

iDigBio Cyberinfrastructure, Partnerships, and Data

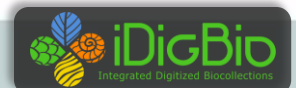
José Fortes

iDigBio Summit 3

11/19/2013



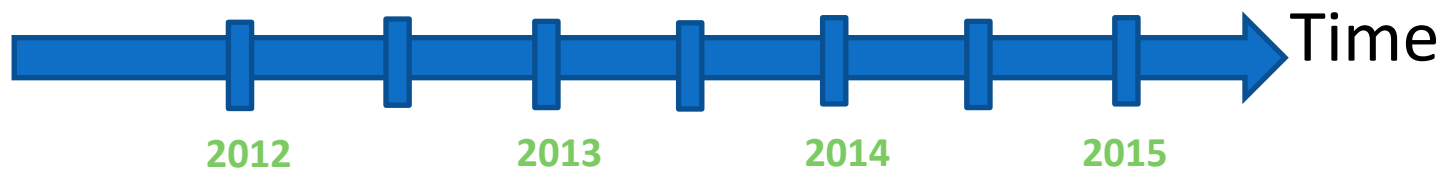
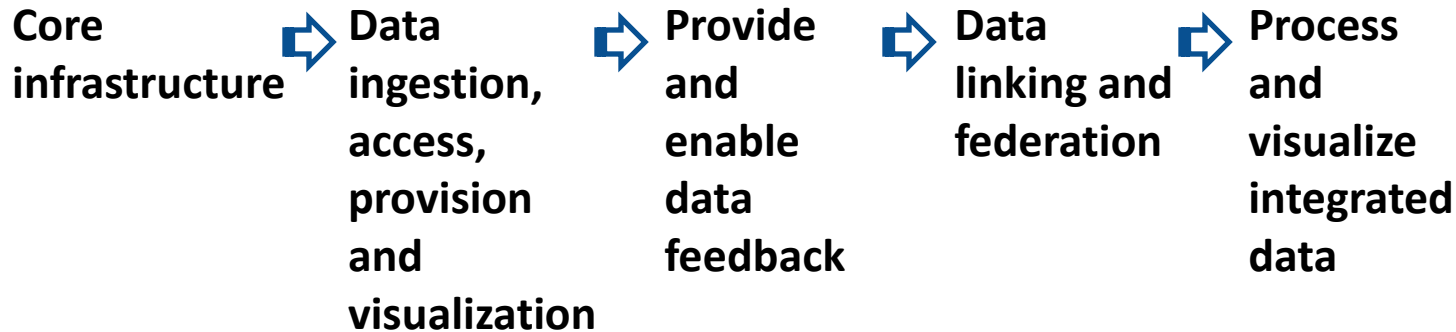
This material is based upon work supported by the National Science Foundation under Cooperative Agreement EF-1115210. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Outline

- Cyberinfrastructure
 - Status and capabilities
 - Futures
- Partnerships
- Data
- Conclusions
- Back-up slides

Evolution of iDigBio capabilities



Increasing storage and server hosting in support of the above

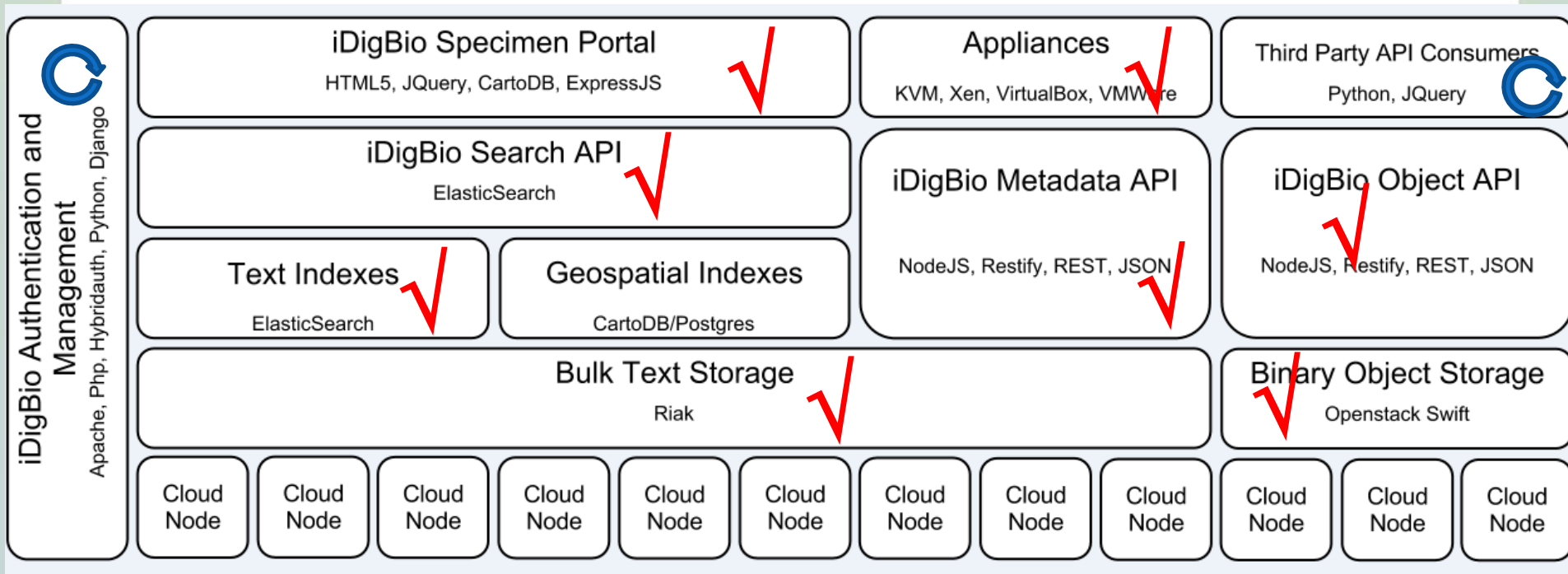
Increasing number of appliances in support of the above

Web site for interaction with public, community, education and above

- Ongoing development and deployment of improvements to the existing base infrastructure, and protocols for data ingestion, data provisioning and data visualization

Current iDigBio cloud architecture

- ✓ - done and deployed (with ongoing extensions)
- - ongoing and not yet deployed



Web presence = web site + portal

The screenshot shows the iDigBio website (https://www.idigbio.org/) with a blue header and a background image of fossilized shells. The header includes the iDigBio logo and three statistics: 3,270,330 Specimen Records, 391,312 Media Records, and 80 Recordsets. A search bar is located on the left, and a navigation menu is at the bottom. The main content area features a large photo of a group of people, a calendar for November, and several article teasers.

iDigBio Home | iDigBio
<https://www.idigbio.org/>

iDigBio
Integrated Digitized Biocollections

3,270,330
Specimen Records

391,312
Media Records

80
Recordsets

Making data and images of millions of biological specimens available in electronic format for the research community, government agencies, students, educators, and the general public

Search the specimen portal:

Home About Collaborators Education Resources News Research Digitization Log In Sign Up

iDigBio's Paleo Digitization Workshop Draws more than 60 Attendees to New Haven

More than 60 paleontologists representing 41 institutions assembled in New Haven, CT the week of September 23rd, 2013 to share ideas, protocols, preferences, and strategies. This was iDigBio's most populous [workshop](#) to date, with an assortment of excellent presentations and ample opportunities for rich discussion.

November

S	M	T	W	T	F	S
		5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

Top Content Tags

Education & Outreach (53) Blog (51)
Workshop (41) Digitization (36)
Featured (32) workflow (16)
Documentation (14) Press Releases (12) Databasing (8) imaging (8)
[More](#)

My Top Resources

- [Contact Us](#)
- [Digitization Documentation](#)
- [iDigBio Forums](#)
- [iDigBio Resources](#)
- [iDigBio Specimen Portal](#)

Florida Museum of Natural History Butterflyfest

10-22-2013

On October 19-20, iDigBio was represented by iDigBio project staff Cathy Bester, Kevin Love, Joanna McCaffrey and David Jennings along with post doc Charlotte Germain-Aubrey and graduate student Claudia Segovia at the [Florida Museum of Natural History](#)

Blog Archives

Welcome Libby Ellwood, New Postdoctoral Scholar with iDigBio
Post date: 09-24-2013

Unlocking the Fossil Cabinet: The Value of Collections in the 21st Century by Austin Hendy, Ph.D., Florida Museum of Natural History
Post date: 09-24-2013

SPNHC 2013 - Special Feature: iDigBio all-day

Upcoming Events

iDigBio Summit III
11-19-2013 to 11-20-2013

Mobilizing Small Herbaria Workshop
12-10-2013 to 12-11-2013

Hackathon to Enable Public Participation in Online Transcription of Biodiversity Specimen Labels
12-16-2013 to 12-20-2013

[more events >>](#)

[feedback](#)

javascript:void(0)

Web presence = web site + portal

Start Searching Specimen Records

Press ESC to close auto-complete suggestions. If no auto-completions match your desired search, close auto-complete and try it anyway. You may still get results from the full-text search of the records.

Welcome to the iDigBio Data Portal

If you're already familiar with our portal's interface, go in and start searching [Specimen Records](#) or [Media Records](#).
If this is your first time here, you might consider browsing [our tutorial](#).

80
Collections

3,270,330
Specimen Records

391,312
Media Records

Specimen Records by Collection Type

Collection Type	Percentage
Plants	33.66 %
Fishes	9.91 %
Arthropods	10.77 %
Invertebrates	15.83 %
Amphibians and Reptiles	5.11 %
Mammals	0.99 %
Fungi	23.73 %

Media Records by Collection Type

Collection Type	Percentage
Plants	37.77 %
Fungi	59.29 %
Arthropods	2.94 %

Our data are based on the [Darwin Core](#) and [Audubon Core](#) standards, and all term definitions can be found on the relevant pages.

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.

Use of this website is subject to iDigBio's [Terms of Use](#) and [Service Level Agreement](#).
This page uses [Google Analytics](#) • [Google Privacy Policy](#) © Copyright 2011

[Like iDigBio on Facebook](#) | [Follow iDigBio on Twitter](#)

• <https://portal.idigbio.org/>

Web presence = web site + portal

The screenshot shows the iDigBio Portal interface. The top navigation bar includes links for Home, Collections, Specimen Records, Media Records, Tutorial, and Feedback, along with a Login button. The main content area displays search results for 'Bathyrāja parmifera'. On the left, there are search filters and a search bar. The search results are displayed in a table with columns for Institution Code, Dataset Name, Collected By, State/Province, and Country. The first result is for 'Bathyrāja parmifera' from the University of Florida (UF), collected by WG Raschi et al. aboard the R/V Alaska. Below the search results, there are sections for 'Top Scientific Name Terms', 'Top Country Terms', and 'Top Collected By Terms'. The 'Top Scientific Name Terms' section lists terms like 'aulacomnium palustre', 'ceratodon purpureus', 'desmognathus brimleyorum', 'dicranum scoparium', 'hypogymnia physodes', 'liguus fasciatus', 'marchantiophyta stotler & crand.-stotl.', 'pardosa moesta', 'polytrichum juniperinum', and 'undetermined'. The 'Top Country Terms' section lists 'america', 'australia', 'canada', 'islands', 'madagascar', 'mexico', 'states', 'u.s.a', and 'united usa'. The 'Top Collected By Terms' section lists 'caloplaca flavovirescens'.

portal.idigbio.org

Home Collections Specimen Records Media Records Tutorial Feedback Login

Clear Download as CSV

Matching All Records.

Last Ten Searches: All Records

Basic Search Advanced Search

Press ESC to close auto-complete suggestions. If no auto-completions match your desired search, close auto-complete and try it anyway. You may still get results from the full-text search of the records.

Search

Change Autocomplete Terms

Displaying 1 to 10 of 3270330

Go To Page# 1

10 per Page

Summary List

Prev 1 2 3 4 5 ... 327032 327033 Next

Bathyrāja parmifera

- Institution Code: UF
- Dataset Name:
- Collected By: WG Raschi et al. aboard the R/V Alaska
- State/Province: Alaska
- Country: USA

Pseudocrossidium replicatum

- Institution Code: MO
- Dataset Name:
- Collected By: R. Zander
- State/Province: Puebla
- Country: Mexico

Micropterus salmoides

- Institution Code: INHS
- Dataset Name:
- Collected By: C.A. Taylor, M.H. Sabaj & T.J. Near
- State/Province: Arkansas
- Country: USA

Lecidea floridensis

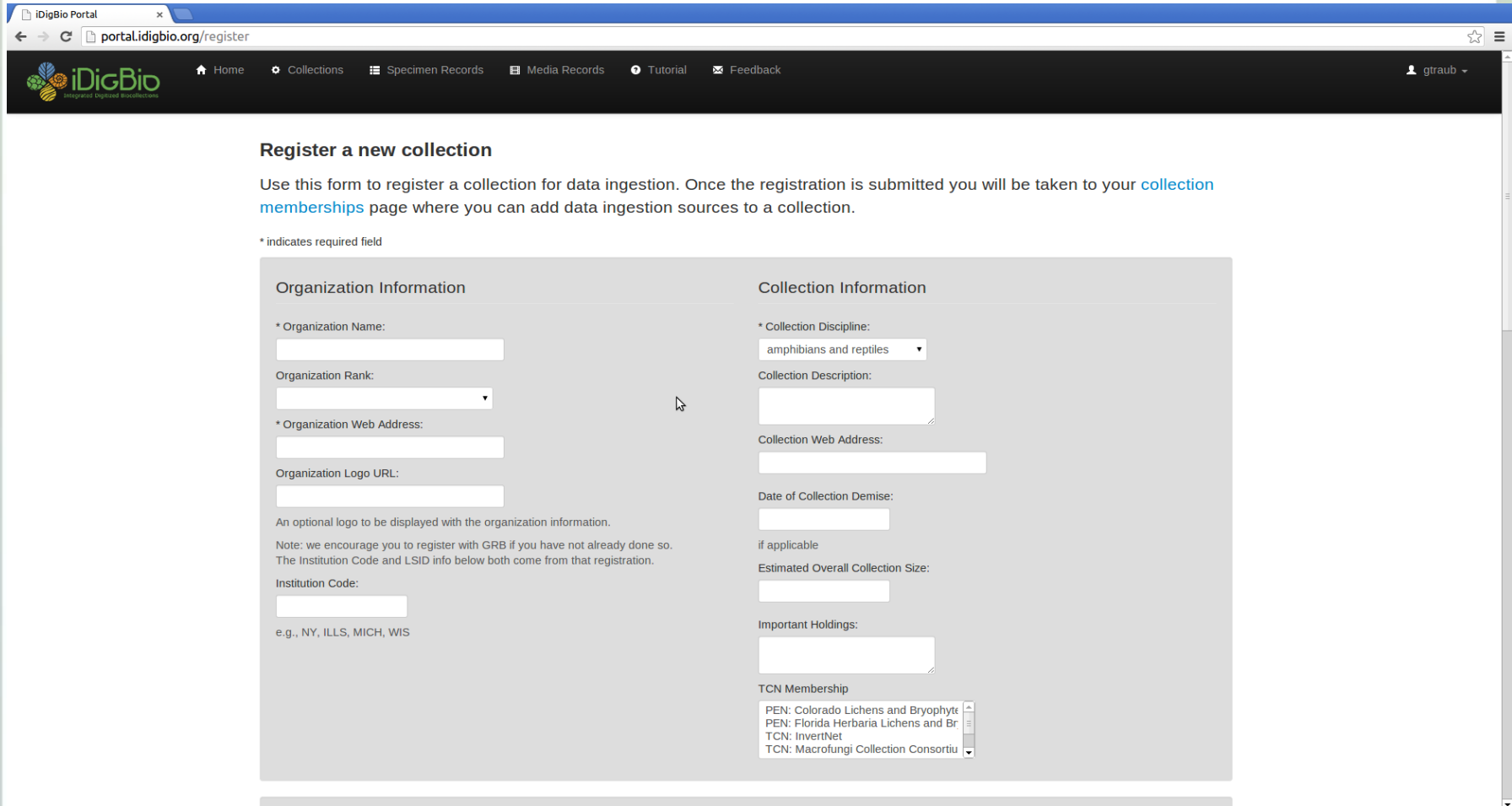
- Institution Code: NY
- Dataset Name:
- Collected By: W. R. Buck
- State/Province: Alabama
- Country: United States of America

Caloplaca flavovirescens

portal.idigbio.org/view/records/da26a2f2-6615-4648-82e8-503805481282

• <https://portal.idigbio.org/>

Web presence = web site + portal



The screenshot shows the iDigBio Portal registration page. The browser address bar displays 'portal.idigbio.org/register'. The navigation bar includes links for Home, Collections, Specimen Records, Media Records, Tutorial, and Feedback, along with a user profile icon for 'gtraub'. The main heading is 'Register a new collection'. Below this, a paragraph explains the purpose of the form. A note indicates that an asterisk (*) denotes required fields. The registration form is divided into two columns: 'Organization Information' and 'Collection Information'. The 'Organization Information' column contains fields for Organization Name, Organization Rank, Organization Web Address, Organization Logo URL, and Institution Code, with a note about the logo and a list of example institution codes. The 'Collection Information' column contains fields for Collection Discipline, Collection Description, Collection Web Address, Date of Collection Demise, Estimated Overall Collection Size, Important Holdings, and TCN Membership, with a note about the date of collection demise. The TCN Membership field is a dropdown menu with options for PEN: Colorado Lichens and Bryophyte, PEN: Florida Herbaria Lichens and Br, TCN: InvertNet, and TCN: Macrofungi Collection Consortiu.

Register a new collection

Use this form to register a collection for data ingestion. Once the registration is submitted you will be taken to your [collection memberships](#) page where you can add data ingestion sources to a collection.

* indicates required field

Organization Information	Collection Information
<p>* Organization Name:</p> <input type="text"/>	<p>* Collection Discipline:</p> <input type="text" value="amphibians and reptiles"/>
<p>Organization Rank:</p> <input type="text"/>	<p>Collection Description:</p> <input type="text"/>
<p>* Organization Web Address:</p> <input type="text"/>	<p>Collection Web Address:</p> <input type="text"/>
<p>Organization Logo URL:</p> <input type="text"/>	<p>Date of Collection Demise:</p> <input type="text"/>
<p>An optional logo to be displayed with the organization information.</p> <p>Note: we encourage you to register with GRB if you have not already done so. The Institution Code and LSID info below both come from that registration.</p> <p>Institution Code:</p> <input type="text"/>	<p>if applicable</p> <p>Estimated Overall Collection Size:</p> <input type="text"/>
<p>e.g., NY, ILLS, MICH, WIS</p>	<p>Important Holdings:</p> <input type="text"/>
	<p>TCN Membership</p> <input type="text" value="PEN: Colorado Lichens and Bryophyte"/>

- <https://portal.idigbio.org/>

New Web presence = web site + portal

The screenshot displays the iDigBio website in a web browser. The browser's address bar shows the URL <https://www-test2.idigbio.org>. The website's header includes the iDigBio logo (Integrated Digitized Biocollections) and navigation links for About iDigBio, Research, Technical Information, and Education. A search bar with a "Google Custom Search" button and a "Search" button is present, along with "Log In" and "Sign Up" links.

The main content area features a large banner with a background image of biological specimens. The banner text reads: "Making data and images of millions of biological specimens available on the web". To the right of the banner, statistics are listed: 3,123,123 Specimen Records, 391,312 Media Records, and 80 Recordsets. A "Search the Portal" button is located below these statistics. 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 banner, there are three columns of links for different user groups: "Researchers" (Browse our specimen portal), "Collections Staff" (Learn how your collection can benefit from our work), and "Teachers & Students" (Download lesson plans about using digitized specimens). Each column has a right-pointing arrow icon.

Below these links is a row of five icons with corresponding text: "Digitization" (Learn, share and develop best practices), "Sharing Collections" (API documentation on data transfer), "Working Groups" (Join in, contribute, be part of the community), "Proposals" (New tool, workshop and working group ideas), and "Citizen Scientists" (How can you help biological collections?).

At the bottom left, there are sections for "Upcoming Events" (iDigBio Summit III, 11-19-2013 to 11-20-2013, with a "more events >>" link) and "Latest Reports" (iDigBio Pamphlet Available, Post date: 05-15-2013; iDigBio Tours New UF CNS Data Center, Post date: 04-08-2013; SPNHC 2013 Registration Now Open, Post date: 03-15-2013).

At the bottom right, there is a large image of a man standing in front of a display titled "National Insect Collection". The display has three panels: "Insects", "flies & mosquitoes", and "moths & butterflies". A "feedback" button is located on the right side of the page.

New Web presence = web site + portal

The screenshot displays the iDigBio Portal interface. At the top, the iDigBio logo is accompanied by navigation links: About iDigBio, Research, Technical Information, Education, Login, and Sign Up. A secondary menu includes Portal Home, Collections, Specimen Records, Media Records, Register a Collection, Tutorial, and Feedback.

Start Searching Specimen Records
Press ESC to close auto-complete suggestions.
[Search Input Field] [Search Button]

Making data and images of millions of biological specimens available on the web
If you're already familiar with our portal's interface, go in and start searching [Specimen Records](#) or [Media Records](#).
If this is your first time here, you might consider browsing [our tutorial](#).
Our data are based on the [Darwin Core](#) and [Audubon Core](#) standards, and all term definitions can be found on the relevant pages.

Specimen Records by Collection Type

Collection Type	Percentage
Plants	33.66 %
Amphibians and Reptiles	5.11 %
Mammals	0.99 %
Fishes	9.91 %
Invertebrates	15.83 %
Fungi	23.73 %
Arthropods	10.77 %

Media Records by Collection Type

Collection Type	Percentage
Plants	37.77 %
Fungi	59.29 %
Arthropods	2.94 %

Summary Statistics:
0 Specimen Records
0 Media Records
125 Collections

Institutional Logos: UF Florida, University of Florida, Florida Museum of Natural History, NSF.

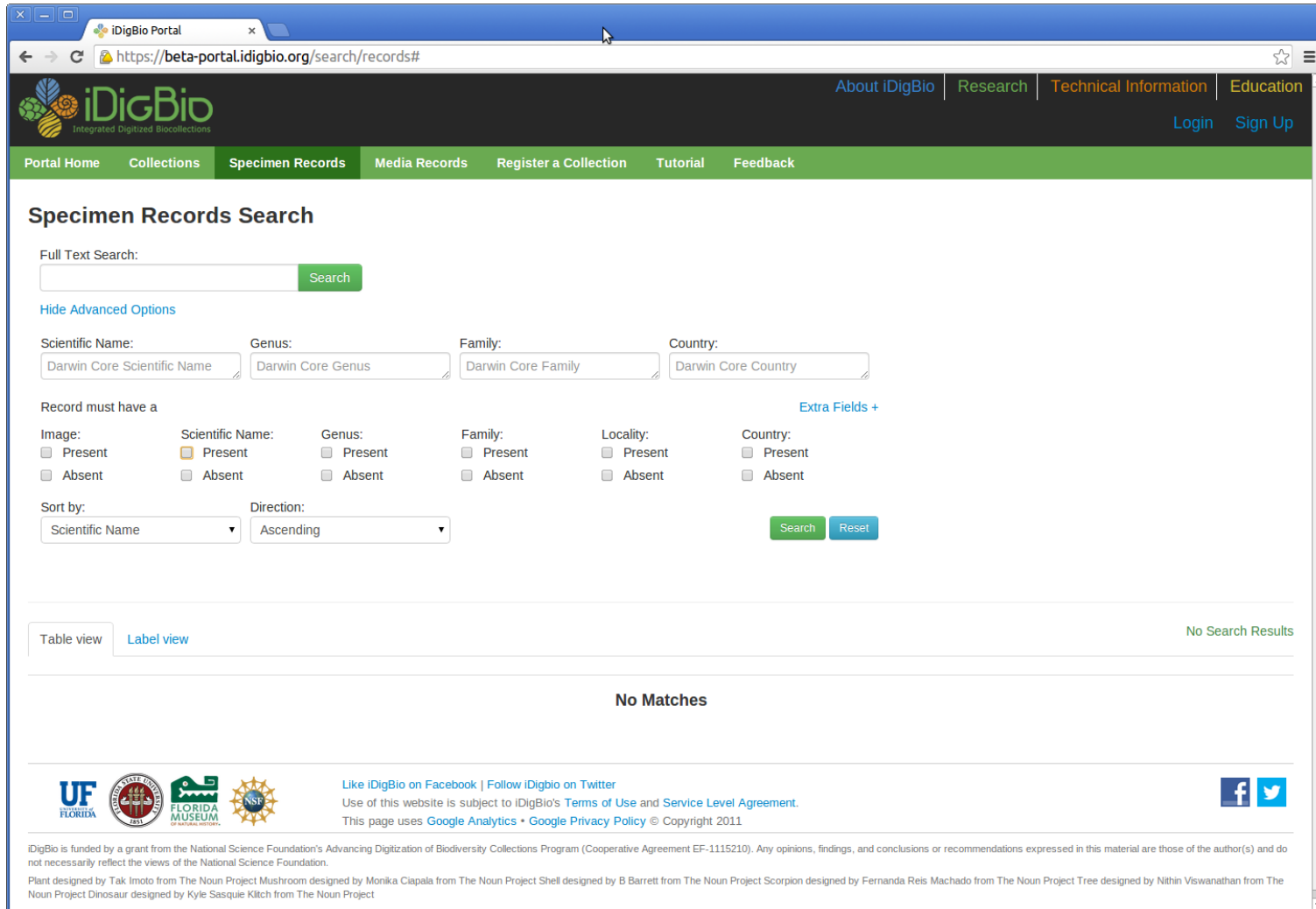
Social Media & Legal: Like iDigBio on Facebook | Follow iDigBio on Twitter. Use of this website is subject to iDigBio's [Terms of Use](#) and [Service Level Agreement](#). This page uses [Google Analytics](#) • [Google Privacy Policy](#) © Copyright 2011.

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.

Plant designed by Tak Imoto from The Noun Project Mushroom designed by Monika Ciapala from The Noun Project Shell designed by B Barrett from The Noun Project Scorpion designed by Fernanda Reis Machado from The Noun Project Tree designed by Nithin Viswanathan from The Noun Project Dinosaur designed by Kyle Sasquie Kilch from The Noun Project

- <http://beta-portal.idigbio.org/>

New Web presence = web site + portal



The screenshot shows the iDigBio Portal search interface. The browser address bar displays <https://beta-portal.idigbio.org/search/records#>. The iDigBio logo is in the top left, and navigation links for About iDigBio, Research, Technical Information, and Education are in the top right. A green navigation bar contains links for Portal Home, Collections, Specimen Records, Media Records, Register a Collection, Tutorial, and Feedback. The main section is titled "Specimen Records Search". It features a "Full Text Search" input field with a "Search" button. Below this is a "Hide Advanced Options" link. The "Advanced Options" section includes fields for Scientific Name, Genus, Family, and Country, each with a "Darwin Core" placeholder. Below these are checkboxes for "Record must have a" Image, Scientific Name, Genus, Family, Locality, and Country, with "Present" and "Absent" options. There are also "Sort by" and "Direction" dropdown menus. "Search" and "Reset" buttons are at the bottom of the search section. The results area shows "Table view" and "Label view" links, with "No Search Results" displayed. At the bottom, there are logos for UF, Florida Museum, and NSF, along with social media links and a footer with funding information and a disclaimer.

Full Text Search:

[Hide Advanced Options](#)

Scientific Name: Genus: Family: Country:

Record must have a [Extra Fields +](#)

Image: ☐ Present ☐ Absent Scientific Name: ☐ Present ☐ Absent Genus: ☐ Present ☐ Absent Family: ☐ Present ☐ Absent Locality: ☐ Present ☐ Absent Country: ☐ Present ☐ Absent

Sort by: Direction:

[Label view](#) No Search Results

No Matches

UF FLORIDA FLORIDA MUSEUM OF NATURAL HISTORY NSF

[Like iDigBio on Facebook](#) | [Follow iDigBio on Twitter](#)

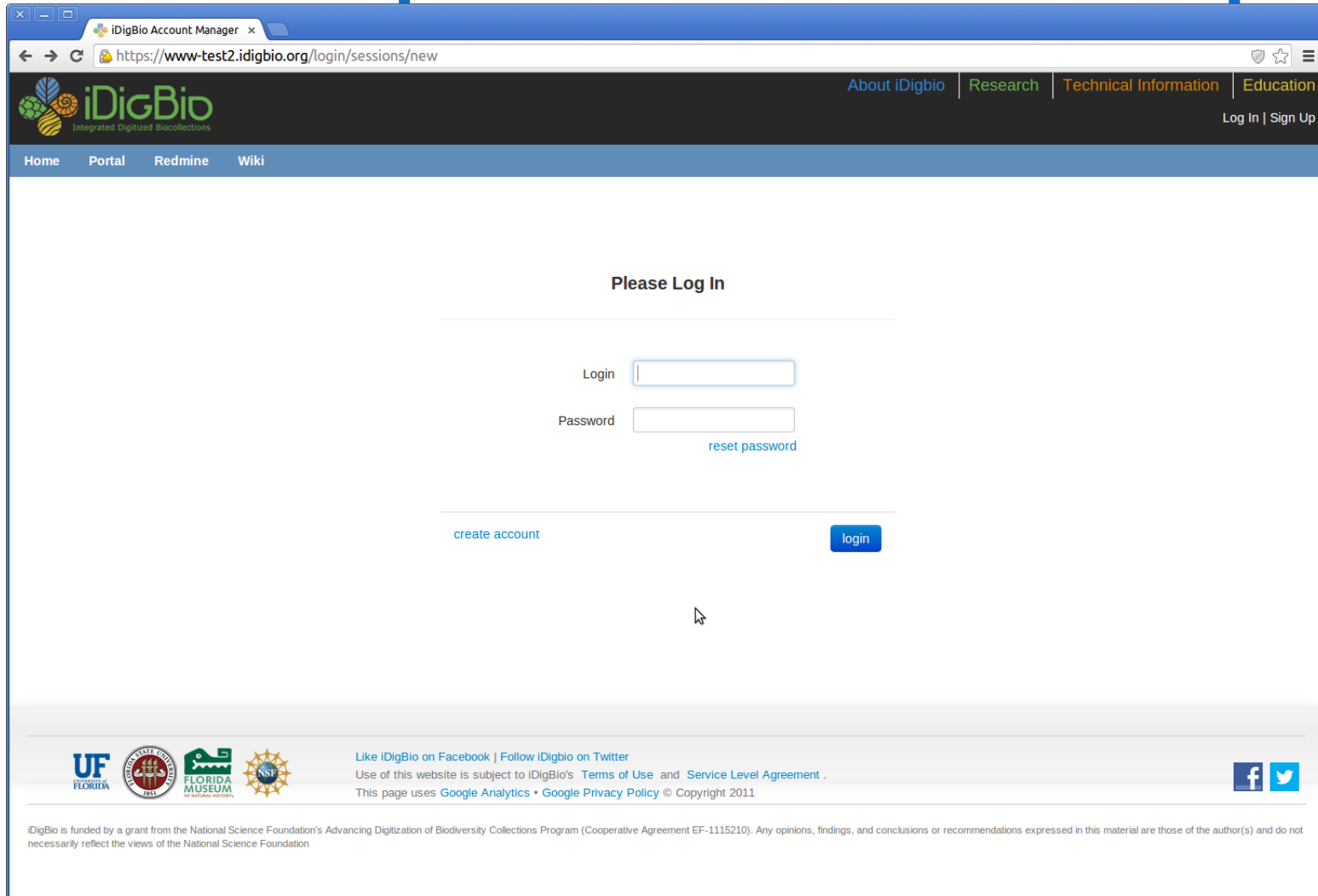
Use of this website is subject to iDigBio's [Terms of Use](#) and [Service Level Agreement](#).
This page uses [Google Analytics](#) • [Google Privacy Policy](#) © Copyright 2011

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.

Plant designed by Tak Imoto from The Noun Project Mushroom designed by Monika Ciapala from The Noun Project Shell designed by B Barrett from The Noun Project Scorpion designed by Fernanda Reis Machado from The Noun Project Tree designed by Nithin Viswanathan from The Noun Project Dinosaur designed by Kyle Sasque Klitch from The Noun Project

- <http://beta-portal.idigbio.org/>

New Web presence = web site + portal



The screenshot shows a web browser window with the address bar displaying <https://www-test2.idigbio.org/login/sessions/new>. The page features the iDigBio logo (Integrated Digitized Biocollections) in the top left. The top navigation bar includes links for [About iDigbio](#), [Research](#), [Technical Information](#), and [Education](#), along with [Log In](#) and [Sign Up](#) options. Below this, a secondary navigation bar contains [Home](#), [Portal](#), [Redmine](#), and [Wiki](#). The main content area is titled "Please Log In" and contains a login form with fields for "Login" and "Password". A "reset password" link is located below the password field. At the bottom of the form, there are links for [create account](#) and a blue "login" button. The footer section includes logos for the University of Florida, the Florida Museum of Natural History, and the National Science Foundation (NSF). It also contains social media links for Facebook and Twitter, and a copyright notice for 2011. A disclaimer at the very bottom states that the content is funded by the National Science Foundation's Advancing Digitization of Biodiversity Collections Program.

Please Log In

Login

Password

[reset password](#)

[create account](#) [login](#)

UF FLORIDA MUSEUM OF NATURAL HISTORY NSF

Like iDigBio on Facebook | Follow iDigBio on Twitter

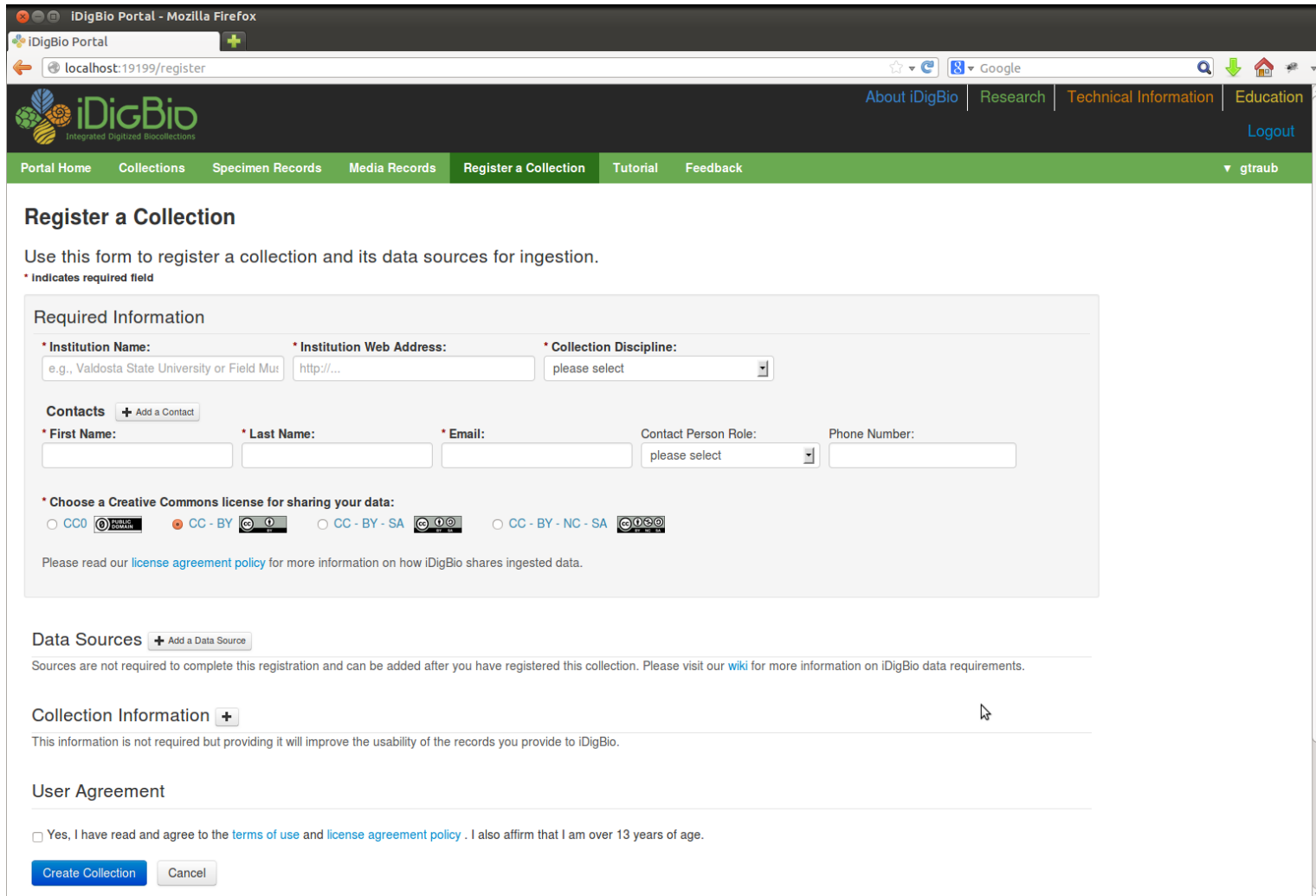
Use of this website is subject to iDigBio's [Terms of Use](#) and [Service Level Agreement](#).

This page uses [Google Analytics](#) • [Google Privacy Policy](#) © Copyright 2011

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

- <http://beta-portal.idigbio.org/>

New Web presence = web site + portal



The screenshot shows the iDigBio Portal registration page in a Mozilla Firefox browser. The browser's address bar shows 'localhost:19199/register'. The page has a green header with the iDigBio logo and navigation links: 'About iDigBio', 'Research', 'Technical Information', 'Education', and 'Logout'. Below the header is a green navigation bar with links: 'Portal Home', 'Collections', 'Specimen Records', 'Media Records', 'Register a Collection' (highlighted), 'Tutorial', and 'Feedback'. A user profile 'gtraub' is visible in the top right corner.

Register a Collection

Use this form to register a collection and its data sources for ingestion.

* Indicates required field

Required Information

* Institution Name: * Institution Web Address: * Collection Discipline:

Contacts [+ Add a Contact](#)

* First Name: * Last Name: * Email: Contact Person Role: Phone Number:

* Choose a Creative Commons license for sharing your data:

☐ CC0 ☒ CC-BY ☐ CC-BY-SA ☐ CC-BY-NC-SA

Please read our [license agreement policy](#) for more information on how iDigBio shares ingested data.

Data Sources [+ Add a Data Source](#)

Sources are not required to complete this registration and can be added after you have registered this collection. Please visit our [wiki](#) for more information on iDigBio data requirements.

Collection Information [+](#)

This information is not required but providing it will improve the usability of the records you provide to iDigBio.

User Agreement

☐ Yes, I have read and agree to the [terms of use](#) and [license agreement policy](#). I also affirm that I am over 13 years of age.

[Create Collection](#) [Cancel](#)

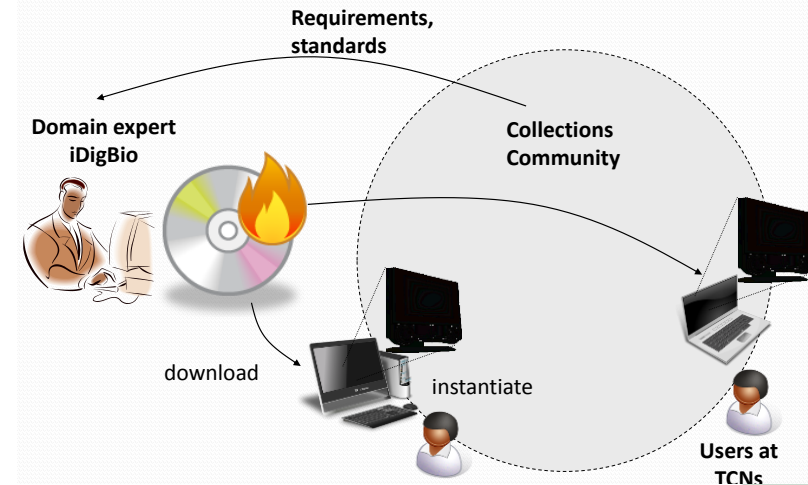
- <http://beta-portal.idigbio.org/>

New Web presence = web site + portal

- Please provide feedback
- iDigBio BETA Website Playground, Wed@3-5pm
- iDigBio BETA Portal Playground, Wed@3-5pm

Appliances

- Appliances complement the cyber-infrastructure core
 - Functionality desired on the client
 - Hide low-level iDigBio APIs, expose user-friendly interface
 - Package tools of general interest to the community in virtual machines for ease of software deployment
 - On resources local to the institution (desktops, servers)
 - On cloud-provisioned resources (e.g. iDigBio cyber-infrastructure)



Appliances progress

- Image ingestion appliance
 - Batch upload of images to iDigBio cloud using its APIs
- Virtual machine appliances
 - Package existing bio-collection tools
 - Use cases: training, technology evaluation
 - Vagrant packaging; desktops, servers (VMware, Virtualbox)
 - Specify, Arthropod Easy Capture, Geo-referencing calculator
- Futures
 - Evolving collection of community-selected tools
 - Built-in interfaces for effortless iDigBio integration
- Details and discussion: Wed 9:00 session (room: tactic 2)

Hosting services: Virtual Private Servers

- Total: 7 VMs, 17 cores, 39GB RAM, 1.7TB storage
 - **Symbiota**: 2VMs
 - 1 production, 2 cores, 8GB RAM, 200GB disk, 1 pub IP, apache, php, java, MySQL, SVN, tomcat, 1user
 - 1 for FP testing/development, 2 cores, 8GB RAM, 200GB disk, 1 pub IP, apache, php, java, MySQL, SVN, tomcat, 3 users
 - **FilteredPush**: 2VMs
 - 1 core, 1024MB RAM, 40GB storage, fp-lite SCAN testbed
 - 2 cores, 4GB RAM, 80 GB storage, mysql, apache, php, tomcat for Symbiota, Morphbank, and FilteredPush
 - **Vertnet**: 1VM
 - 2 cores, 2 GB RAM, 500 GB storage, 1 pub IP, CentOS6, 5 users, Tomcat, IPT
 - **Biogeomancer**: 1VM
 - 4 core, 8GB RAM, 500GB storage, 1 public IP, apache, tomcat, postgres and postgis, 3 users
 - **aOCR hackathon**: 1VM
 - 4 cores, 8 GB RAM, 250 GB storage, Linux (Ubuntu 12.04), Java, PHP, Python, Perl, MySQL, Apache HTTP server, FTP server, ImageMagick, Tesseract, OCRopus, GOCR/JOOCR , ZBar

iDigBio Partnerships: meetings summary

Partner	VPS	Ingestion	Proposal/Expected/Current	Comments
TCN TTD-AMNH	-	AECD ✓	1.6M*/300k/373k specimens	<ul style="list-style-type: none"> Valuable meeting to clarify ingestion mechanism Contact TTD participants individually
TCN TTD-TAMU	-	✓		<ul style="list-style-type: none"> Will send data to SCAN
TCN InvertNet	-	✓	56M/41M/0 specimens 890k/890k/10k images	<ul style="list-style-type: none"> Only images for now
TCN SCAN	✓	Symbiota ✓	740k/740k/505k specimens 16k/16k/50k images	<ul style="list-style-type: none"> Plans in the future for FilteredPush
TCN Paleoniches	-	Specify ✓	450k/1M**/0 specimens	<ul style="list-style-type: none"> Includes TTD Important features: usage tracking, citation, attribution, feedback, improve visibility and loans
TCN NEVP	✓	Symbiota ✓	1.3/1.3M/0 specimens	<ul style="list-style-type: none"> Raw images at iPlant Rate of ~8k/week
TCN LBCC	✓	Symbiota ✓	2.3M/2.3M/2.3M specimens 2.4M/2.4M/700k images	<ul style="list-style-type: none"> 1.6M specimens previously digitized
TCN MaCC	✓	Symbiota ✓	1.4M/1.4M/0 specimens 1.3M/1.3M/0 images	<ul style="list-style-type: none"> Rate of 15~25k/month
BISON				<ul style="list-style-type: none"> iDigBio → GBIF → BISON
EOL				<ul style="list-style-type: none"> EOL Media extension to share media Would prefer to distinguish type of media
COLLABIT/SESYNC/ iPlant				<ul style="list-style-type: none"> Atmosphere, NotesForNature, Distributed object storage , RSS feeds to other bio repositories

*Only insects half of TTD

**All KU (Kansas University), includes TTD

✓ = satisfaction

Data expected/ingested

	TCN	Specimens			Images	
		Proposal	Expected	Ingested	Proposal	Ingested
2011 TCNs	TTD <i>Tri-Trophic Databasing</i>	7.7M (1.6M insects) (6.1M plants)	1.3M*	0	600k	0
	InvertNet	56M	41M	0	890k	0
	LBCC <i>Lichens, Bryophytes and Climate Change</i>	2.3M	2.3M	2.3M	2.4M	700k
2012 TCNs	SCAN <i>Southwest Collections of Arthropods Network</i>	740k	740k	505k**	16k	50k
	Paleoniches	450k	78k	0	3.6k	0
	NEVP <i>New England Vascular Plant</i>	1.3M	1.3M	0	1.3M	0
	MaCC <i>Macrofungi Collection Consortium</i>	1.4M	1.4M	0	1.3M	0
2013 TCNs	FIC <i>Fossil Insect Collaborative</i>	220k	220k	-	33k	-
	VACS <i>Vouchered Animal Communication Signals</i>	58k	58k	-	23k	-
	MHC <i>Macroalgal Herbarium Consortium</i>	1.1M	1.1M	-	1.1M	-
	Total	71.3M	49.4M	3.3M	7.6M	760k

*AMNH insects and all KU (Kansas University) except paleo invertebrates

** Includes TTD-TAMU

Non-TCN Data Expected/Ingested

Non-TCN	Specimens		Images	
	Ingested	Ready	Ingested	Ready
INHS	105,742	105,742	-	-
Ohio State University	-	2,388	-	-
FLMNH	927,059	946,165	-	-
SEINet	341,562	341,562	149,737	149,737
Intermountain	172,352	172,352	59,160	59,160
TTRS	-	6,289	-	-
FSU	-	81,902	-	-
Morphbank*	27,946	28,349	7,247	40,450
Harvard	-	1,994,917	-	-
VertNet**	-	5,608,140	-	-
Total Non-TCN	1,574,661	9,287,806	216,144	249,347

* Dwc-A at source needs to be updated

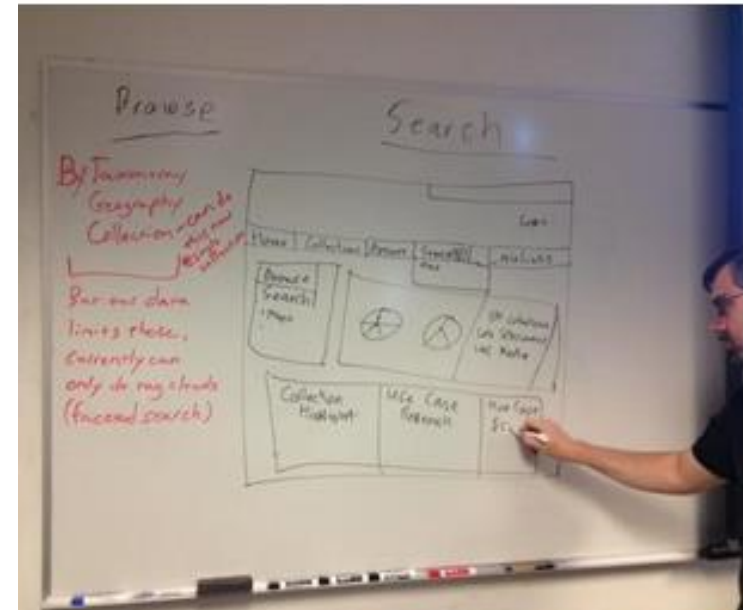
** VertNet requires approval of each individual source

Conclusions

- Deployed cyberinfrastructure for
 - Ingestion and access to digitized specimen records and media
 - Appliances and interfaces to support above
 - Web site and portal for useful information and resources
 - Hosting for databasing, OCR, and other IT services
- Development informed by experience + stakeholders' feedback
 - Version 2 of iDigBio website and cloud
 - Please learn more and provide feedback
 - iDigBio Nuts & Bolts: How to Submit Data, Tue@3:30pm
 - iDigBio Nuts & Bolts: Appliances Present & Future, Wed@9am
 - iDigBio BETA Website Playground, Wed@3-5pm
 - iDigBio BETA Portal Playground, Wed@3-5pm

Thanks are due to ...

- **Website team:** Jeremy Spinks, Greg Traub, Kevin Love, Joanna McCaffrey, Reed Beaman, Betty Dunckel, Shari Ellis and 29 user-experience volunteers
- **Portal team:** Alex Thompson, Larry Page, Pam Soltis, Joanna McCaffrey, Reed Beaman, Kevin Love, Greg Traub, Andrea Matsunaga
- **Appliances team:** Renato Figueiredo, Kyuho Jeong, Yonggang Liu, Alex Thompson, Matthew Collins, Andrea Matsunaga



Extras

Futures

- Protocols for data feedback to flow back to providers
- Linking data
 - within iDigBio data concepts and between
 - across iDigBio and other biodiversity data
 - e.g., genetic material, scientific publications, mapping information and ecological information
- Virtual appliances to use bio-collection databases
- **Strengths**: able to deploy and take advantage of state-of-the-art cyberinfrastructure elements
- **Weaknesses**: need to accommodate to heterogeneous data management/digitization provider strategies

Integration with community tools

- Past
 - GBIF/IPT
- Ongoing
 - Taxonomic tools
 - iPlant's TNRS, GBIF's Checklist Bank, the global names architecture, EOL's name resolution services
 - custom iDigBio hosted version of iPlant's TNRS software, loaded with authority files from the TCNs
- Futures
 - geolocation, reverse geolocation (coordinates to administrative boundaries), and location validation tools
 - GBIF, BioGeomancer, GeoLocate, SpeciesLink, Google, Microsoft, etc

Integration with other projects

- BISON, BiSciCol, DataONE, EOL, FilteredPush, iPlant, Kurator, Specify, VertNet...
- Virtual Private Servers for VertNet to serve as an IPT server, and FilteredPush test bed and as single node FilteredPush network, along with Morphbank and Symbiota clients
- BioSciCol, a solution for Globally Unique IDentifiers (GUIDs) based on a central permanent registry is being investigated
- Commercial solutions:
 - ABBYY, a successful OCR application, tested at hackathon,
 - EMu, a museum data management system, will add GUID
- General purpose: developed as open-source components
 - OpenStack Swift, Drupal, Riak, MediaWiki, Postgres, ElasticSearch, Xen, Python
- **Weaknesses:** resource/personnel constraints

Support of tool development

- Filtered Push
 - (ongoing) hosting Filtered Push annotation stores, prototype Symbiota deployments, and other hosting resources.
 - (future) integrate with iDigBio with Filtered Push network, as an annotation viewer, and as an annotation generator.
- BiSciCol
 - (ongoing) prototype linked iDigBio+BioSciCol data integration.
 - (ongoing) global identifier resolution services via EZID project.
- Specify:
 - Plugin to mobilize Specify data to iDigBio
 - Appliance
- Hackaton
 - To accelerate tool/adoption and integration

Setting priorities

- Prioritization procedures in place involving Internal Advisory Committee (IAC), External Advisory Committee (EAC), and working groups (WG).
- iDigBio IT Standards Workshop
- Cyberinfrastructure Working Group and other groups with community representation identify needs
- iDigBio IT identifies approaches to meet needs
- Steering Committee decides on high-level directions

Cyberinfrastructure design

- **Drivers:** architecture derived in consultation with stakeholders; implementation determined internally
 - feedback from interested parties during development,
 - policies and standards submitted for public comment,
 - developments announced on mailing list + newsletters.
 - prototypes through focus groups at FLMNH + feedback from other parties and cyber-infrastructure working group.
 - beta versions with changes and functionality (6 months)
- **Strengths:** sound IT designs for identified requirements
- **Weaknesses:** incomplete and conflicting requirements from diverse stakeholders

Kinds of iDigBio data

- **Currently:** primarily focused on specimen and image metadata, and images
 - secondary: determination histories, locality data, and geology data (possibly transmitted as specimen metadata)
- **Future:**
 - specimen info (e.g., taxa, date and location of existence, collector),
 - media objects that capture additional information about the specimen (e.g., specimen or habitat images, vocal recordings), and
 - auxiliary information (e.g., lists of known taxa, geographic locations, geological terms).
 - full list at wiki pages of the Minimum Information for Scientific Collections/Authority-File (MISC/AF) working group.

Data storage needs

- Diverse parameters (size, total storage size, access performance, availability, reliability, and longevity)
- Representative patterns
 - small objects (KBs to MBs), medium (few TBs), fast, highly-available, minimally reliable, temporary traditional primary storage (e.g. compressed media objects that need to be centralized and shared among collaborators) (strength)
 - medium objects (MBs to GBs), large (10s-100s TBs), slow, minimally-available, highly-reliable, long-term storage for archival of full size media objects (weakness)
 - large objects (GBs to TBs), medium (few TBs), fast, highly-available, minimally reliable, temporary storage for virtual machine images, applications, and minimum storage (strength)

TCN Data ingestion progress

Sophomore TCNs

- LBCC – Data flowing into iDigBio
- TTD – Data ready, preparing export format
- InvertNet – Data ready, preparing export format

Freshman TCNs

- SCAN – Data and software ready, to be ingested soon
- Paleoniches – Data ready, waiting on Specify release
- NE Vascular Plants – Data not ready
- MFCC - Data and software ready, to be ingested soon



Cyberinfrastructure WG

- [https://www.idigbio.org/wiki/index.php/Cyberinfrastructure Working Group](https://www.idigbio.org/wiki/index.php/Cyberinfrastructure_Working_Group)
- Established as an outcome of the iDigBio IT Standards Workshop
- The focus of the initial group will be on the iDigBio data ingestion procedures related to Application Programming Interface (API) or appliance specification, implementation and test
- Produce material with concrete data ingestion use cases from TCNs, provide input about the existing cyberinfrastructure, produce data ingestion requirements, and help evaluate iDigBio services and appliances implementation

CYWG Current Members

- Andréa Matsunaga (Co-Lead), iDigBio IT
- Joanna McCaffrey (Co-Lead), iDigBio Program Manager
- Reed Beaman, iDigBio IT
- Renato Figueiredo, iDigBio IT
- Alex Thompson, iDigBio IT
- Greg Traub, iDigBio IT
- Yonggang Liu, iDigBio IT
- Kyuho Jeong, iDigBio IT
- Casey McLaughlin, iDigBio IT
- James Beach, Paleoniches TCN IT/Co-PI, University of Kansas, Specify
- Andrew Brown, KE Software
- Edward Gilbert, Lichens & Bryophytes and SCAN TCNs IT, Symbiota
- Corinna Gries, Lichens & Bryophytes TCN, PI
- Paul L Heinrich, SCAN TCN IT
- Tony Kirchgessner, Tri-trophic TCN IT, NYBG
- Derek Masaki, BISON
- Katja Seltsmann, Tri-trophic TCN IT, AMNH
- Nahil Sobh, InvertNet TCN IT, Co-PI
- Omar Sobh, InvertNet TCN IT

- Ex Officio: José Fortes, iDigBio Director for Computational Activities

iDigBio Data Portal v0 API

- Retrieval-only operations (REST GET operations)
- Endpoints:
 - List all endpoints: <http://api.idigbio.org/v0>
 - List collections: <http://api.idigbio.org/v0/recordsets>
 - List specimens: <http://api.idigbio.org/v0/records>
 - List media metadata: <http://api.idigbio.org/v0/mediarecords>
 - List media objects: <http://api.idigbio.org/v0/mediaaps>
- Individual records:
 - Example: <http://api.idigbio.org/v0/records/eac2e4ec-5dbb-4c34-b56f-231ed28a5bca>
 - ```
{"idigbio:data":{
 "dwc:county":"Liberty",
 "dwc:recordedBy":"Loran C. Anderson",
 "dwc:scientificNameAuthorship":"(Nees) Small",
 "id":"http://www.morphbank.net/586214",
 "dwc:eventDate":"2009-06-30 00:00:00.0",
 "dwc:scientificName":"Yeatesia viridiflora"},
 "idigbio:etag":"c3113b3aa2612ce8af46cde267c355ba18325719",
 "idigbio:links":{
 "thumbnailurl":"http://api.idigbio.org/v0/mediaaps/f43df2a6-22e8-4783-a998-39a12d7784ef/media",
 "mediarecord":"http://api.idigbio.org/v0/mediarecords/3409722c-9c23-4a62-808b-7ae684ad2046",
 "recordset":"http://api.idigbio.org/v0/recordsets/b4372b49-c7cc-42db-b1a3-f1c001de0f18"},
 "idigbio:uuid":"eac2e4ec-5dbb-4c34-b56f-231ed28a5bca"}
```

# iDigBio Data Portal v0 Presentation

- <http://portal.idigbio.org/record-view.shtml#eac2e4ec-5dbb-4c34-b56f-231ed28a5bca>

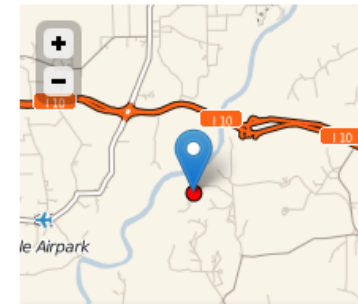
[Home](#)[Specimen Records](#)[Media Records](#)[Tutorial](#)[Feedback? Need Help? Contact Us!](#)

## Specimen Record

iDigBio ID: eac2e4ec-5dbb-4c34-b56f-231ed28a5bca

|                         |                                                                               |
|-------------------------|-------------------------------------------------------------------------------|
| dcterms:language        | en                                                                            |
| dcterms:modified        | 2011-02-09 12:43:07.0                                                         |
| dcterms:type            | Collection                                                                    |
| dwc:basisOfRecord       | Specimen                                                                      |
| dwc:catalogNumber       | 000059955                                                                     |
| dwc:collectionCode      | Florida State University                                                      |
| dwc:continent           | North America                                                                 |
| dwc:country             | United States of America                                                      |
| dwc:county              | Liberty                                                                       |
| dwc:eventDate           | 2009-06-30 00:00:00.0                                                         |
| dwc:institutionCode     | FSU                                                                           |
| dwc:kingdom             | Plantae                                                                       |
| dwc:locality            | SW of old Aspalaga in extreme NW corner of the county.                        |
| dwc:locationID          | <a href="http://www.morphbank.net/586215">http://www.morphbank.net/586215</a> |
| dwc:nomenclaturalStatus | accepted                                                                      |

## Georeference Data



Powered by [Leaflet](#) — Map data © 2011  
OpenStreetMap contributors, Imagery ©  
2011 CloudMade, CartoDB

The blue marker indicates the location of the current record, the red points are locations of similar specimens in the idigbio system.

## Record Image





# Virtual Private Server (VPS)

- Total: 7 VMs, 17 cores, 39GB RAM, 1.7TB storage
  - **Symbiota**: 2VMs
    - 1 production, 2 cores, 8GB RAM, 200GB disk, 1 pub IP, apache, php, java, MySQL, SVN, tomcat, 1user
    - 1 for FP testing/development, 2 cores, 8GB RAM, 200GB disk, 1 pub IP, apache, php, java, MySQL, SVN, tomcat, 3 users
  - **FilteredPush**: 2VMs
    - 1 core, 1024MB RAM, 40GB storage, fp-lite SCAN testbed
    - 2 cores, 4GB RAM, 80 GB storage, mysql, apache, php, tomcat for Symbiota, Morphbank, and FilteredPush
  - **Vertnet**: 1VM
    - 2 cores, 2 GB RAM, 500 GB storage, 1 pub IP, CentOS6, 5 users, Tomcat, IPT
  - **Biogeomancer**: 1VM
    - 4 core, 8GB RAM, 500GB storage, 1 public IP, apache, tomcat, postgres and postgis, 3 users
  - **aOCR hackathon**: 1VM
    - 4 cores, 8 GB RAM, 250 GB storage, Linux (Ubuntu 12.04), Java, PHP, Python, Perl, MySQL, Apache HTTP server, FTP server, ImageMagick, Tesseract, OCRopus, GOCR/JOOCR , ZBar

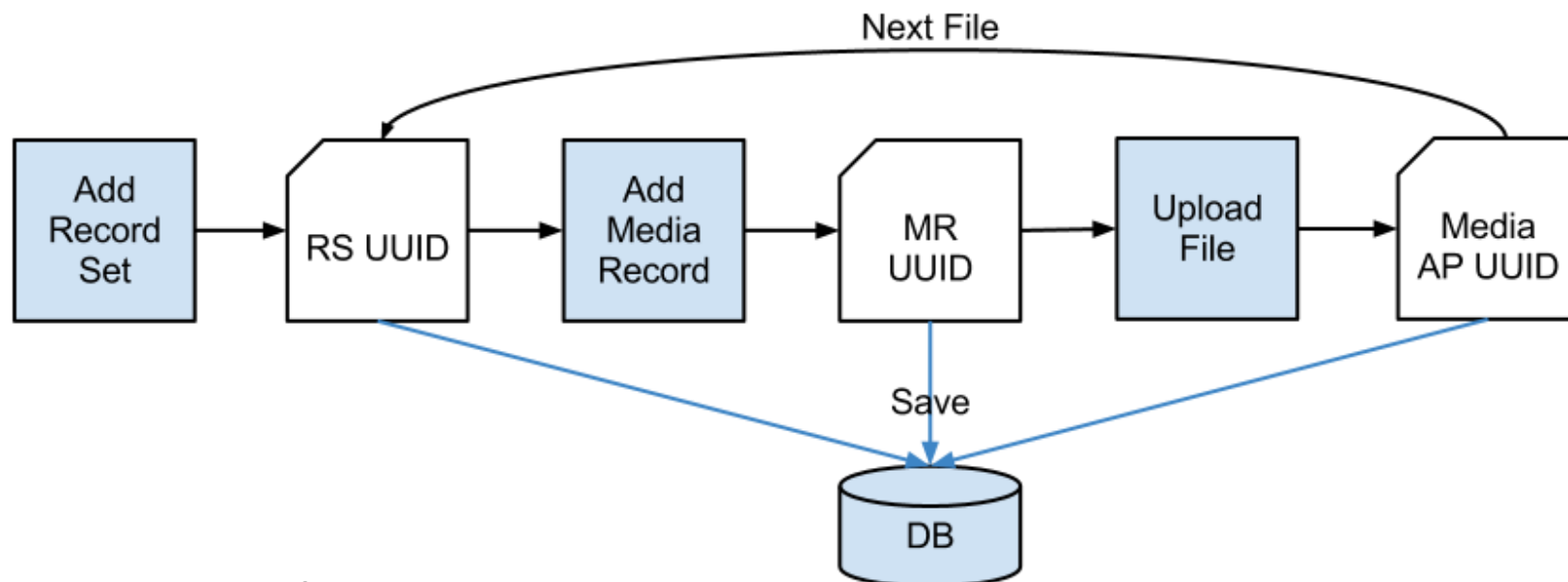
# Databases/DwC-A Examined

| Dataset           | Date        | Format         | Occurrences      | Media          | Taxon          |
|-------------------|-------------|----------------|------------------|----------------|----------------|
| TCN-Bryophytes    | Jun/01/2012 | Symbiota-MySQL | 961881           | 56217          | 49882          |
| TCN-Lichens       | Jun/01/2012 | Symbiota-MySQL | 691967           | 59438          | 10647          |
| TCN-Mycology      | Jun/01/2012 | Symbiota-MySQL | 279529           | 1179           | 415812         |
| TCN-InvertNet     | Mar/14/2012 | DwC-A          | 631388           | 0              | 0              |
| TCN-TTD-AMNH      | Jun/21/2012 | AMNH-MySQL     | 785134           | 4195           | 61655          |
| TCN-TTD-NYBG      | Apr/26/2012 | CSV            | 1469089          | 905            | 0              |
| TCN-PALEONICHES   | Jul/12/2012 | Specify-MySQL  | 96079            | 0              | 6128           |
| FLMNH-Ichthyology | Dec/19/2011 | DwC-A          | 213361           | 0              | 0              |
| FLMNH-Ichthyology | Apr/27/2012 | DwC-A          | 214487           | 0              | 0              |
| Valdosta          | Apr/16/2012 | Specify-MySQL  | 14827            | 12291          | 96817          |
| Morphbank         | Nov/22/2011 | DwC-A          | 193704           | 250442         | 0              |
| Morphbank         | Jun/29/2012 | DwC-A          | 194015           | 252303         | 0              |
| <b>Total</b>      |             |                | <b>5,338,396</b> | <b>386,528</b> | <b>640,941</b> |

# Image ingestion appliance

- First instance of an application built upon the iDigBio APIs
  - Enables easy, reliable bulk-ingestion of media records
- User selects image directory folder; appliance takes care of
  - Traversing sub-directories
  - Uploading individual images through iDigBio API
  - Transparently recovering from various failure conditions
  - Providing to user the mapping of file name to iDigBio URLs
  - Ingested images can be accessed by the URLs
- Appliance runs a lightweight Web server to expose a Web-based UI to users

# Image Ingestion APIs Used By Appliance



## Create RecordSet

URL: RecordSet collection level endpoint

e.g. POST <http://idb-websrv1-dev.acis.ufl.edu:9197/v1/recordsets>

Request content: JSON

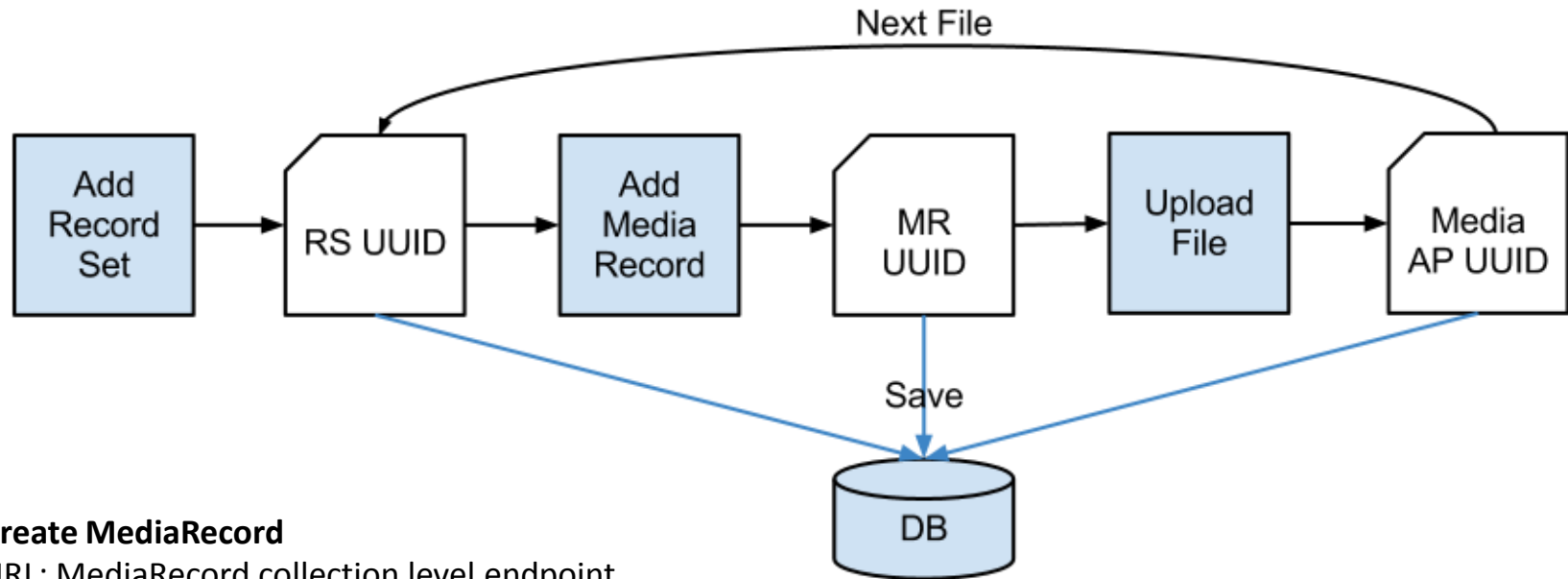
`["idigbio:data"]["ac:variant"]: "IngestionTool"`

`["idigbio:providerId"]: "Currently, client generated random UUID"`

Response content: JSON

`["idigbio:uuid"]: "RecordSet iDigBio UUID"`

# Image Ingestion APIs Used By Appliance



## Create MediaRecord

URL: MediaRecord collection level endpoint

e.g. POST <http://idb-websrv1-dev.acis.ufl.edu:9197/v1/mediarecords>

Request content: JSON

`["idigbio:data"]["ac:variant"]: "IngestionTool"`

`["idigbio:data"]["dc:rights"]: One of {"cc0", "cc-by", "cc-by-sa", "cc-by-nc", "cc-by-nc-sa"}`

`["idigbio:data"]["idigbio:localpath"]: Full local path`

`["idigbio:data"]["idigbio:relationships"]["recordset"]: The iDigBio UUID of the RecordSet the media record belongs to, from the previous step.`

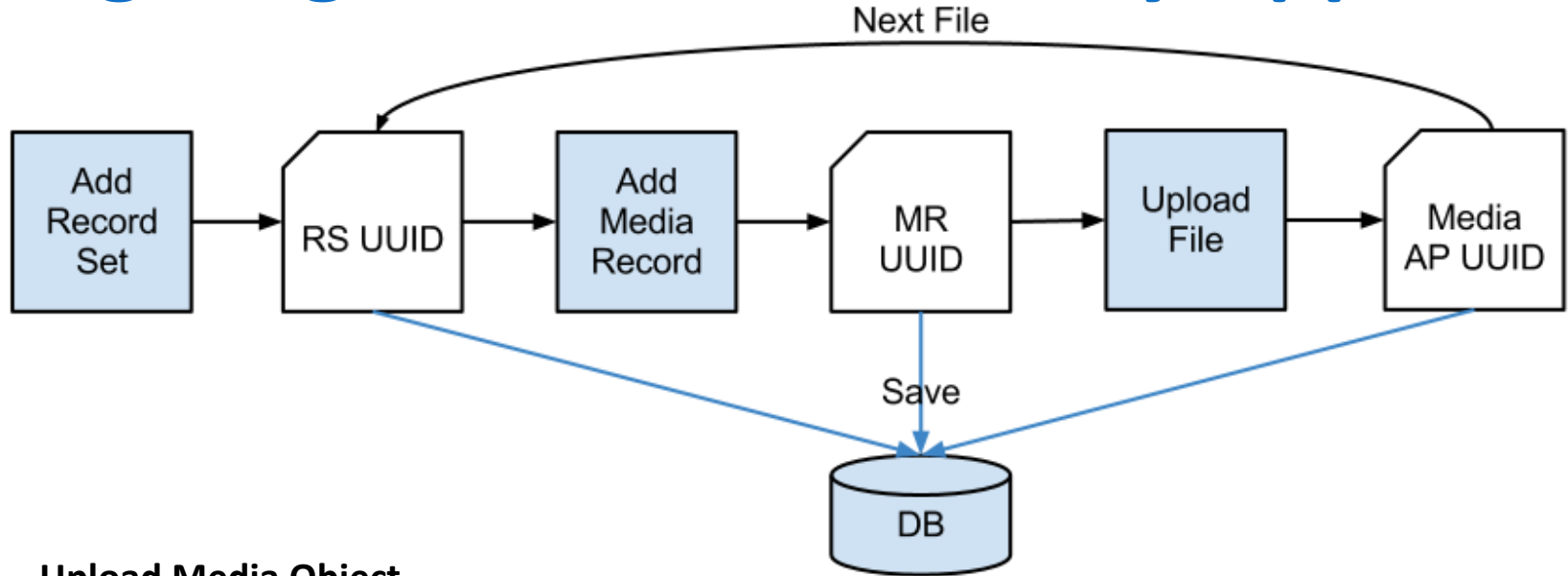
`["idigbio:providerId"]: User defined GUID prefix + (full local path or file name)`

`["idigbio:data"]["idigbio:relationships"]["owner"]: Organizational owner of the record, otherwise the signed-in user is saved as the owner. (Optional)`

Response content: JSON

`["idigbio:uuid"]: iDigBio MediaRecord UUID`

# Image Ingestion APIs Used By Appliance



## Upload Media Object

URL: API sub-collection level endpoint

e.g. POST <http://idb-websrv1-dev.acis.ufl.edu:9197/v1/mediarecords/8c7ae0c3-a3b8-4ddd-b433-ab99141ed405/media> (the UUID in the middle is the "iDigBio MediaRecord UUID" returned in the previous step)

Request content: Binary multipart/form-data with the image as the "file"

Response content: JSON

["idigbio:links"]["media"]: The URL where the image is accessible online

['idigbio:uuid']: The Media Access Point UUID

['idigbio:data']['idigbio:imageEtag']: The hash (MD5) of the image stored at the server, which is compared with the MD5 of the local image to verify the success of the upload

# Image Ingestion v1 – Call for beta-testers

- **Features already implemented**

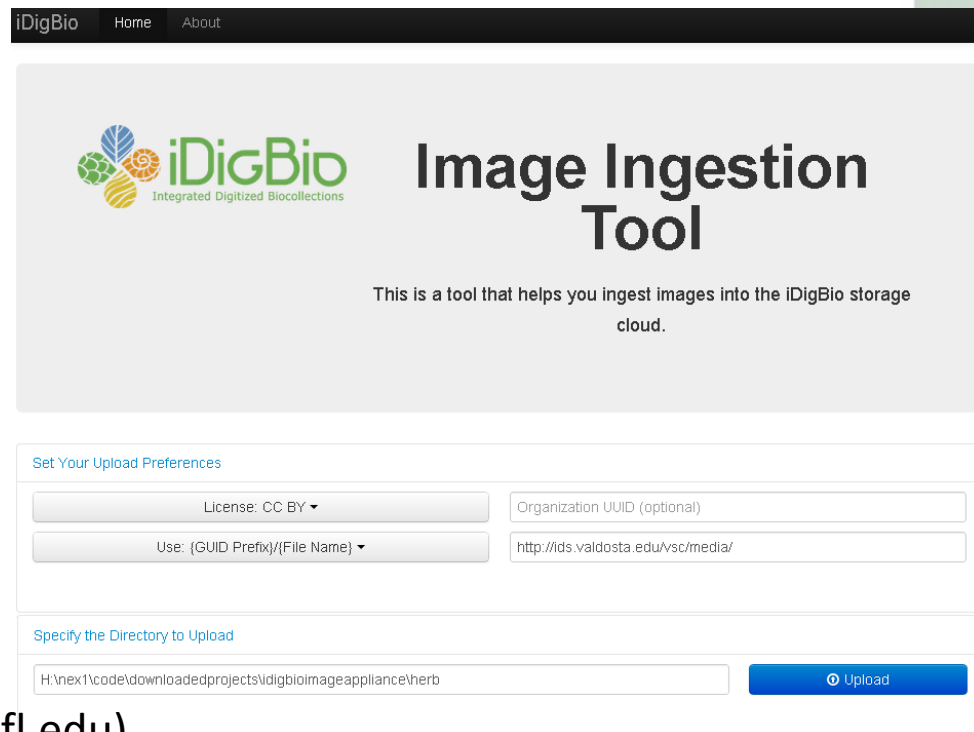
- Reliable uploads
  - Automatically retry of failed upload of individual files
  - Keep local record of unsuccessful transfers ; resume after failure of network or service
  - Skip already uploaded files when the same directory is uploaded multiple times
- Allow user to specify a license for the media object being uploaded
- Save, Export (local path : URL) mappings for individual files in the batch upload

- **Features to add**

- Integrated user authentication
  - Currently , application keys can be provided to beta testers
- Improved UI, error reporting, encoded best practices on UI

- **Feedback from early adopters**

- Help guide UI improvements and prioritize features to be incorporated
- Help fine-tune performance (e.g. parallel uploads) and failure handling



The screenshot shows the iDigBio Image Ingestion Tool web interface. At the top, there is a navigation bar with 'iDigBio', 'Home', and 'About' links. The main header features the iDigBio logo (a stylized flower with four colored petals) and the text 'iDigBio Integrated Digitized Biocollections'. Below the logo, the title 'Image Ingestion Tool' is displayed in large, bold letters. A subtitle reads: 'This is a tool that helps you ingest images into the iDigBio storage cloud.' The interface is divided into two main sections: 'Set Your Upload Preferences' and 'Specify the Directory to Upload'. The 'Set Your Upload Preferences' section contains two rows of input fields. The first row has a 'License' dropdown menu set to 'CC BY' and an 'Organization UUID (optional)' text input field. The second row has a 'Use:' dropdown menu set to '{GUID Prefix}/{File Name}' and a text input field containing 'http://ids.valdosta.edu/vsc/media/'. The 'Specify the Directory to Upload' section has a text input field containing 'H:\nex1\code\downloadedprojects\idigbioimageappliance\herb' and a blue 'Upload' button with a circular arrow icon.

Contact: Andréa Matsunaga (ammatsun@ufl.edu)

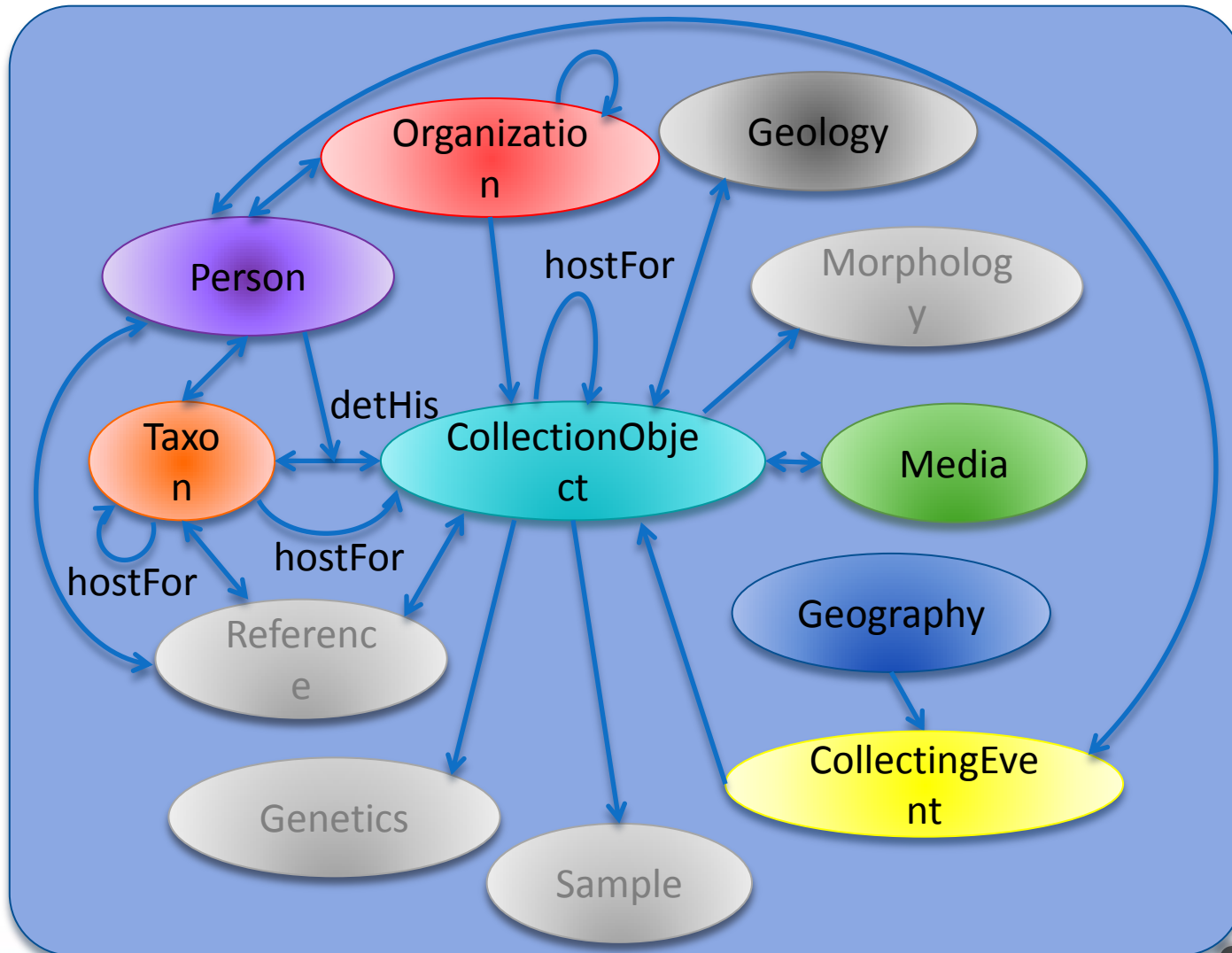
# GUID Standards

- Universally Unique Identifier (UUID): unique id
  - 128-bit number displayed 32 hexadecimal digits
  - 550e8400-e29b-41d4-a716-446655440000
- Digital Object Identifier (DOI): unique id + resolution + business model
  - ISO standard
  - doi:10.1000/182
- Life Science Identifier (LSID): unique identifier for biological information + resolution, including (but not limited to) taxon names
  - Represented as a [Uniform Resource Name](#) (URN) with the following format.
  - Urn:lsid:<Authority>:<Namespace>:<ObjectID>[:<Version>]
- iDigBio GUID: unique id
  - Choose a URI scheme
  - http://ids.flnmh.ufl.edu/herb/abcd12345678





# MISC WG – Data Model Concepts



# Relational Databases

Symbiota

Organization

*omcollections*

Dataset

*Exsiccati title,  
number, links*

Person

*omcollectors*

Taxon

*taxa*

CollObject

*omoccurrences*

Media

*Images*

detHis

TTD-AMNH

CollectingEvent

*colevent*

parent

Person

*collector*

Geography

*locality*

parent

hostFor

CollObject

*omoccurrences*

Taxon

*mnl U  
flora\_mnl*

subfamily  
tribe  
genus  
species

Organization

*institution*

Media

*images*

Specify

CollectingEvent

*colevent*

parent

Person

*agent*

Geography

*Locality U  
geography*

CollObject

*collectionobject*

Media

*attachments*

Taxon

*taxon*

detHis

Organization

*collection*

Geology

*paleocontext*

EMu

CollectingEvent

*ecollectionevents*

Person

*eparties*

Geography

*esites*

CollObject

*ecatalogue*

Geology

*esites*

Taxon

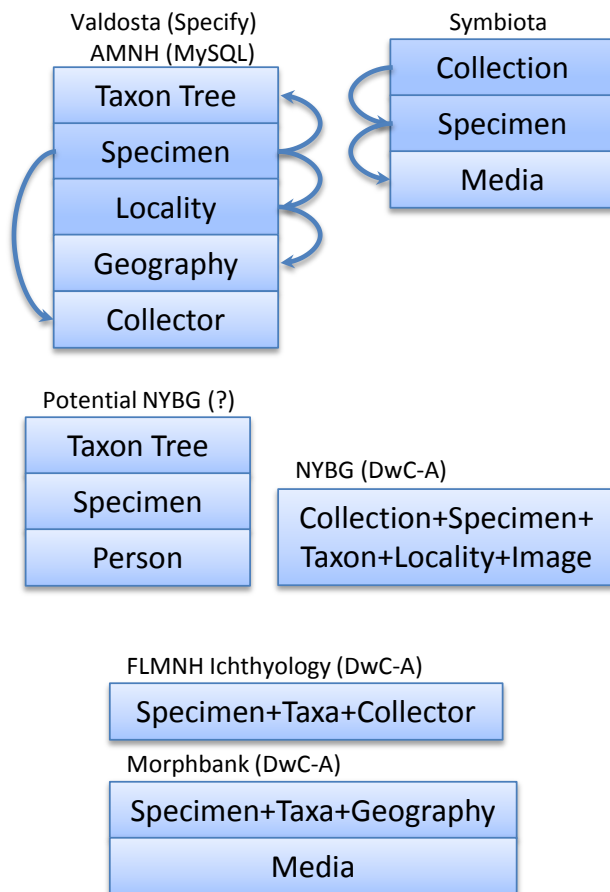
*etaxonomy*

Media

*emultimedia*

# Potential MISC

## TCN and other data providers



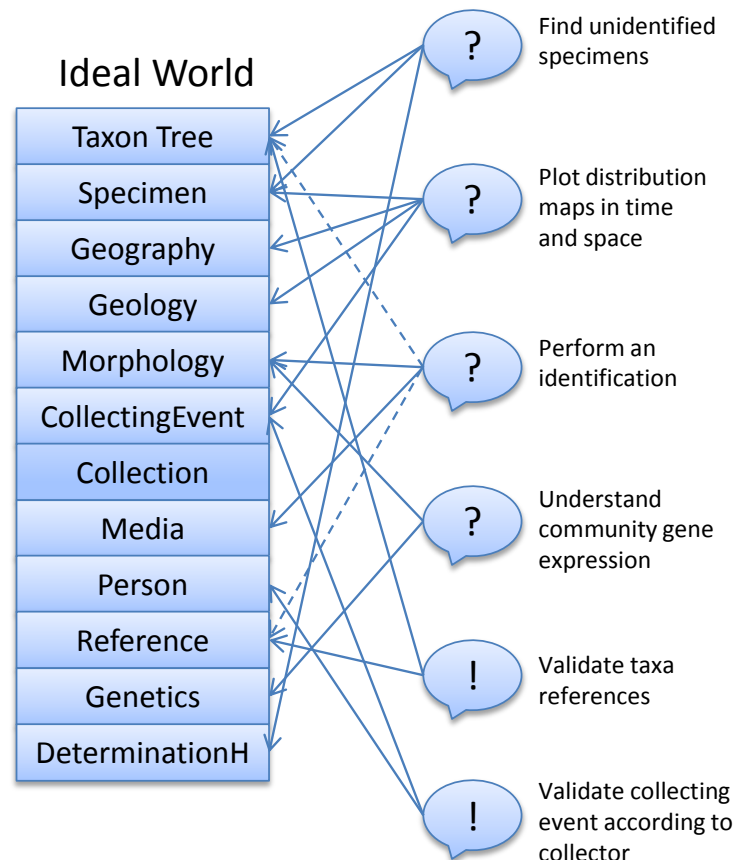
## MISC



Missing information:

- Sampling effort
- Absence / abundance
- Precise Time
- Habit
- Host (specimen-specimen; specimen-taxa; taxa-taxa)
- Locality security
- Duplicates (Exsiccati)
- Copyright controlled vocabulary
- Elevation Source

## TCN research questions and digitization process



# Final remarks and conclusions

- Significant progress in building iDigBio cyberinfrastructure
  - Basic architecture
  - Technology foundations
  - Human and physical resource
- Data ingestion harder than expected
  - It takes a village ...
  - Slowed by, but also driving, community-wide practices regarding data models and bio-collection practices
  - Human interaction still needed on our way to automation
    - Expected to accelerate and grow in immediate future
  - Current portal already shows that our approach works