iDigBio
Integrated Digitized Biocollections

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

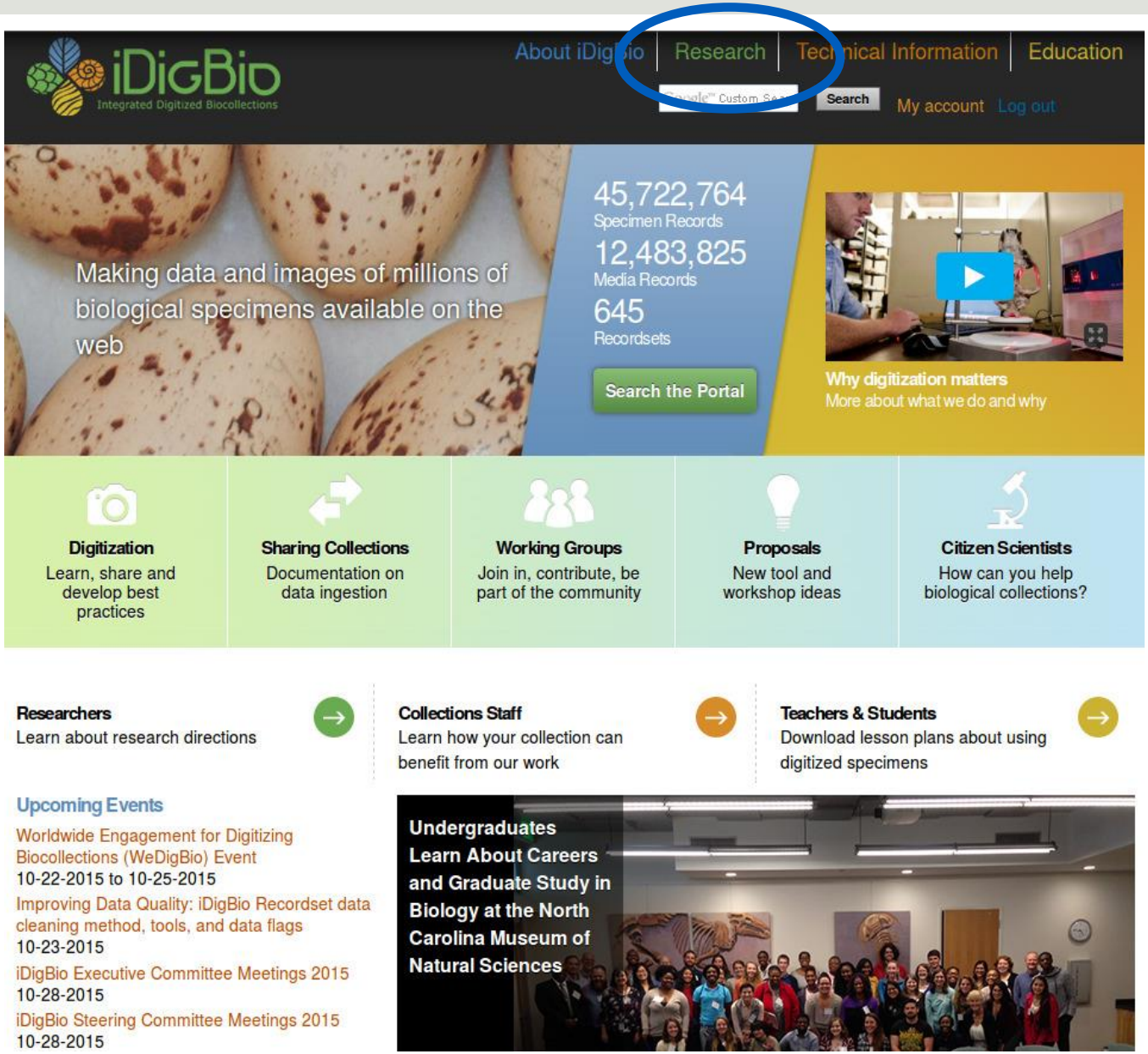
iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect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

Outline

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

http://datasupport.researchdata.nl/en/start-de-cursus/iv-gebruiksfase/data-reuse/

iDigBio Website



iDigBio Research Section

Research | Portal Home | Research Tools | Genetic Resources

Researchers
Browse our specimen portal

Collections Staff
Learn how your collection can benefit from our work

Teachers & Students
Learning resources & opportunities to engage

Research

Looking for research ideas?

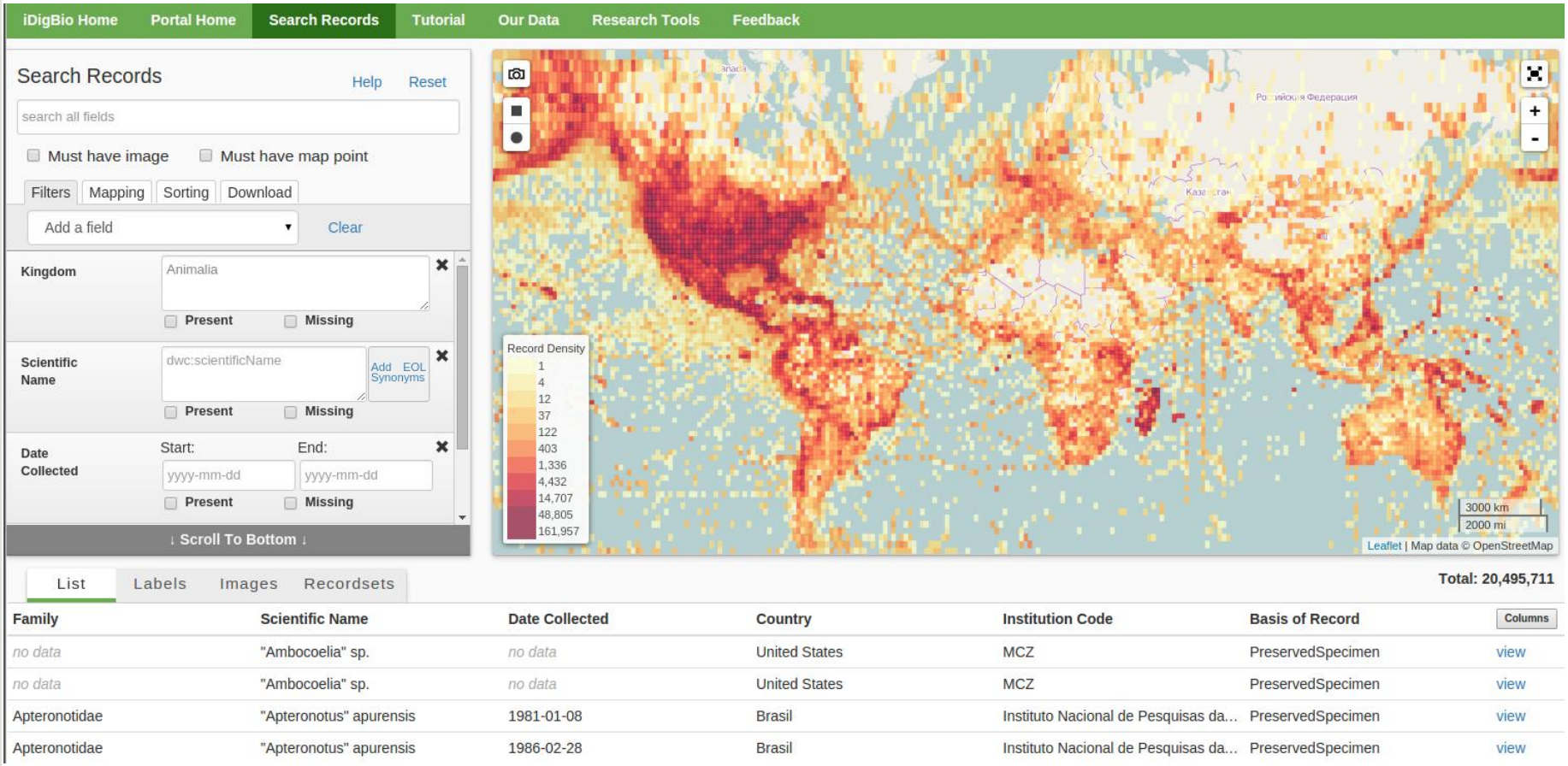
Read about the research questions proposed by each National Science Foundation funded Thematic Collections Network (TCN):

2011	• InvertNet - Watch highlights • Tri-trophic - Watch highlights • Lichens & Bryophytes (LBCC) - Watch highlights
2012	• 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	• Fossil Insect Collaborative (FIC) - Watch highlights • Vouchered Animal Communication Signals (VACS) - Watch highlights • Macroalgal Herbarium Consortium (MHC) - Watch highlights
2014	• Great Lakes Invasives (GLI) - Watch highlights • InvertEBase - Watch highlights • SouthEast Regional Network of Expertise and Collections (SERNEC)
2015	• 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 autocompletion, synonyms,…



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	Abacetus rufitarsis	1934-07-30	India	MCZ	PreservedSpecimen	view
Carabidae	Abacetus straneoi	1944-03-01/1944-07-31	Papua New Guinea	MCZ	PreservedSpecimen	view
Carabidae	Abacidus permundus	1878-05-01/1878-05-31	United States	MCZ	PreservedSpecimen	view
Abacionidae	Abacion	2014-09-01	United States	MCZ	PreservedSpecimen	view

List | Labels | Images | Recordsets

Total: 280,461

Abacetus rufitarsis Straneo — *Carabidae*
India, Puducherry Territory, Karaikal District, Nedungadu
MCZ, Ent, 29281, P. S. Nathan
Animalia, Arthropoda, Insecta, Coleoptera — 1934-07-30

Abacetus straneoi Darlington (1962) — *Carabidae*
Papua New Guinea, Oro Province, Dobodura
MCZ, Ent, 30218, Philip Jackson Darlington, Jr.
Animalia, Arthropoda, Insecta, Coleoptera — 1944-03-01/1944-07-31

Abacidus permundus — *Carabidae*
United States, Illinois, Richland & Lawrence Co., Wabash Valley
MCZ, Ent, 33042, [no agent data]
Animalia, Arthropoda, Insecta, Coleoptera — 1878-05-01/1878-05-31

Abacion (Loomis, 1937) — *Abacionidae*
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 Garcia
Animalia, Arthropoda, Diplopoda, Callipodida — 2014-09-01

Abaeis nicippe (Cramer, 1779) — *Pieridae*
Mexico, Tamaulipas, Victoria
MCZ, Ent, 27382, Turner, Donald B. Stallings, Sr.
Animalia, Arthropoda, Insecta, Lepidoptera — 1941-06-10

Abaeis nicippe (Cramer, 1779) — *Pieridae*
USA, Arizona, Maricopa County, Phoenix
Lat: 33°26' 53" **Lon:** -112°4' 23"
ASU, ASUHIC, ASUHIC0079342, Roman S. Wielgus
Animalia, Arthropoda, Insecta, Lepidoptera — 1963-09-01

List | Labels | Images | Recordsets

Total: 280,461

1 of 3 — ***Abacetus rufitarsis*** MCZ, Ent
2 of 3 — ***Abacetus rufitarsis*** MCZ, Ent
3 of 3 — ***Abacetus rufitarsis*** MCZ, Ent
1 of 3 — ***Abacetus straneoi*** MCZ, Ent
2 of 3 — ***Abacetus straneoi*** MCZ, Ent

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

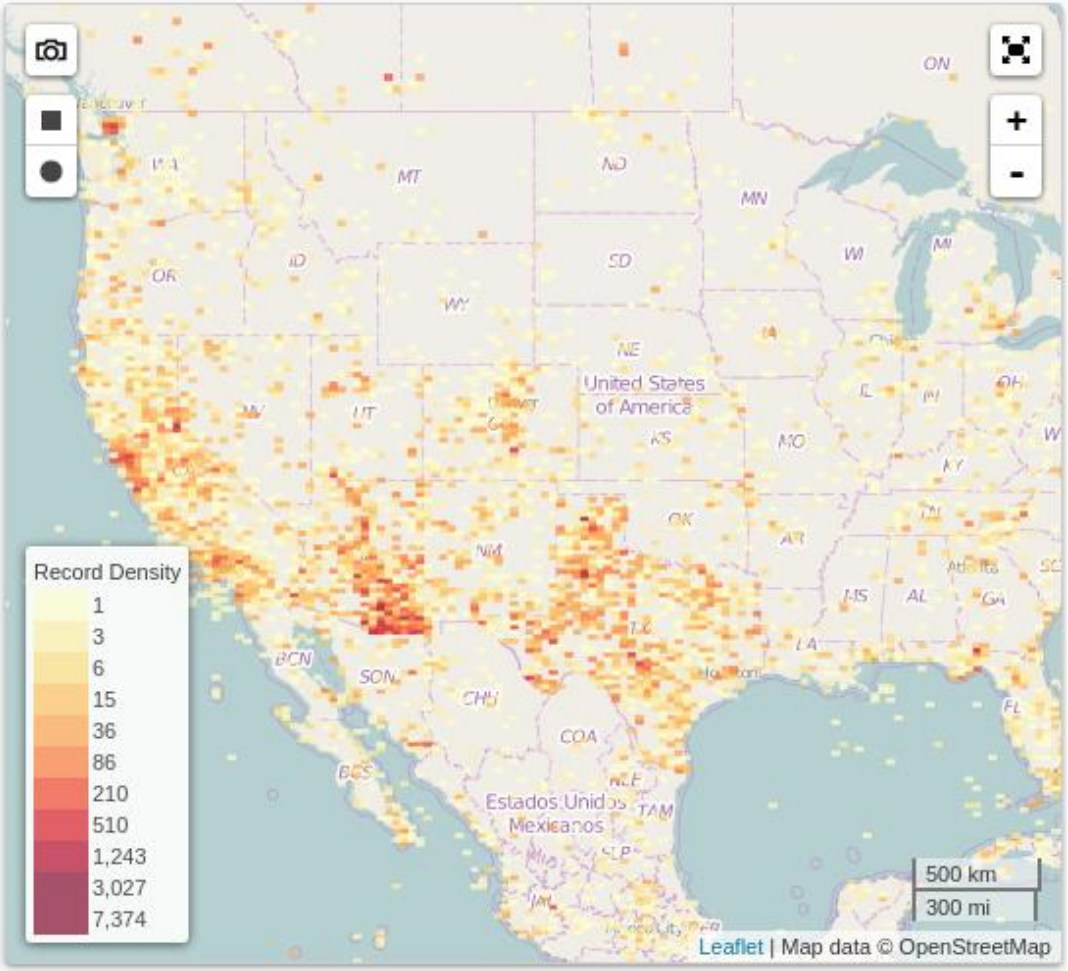
Image is present. Kingdom = Animalia. S

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

Email godfoder@acis.ufl.edu

Downloads

Search	Status
is present. Kingdom = Animali…	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	Institution Code	FLMNH
Country	United States	Collection Code	Invertebrate Zoology
State/Province	Washington	Catalog Number	952 Annelida
County/Parish	San Juan County	Collected By	G Paulay
Latitude	48.483333333333334		
Longitude	-123.06944444444444		

Media

Contents

- Summary
- Map
- Media
- Attribution
- All Data

Data | Flags | Raw

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

	Records			Media		
Publisher Name	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 -20ºC 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 rhrobins@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

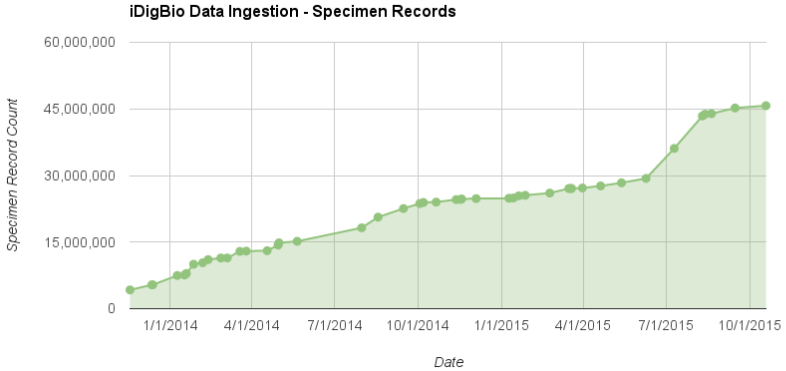
Data Corrected | Data Use | Raw

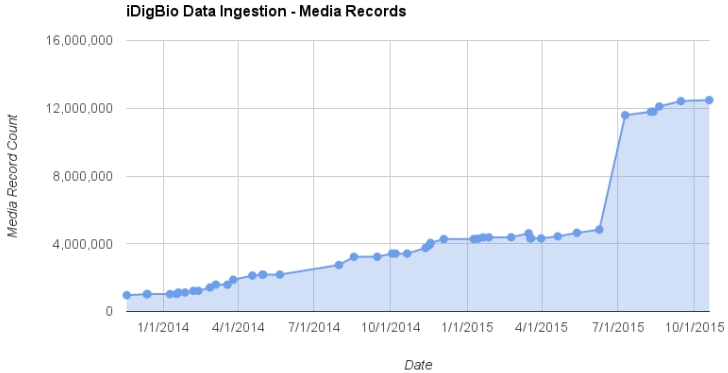
Month of	Search	Download	Seen	Records Viewed	Media Viewed
01 / 2015	52,169,787	224,080	733	1,877	0
02 / 2015	160,296,710	223,895	2,372	1,035	0
03 / 2015	84,959,293	440,641	455	1,173	0
05 / 2015	257,737,259	213,932	2,486	7,539	0
06 / 2015	886,140,576	986,892	5,521	16,490	0
07 / 2015	2,446,086,020	221,546	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

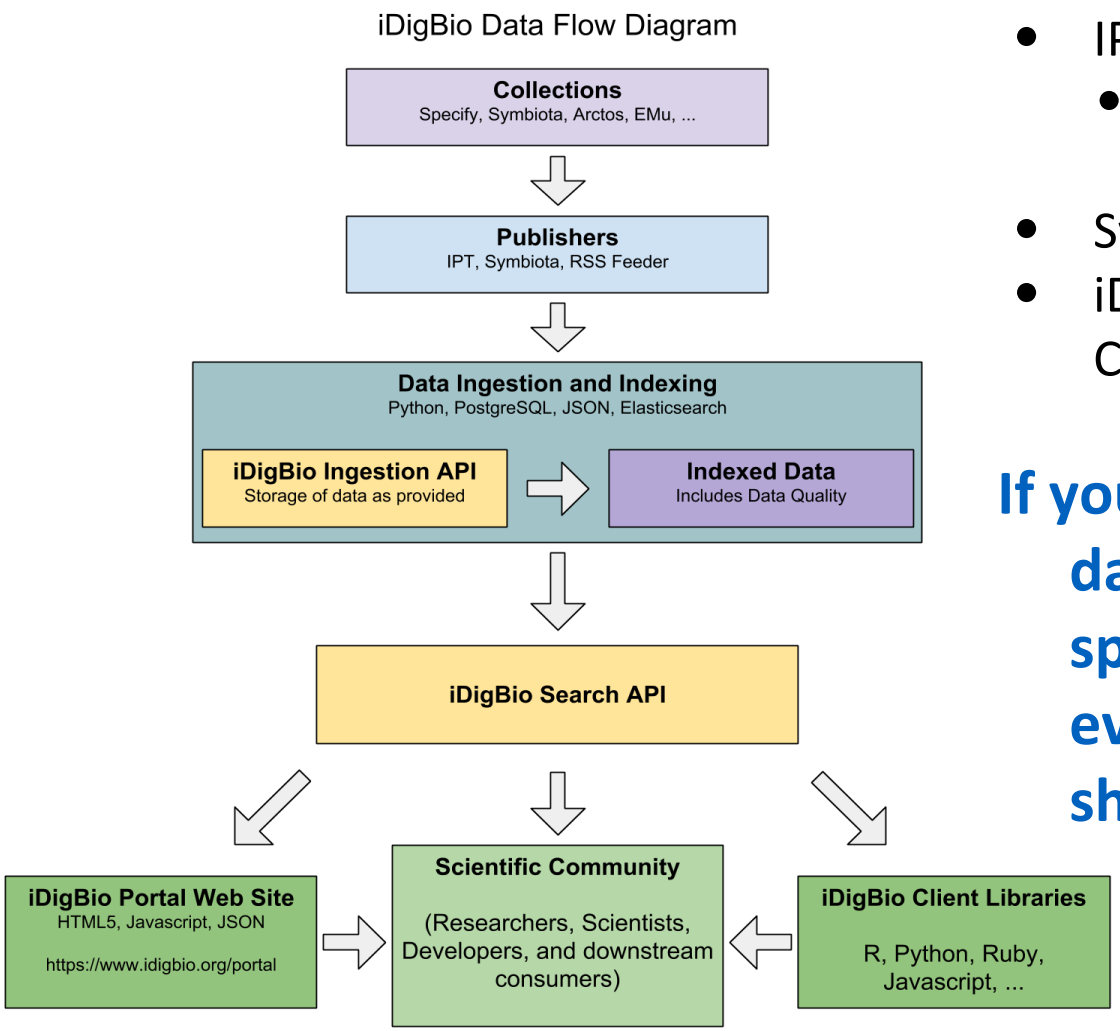
Publishing technologies: IPT, Symbiota, RSS (DwC-a, CSV)

Media data using Audubon Core terms





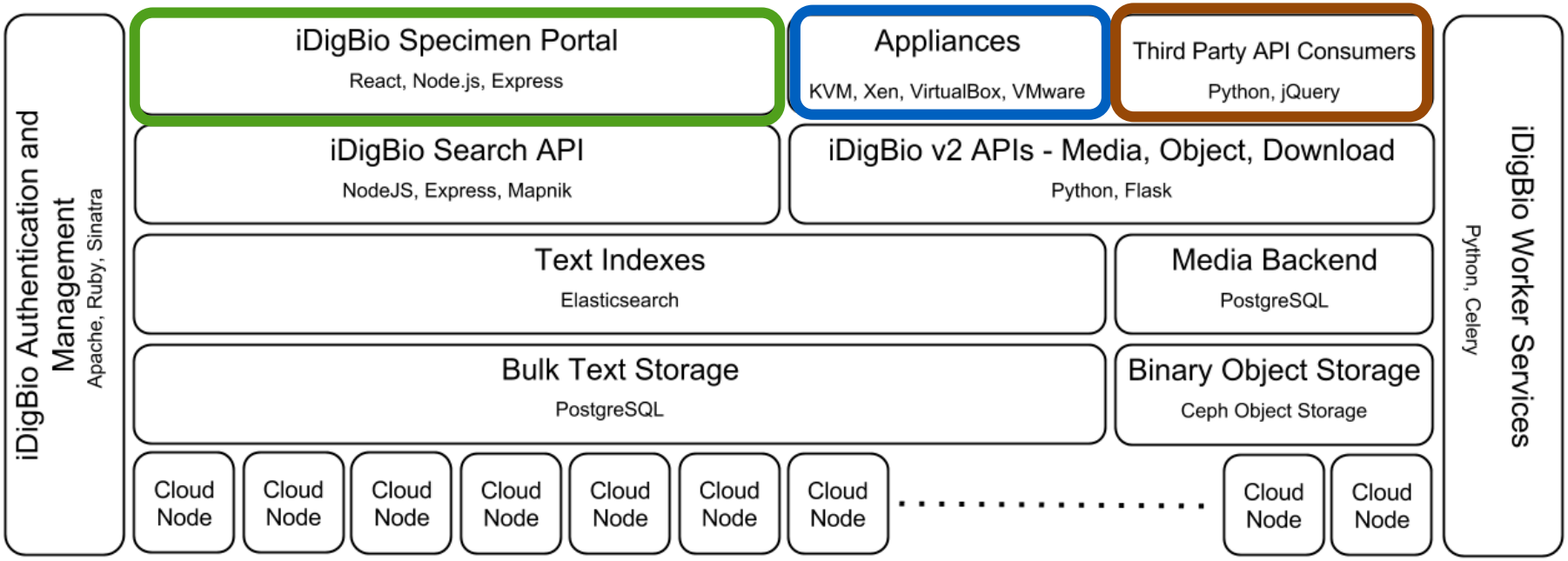
The what and how of data ingestion



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

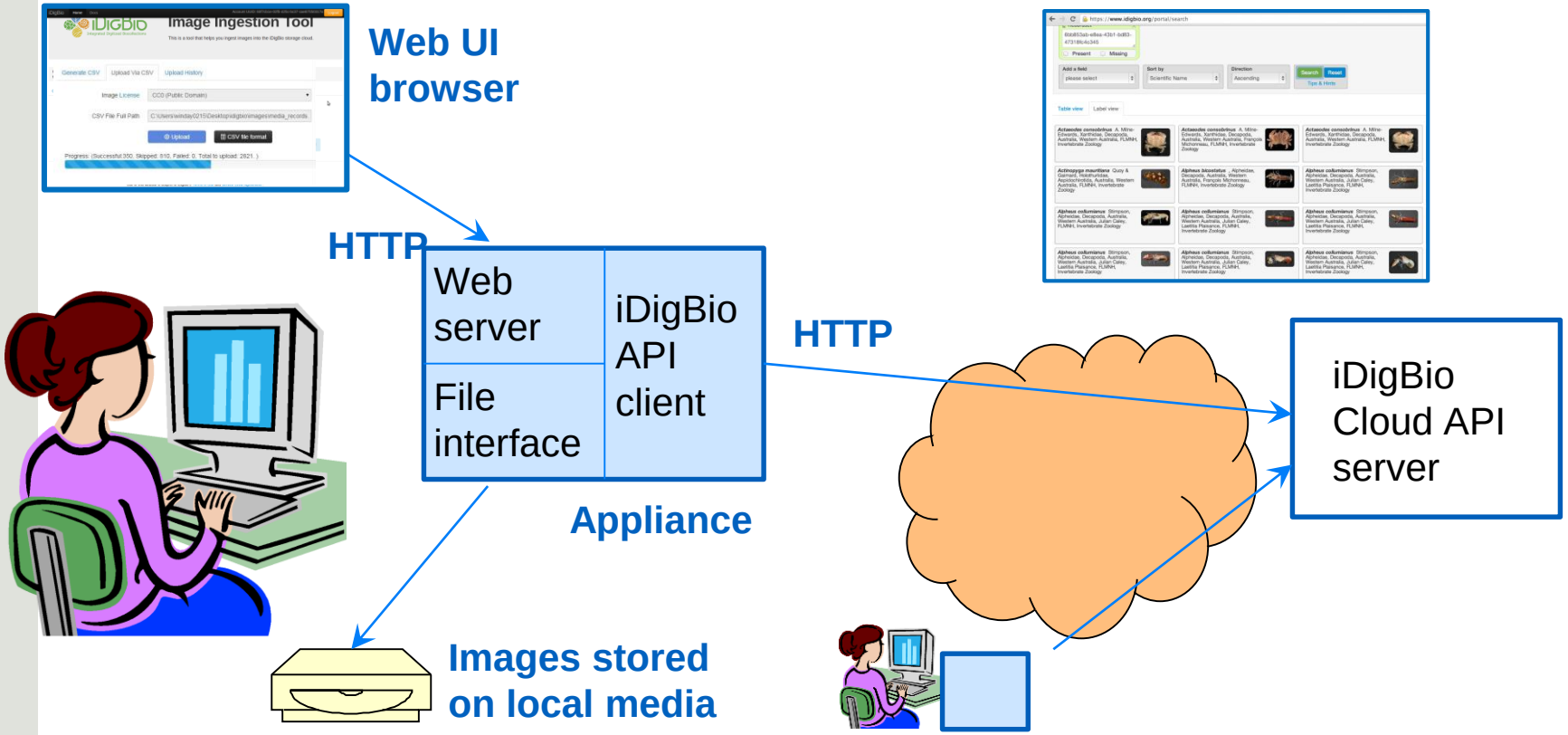
If you can export specimen data from your database/ spreadsheet into DwC-A (or even CSV), then you can share data with iDigBio.

Architecture Components



Appliances

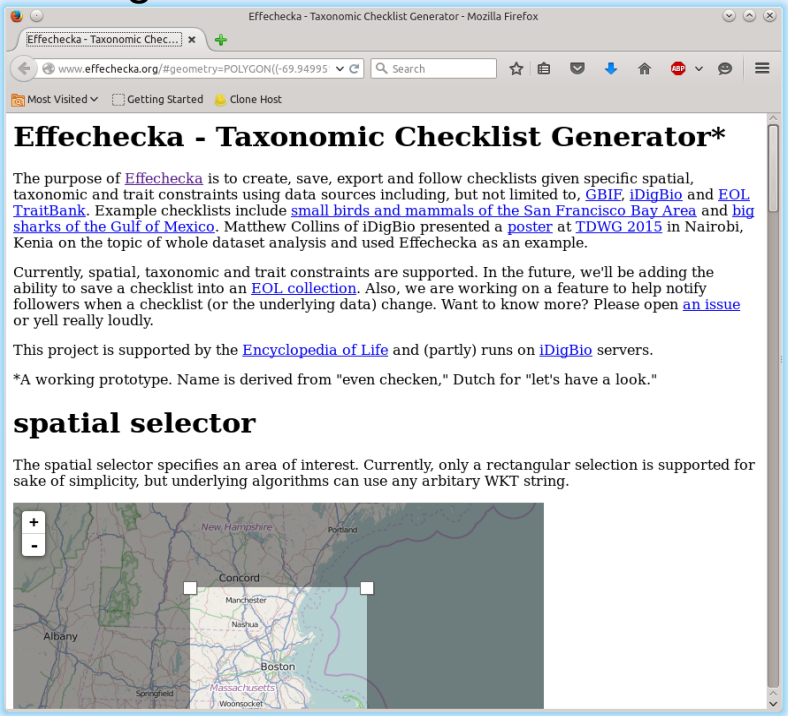
- Image upload appliance - reliable upload of images+metadata batches
  - users include NYBG and SCAN (Symbiota) - NAU, UHIM, UC Boulder
- Specify appliance
  - used in 10 training workshops, 4 countries; ~200 people/~25 institutions

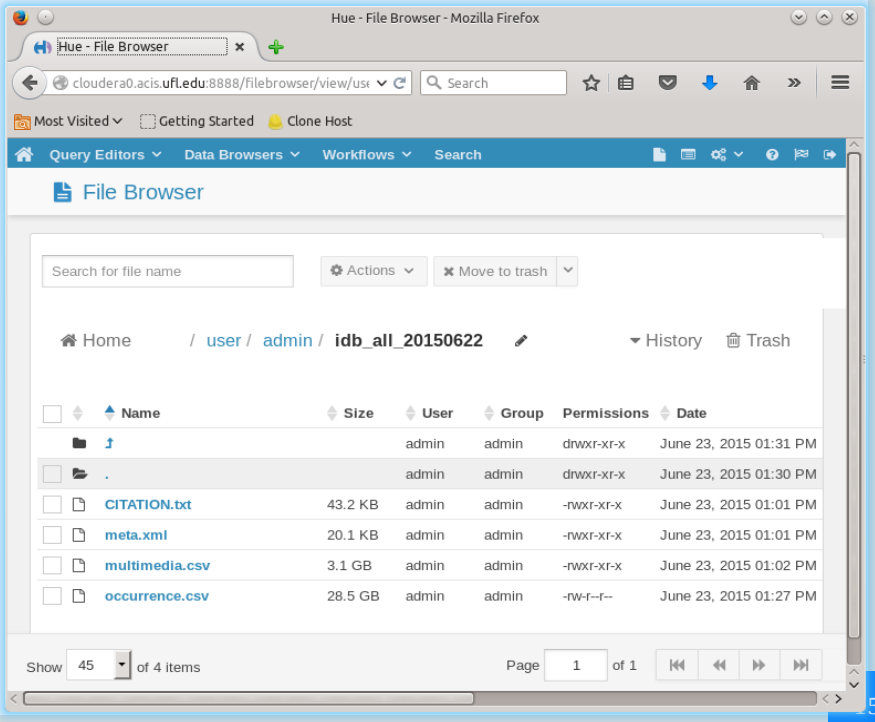


Emerging Research Tools from iDigBio and the Community

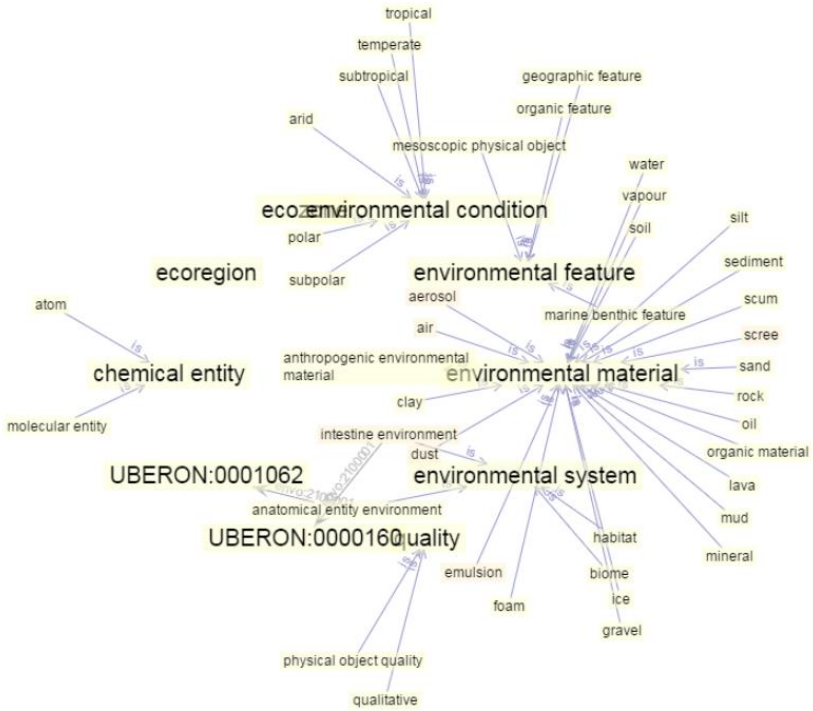
Checklist generation & "Spark" data processing

Jorrit Poelen (GLOBI) &
Jen Hammock
working with EOL

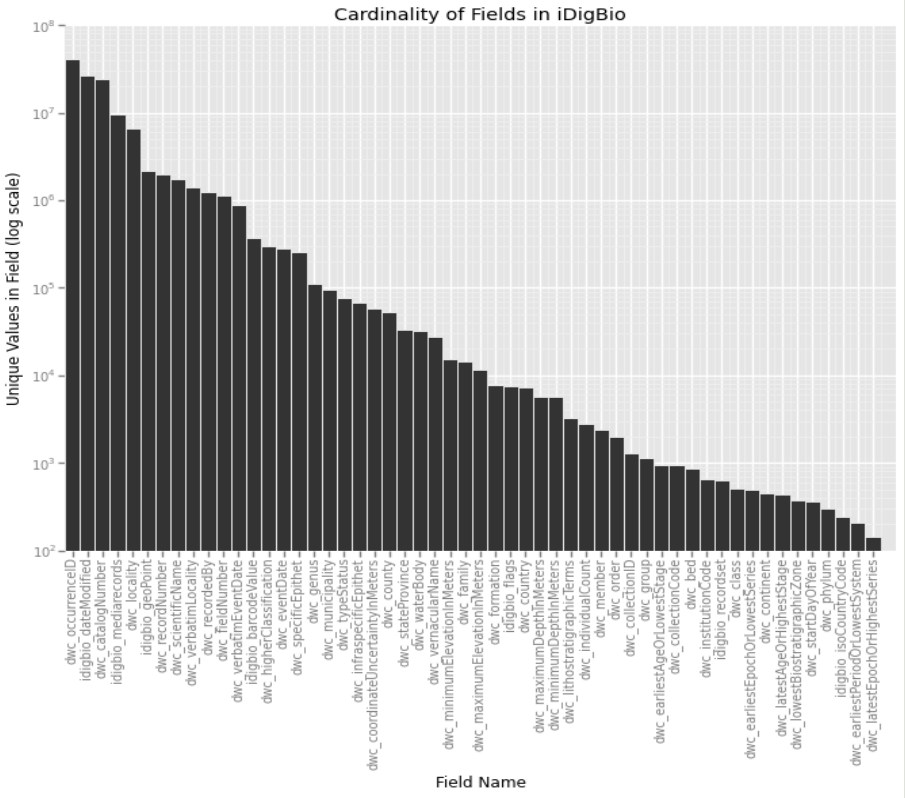




Data Characterization - Examining iDigBio Data



Grant Godden and **Pier Luigi Buttigieg** (Phenotype RCN) processed label data from over one million plant specimen records hosted by iDigBio …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

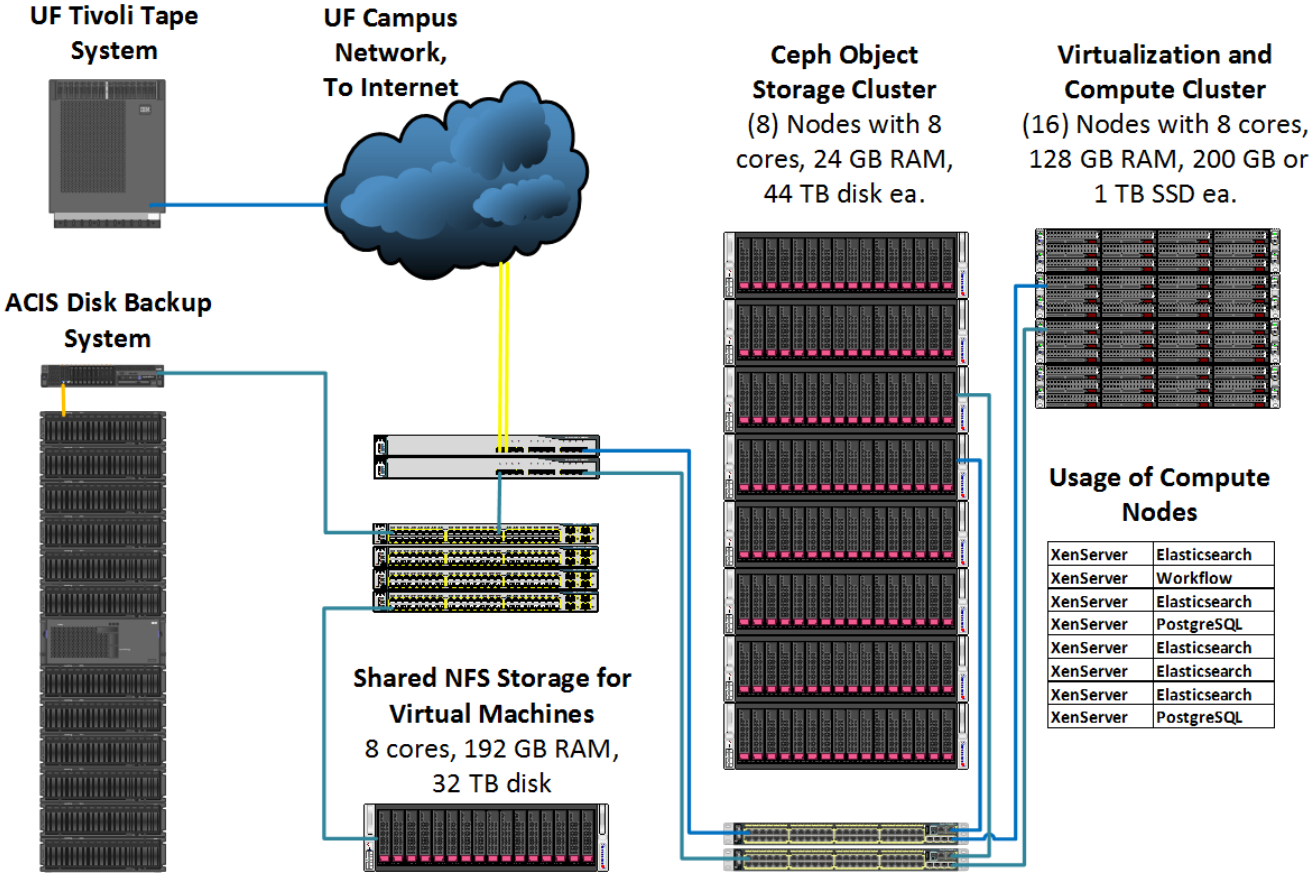
Demonstrated integrations of research tools with iDigBio

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

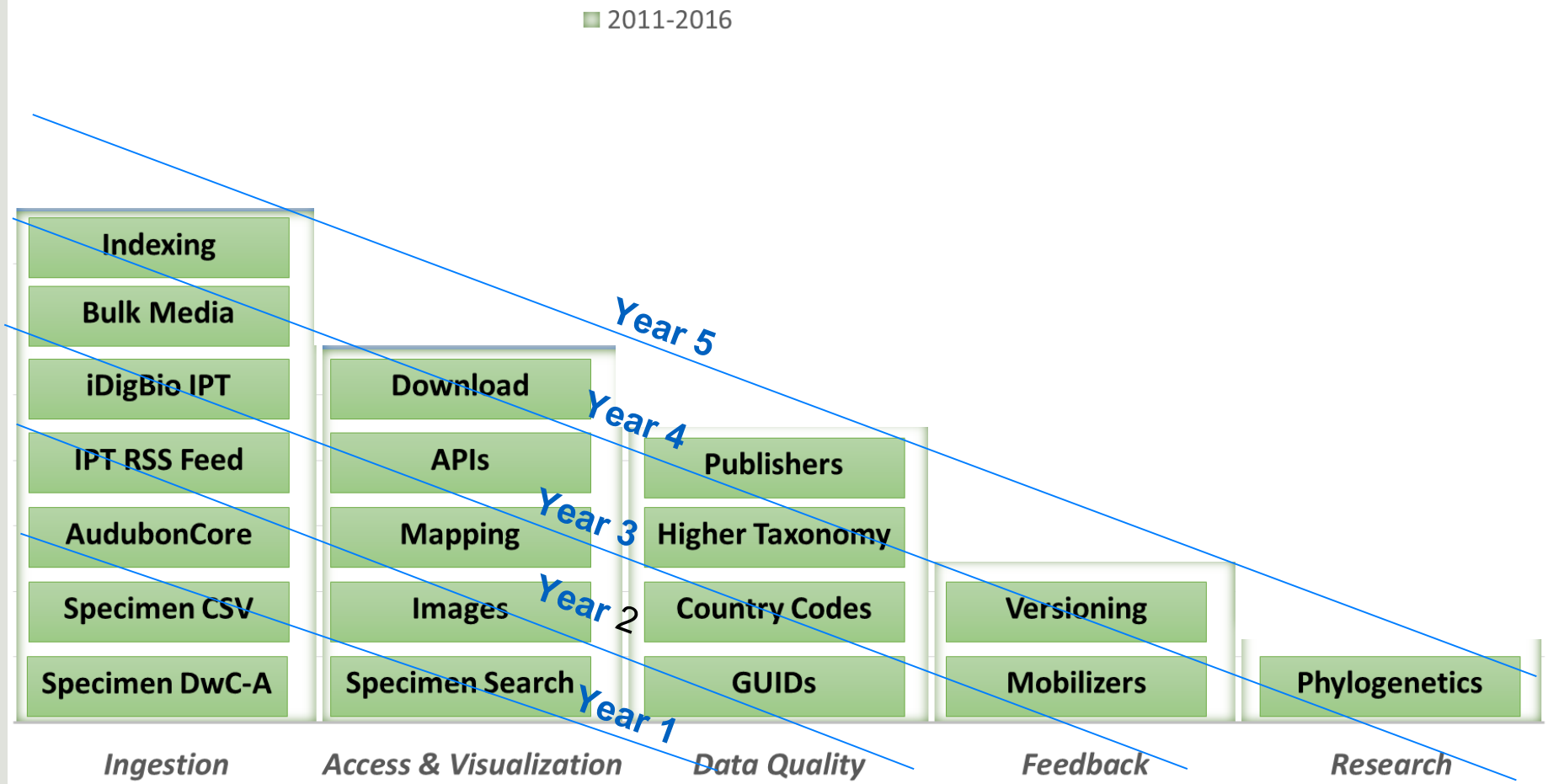
- Presented at the last Summit and SPNCH 2015
- Contact idigbio@acis.ufl.edu if you are interested in integration of your research tool(s)

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



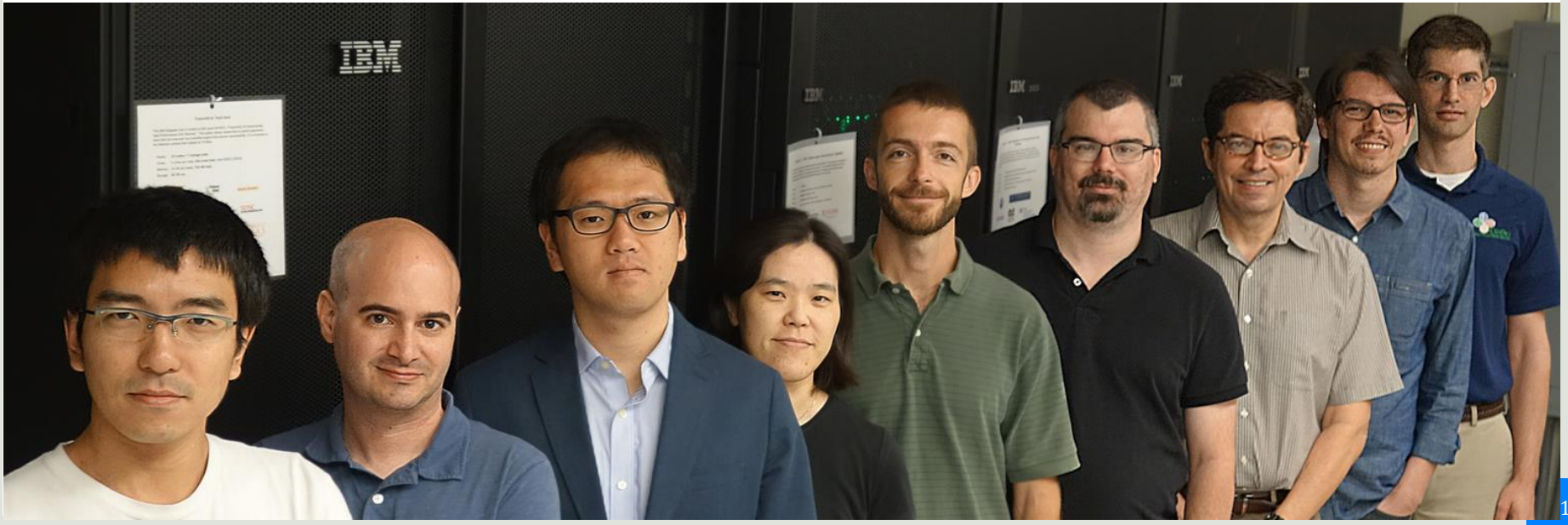
iDigBio ACTIVITIES TO BE SUPPORTED BY CYBERINFRASTRUCTURE

2011-2016 2016-2021

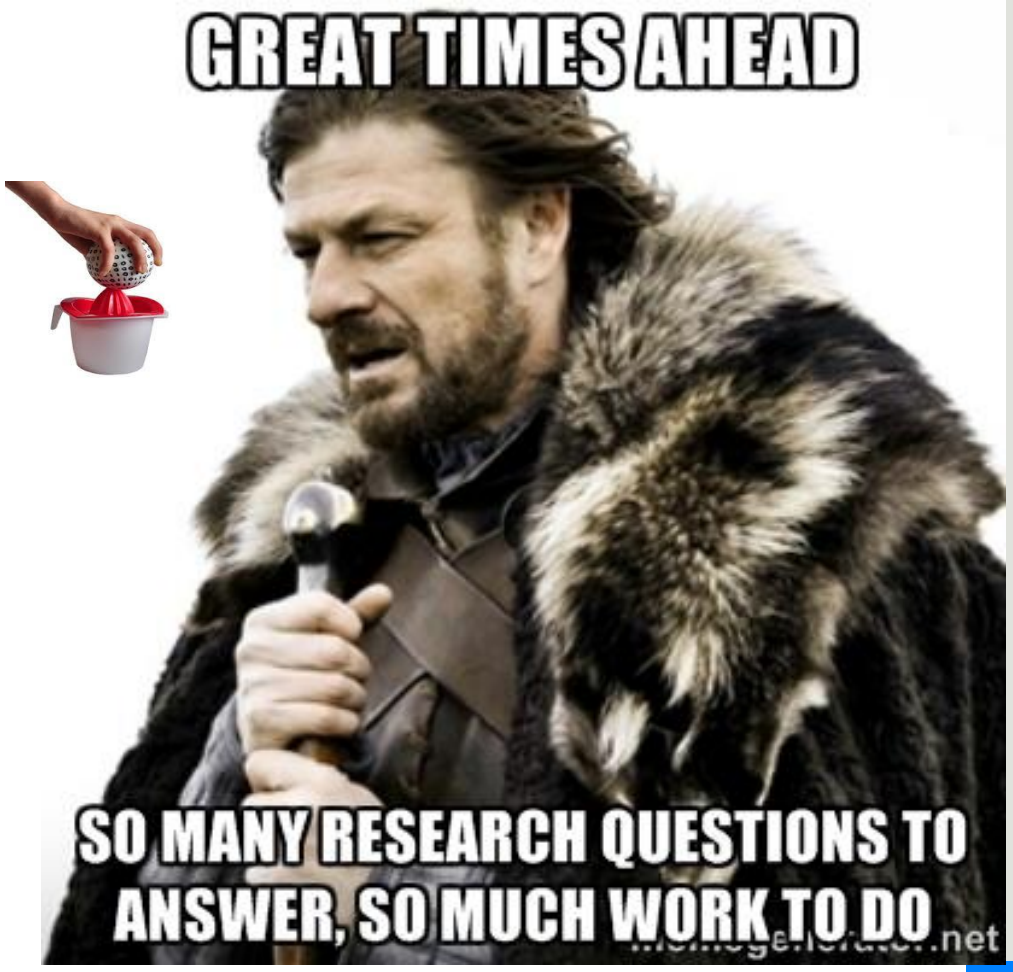
Ingestion	Access & Visualization	Data Quality	Feedback	Research
Relationships	Trees	Metrics	Difference Viewer	Agriculture
Self-service	3D-images	Correlation	Filters	Functional
Indexing	Sound	Outliers	Summary	Molecular
Bulk Media	Statistics	Pattern Analysis	Bulk updates	Ecosystem
iDigBio IPT	Download	Vocabularies	Annotation	Trait Evolution
IPT RSS Feed	APIs	Publishers	Networking	Communities
AudubonCore	Mapping	Higher Taxonomy	Adapters	Medicinal
Specimen CSV	Images	Country Codes	Versioning	Niche Modeling
Specimen DwC-A	Specimen Search	GUIDs	Mobilizers	Phylogenetics

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

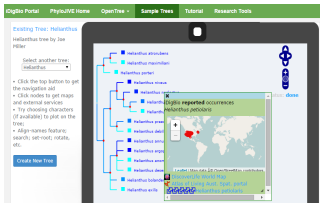


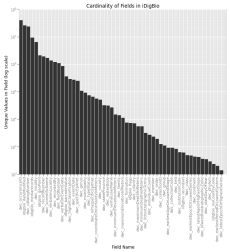
http://memegenerator.net/instance/56576771

Acknowledgements

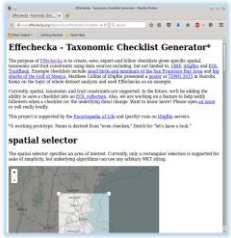
APIs and Client Libraries Under it All

Appliances

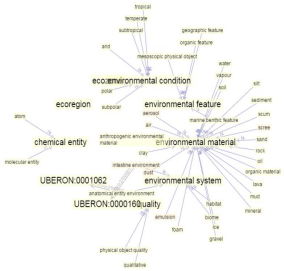




Portal



Original Research



Tools

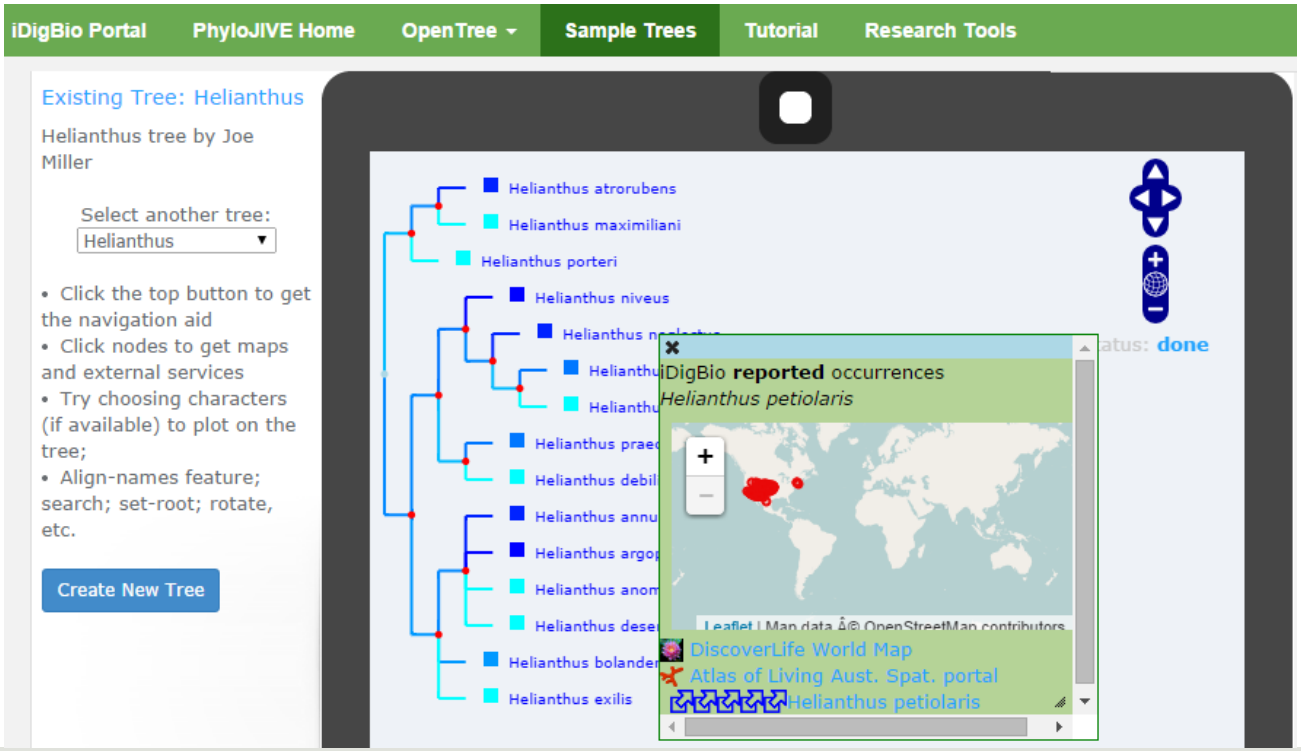

```
{
    "items" : [
        {
            "data" : {
                "dwc:eventDate" : "1968-04-11",
                "dwc:recordedBy" : "Harris, Arthur H. (Van Goebel)",
                "dwc:occurrenceStatus" : "present",
                "dwc:catalogNumber" : "5-298",
                "dwc:maximumElevationInMeters"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

LifeMapper instance at iDigBio

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

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,…



iDigBio Research Tools

- 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