Integrated Digitized Biocollections (iDigBio)
Cyberinfrastructure Status and Futures

Andréa Matsunaga, Renato Figueiredo, Alex Thompson, Gregory Traub, Reed Beaman, José A.B. Fortes

Supported by NSF Award EF-1115210

TDWG 2013 Annual Conference
Florence, Italy
Monday 28th of October - Friday 1st of November 2013
Integrated Digitized Biocollections (iDigBio)

- 10-year effort to digitize and mobilize the scientific information associated with vouchered specimens held in U.S. neontological and paleontological research collections.

Diagram:
- InvertNet
- Paleoniches
- Southwest Collections of Arthropods Network
- New England Vascular Plant
- Macrofungi Collection Consortium
- Future TCN
- FIC
- MHC
- VACS
- HUB (Home Uniting Biocollections)
- TCN (Thematic Collections Network)
- PEN (Partners to Existing Network)
- Tri-Trophic Associations
- North American Lichens and Bryophytes
- Colorado Lichens and Bryophytes
- Florida Herbaria
- PEN
Thematic Collections Network

- Lichens and Bryophytes
- Tri-Trophic Relationships

Environmental Change

Feed on plants

Agriculture

Lay eggs on aphids
TCN Institutions Distribution

- 3 TCNs (2011), 7 TCNs+2PENs (2012), 10 TCNs+7 PENs (2013)

<table>
<thead>
<tr>
<th>Year</th>
<th>Collections</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>118</td>
</tr>
<tr>
<td>2012</td>
<td>193</td>
</tr>
<tr>
<td>2013</td>
<td>256</td>
</tr>
</tbody>
</table>
Cloud Computing Environment

- **Flexible** to meet the diverse needs of TCNs
- Horizontally **scalable** to meet future demands to access the data
- **Agile** in taking advantage of and integrating proven open-source technologies, thus minimizing system development and maintenance risk
- **Resilient** to certain types of failures
- Based on **standards** where they exist to enable interoperability and reuse of tools, libraries, and services
  - Estimated 1 billion specimens in 1,600 collections in the US
iDigBio Cyberinfrastructure

- Developed in consultation with stakeholders
- Implementation determined internally
- Feedback on prototypes solicited from community

![Diagram of iDigBio Cyberinfrastructure]

- User-facing Application Layer
  - Community Portal
  - Specimen Data Portal
  - Appliances
  - Project Management
  - Web Conferencing
  - Source Code Management
  - Web Map Service
  - Social Media

- Cloud Service Layer
  - Data API
  - Media API
  - Text Indexing
  - Geospatial Indexing
  - Authentication
  - HTTP High Availability
  - HTTP Load Balancing
  - Relational Database
  - File System

- Hardware Layer
  - System Provisioning
  - System Monitoring and Alerting
  - Machine Virtualization
  - Distributed Document Store
  - Distributed Object Store

- Storage
- Network
- Servers

- iDigBio Hosting
- External Hosting
Supporting iDigBio Activities

- Workshops and Hackathons
- Community Portal and Third Party Services

1. New event Notification
   - Social Media Services (Facebook, Twitter)

2. Real-time event broadcast and remote participation
   - Web Conferencing (AdobeConnect)

3a. Though/idea/work generation
   - Collaborative Editing (Google docs, wiki)

3b. Code generation during Hackathon
   - Collaborative Coding (GitHub)

4. Successful?
   - Event Assessment (Qualtrics)

5. Polish outcomes
   - Final Product Publication (Drupal, AdobeConnect)

6. Disseminate
Object and NoSQL Storage

- Horizontally scalable, replicated (x3)

```json
{
  "idigbio uuid": "b4a0382b-2102-4a2a-8111-1fed3eca848b",
  "idigbio etag": "486912e7ec555880eb26443d4d861cb6d356f8b9",
  "idigbio dateModified": "2013-04-07T16:56:28.466Z",
  "idigbio version": 0,
  "idigbio data": {
    "dwc.kingdom": "Plantae",
    "dwc.recordedBy": "Robert K. Godfrey, H. Kurz",
    "dwc.scientificNameAuthorship": "Michx.",
    "dctterms.language": "en",
    "id": "ark:/87286/B2ab16e044-33a0-464b-8dce-afbf7115b20c7",
    "dwc.stateProvince": "Florida",
    "dwc.eventDate": "1957-06-27 23:00:00.0",
    "dwc.country": "UNITED STATES",
    "dwc.collectorCode": "Florida State University",
    "dwc.state": "Leon",
    "dwc.occurrenceRemarks": "Quercus falcata x Quercus hemisphaerica.",
    "dwc.locationID": "http://www.morphbank.net/528431",
    "dwc.basisOfRecord": "Specimen",
    "dwc.continent": "North America",
  }
}
```
iDigBio Specimen Portal


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.

**Specimen Records by Collection Type**

- 80 Collections
- 3,270,330 Specimen Records
- 391,312 Media Records

**Media Records by Collection Type**

Our data are based on the Darwin Core and Audubon Core standards, and all term definitions can be found on the relevant pages.
Search Interfaces

- **Basic Search:** suggestion by auto-completion
- **Advanced Search:** term-based
- **Dynamic Tag Cloud**
  - Top Scientific Name Terms:
    - eptesicus fuscus lasiurus borealis
    - lasiurus intermedius lasiurus seminolus
    - myotis australis
    - nycticebus humeralis
    - peromyscus polionotus
    - pipistrellus subflavus
    - scalopus aquaticus tadarida brasiliensis
  - Top Country Terms:
    - states
    - united

**Taxonomy**

- **Scientific Name**
- **Genus**
- **Species**
- **Tribe**
- **Subfamily**
- **Family**
- **Order**

---

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.

**Basic Search**
- quercus laurifolia (112) - dwc_scientificname_k
- quercus hemiphaerica (78) - dwc_scientificname_k
- quercus myrtifolia (71) - dwc_scientificname_k
- quercus virginiana (69) - dwc_scientificname_k
- quercus chapmanii (65) - dwc_scientificname_k
- quercus margarettiae (57) - dwc_scientificname_k
- quercus geminata (53) - dwc_scientificname_k
- quercus incana (42) - dwc_scientificname_k
- quercus nigra (41) - dwc_scientificname_k
- quercus minima (37) - dwc_scientificname_k
- quercus (218) - dwc_recordedby

**Advanced Search**
- term-based
Search Result Interfaces

- Search results as lists
  Displaying 11 to 20 of 177
  Go To Page# 1
  Prev 1 2 3 4 5 ... 17 18 Next

Cryptotis parva
- Institution Code: TTRS
- Dataset Name:
- Collected By: W.W. Baker
- State/Province:
- Country: United States

Peromyscus polionotus
- Institution Code: TTRS
- Dataset Name:
- Collected By: W.W. Baker
- State/Province:
- Country: United States

Lasiurus seminolus
- Institution Code: TTRS
- Dataset Name:
- Collected By: W.W. Baker
- State/Province:
- Country: United States

Detailed record view

Specimen
- Basis of Record: Specimen
- Catalog Number: 240
- Institution Code: TTRS
- Sex: Male

Taxonomy
- Kingdom: Animalia
- Nomenclatural Status: invalid
- Scientific Name: Pitymys pinetorum
- Scientific Name Authors: (Le Conte, 1830)
- Taxon Rank: Species

Collection Event
- Occurrence Remarks: no skull
- Collected By: W.W. Baker
- Verbatim Event Date (Verbatim Collection Date): 1966-11-30

Georeference

Record Image

The blue marker indicates the location of the current record, the red points are locations of similar specimens in the idigbio system.
Virtual Private Server (VPS)

- **Total:** 8 VMs, 18 cores, 41GB RAM, 1.8TB storage
  - **Symbiota:** 2VMs
    - 1 production, 2 cores, 8GB RAM, 200GB disk, 1 pub IP, apache, php, java, MySQL, SVN, tomcat, 1 user
    - 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, 2GB RAM, 500GB 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/JOCR, ZBar
  - **SCAN:** 1VM
    - 1 core, 2GB RAM, 20GB storage, Drupal environment
Media Ingestion Appliance

- iDigBio-developed appliance
- iDigBio API client

Load CSV File ➔ Upload MediaRecord ➔ Upload Media

8 retries with x2 backoff time

Number of images: 19,266
Total Size: 124GB
Each Image Size: 4.8 ~ 11.4MB
iDigBio Custom Appliances
Biodiversity Research

- Potential adaptation to climate change
- Future changes
- Ecological drivers of change
- Phylogenetic distribution
- Phylogenetic uniqueness
- Evolutionary signal to response to climate change
- Phylogenetic signal
- Phylogenetic communities
- Trait evolution
- Response of traits to historical changes

Geospatial layers (WorldClim, remote sensing data, etc.)
Ecological data (physiology, morphology, etc.)
Georeferenced collections data (iDigBio, GBIF, etc.)
Genetic data (GenBank)

Increasingly interdisciplinary

Ecological data
Niche modeling
Regional Phylogeny
Phylogenetic and functional traits analyses
Future Work – Data Model

TCN and other data providers

TCN research questions and digitization process services

Missing information:
- Sampling effort
- Absence / abundance
- Precise Time
- Habit
- Host (specimen-specimen; specimen-taxon; taxon-taxon)
- Water-specific locality
- Locality security
- Duplicates (Exsiccata)
- Copyright controlled vocabulary
- Elevation Source
- Funding

Find unidentified specimens
Determine host relations among species
Plot past and future distribution maps in time and space
Perform an identification
Understand community gene expression
Find effect of climate change on biodiversity
Validate taxa references
Validate collecting event according to collector
Future Work – Interface Improvements

- Presence/absence of values; sorting

Specimen Records Search

Full Text Search:

Search

Hide Advanced Options

Scientific Name: 

Darwin Core Scientific Name

Genus:

use a new line for each term

Family:

Darwin Core Family

Country:

Darwin Core Country

State/Province:

County:

Institution Code:

Date Collected:

Record must have a

Image:

☐ Present ☐ Absent

Scientific Name:

☐ Present ☐ Absent

Genus:

☐ Present ☐ Absent

Family:

☐ Present ☐ Absent

Locality:

☐ Present ☐ Absent

Country:

☐ Present ☐ Absent

State/Province:

☐ Present ☐ Absent

County:

☐ Present ☐ Absent

Institution Code:

☐ Present ☐ Absent

Date Collected:

☐ Present ☐ Absent

Sort by:

Scientific Name

Direction:

Descending
iDigBio Geospatial Mapping

- With and/or without heat layer
References


- http://portal.idigbio.org

- http://api.idigbio.org

- http://www.idigbio.org
Supported by NSF Award EF-1115210