Evolution of iDigBio capabilities

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

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
Current iDigBio cloud architecture

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

<table>
<thead>
<tr>
<th>iDigBio Specimen Portal</th>
<th>Appliances</th>
<th>Third Party API Consumers</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTML5, JQuery, CartoDB, ExpressJS</td>
<td>KVM, Xen, VirtualBox, VMware</td>
<td>Python, JQuery</td>
</tr>
<tr>
<td>iDigBio Search API</td>
<td>iDigBio Metadata API</td>
<td>iDigBio Object API</td>
</tr>
<tr>
<td>ElasticSearch</td>
<td>NodeJS, Restify, REST, JSON</td>
<td>NodeJS, Restify, REST, JSON</td>
</tr>
<tr>
<td>Text Indexes</td>
<td>Geospatial Indexes</td>
<td></td>
</tr>
<tr>
<td>ElasticSearch</td>
<td>CartoDB/Postgres</td>
<td></td>
</tr>
<tr>
<td>Bulk Text Storage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Riak</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Cloud Node

Advanced Computing and Information Systems laboratory
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
- BiSciCol, 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).
- 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 and supporting 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
- Weakness: 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)
Relational Databases

Symbiota
- Person omcollectors
- Taxon taxa
- CollObject omoccurrences
- Media Images
- Dataset Exsiccati title, number, links

TTD-AMNH
- Person collector
- Taxon mnl U flora_mnl
- CollObject omoccurrences
- Organization institution
- Media images
- Geography locality

Specify
- Person agent
- CollObject collectionobject
- Media attachments
- Organization collection
- DetHis
- Geology paleocontext

EMu
- Person eparties
- CollObject ecatalogue
- Media emultimedia
- Taxon etaxonomy
- Geography esites
- Geology esites
Potential MISC

TCN and other data providers

- Validosta (Specify)
- AMNH (MySQL)
- Taxon Tree
- Specimen
- Locality
- Geography
- Collector
- NYBG (DwC-A)
- Collection+Specimen+Taxon+Locality+Image
- Morphbank (DwC-A)
- Specimen+Taxa+Collector
- FLMNH Ichthyology (DwC-A)

Symbiota

- Collection
- Specimen
- Media

MISC

- Taxon Tree
- Specimen
- Collection
- Media
- Person

Ideal World

- Taxon Tree
- Specimen
- Geography
- Geology
- Morphology
- CollectingEvent
- Collection
- Media
- Person
- Reference
- Genetics
- DeterminationH

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

- Find unidentified specimens
- Plot distribution maps in time and space
- Perform an identification
- Understand community gene expression
- Validate taxa references
- Validate collecting event according to collector