iDigBio: data management systems and opportunities for collaboration

José Fortes
on behalf of the iDigBio team
Outline

• Introduction to iDigBio
• Data management
• Conclusions
  – Opportunities for collaboration
What is iDigBio

• Coordinating center for the national effort to digitize non-federal U.S. natural history collections
  – Based at the University of Florida and Florida State University
• Central resource for NSF’s Advancing Digitization of Biodiversity Collections (ADBC) program
  – 10 year, $100 million nationwide effort
  – National network of institutions organized by theme to focus research, drive digitization efforts, & build community
  • Thematic Collection Networks (TCNs)
  • Partners to Existing Networks (PENs)

“To advance scientific knowledge by improving access to digitized information in vouchered scientific collections across the US.”
What does iDigBio do?

• Enable digitization of biodiversity collections data
  – Develop efficient & effective standards & workflows
  – Workforce education & training

• Provide portal access to biodiversity data in a cloud computing environment
  – Respond to cyberinfrastructure needs
  – Enable access & discoverability

• Facilitate use of biodiversity data to address key environmental and economic challenges
  – Researchers, educators, general public, policy-makers, ...

• Plan for long-term sustainability of the national digitization network & effort
  – Expand participation: partners, data sources, public, ...
  – Proliferate and broaden uses of biodiversity data
Why?

Estimates suggest there are between **500 million and 1 billion** biological and paleobiological specimens in the United States and potentially **3-4 billion worldwide**. Many are digitized, but most are not. **An untapped trove of information!**
Specimen data: what are they?

- **Records**
  - Specimen
  - Labels
  - Thumbnails
  - 2D, 3D, ...
- **Sounds**
- **Video**
- **Other**
  - Metadata
  - Providers
  - ...
iDigBio Portal has 109M records for 327M specimens with 23M associated media records

iDigBio is working with 708 collections in 396 institutions

Blue = NSF-funded, green = other.
Specimen data: qualitative/functional effectiveness

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

View search results as table, pseudolabels, or images

Results mapped, rendered and downloadable

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

Above actions, downloads and more can also be done via API (i.e. programmatically)
Specimen data: quantitative effectiveness

(www.idigbio.org/portal/portalstats)

109M records for 327M specimens with 23M associated media records
Specimen data: software and hardware

64 VMs/Containers: Proxy/load balance (2); Portal (5); API (5); Search (5); Media API (10); Celery task (5); Ceph manager (3); Ceph gateway (3); Redis cache (3); Applications & VPS (18);

@ACIS since 2011

Getting data ingested into iDigBio
1. Data in Darwin Core format, media in Audubon Core format.
2. Each record with a GUID.
3. Packaged in a DwC archive, one collection per archive.
4. Published on an IPT with an RSS feed (or via Symbiota)
Search across all data, all/individual fields, customize, use autocompletion, synonyms, ...
Data on specimens collected in Africa
Zooming in

South Africa IPT (part 2)
Agricultural Research Council, Plant Protection Research Institute, South Africa - ARC-PPRI
Albany Museum South Africa - AM / AMAI
Albany Museum South Africa - Albany Mus. / AMGS
Zooming in further
Conclusions

• Collaboration opportunities: cyberinfrastructure and data
  – Data provision and data usage
  – Distributed data hosting
  – Resource sharing

• People are important
  – Large community involvement
  – Community practices and their evolution
  – IT staff with different skills

• Cyberinfrastructure iDigBio has been successful but is never done
  – New requirements (user wishes, research questions, ...)
  – New technologies, limited life of hardware, software, ...
Pre-ingestion workflows

Collection Data
- EMu
- MySQL
- MS-SQL
- Access db
- CSV
- Excel Sheet
- Filemaker Pro
- Others

MySQL
- EMu
- MySQL
- MS-SQL
- Others

Data Publishers
- ALA
- NHM London
- iDigBio Feeder
- iDigBio IPT
- VertNet IPT
- Dropbox
- Migrator
- Ingestion appliance

Specified
- EMu
- Tropicos
- Biocase
- EMu, Arctos

Ingestion
- Symbiota server
- Symbiota
- Symbiota server
- Ingestion appliance

iDigBio
- Various IPT
- Pre-ingestion workflows
Cyberinfrastructure people

Present

M. Collins  
IT Developer

R. Figueiredo  
Professor, Senior Personnel

J. Fortes  
Professor, co-PI

R. Garand  
IT Developer

K. Love  
IT Developer

D. Stoner  
IT Developer
iDigBio is funded by grants from the National Science Foundation’s Advancing Digitization of Biodiversity Collections Program. 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. All images used with permission or are free from copyright.