iDigBio Cyberinfrastructure, Portal and Data

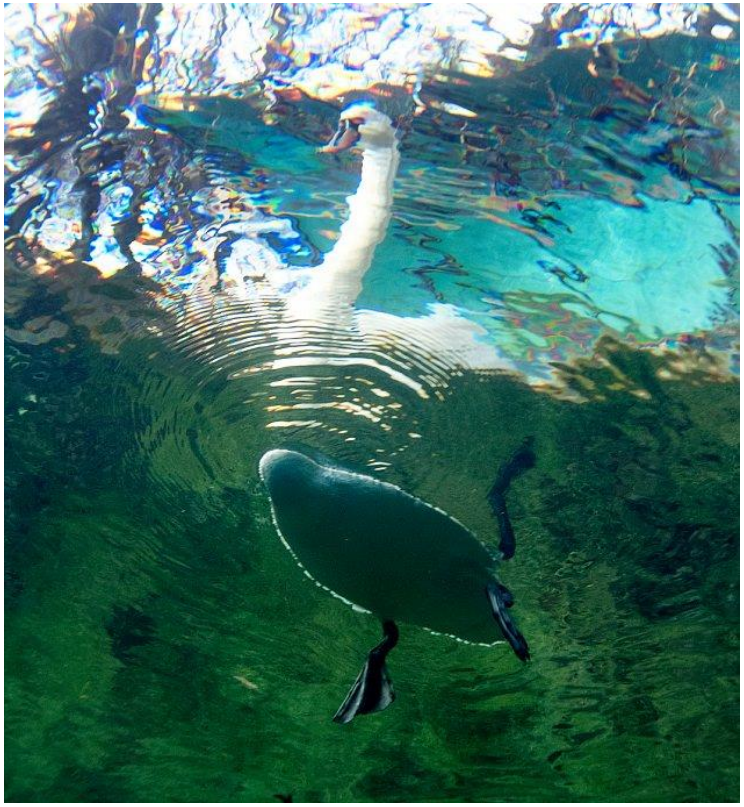
José Fortes
Advanced Computing and Information Systems Laboratory (ACIS)
University of Florida

fortes@acis.ufl.edu

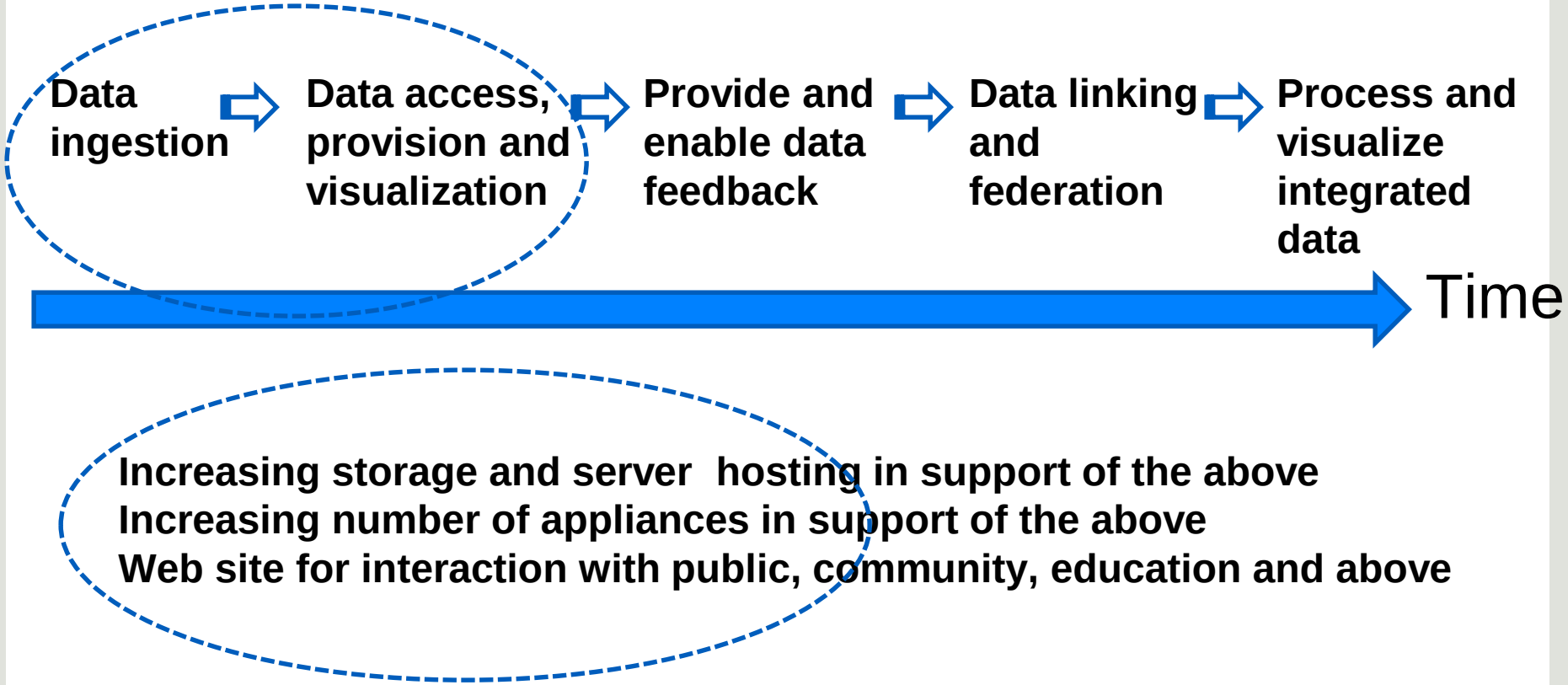
iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect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

Outline

- Cyberinfrastructure
  - Web portal
  - Portal
  - Appliances
  - Research applications
- Data
  - Ingestion
  - Use
  - Integration

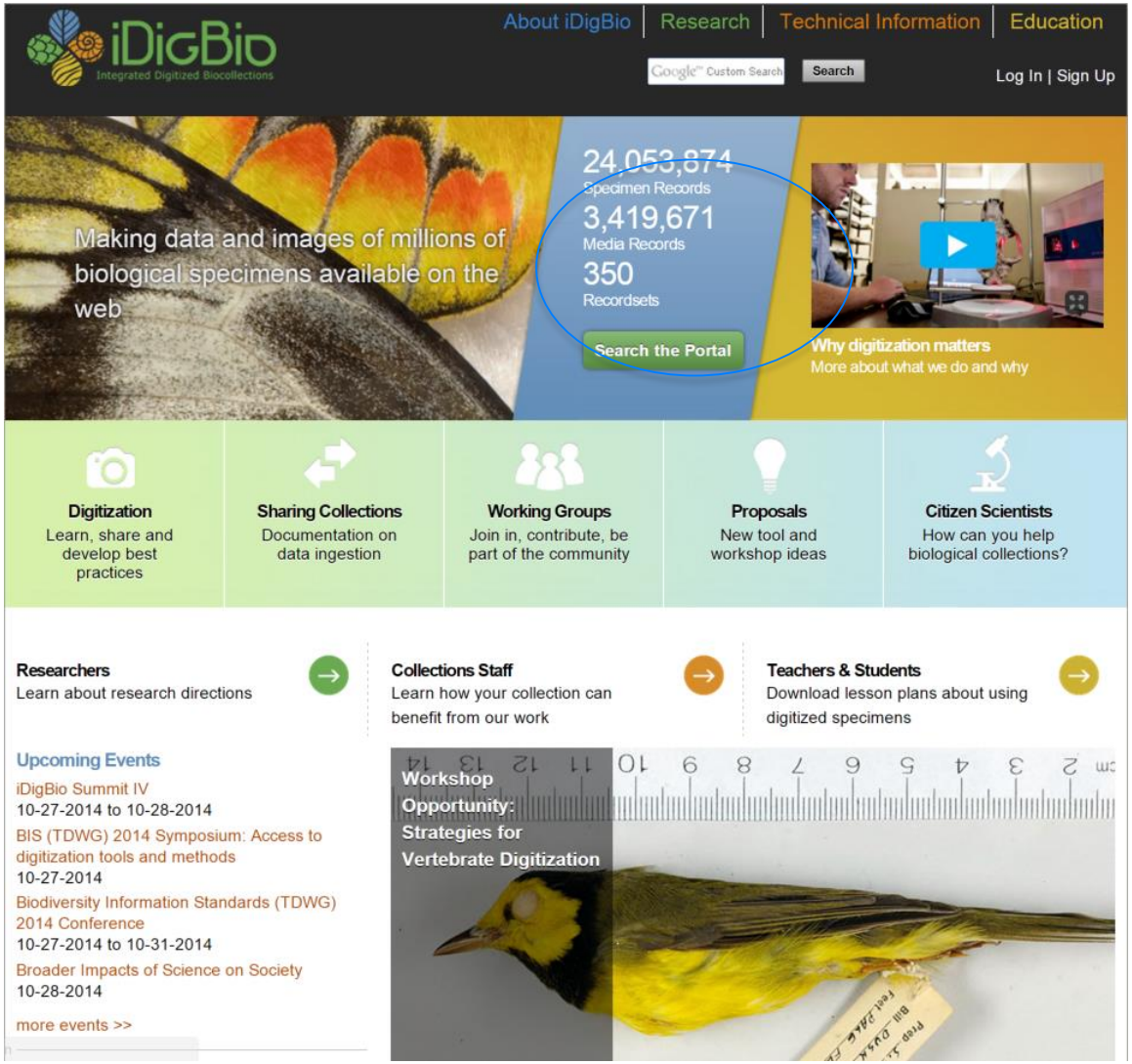


Evolution of iDigBio capabilities

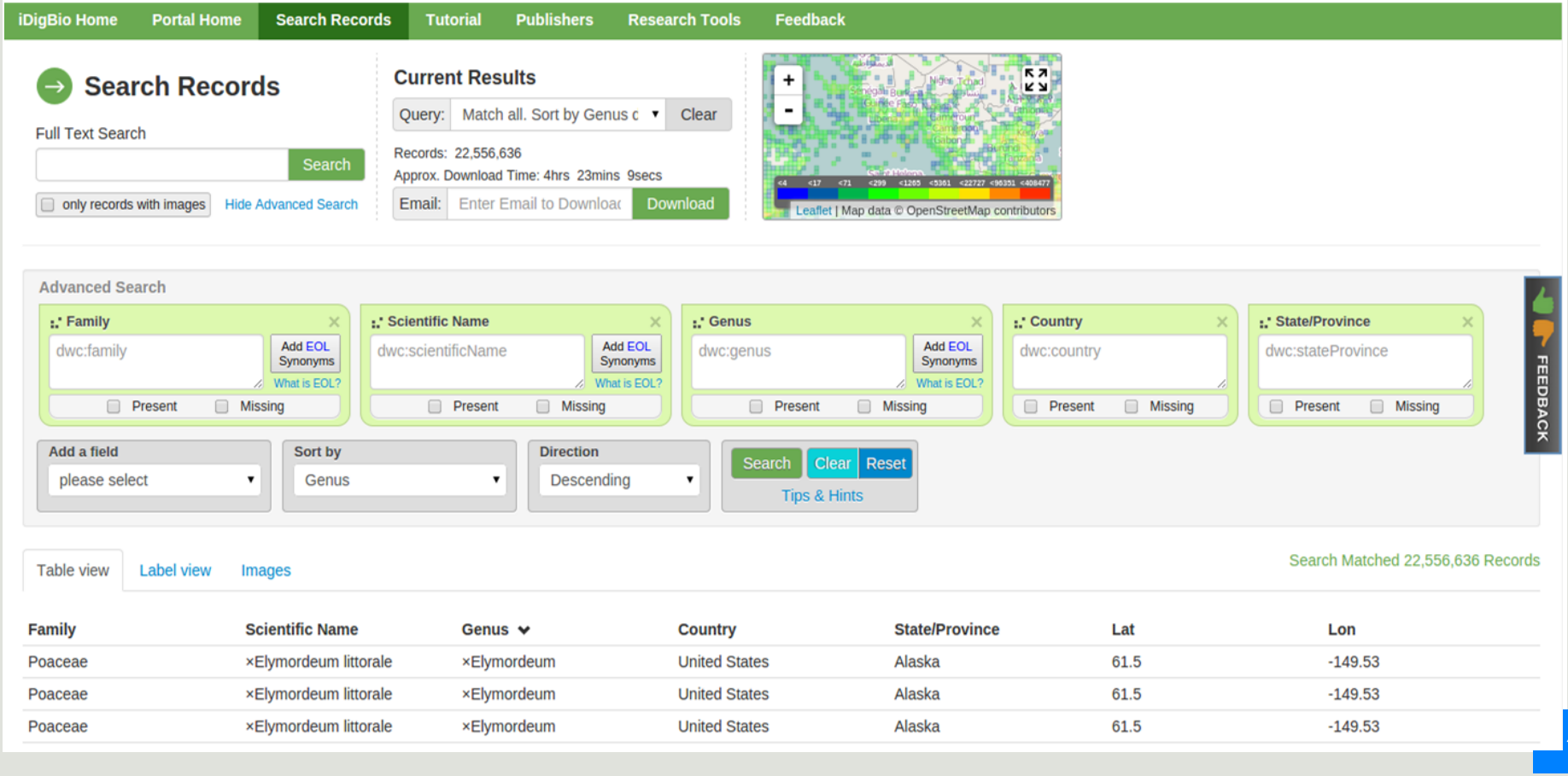


2

iDigBio Website



Search across all data, all/individual fields, customize, use autocompletion, synonyms,…



View search results as table, labels, images…

Table view | Label view | Images

Search Matched 65 Records

Family	Scientific Name ▲	Genus	Country	State/Province
Endodontidae	Aaadonta	Aaadonta	Palau	Ulebsechel Island
Endodontidae	Aaadonta	Aaadonta	Palau	
Endodontidae	Aaadonta	Aaadonta	Palau	

Table view | Label view | Images

Search Matched 3,512,348 Records

"Adlumia fungosa Greene ex Britton, Stern & Poggenb.", Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Lower slope of north side of Bear Mountain, Salisbury, Mt. Riga, Leslie J. Mehrhoff, UConn, CONN

"Adlumia fungosa Greene ex Britton, Stern & Poggenb.", Dicotyledonae, Papaverales, Papaveraceae, USA, Vermont, Devil's Den, Unknown, UConn, CONN

"Adlumia fungosa Greene ex Britton, Stern & Poggenb.", Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Mount Tom Pond, E.H. Eames, UConn, CONN

"Adlumia fungosa Greene ex Britton, Stern & Poggenb.", Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Pistapaug Mountain, Durham, Leslie J. Mehrhoff, UConn, CONN

"Adlumia fungosa Greene ex Britton, Stern & Poggenb.", Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Wooster Mountain, Leslie J. Mehrhoff, UConn, CONN

"Adlumia fungosa Greene ex Britton, Stern & Poggenb.", Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Mrs. G. J. Mendel 111 Highland Ave., So. Norwalk, "Wilkinson, A.E.;", UConn, CONN

"Adlumia fungosa Greene ex Britton, Stern & Poggenb.", Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Ashford, Boston Hollow Road, Leslie J. Mehrhoff, UConn, CONN

"Adlumia fungosa Greene ex Britton, Stern & Poggenb.", Dicotyledonae, Papaverales, Papaveraceae, USA, Connecticut, Roxbury Iron Mine; area Mine Hill, Leslie J. Mehrhoff, UConn, CONN

Table view | Label view | Images

Search Matched 749,373 Records

Results mapped/rendered and downloadable

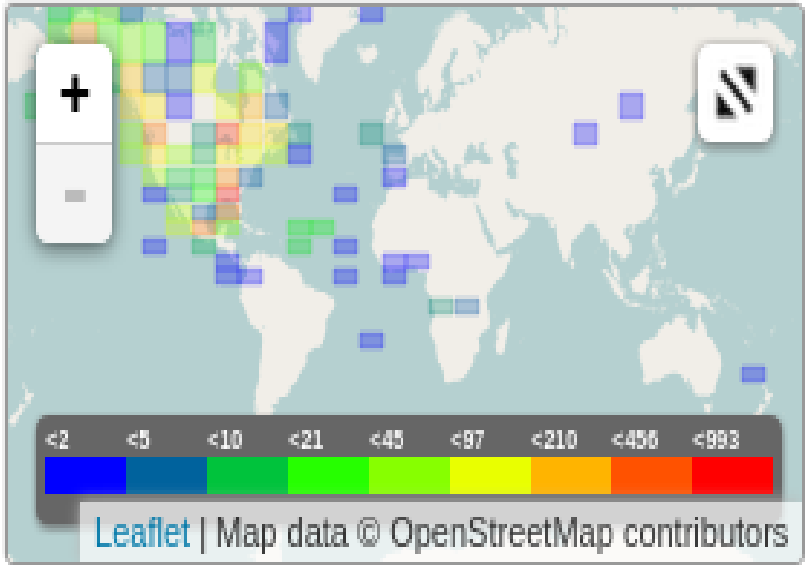
Current Results

Query: Recordset = fccc3c1d-d9d Clear

Records: 22,636

Approx. Download Time: 0hrs 0mins 15secs

Email: Enter Email to Downloac Download



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

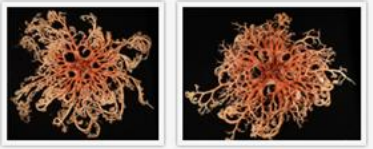
***Astrophyton* FLMNH, Invertebrate Zoology, 11905-Echinodermata**

Taxonomy | Specimen | Collection Event | Locality | Other

Class	Ophiuroidea
Order	Euryalida
Family	Gorgonocephalidae
Genus	Astrophyton

Go To Recordset | View Raw Data

Associated Media



Specimen Georeference

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
paulay@flmnh.ufl.edu

William Paine
IT Director
netadmin@flmnh.ufl.edu

FEEDBACK

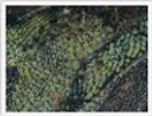
Media records with metadata, other media, provider, links to specimen record, data set …

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

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

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

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

Search This Recordset | View Raw Data

Website
http://www.washington.edu/burkemuseum/collections/herbarium/index.php

Contacts
egbot@asu.edu
David Giblin, Collections Manager
wtu@u.washington.edu

FEEDBACK

Specimen Fields Used for Search

Total Records: 74,141

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.

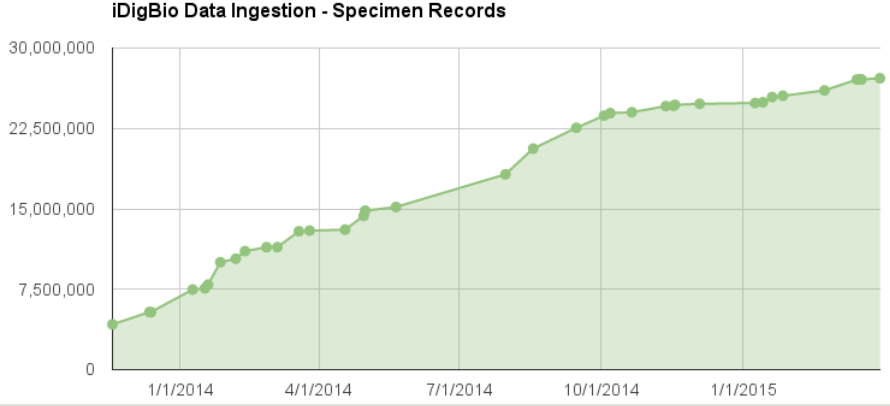
Field	Records With This Field	(%) Percent Used
Kingdom (dwc:kingdom)	0	0
Phylum (dwc:phylum)	0	0
Class (dwc:class)	0	0
Order (dwc:order)	0	0
Family (dwc:family)	66,675	89.93
Scientific Name (dwc:scientificName)	66,011	89.034
Genus (dwc:genus)	66,004	89.025
Specific Epithet (dwc:specificEpithet)	64,589	87.116
Infraspecific Epithet (dwc:infraspecificEpithet)	3,769	5.084
Higher Taxon (dwc:higherClassification)	0	0
Common Name (dwc:vernacularName)	0	0
Lat (dwc:decimalLatitude)	13,301	17.94
Lon (dwc:decimalLongitude)	13,110	17.683

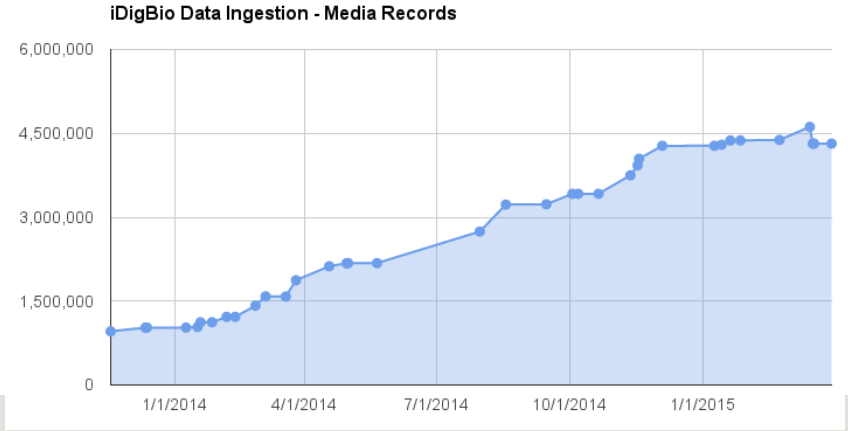
Over 300 publishers, 27M specimen records, 4.3M media records

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

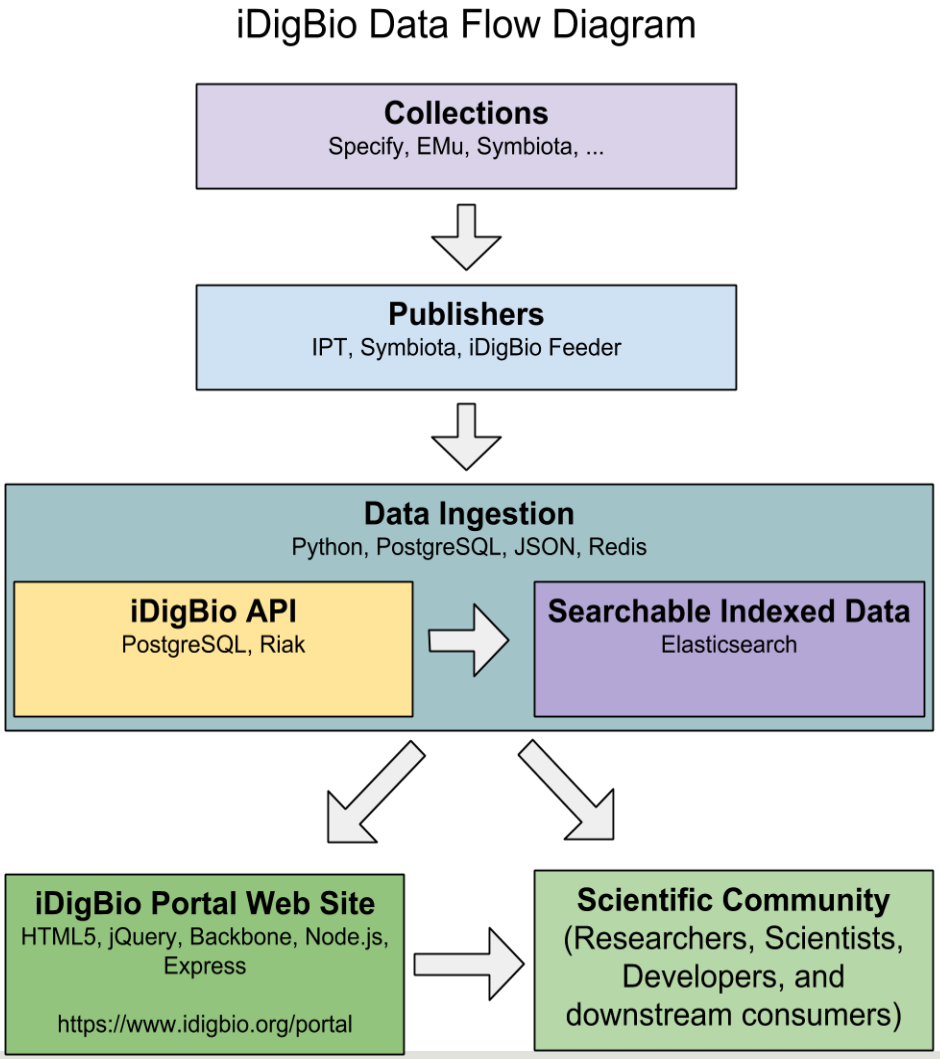
Media data using Audubon core terms

... and many more.





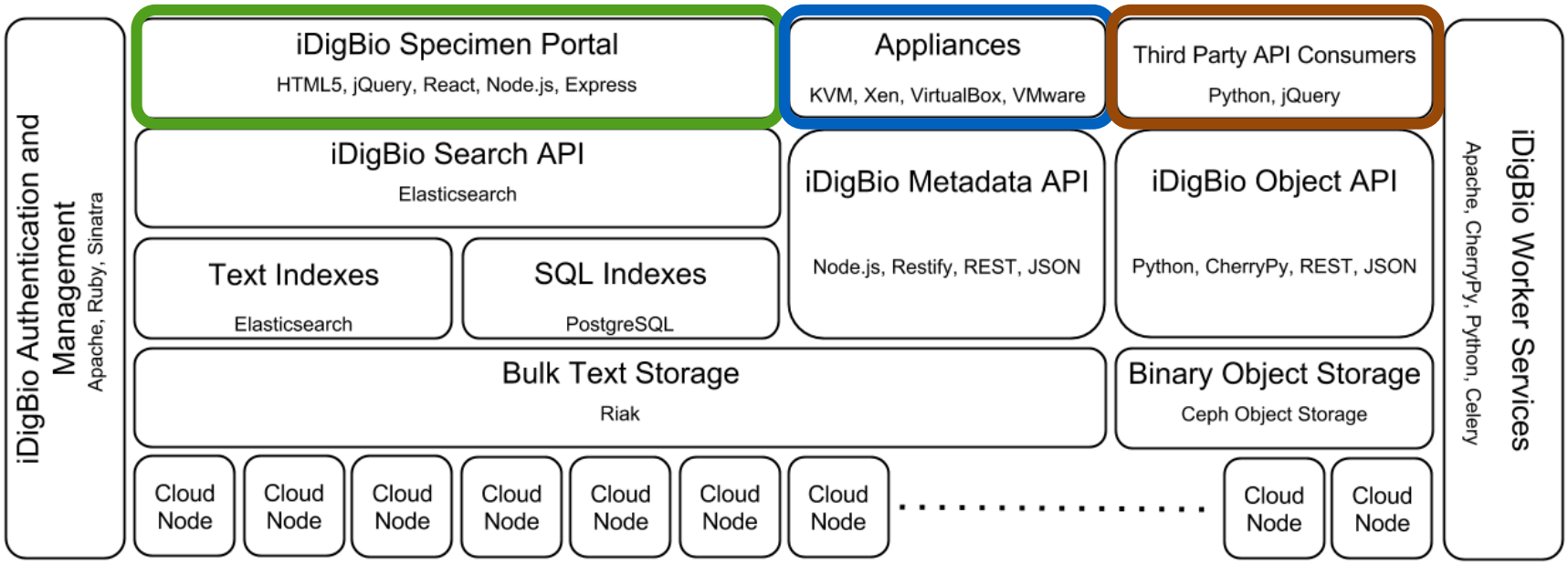
The what and how of data ingestion



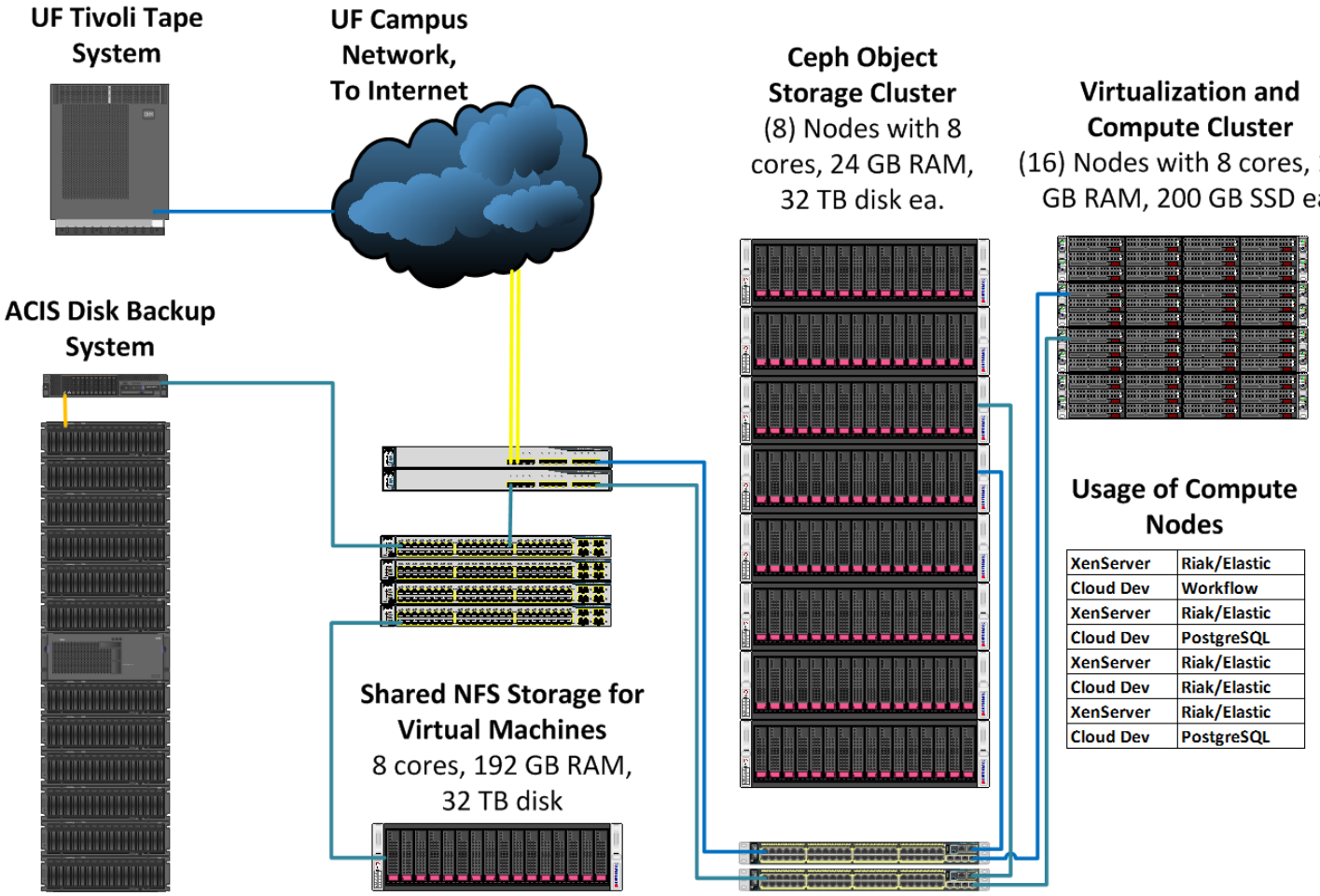
- IPT – RSS of DwC-A
  - Specify, EMu, Arctos, VertNet Migrator, etc.
- Symbiota portals – RSS of DwC-A
- iDigBio Feeder – DwC-A, CSV, …

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)

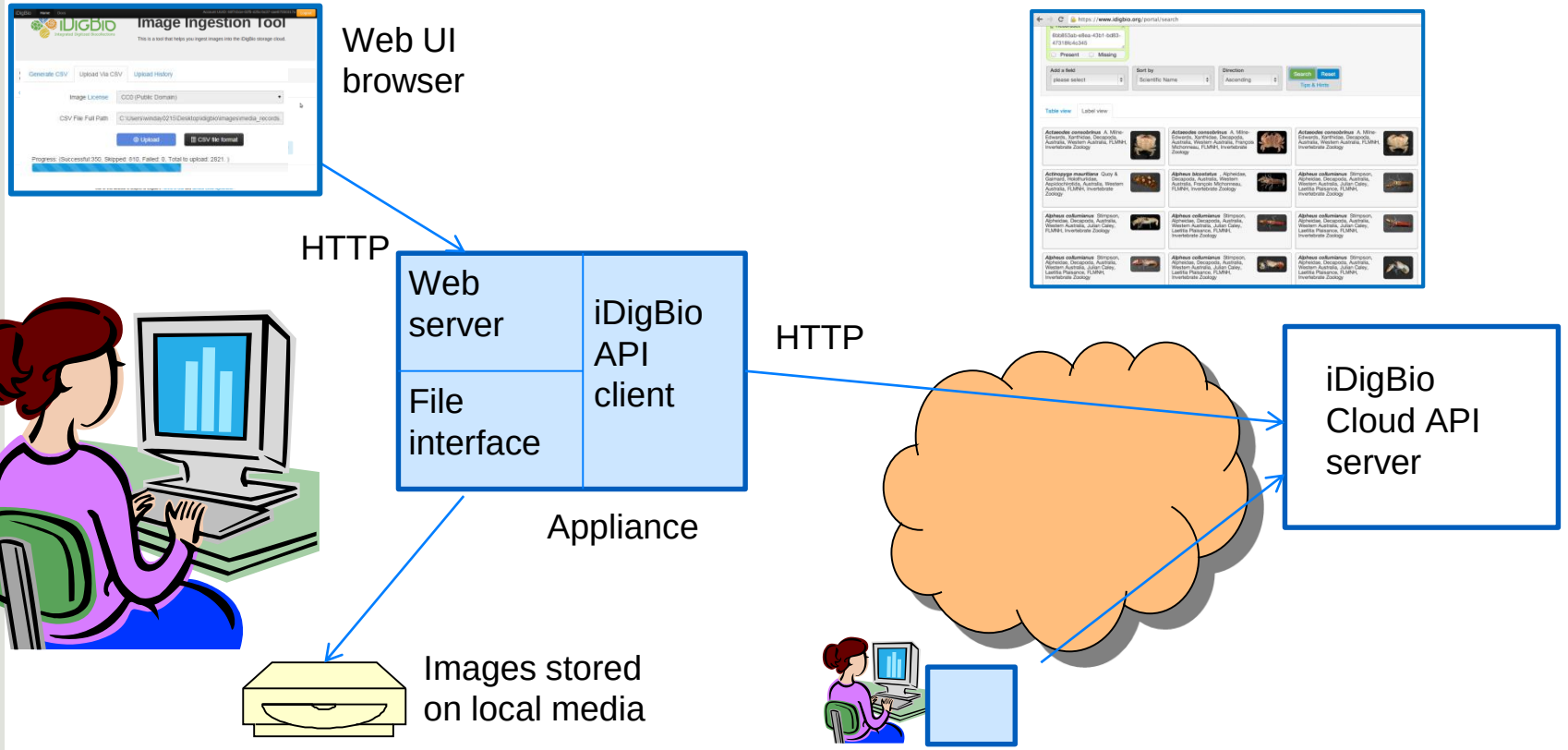


Usage of Compute Nodes

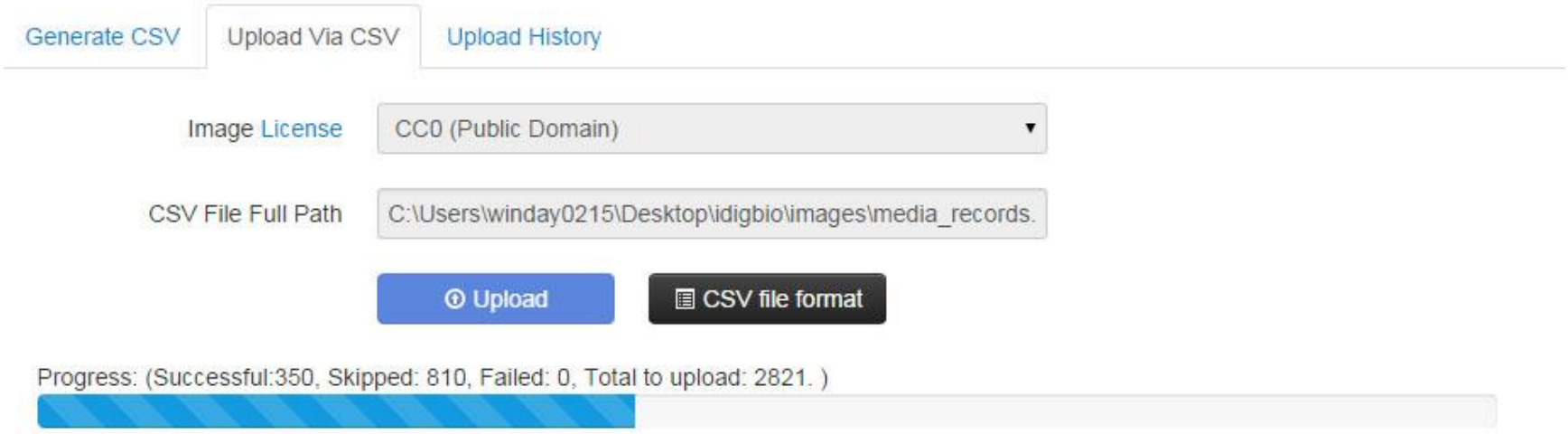
XenServer	Riak/Elastic
Cloud Dev	Workflow
XenServer	Riak/Elastic
Cloud Dev	PostgreSQL
XenServer	Riak/Elastic
Cloud Dev	Riak/Elastic
XenServer	Riak/Elastic
Cloud Dev	PostgreSQL

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.

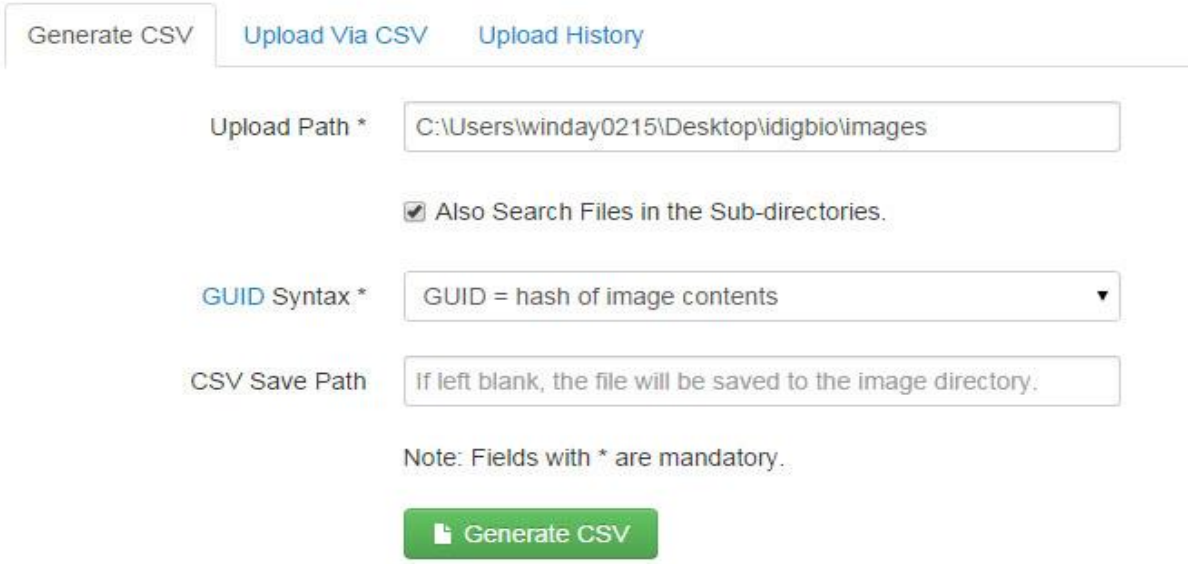


Upload Images with CSV File



- 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

Generate CSV | Upload Via CSV | Upload History

Batch Information Table

(Click on each row to see the details)

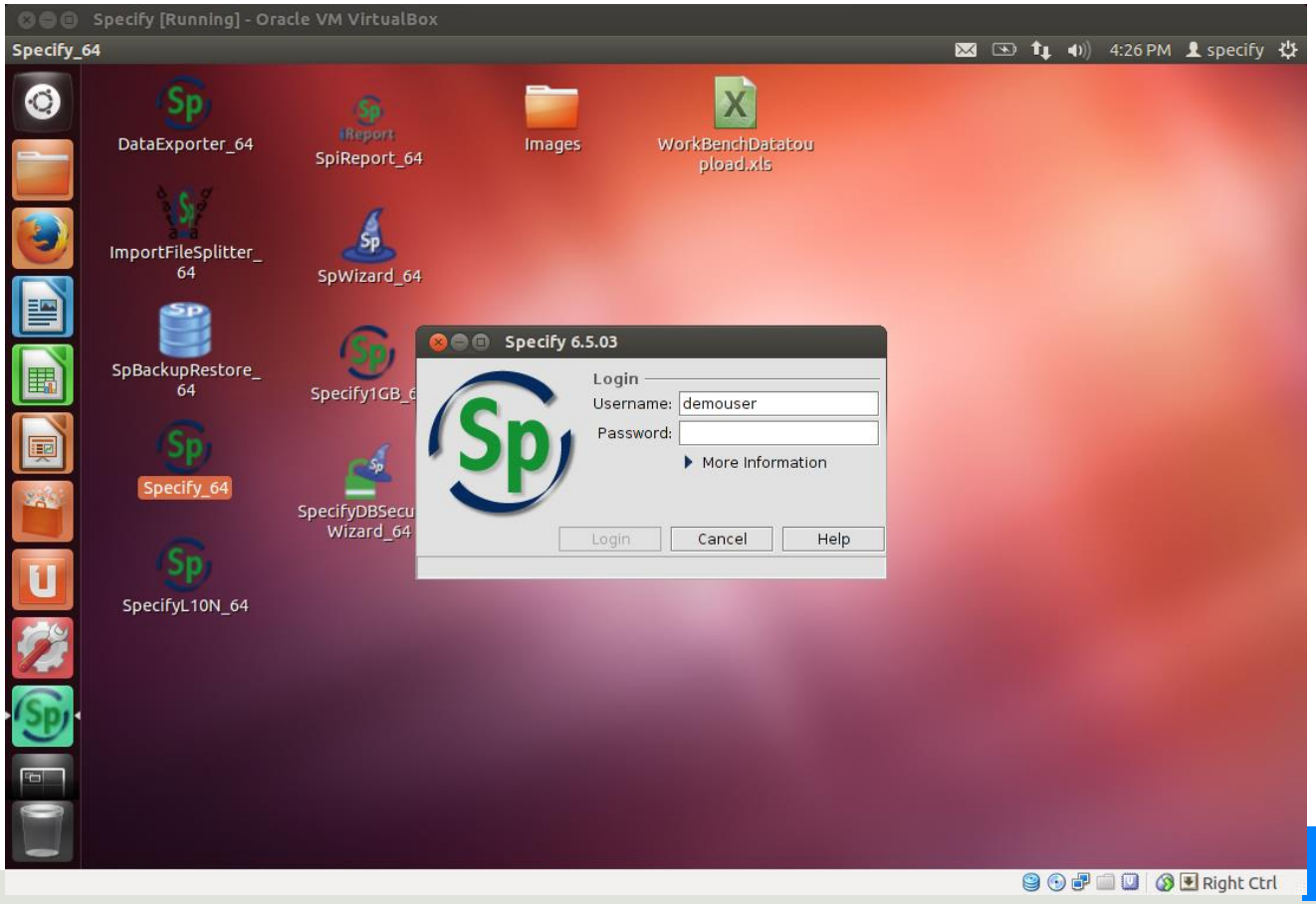
Show 10 entries ← Previous 1 Next →

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

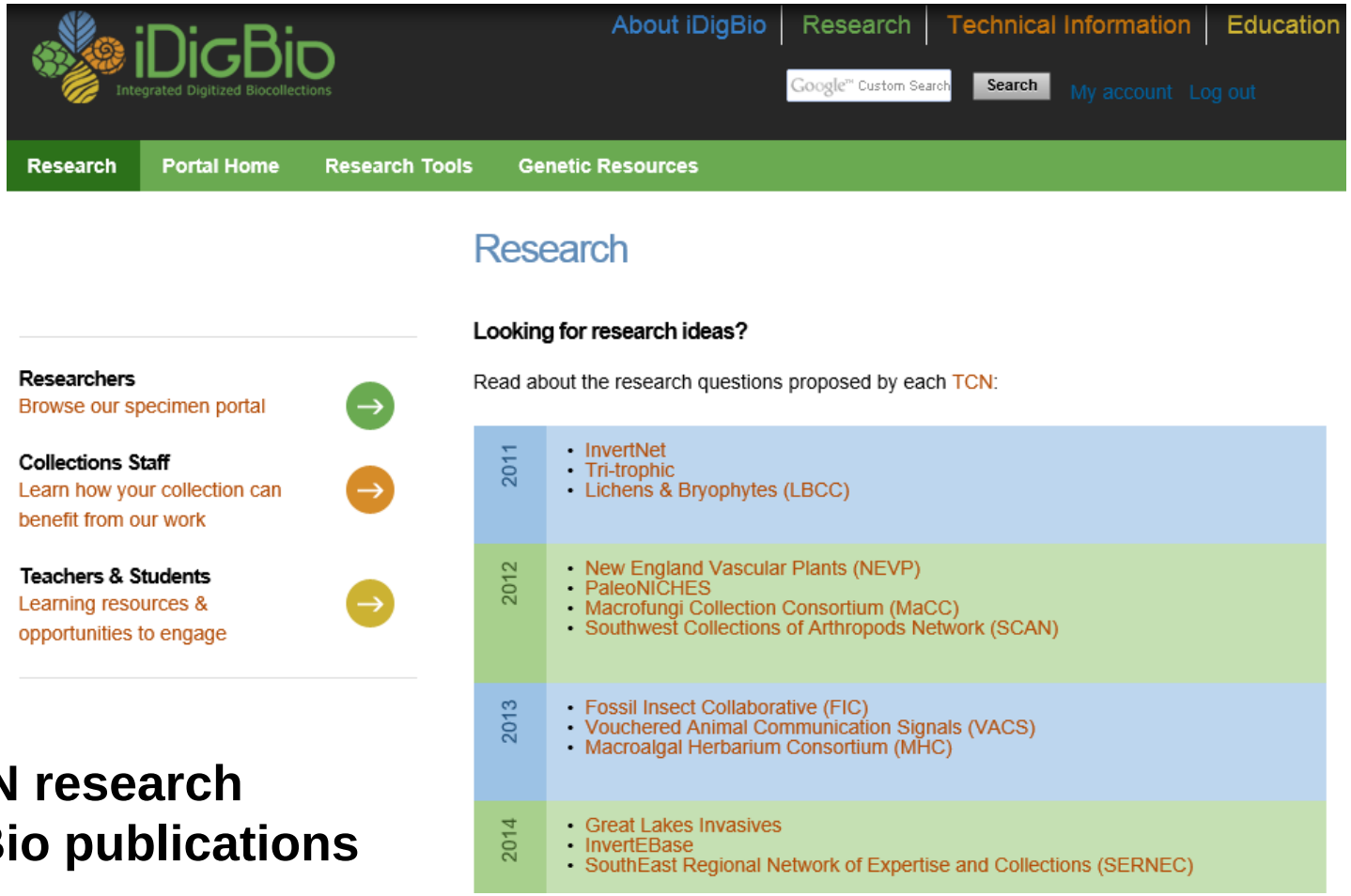
- The local upload history can be viewed/saved to CSV files
- Current upload results are also shown after each upload

Specify Appliance

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



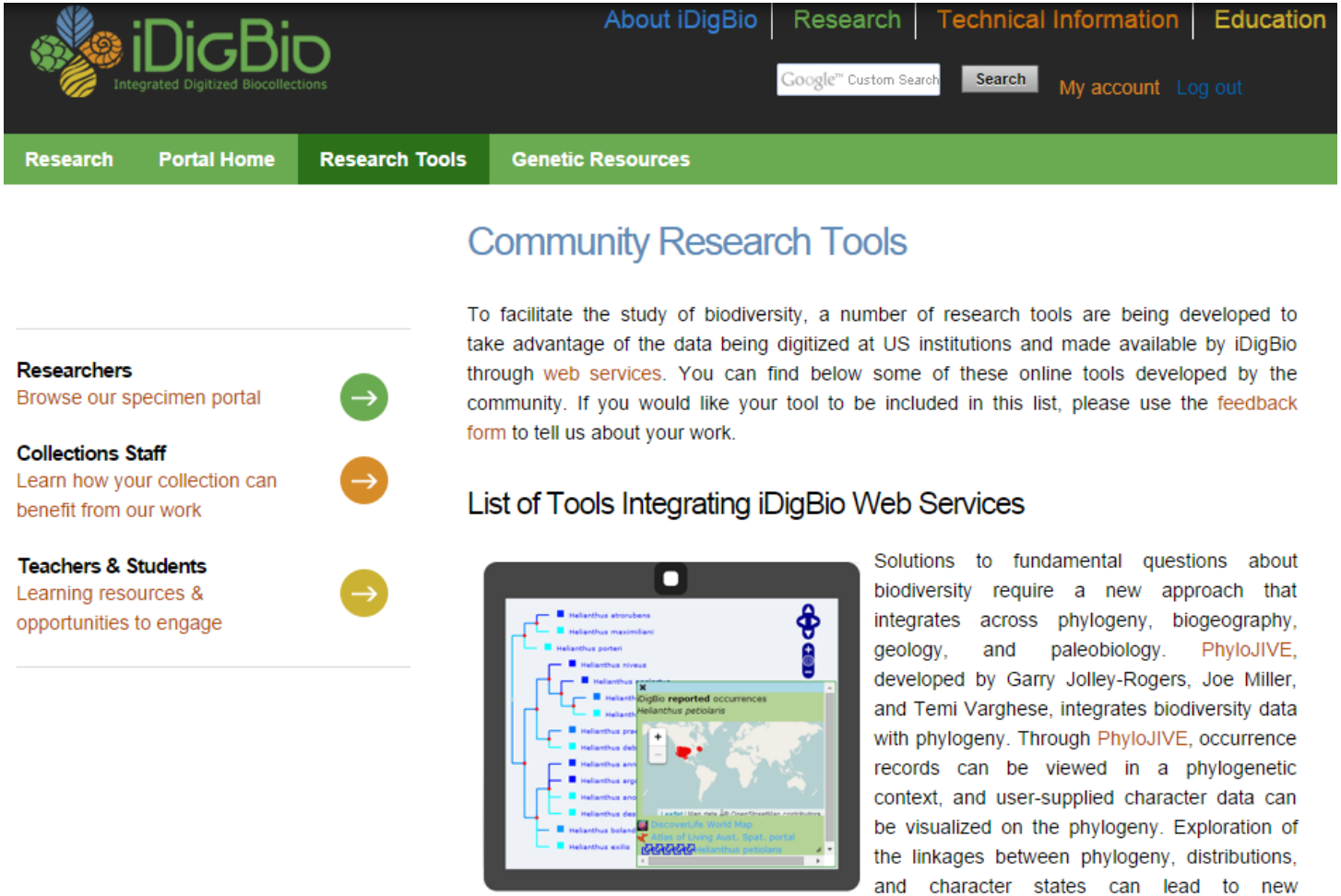
iDigBio Research Section



Links to TCN research
List of iDigBio publications

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

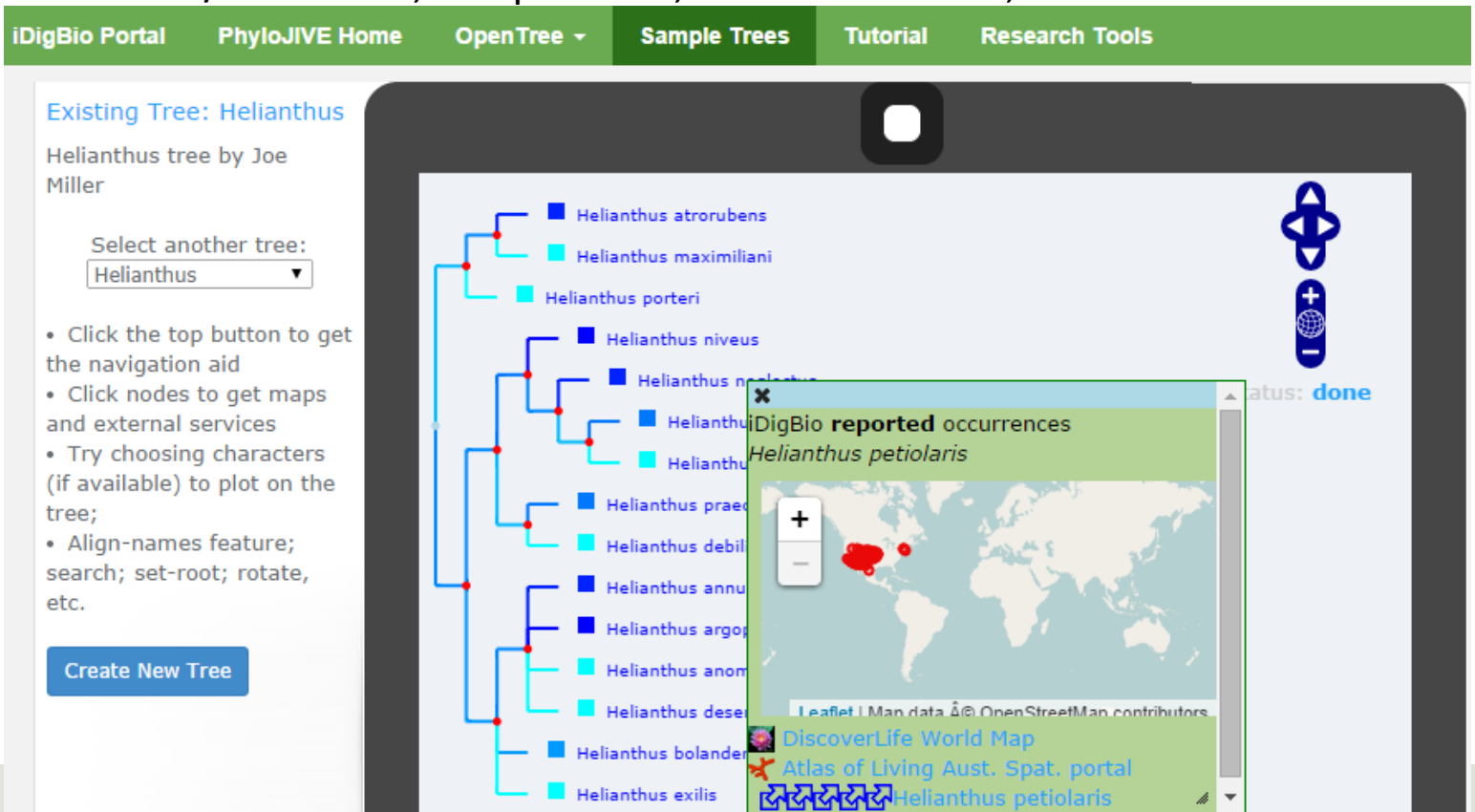
iDigBio Research Tools



- 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,…



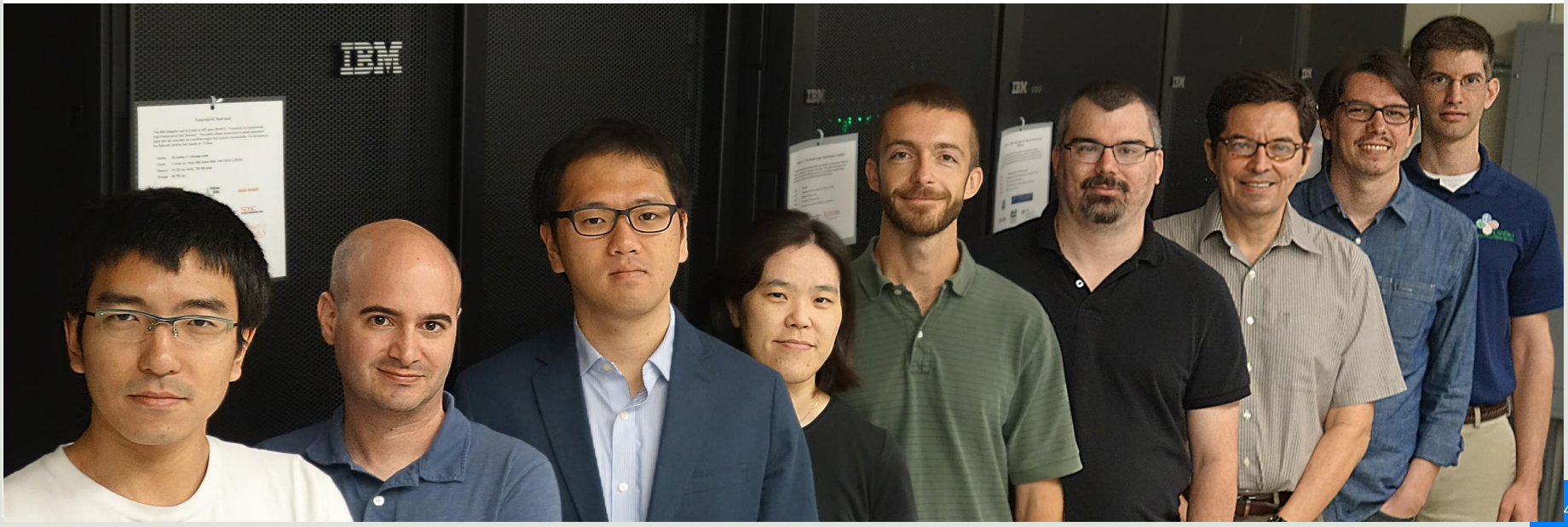
Research tools integrated with iDigBio

- 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