Trends in Digitization
Introduction to iDigBio
(Integrated Digitized Biodiversity Collections)

Gil Nelson
Institute for Digital Information and Scientific Communication
Integrated Digitized Biocollections
Florida State University

12th Pacific Science Inter-Congress
10 July 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.
The U.S. National Science Foundation estimates there may be as many as 1.8 billion biological and paleontological specimens stored in U.S. museums and academic institutions (perhaps as many as 2.5 billion representing about 1.8 million species worldwide). But, no one really knows!

In an effort to make these collections universally accessible to taxonomists, ecologists, researchers, and the general public, in 2011 NSF launched a $100 million, 10-year Advancing Digitization of Biodiversity Collections program and named Florida State University and University of Florida jointly as the national resource for digitization.
The goal is to digitize and make available via the Web at least 1 billion biological and paleontological records over the 10-year life of the project.
Mandate and Responsibility

- Provide/facilitate portal access to collections data
  - Make information available and discoverable
  - Label data and images
- Enable digitization and research
  - Facilitate digitization workflows
  - Oversee implementation of standards and best practices for digitization
  - Allow for data discovery across organismal groups
- Be a client of digitization projects/networks
  - Actively seek partners and data sources
  - Respond to cyberinfrastructure needs
- Engage communities
  - Collections
  - Research
  - Citizen science and education
- Support ADBC goals
  - Access to information
  - Support for collections
  - Sustainability
Mandate and Responsibility

- Provide/facilitate portal access to collections data
  - Make information available and discoverable
- Enable digitization and research
  - Facilitate digitization workflows
  - Oversee implementation of standards and best practices for digitization
- Be a client of digitization projects/networks
  - Actively seek partners and data sources
  - Respond to cyberinfrastructure needs
- Engage communities
  - Collections
  - Research
  - Citizen science and education
- Support ADBC goals
  - Access to information
  - Support for collections
  - Sustainability

Grand Challenge

Develop a cloud computing infrastructure that links biological data from collections across the U.S. through one or more unified web interfaces to overcome the limitations of “data silos.”
Mandate and Responsibility

- Provide/facilitate portal access to collections data
  - Make information available and discoverable
- Enable digitization and research
  - Facilitate digitization workflows
  - Oversee implementation of standards and best practices for digitization
- Be a client of digitization projects/networks
  - Actively seek partners and data sources
  - Respond to cyberinfrastructure needs
- Engage communities
  - Collections
  - Research
  - Citizen science and education
- Support ADBC goals
  - Access to information
  - Support for collections
  - Sustainability

Grand Challenge

More recently, we have been encouraged by NSF to establish international collaboration and sharing to overcome the limitations of “data silos.”
The challenges being pursued by iDigBio are reflective of worldwide trends in digitization

- Global Biodiversity Informatics Facility (GBIF)
- OpenUp! (European Union)
- Atlas of Living Australia (ALA)
- SYNTHESYS (20 European natural history museums)
- IDigBio (U.S.)
- CRIA (Brazil)
Ten Thematic Collections Networks (TCNs) plus 2 Partner to Existing Networks (PENs)


- 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*)


- ???

- ???

- ???

New as of 1 July 2013
National Resource (iDigBio), Thematic Collection Networks (TCNs)

To date: 10 TCNs, 2 PENs, 160+ participating institutions, 49 states
TCN: InvertNet
PEN: Lichens & Bryophytes
TCN: MacroFungi
TCN: New England Vascular Plants
TCN: Tri-Trophic
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, interest groups, workshops, person-to-person
Key Features of iDigBio

- Ingest all contributed data with emphasis on use of GUIDs, no restrictions
- Maintain persistent datasets and versioning, allowing new and edited records to be uploaded as needed while preserving existing records
- Ingest textual specimen records, plus associated still images, video, audio, and other media (or links to these resources as determined by the provider)
- 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 (collaborating with FilteredPush)
- Facilitate sharing and integration of data relevant to biodiversity research
- Provide computational services for biodiversity research
Identifying Objects

Add column to data record for a globally unique, persistent identifier.

http://www.talltimbers.org/museum.html#Birds:279
urn:uuid:3Ab1495230-ac34-42ea-b6b7-7af8b9f1b212

Resolver

UUID or GUID does not have to appear on the specimen itself.
Recent, Ongoing, Upcoming Activities

- Assessment of common and effective digitization practices (paper in *ZooKeys*)
- Working groups
  - Minimum information for scientific collections working group (MISC)
  - Digitization workflows working groups
  - Georeferencing
  - Optical character recognition (OCR)
  - Biodiversity Informatics Manager working group
- Workshops - year 2:
  - > 150 institutions, 9 workshops, 3 symposia
  - 368 sponsored participants
  - Video archives on Vimeo, live streaming for remote participation
  - New model this year: train the trainer
  - Series of digitization training workshops (herbaria, wet collections, entomology, paleontology, fluid-preserved invertebrate imaging, small herbaria, )
- Server hosting: 8 virtual machines, TCN support
- Specimen data portal and website – continuous improvements
- Call for appliances, frequent opinion surveys
Launched the **Biodiversity Informatics Managers Working Group** to focus on the role of biodiversity informatics manager as an essential component underpinning the successful digitization enterprise, including the definition and delineation of career path dimensions, skill sets, academic training requirements, and recommendations about the placement of this role within the organizational structure of museums and academic institutions.

Launched the **International Whole-Drawer Digitization Interest Group** in collaboration with partners at CSIRO, with representatives from Australia, Germany, The Netherlands, the United Kingdom, and the United States.
Trends in Digitization
Essential Components of Successful Digitization Programs

Gil Nelson
iDigBio
Institute for Digital Information and Scientific Communication
Florida State University

12th Pacific Science Inter-Congress
10 July 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.
Digitization

Converting specimens and specimen-related information to digital format

- Label data to digital records
- Specimens to images
- Ancillary materials to digital records or images
  - field notes, field books, catalogs, ledgers, monographs, journal articles, white papers, etc.
- Audio to digital
- Video to digital

Essential components:
- Data standards
- Data/Image capture
- Workflows/protocols
Two things we recognized from the outset:

1. The importance of clear, biologically relevant standards to guide data acquisition and distribution.

1. The importance of effective, community-based digitization workflows and practices.
The product of the MISC working group was designed to:

• provide detail to the data model,
• reflect a biologist’s or collection manager’s perspective,
• ensure that all data currently or potentially stored in collections databases are accounted for,
• evolve over time,
• Prioritize data elements as required, highly desired, or supplementary,
• take a scientific perspective on data fitness,
• start with Darwin Core as a foundation and augment from other schemas where necessary,
• map MISC to existing schemas.
Assessing Digitization Practices in Biological and Paleontological Collections

28 Collections
10 Museums
Spanning biological and paleontological collections
Insects and other invertebrates, plants, birds, mammals
Wet, dry
Five task clusters that enable efficient and effective digitization of biological collections

Gil Nelson¹, Deborah Paul¹, Gregory Riccardi¹, Austin R. Mast²

¹ Institute for Digital Information, Florida State University, Tallahassee, FL 32306-2100, United States ² Department of Biological Science, Florida State University, Tallahassee, FL 32306-4295, United States

Corresponding author: Gil Nelson (gnelson@bio.fsu.edu)
Composite of Workflows Observed

Dominant Digitization Patterns Observed

a. Data to Occasional or Optional Image to Distribution

- Predigitization curation
- Data capture
- Capture Exemplar images
- Process images
- Store images
- World Wide Web
- Data aggregators
- In-house usage

b. Parallel Data/Image to Distribution

- Predigitization curation
- Data capture
- Capture images
- Process images
- Store images
- World Wide Web
- Data aggregators
- In-house usage

c. Image to Data to Distribution

- Predigitization curation
- Capture images
- Process images
- Store images
- Data capture
- World Wide Web
- Data aggregators
- In-house usage
Developing Robust Object to Image to Data Workflows (DROID)

Scientific Software Innovation Institutes
Yale Peabody Museum
Biodiversity Institute, KU
iDigBio

Digitization Workflow Workshop Report

Developing Robust Object-to-Image-to-Data (DROID) Workflow Workshop
30-31 May 2012, Florida Museum of Natural History, University of Florida (FLMNH)

Biological specimens document the historical and modern occurrence of plant and animal species - and most of what we know about the diversity and distribution of life on earth. The majority of collected specimens have yet to be digitized, but at the same time, current biodiversity digitization processes and technologies are often inefficient and uncoordinated, preventing timely and cost-effective digitization of these specimens. This research workshop focused on the design, documentation, and optimization of workflows necessary to transform physical specimens collected in the field into useful, sharable, and manageable digital objects within a collection. Approximately twenty hands-on collections experts provided input during the workshop.

Why document workflows?
Workflow documentation is a powerful tool both within a collection and across the entire collections community. Internally, effective workflow documentation for a collection can highlight inefficiencies, identify bottlenecks that hinder throughput, and expose opportunities for automation. Workflow documentation also serves as initial input into the development of collections digitization training materials and checklists that improve quality and consistency. Collectively, the documentation and sharing of effective digitization workflows 1) enables collections to test and compare results in order to identify optimal processes, 2) prevents collections from investing resources in (re)designing a process that already exists within the community, 3) enhances communication and standardization by enabling agreement on a common workflow vocabulary for each task, and 4) exposes new innovations to the entire community. Additionally, comprehensive workflow documentation enables the natural history collections community to approach digitization and technology innovators from other domains, such as library sciences, robotics development, industrial workflow design, or software development, for assistance. This includes the ability to present documented workflows to collaborators to learn about improved methods as well as innovative or re-purposed tools.

But we are unique!
The workshop participants recognized that various factors impact the design of appropriate workflows for a particular collection.

- Tradeoffs must be determined at a high level (e.g., volume of objects digitized vs. cost vs. completeness of each record). These decisions may be dependent upon grant requirements or other externally imposed requirements.
- Local decisions and policies may impact a digitization workflow, including institutional or collection policies.
- Specific workflow decisions within a collection will be based upon constraints such as the quantity of personnel, available expertise, available funds, physical layout of the collection space, the method of specimen preservation, and other factors.

To overcome these issues, the DROID workshop participants produced two recommendations. The first was to approach the challenge by developing workflows specific to these broad presentation types, including 1) objects on flat sheets (typically plant specimens), 2) objects on pins (primary insects), and 3) larger three-dimensional objects (fossils, mammals, reptiles, etc). Each high-level presentation type has enough similarity that workflows can be developed that have a reasonable number of common tasks. Participants then divided into groups, each focused on the requirements for a specific type.

A second recommendation was to develop more generalized, flexible workflows, with common tasks grouped into "modules" that could be inserted, removed, or modified within a collection’s workflow based on the features described above. Workshop participants were actively able to...
• The Flat Sheets and Packets Working Group completed modules and associated tasks for herbarium and related collections (October 2012).
• The Pinned Things in Trays and Drawers finished and posted its work for entomology (January 2013).
• 3D Objects in Spirits in Jars and Vials completed and posted its workflows for fluid-preserved specimens (May 2013).
• 3D Objects in Drawers and Trays workflows group started work in June 2013.
• Preparation-independent workflows to follow (2013).
Digitization Workflows

Efficient and effective workflows are at the heart of successful biological and paleontological collections digitization. Much work has been done with developing workflows and protocols at the museum and collections level, but few of those workflows have been documented or made available to the larger collections community. iDigBio, through its Documentation pages, is establishing an online repository for sharing existing customized workflows from as many collection types and institutions as possible, an idea that stems largely from the Developing Robust Object-to-Image-to-Data (DROID) workshop held May 30-31, 2012. We have assembled an initial set of workflows, including selected examples from the DROID workshop, as well as those developed by iDigBio staff. Here we offer the beginnings of the repository and encourage those in the community to both discuss the workflows via the forum links, and to contribute to this resource by adding new workflows and updating existing workflows. If you would like to submit a workflow for inclusion on this page, please contact DigBio for instructions. We are also assembling detailed modules of tasks to be performed at each stage of the workflow, accessible on our Workflow Modules and Tasks page.

<table>
<thead>
<tr>
<th>Workflow</th>
<th>Contributor</th>
<th>Workflow Documentation</th>
<th>Link to Public Comments (Forums)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dominant Digitization Workflows</td>
<td>iDigBio</td>
<td>Dominant Digitization Workflows Documentation</td>
<td>Dominant Digitization Workflows Forum</td>
</tr>
<tr>
<td>Field Notes-to-Data-to-Image</td>
<td>iDigBio</td>
<td>Field Notes-to-Data-to-Image Documentation</td>
<td>Field Notes-to-Data-to-Image Forum</td>
</tr>
<tr>
<td>Specimen-to-Data-to-Exemplar Image</td>
<td>iDigBio</td>
<td>Specimen-to-Data-to-Exemplar Image Documentation</td>
<td>Specimen-to-Data-to-Exemplar Image Forum</td>
</tr>
<tr>
<td>Object-to-Image to Data (1)</td>
<td>iDigBio</td>
<td>Object-to-Image to Data (1) Documentation</td>
<td>Object-to-Image to Data (1) Forum</td>
</tr>
<tr>
<td>Object-to-Image to Data (2)</td>
<td>iDigBio</td>
<td>Object-to-Image to Data (2) Documentation</td>
<td>Object-to-Image to Data (2) Forum</td>
</tr>
<tr>
<td>University of Vermont Herbarium</td>
<td>Dorothy Allard</td>
<td>University of Vermont Herbarium Documentation</td>
<td>University of Vermont Herbarium Forum</td>
</tr>
<tr>
<td>Southwest Collections of Arthropods Network</td>
<td>Paul Heinrich</td>
<td>Southwest Collections of Arthropods Network Documentation</td>
<td>Southwest Collections of Arthropods Network Forum</td>
</tr>
</tbody>
</table>

https://www.idigbio.org/content/digitization-workflows

Posted To Collaborative Workflows Page Linked to the Digitization Resources Wiki
Workflow Modules and Task Lists

One outcome of the [DECO](#) (Developing Robust Object-to-
Image-to-Data) workflow workshop held in May 2012 was the
establishment of a series of working groups, each focused on
workflow modules and tasks for various preparation types.
The first of these groups, informally called the [Flat Sheets
and Packets Working Group](#), was charged with fleshing out
task lists for digitizing vascular and non-vascular plant
collections. The second working group, [Placed Specimens in Trays and Drawers](#), invested its time developing modules to support effective
eumological digitization workflows. Other preparation types will follow, including fluid collections and other 3-dimensional objects, concluding
with the development of an overall project management module designed to provide guidance for developing and managing digitization projects
across disciplines and preservation types.

We have chosen a modular approach for presenting our results in order to accommodate the broad range of workflow implementations within the
collections community. We recognize that there is no consensus workflow that fits all situations, even within a single preservation type. In light
of this, we have attempted to assemble a series of comprehensive task lists to serve as foundations from which institutionally specific workflows can be
created. Not all institutions will use every task, but we hope that the lists we have developed encompass all relevant digitization tasks. We also
hope that those in the collections digitization community will provide feedback on these lists; either through forum posts or e-mails to Gil Nelson,
alerting us to deficiencies and oversights.

Links to published modules as they are completed are provided below:

**Flat Sheets and Packets Working Group - Vascular and Non-vascular Plants**

- Module 1: Pre-digitization Curation Tasks
- Module 2: Imaging Station Setup Camera
- Module 3: Imaging Station Setup Scanner
- Module 4: Imaging Tasks
- Module 5: Image Processing Tasks (Rev 2012-11-07)
- Module 6: Data Capture Tasks

**Placed Things in Trays and Drawers Working Group - Insects**

- Module 9: Generic Tasks Applicable to Two or More Modules
- Module 10: Pre-digitization Curation Tasks
- Module 11: Insect Imaging Tasks
- Module 12: Whole-drawer Imaging Tasks
- Module 13: Label Imaging Tasks
- Module 14: Image Processing Tasks
- Module 15: Data Capture From Specimen Tasks
- Module 16: Event Data Capture Tasks
- Module 17: Quality Assurance Tasks

---

**Workflow Detail: Specimen Image Processing (Pinned Things)**

<table>
<thead>
<tr>
<th>Task ID</th>
<th>Task Name</th>
<th>Explanations and Comments</th>
<th>Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>Transfer images from camera to immediate image processing storage.</td>
<td>This task varies by institution. Some institutions record images to a card within the camera, others download directly to the imaging computer or an external or network drive as images are recorded. Transfer to the image processing storage should be periodic, at least daily.</td>
<td>Ample storage space with backup procedures (also see T8-T9).</td>
</tr>
<tr>
<td>T2</td>
<td>Adjust orientation and crop images, as necessary.</td>
<td>Images should be framed and recorded as precisely as possible to prevent the need for cropping. In cases where cropping is required, batch crop routines for processing multiple images to identical parameters are preferable. Where batch cropping is not possible due to random variation of exemplar image files, individual cropping may Image management or processing software (e.g., Photoshop, Lightroom, ImageMagick, Gimp, or similar).</td>
<td></td>
</tr>
</tbody>
</table>
Community-based Workflow Wiki for sharing workflows across prep types and institutions.
Documentation and Instructions

- **Written Protocols**
  - Essential!
  - Include pictures.
  - Attention to detail (leave nothing to the imagination).
  - Express limits on technician authority.

- **Feedback Loops**
  - Technicians: best source of efficiency adaptations, either by show or tell.
  - Easy methods for receiving feedback.
  - Personal copies of the protocol.
  - Master copy available via Google docs or other shared storage for updates and suggestions.
Trends in Digitization: Getting Started

Gil Nelson
iDigBio
Institute for Digital Information and Scientific Communication
Florida State University

12th Pacific Science Inter-Congress
10 July 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.
Biodiversity Digitization:
Ultimate Goals

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 parameters guiding digitization

Local decisions and policies

Specific workflows

Emphasis in

Implementation in
Tracks to Digitization

• **Taking the inside track [short view]** is often based on stretching the institution’s resources. Decisions are made to maximize resources available for user-initiated digitization by using solid baseline practices. The primary focus on the inside track is to get the job done quickly and to fill the user’s request.

• **Taking the middle track** has the widest range of options, standards, and results. This is the most flexible of the tracks, where decisions often fall in gray areas.

• **Taking the outside track [long view]** focuses on the collections themselves. While users may initiate digitization, it is undertaken to deliver materials to a greater public. These decisions may lead to comprehensive digitization, such as an entire book, series, or collection. The goal is to create maximum access to special collections, using preservation and archival standards. This track usually involves a level of thought and planning that is more in-depth than the fulfillment of day-to-day digitization requests.
The Long View

Fully populated collection object records searchable across all embedded label data and linked to specimen images, field images, related audio and video recordings, duplicate collections, white papers, grey literature, published works, collecting locality, georeferences, nomenclatural histories, collector information, generic habitat descriptions, taxonomic trees, phylogenies…and anything else related to the specimen that might help scientists and others better understand the collection object in question.
Long view can be daunting!

Balancing the long view with the short view: The local decision

How does an institution develop doable, effective, and sustainable strategies for balancing long term goals with short term constraints, while maintaining a commitment to implementing future enhancements?

Pressures mitigating the long view
  So much data, so little time.
  Collections are not getting smaller (proactive vs. legacy).
  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!
  We’ll do the minimum now and enhance it later.
Digitization Continua/Decision Points

Current Tools ⊲ Potential Future Tools

Maximum fitness ⊲ Maximum output

High cost/specimen ⊲ Low cost/specimen

Digital protocols ⊲ Traditional practices

Image everything ⊲ Image exemplars

Ancillary materials ⊲ Specimens only
Future Tools Favoring the Inside/Middle Tracks

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
Long view

Facilitators

- Emphasize fitness for use
- Robust datasets
- Data validation/cleaning
- Integrated quality control
- Integrated georeferencing
- Intensive curation
- Record historical annotations
- Staff specialization
- Emphasize images
- High quality images
- Small collection

Short View

Facilitators

- Emphasize output
- Spartan datasets
- Defer validation/cleaning
- Deferred/minimum quality control
- Deferred georeferencing
- Deferred or cursory curation
- Record current determination
- Staff generalization
- Emphasize data
- Low quality images
- Large collection
Metrics

Issues in Productivity and Use

Comparability: What is being measured?
- What is included in the output?
- Are all steps in the process accounted for?
- Are all expenditures of time accounted for?
- How do we arrive at a true per specimen cost?

Measuring productivity (comparability across collections):
- Unit (output per unit time vs. expenditure/project totals)
- Data fitness (should data robustness be factored in the calculus?)

Measuring use:
- Number of virtual visitors to the collection?
- Types of visitors?
- Average time per visitor?
- Type of data accessed?