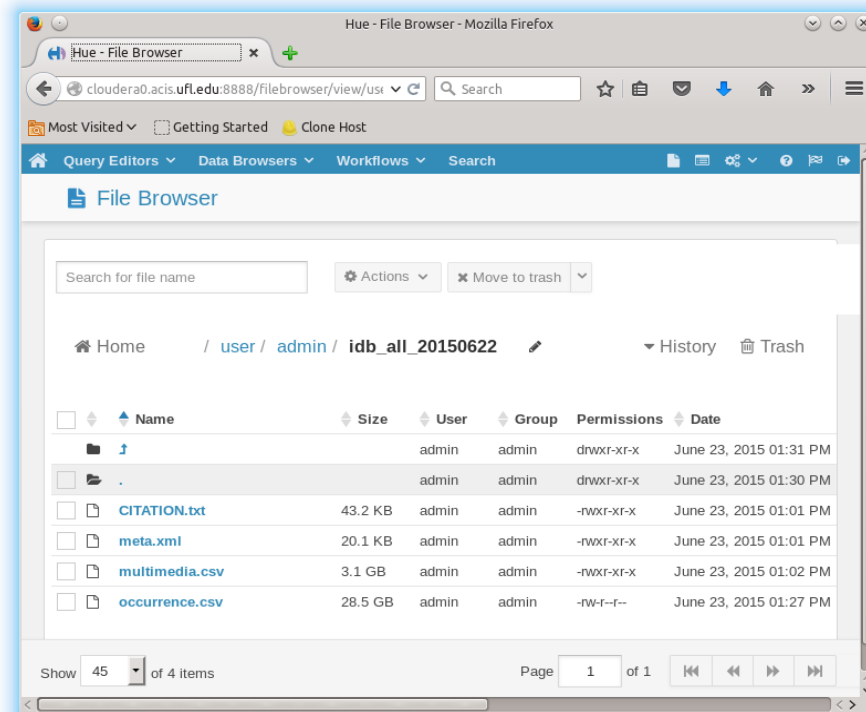
Currently, spatial, taxonomic and trait constraints are supported. In the future, we'll be adding the ability to save a checklist into an [EOL collection](#). Also, we are working on a feature to help notify followers when a checklist (or the underlying data) change. Want to know more? Please open [an issue](#) or yell really loudly.

This project is supported by the [Encyclopedia of Life](#) and (partly) runs on [iDigBio](#) servers.

*A working prototype. Name is derived from "even checken," Dutch for "let's have a look."

spatial selector

The spatial selector specifies an area of interest. Currently, only a rectangular selection is supported for sake of simplicity, but underlying algorithms can use any arbitrary WKT string.

Hue - File Browser

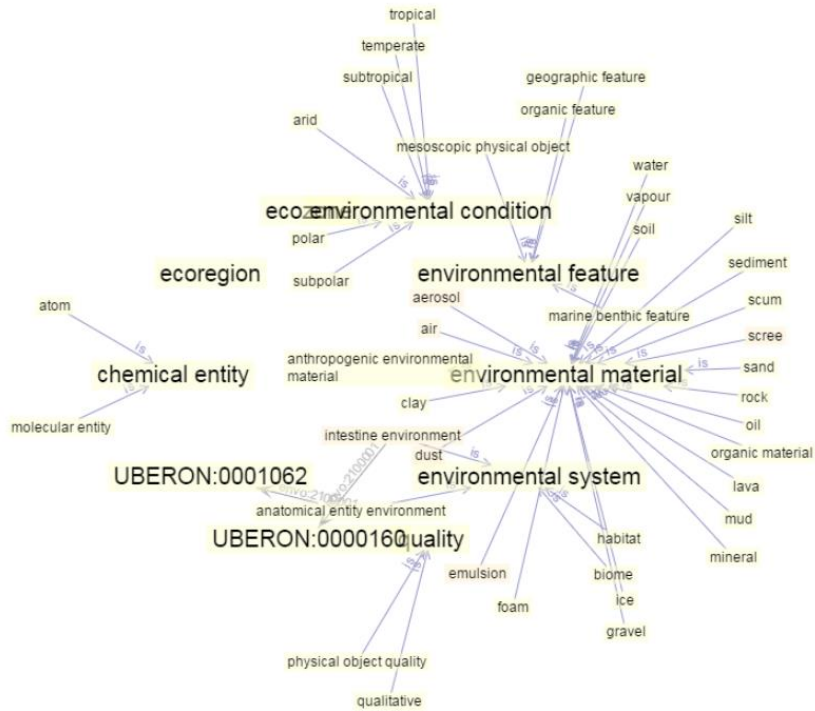
Search for file name Actions Move to trash

Home / user / admin / idb_all_20150622 History Trash

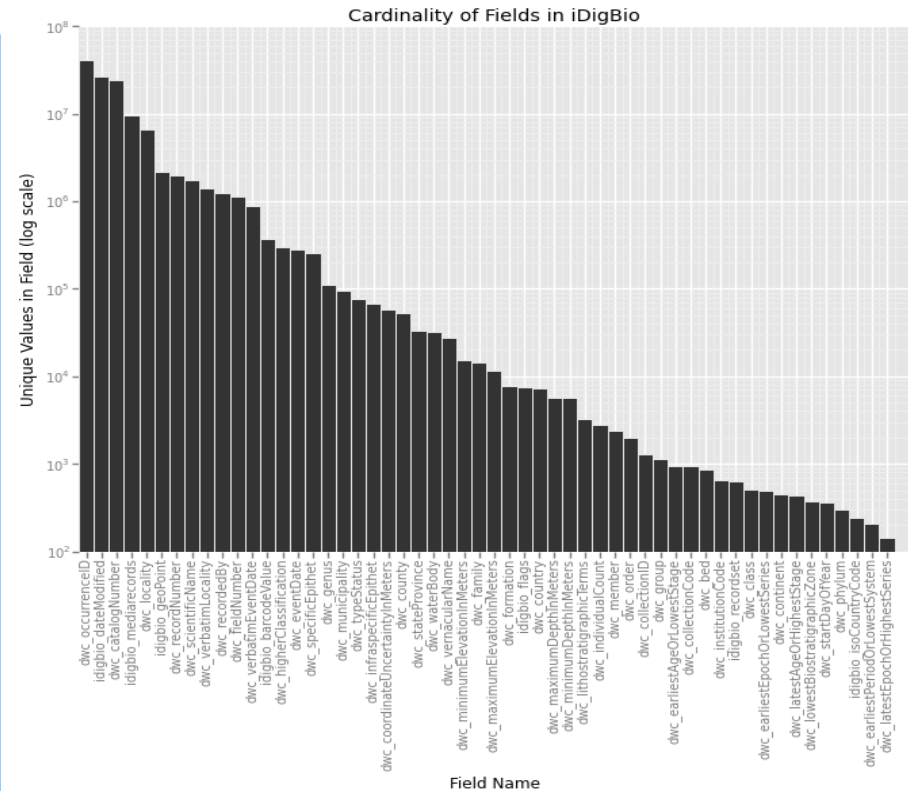
<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		admin	admin	drwxr-xr-x	June 23, 2015 01:31 PM
<input type="checkbox"/>	.		admin	admin	drwxr-xr-x	June 23, 2015 01:30 PM
<input type="checkbox"/>	CITATION.txt	43.2 KB	admin	admin	-rwxr-xr-x	June 23, 2015 01:01 PM
<input type="checkbox"/>	meta.xml	20.1 KB	admin	admin	-rwxr-xr-x	June 23, 2015 01:01 PM
<input type="checkbox"/>	multimedia.csv	3.1 GB	admin	admin	-rwxr-xr-x	June 23, 2015 01:02 PM
<input type="checkbox"/>	occurrence.csv	28.5 GB	admin	admin	-rwt--r--	June 23, 2015 01:27 PM

Show 45 of 4 items Page 1 of 1

Data Characterization - Examining iDigBio Data



Grant Godden and Pier Luigi Buttigieg (Phenotype RCN) processed label data from over one million plant specimen records hosted by [iDigBio](https://www.idigbio.org/) ...preliminary results of the analyses were immediately informative, revealing gaps in the current coverage of the [Environment Ontology](#)



Unique field values blog post
at <https://www.idigbio.org/>

<https://www.idigbio.org/content/exploring-unique-values-idigbio-using-apache-spark>

Demonstrated integrations of research tools with iDigBio

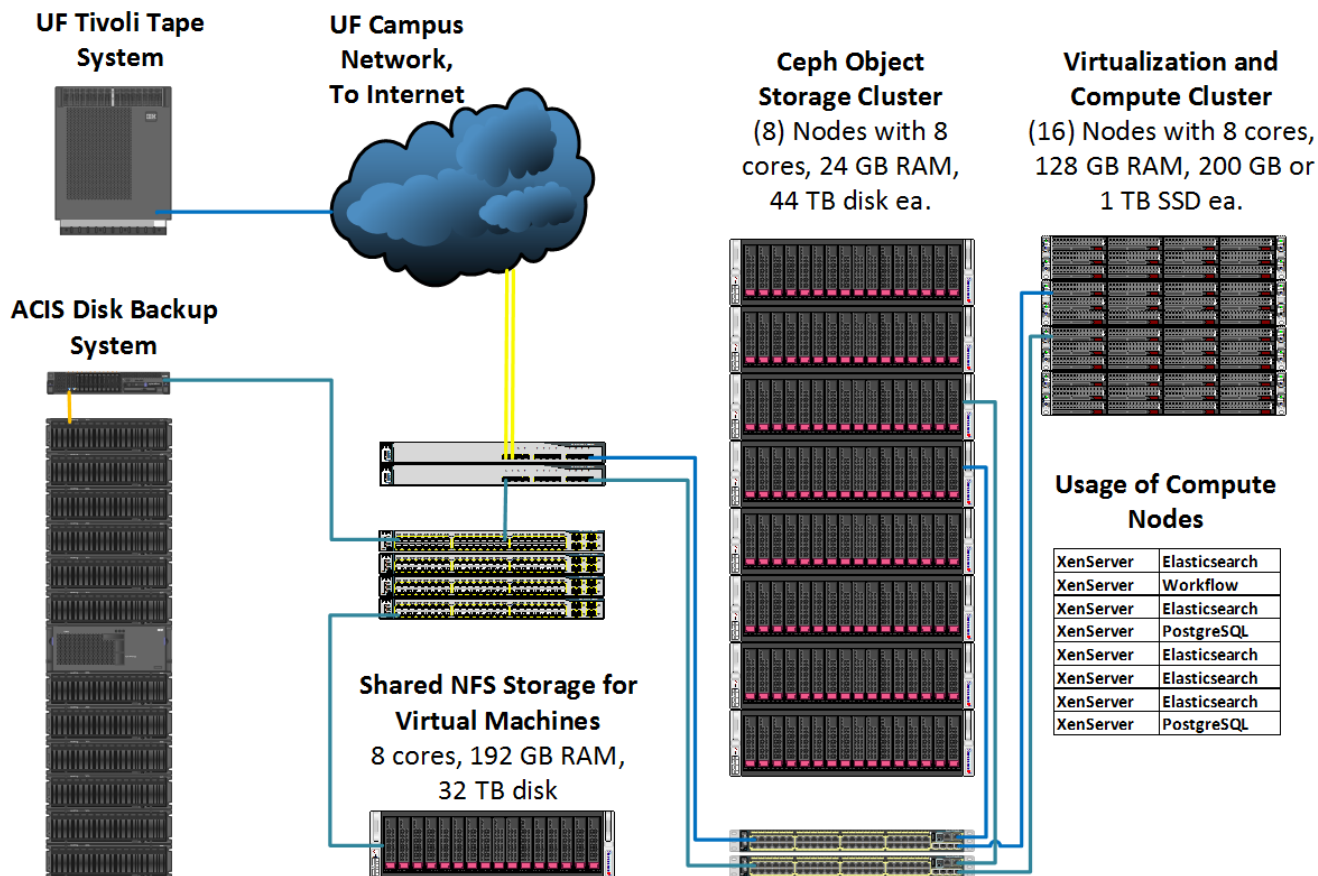
- PhyloJIVE + OpenTree + iDigBio
- Arbor + OpenTree + iDigBio
- OpenRefine + OpenTree + iDigBio
- Lifemapper



- Presented at the last Summit and SPNCH 2015
- Contact idigbio@acis.ufl.edu if you are interested in integration of your research tool(s)

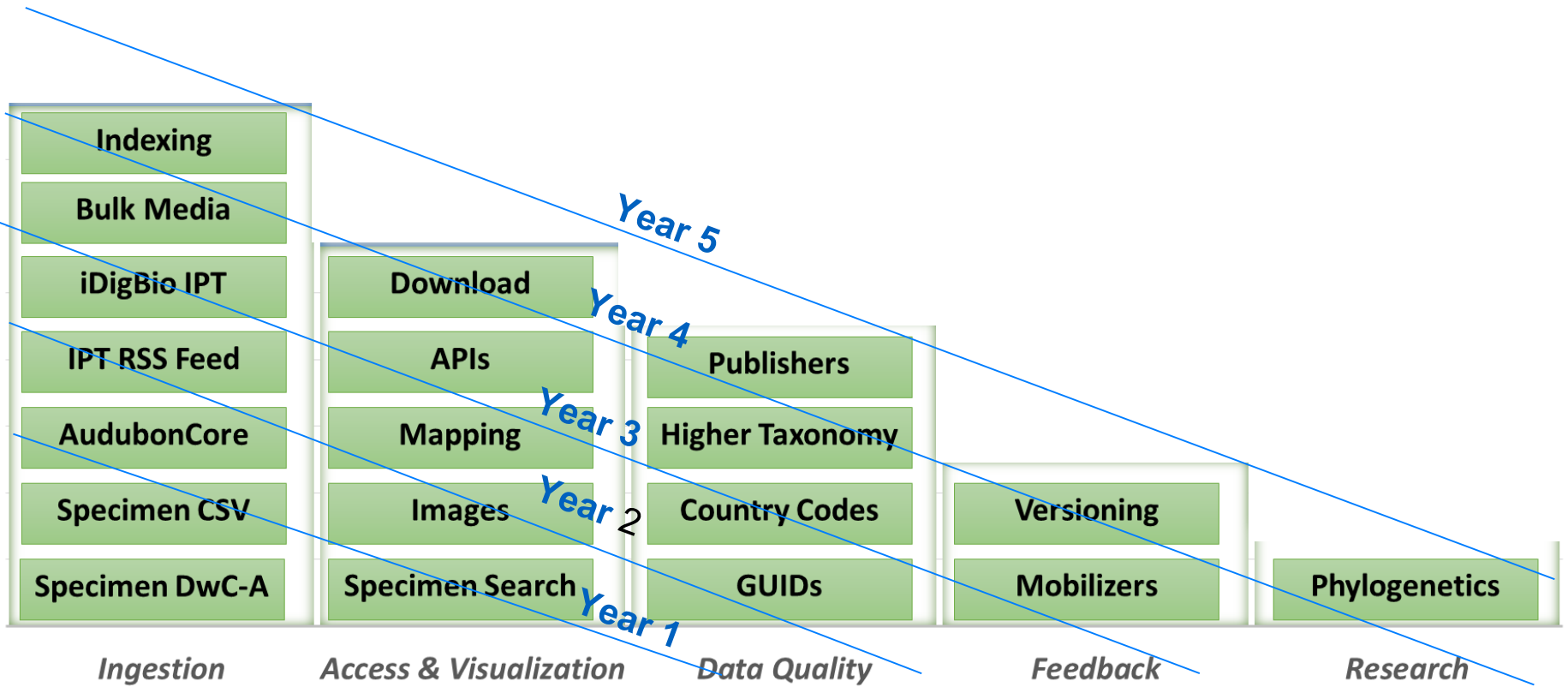
iDigBio infrastructure (48 servers): Proxy/load balance (2); Portal (5); API (5); Media API (10); Celery task (5); Ceph Object Storage (3); Rabbit queue (2); Application and database (18)

iDigBio Infrastructure at ACIS



iDigBio ACTIVITIES SUPPORTED BY CYBERINFRASTRUCTURE

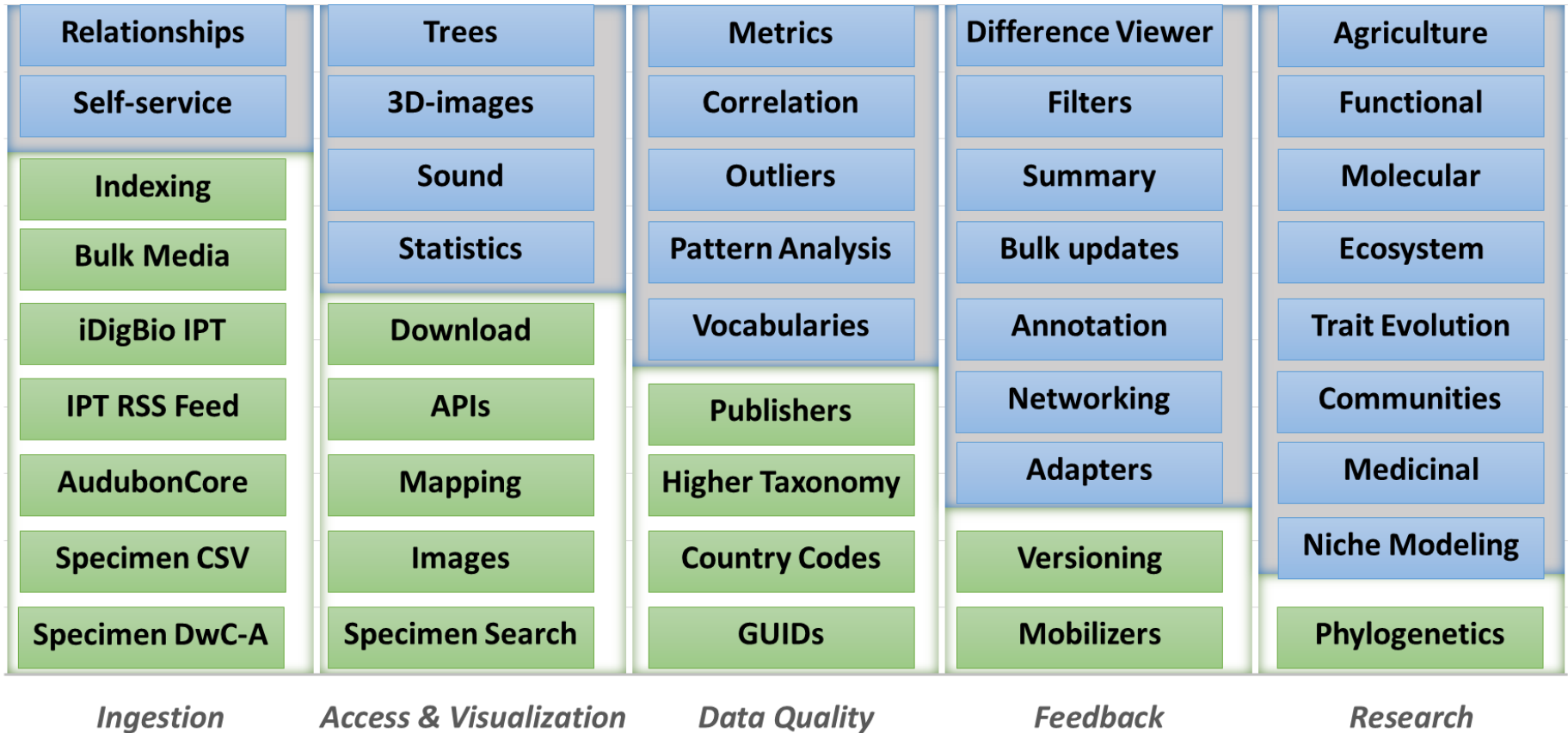
■ 2011-2016



TO BE

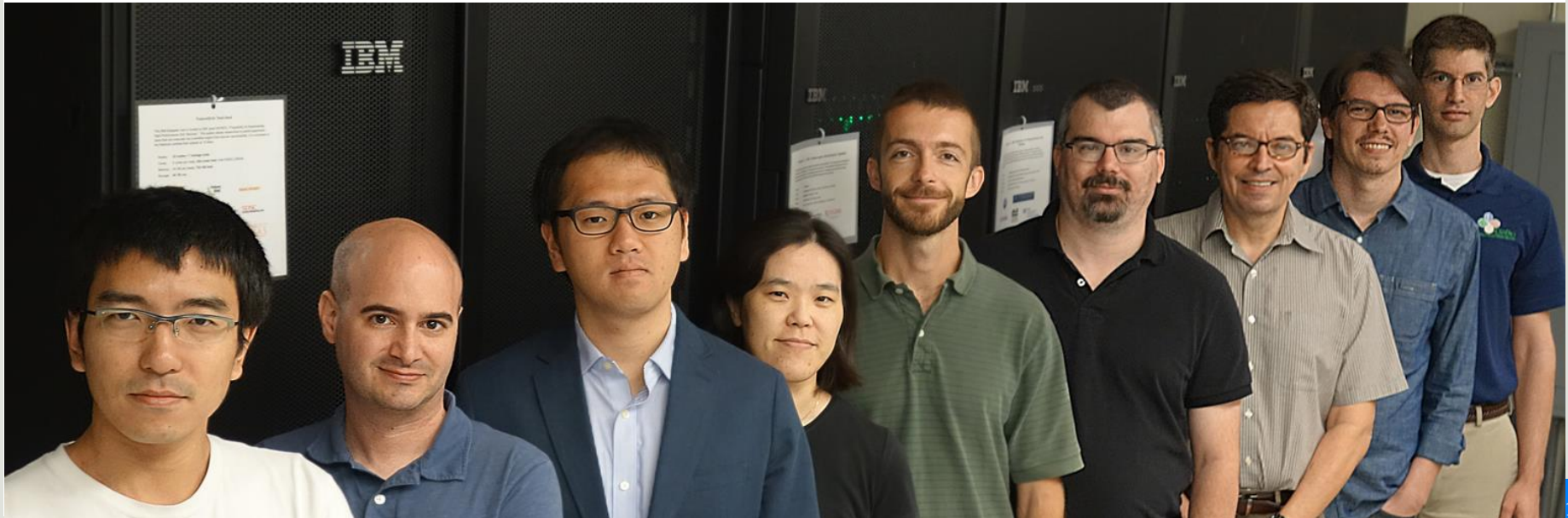
iDigBio ACTIVITIES SUPPORTED BY CYBERINFRASTRUCTURE

■ 2011-2016 ■ 2016-2021



Acknowledgements

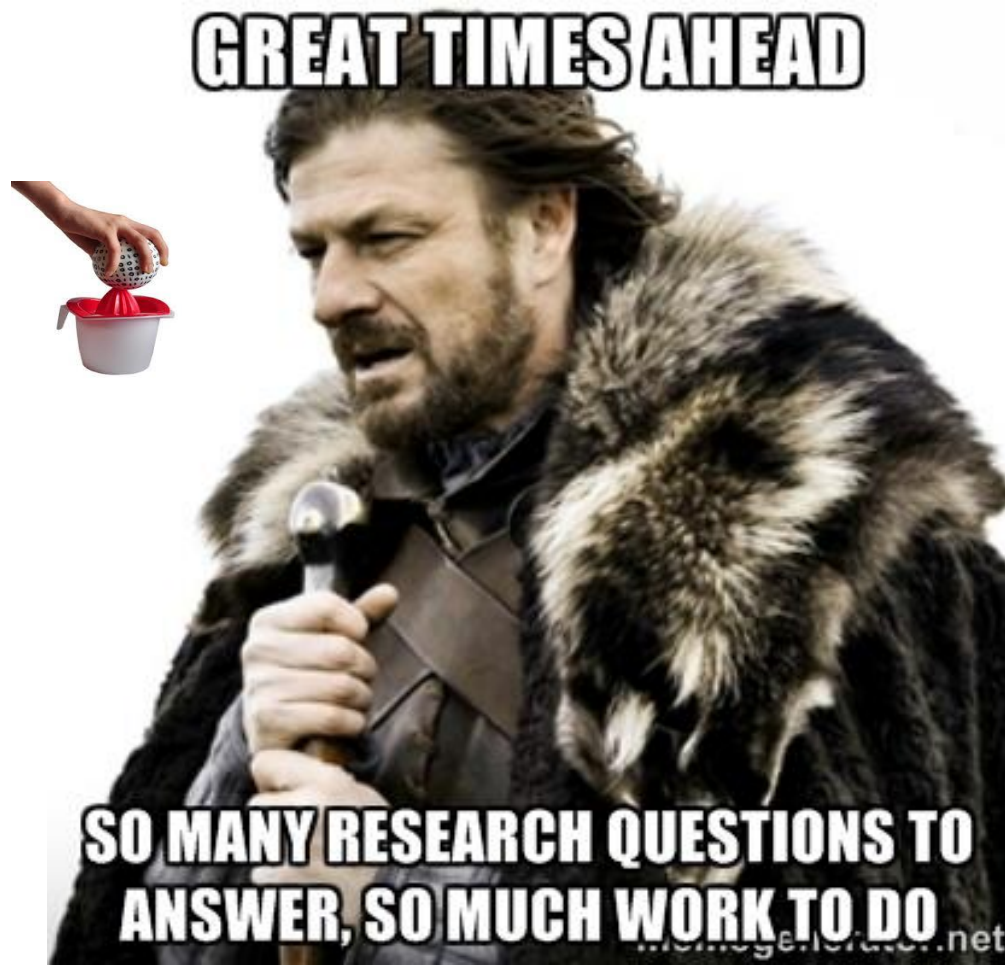
- National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210)
- Dr. Anne Maglia, Dr. Roland Roberts and Dr. Judith Skog @NSF
- The ADBC/collections community for the privilege of hosting their data
- All iDigBio faculty, students and staff at UF and FSU
 - in particular, the iDigBio IT team
 - in particular, the iDigBio IT team members at ACIS



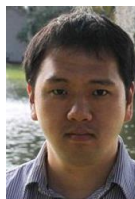
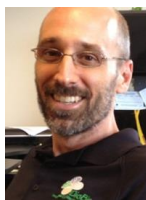
Parting messages



<http://memegenerator.net/instance/56576771>

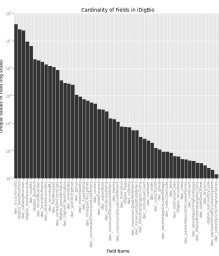
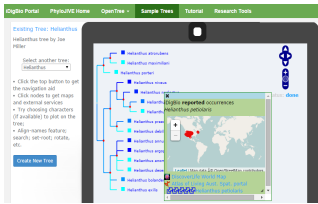


Acknowledgements

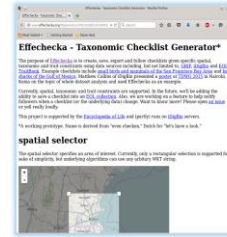


APIs and Client Libraries Under it All

Appliances



Portal



Original Research



Tools

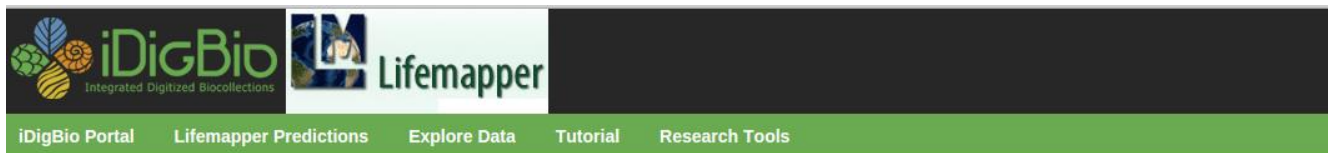
```

{
  "items" : [
    {
      "dwc:eventDate" : "1968-04-11",
      "dwc:recordedBy" : "Harris, Arthur H. (Van Goebel)",
      "dwc:occurrenceStatus" : "present",
      "dwc:catalogNumber" : "5-298",
      "dwc:maximumElevationInMeters" : "1280",

```

LifeMapper instance at iDigBio

- Currently a proof of concept prototype.
- Further development funded under the BiotaPhy project.



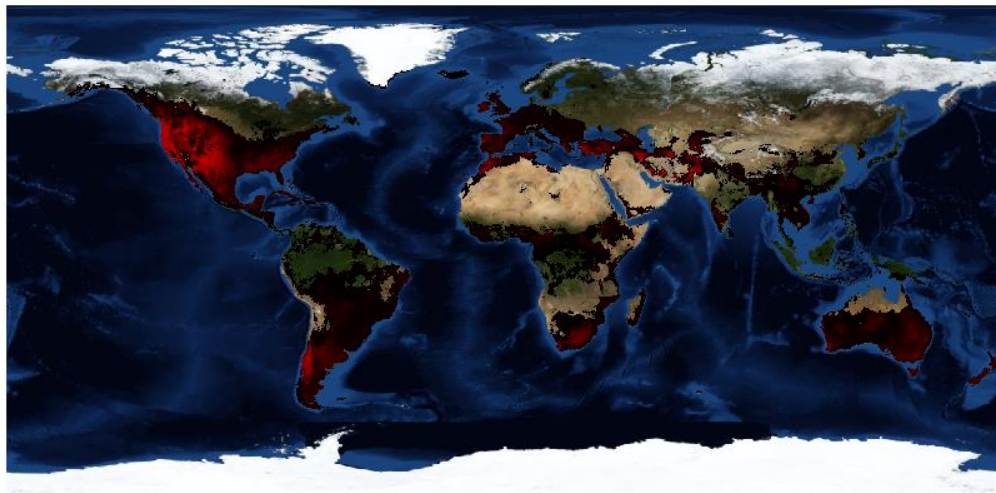
Choose a species:

LifeMapper Species Distribution Modeling for *Puma Concolor*

Model

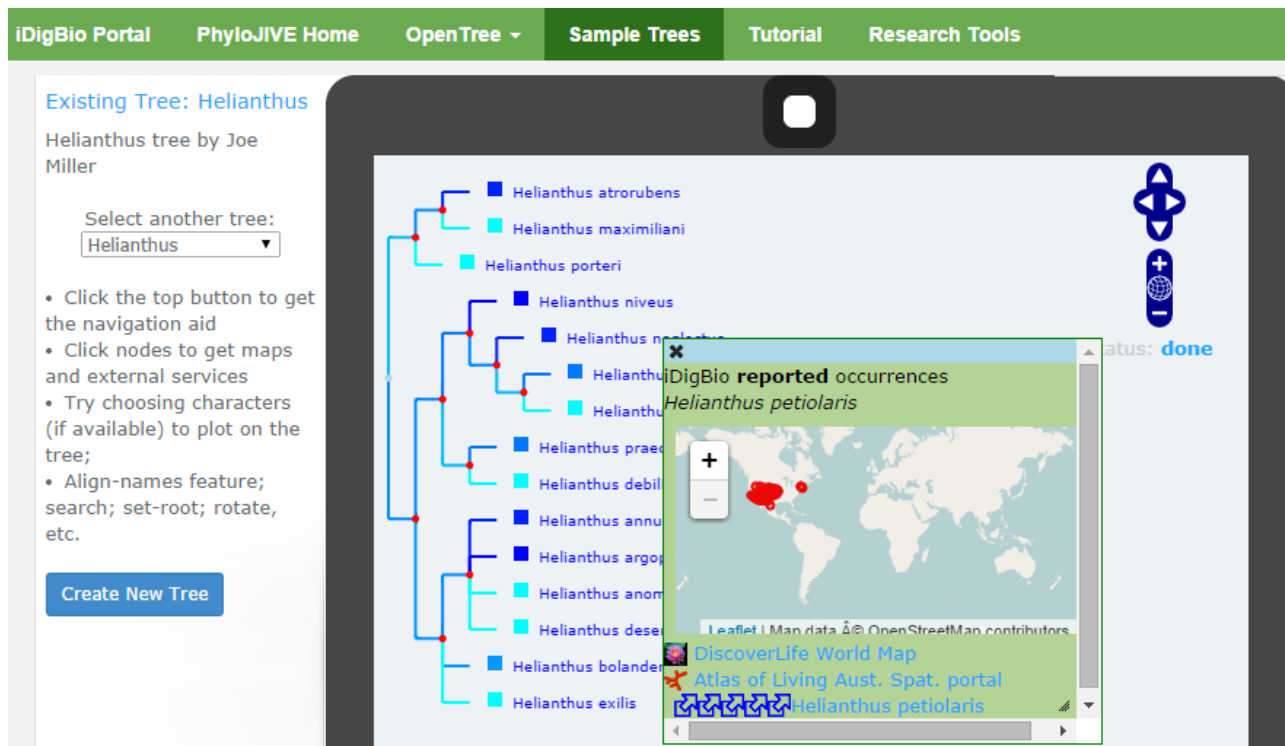
Projection 77672 was built by applying the model results to scenario [WC-5MIN](#)

Distribution



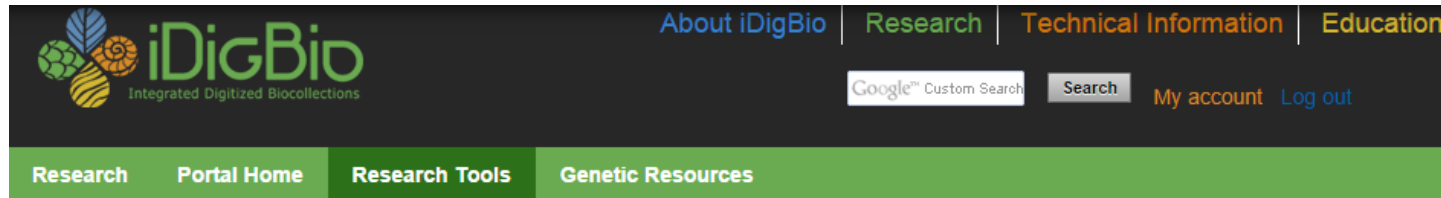
PhyloJIVE instance in iDigBio (biodiversity data + phylogeny)

- 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,...



The screenshot displays the iDigBio Portal interface for PhyloJIVE. The navigation bar includes links for 'iDigBio Portal', 'PhyloJIVE Home', 'OpenTree', 'Sample Trees', 'Tutorial', and 'Research Tools'. The main content area shows an 'Existing Tree: Helianthus' by Joe Miller. A dropdown menu allows selecting another tree, currently set to 'Helianthus'. A list of species is shown with colored squares indicating their predicted first character: Helianthus atrorubens (blue), Helianthus maximiliani (cyan), Helianthus porteri (magenta), Helianthus niveus (dark blue), Helianthus mollis (cyan), Helianthus mollis (dark blue), Helianthus mollis (cyan), Helianthus praecox (dark blue), Helianthus debilis (cyan), Helianthus annuus (dark blue), Helianthus argenteus (cyan), Helianthus anomus (dark blue), Helianthus descurainii (cyan), Helianthus bolanderi (dark blue), and Helianthus exilis (cyan). A map window titled 'iDigBio reported occurrences Helianthus petiolaris' is overlaid on the tree, showing a world map with red dots indicating reported occurrences in North America. The map includes a navigation aid and a 'status: done' indicator. Below the map are links to 'DiscoverLife World Map' and 'Atlas of Living Aust. Spat. portal'.

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

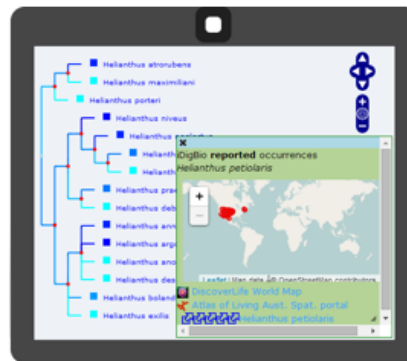


Teachers & Students

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!

Virtual appliance: Specify

- Appliance packages: Ubuntu 12.04 LTS, MySQL, Java 7, Specify 6.5, Demo database
- User installs free software and appliance from iDigBio
- Reduces training session setup times

