Digitization Symposium
Association of Southeastern Biologists

- Gil Nelson (gnelson@bio.fsu.edu)
- Deb Paul (dpaul@fsu.edu)

(Florida State University)

12 April 2013
Charleston, WV

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.
Introducing iDigBio: An NSF Collaborative of the University of Florida and the Florida State University
Advancing Digitization of Biodiversity Collections

- Facilitate use of biodiversity specimen data to address environmental, scientific, and economic challenges
  - Biodiversity researchers and scientists
  - Educators
  - General public
  - Policy-makers

- Enable digitization of biodiversity collections data
  - Develop efficient and effective digitization standards and workflows
  - Respond to cyberinfrastructure needs

- Provide access to biodiversity data in a cloud-computing environment

- Plan for long-term sustainability of the national digitization effort
  - Expand participation: partners and data sources
NSF’s Grand Challenge

Digitize (text + images) and link one billion specimen records from collections across the U.S.
Seven Thematic Collections Networks (TCNs)


- Digitizing Fossils to Enable New Syntheses in Biogeography-Creating a PALEONICHES-TCN (University of Kansas)


- Mobilizing New England Vascular Plant Specimen Data to Track Environmental Change (Yale University)

National Resource (iDigBio), Thematic Collection Networks (TCNs), and Collaborators

7 TCNs, 130+ participating institutions, 49 states
Building the iDigBio Cloud

- Cloud-based strategy
  - Providing useful services/APIs (programmatic and web-based Application Programming Interface)
  - Federated scalable object storage and information processing
  - Digitization-oriented virtual appliances
  - Reliance on standards, proven solutions and sustainable software
- Continuous consultation with stakeholders
  - Surveys, working groups, workshops, person-to-person
What Makes iDigBio Unique?

- Ingest all contributed data with emphasis on GUIDs, not only a restricted set of selected data elements
- Maintain persistent datasets and versioning, allowing new and edited records to be uploaded as needed
- Ingest textual specimen records, associated still images, video, audio, and other media
- Ingest linked documents and associated literature, including field notes, ledgers, monographs, related specimen collections, etc.
- Provide virtual annotation capabilities and track annotations back to the originating collection
- Facilitate sharing and integration of data relevant to biodiversity research
- Provide computational services for biodiversity research
Recent and Ongoing Activities

- Assessment of common and effective practices (paper in *ZooKeys*)
- Minimum information for scientific collections working group
- Collaborative georeferencing pilot project at Godfrey Herbarium
- Digitization workflows working groups
- Biodiversity Informatics Manager working group
- Public Participation in Digitization of Biodiversity Specimens workshop
- Georeferencing working group & train-the-trainers workshop
- OCR/natural language processing working group & Hackathon
- ASB symposium and workshop
- Series of digitization training workshops
- Call for appliances
- Specimen data portal implementation
- Server hosting
Getting Started with Digitization

Gil Nelson
Florida State University
Ultimate Goals of Biological Collections Digitization

Output level: An abundance of scientifically useful and accessible data.

Constituency level: High quality exposure of the content and value of scientific collections.

Improvement level: Collaboration and workflow sharing across the collections community.
Global continua

guiding digitization

Local decisions and policies

Specific workflows

Emphasis in

Implementation in
Taking the long view means developing doable, effective, and sustainable strategies for robust digitization NOW.

Taking the short view means balancing long term goals with short term constraints, including a commitment to implementing future enhancements.

**Pressures mitigating the long view**

- So much data, so little time.
- Our collections are not getting smaller.
- The funding agencies have high output expectations.
- We only have 3 years to get this done.
- All of our data and all of our specimens are important.
- Let’s just use the images!
- Do the minimum now and enhance it later.
Future Tools Favoring the Short View

- OCR, NLP, and ICR (handwriting analysis) improvements.
- Automated image analysis for data extraction.
- Data mining of labels.
- Robotic technologies, conveyor belts, etc.
- Improvements in discovery/capture/use of duplicates.
- Improvements in voice recognition and other data entry technologies.
- Post-digitization tools for curation and quality control.
- Field data capture.
Digitization Decision Continua


Maximum Fitness ← Quantity ← Low cost/specimen ← Speed ← Traditional practices ← Image exemplars ← Image nothing ← Specimens only ← Static workflows

Maximum Fitness ← Quantity ← Low cost/specimen ← Speed ← Traditional practices ← Image exemplars ← Image nothing ← Specimens only ← Static workflows
<table>
<thead>
<tr>
<th>Robust Facilitators</th>
<th>Spartan Facilitators</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Emphasize immediate fitness for use</td>
<td>- Emphasize output</td>
</tr>
<tr>
<td>- Robust datasets</td>
<td>- Skeletal datasets</td>
</tr>
<tr>
<td>- Data validation/cleaning</td>
<td>- Defer validation/cleaning</td>
</tr>
<tr>
<td>- Integrated quality control</td>
<td>- Deferred quality control</td>
</tr>
<tr>
<td>- Integrated georeferencing</td>
<td>- Deferred georeferencing</td>
</tr>
<tr>
<td>- Intensive physical curation</td>
<td>- Deferred digital curation</td>
</tr>
<tr>
<td>- Record historical annotations</td>
<td>- Record current determination</td>
</tr>
<tr>
<td>- Staff specialization</td>
<td>- Staff generalization</td>
</tr>
<tr>
<td>- Small collection</td>
<td>- Large collection</td>
</tr>
<tr>
<td>- Emphasize images</td>
<td>- Emphasize data</td>
</tr>
<tr>
<td>- High quality images</td>
<td>- Low quality images</td>
</tr>
</tbody>
</table>

**Notes:**
- Robust datasets prioritize immediate fitness and high-quality images.
- Spartan datasets focus on output and low-quality images.
Thank you!