Digitization Workshop
Association of Southeastern Biologists

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

(Florida State University)

13 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.

iDigBio
Integrated Digitized Biocollections

Workshop Goals

- Introduce Integrated Digitized Biocollections (iDigBio).
- Review existing workflows from workshop participants.
- Detail important principles for digitization workflow design and development.
- Outline examples of major workflow patterns.
- Consider the cultural/social issues that underpin a digitization program.
- Present strategies for serving data and images on the web.
- Develop an understanding of the importance of providing identifiers for your specimens and data.
- Offer methods for improving data in Excel spredsheets.
- Provide for plenty of time for discussion, contributions, and questions!

Advancing Digitization of Biodiversity Collections

- **Facilitate use of biodiversity data to address environmental and economic challenges**
  - Researchers
  - Educators
  - General public
  - Policy-makers

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

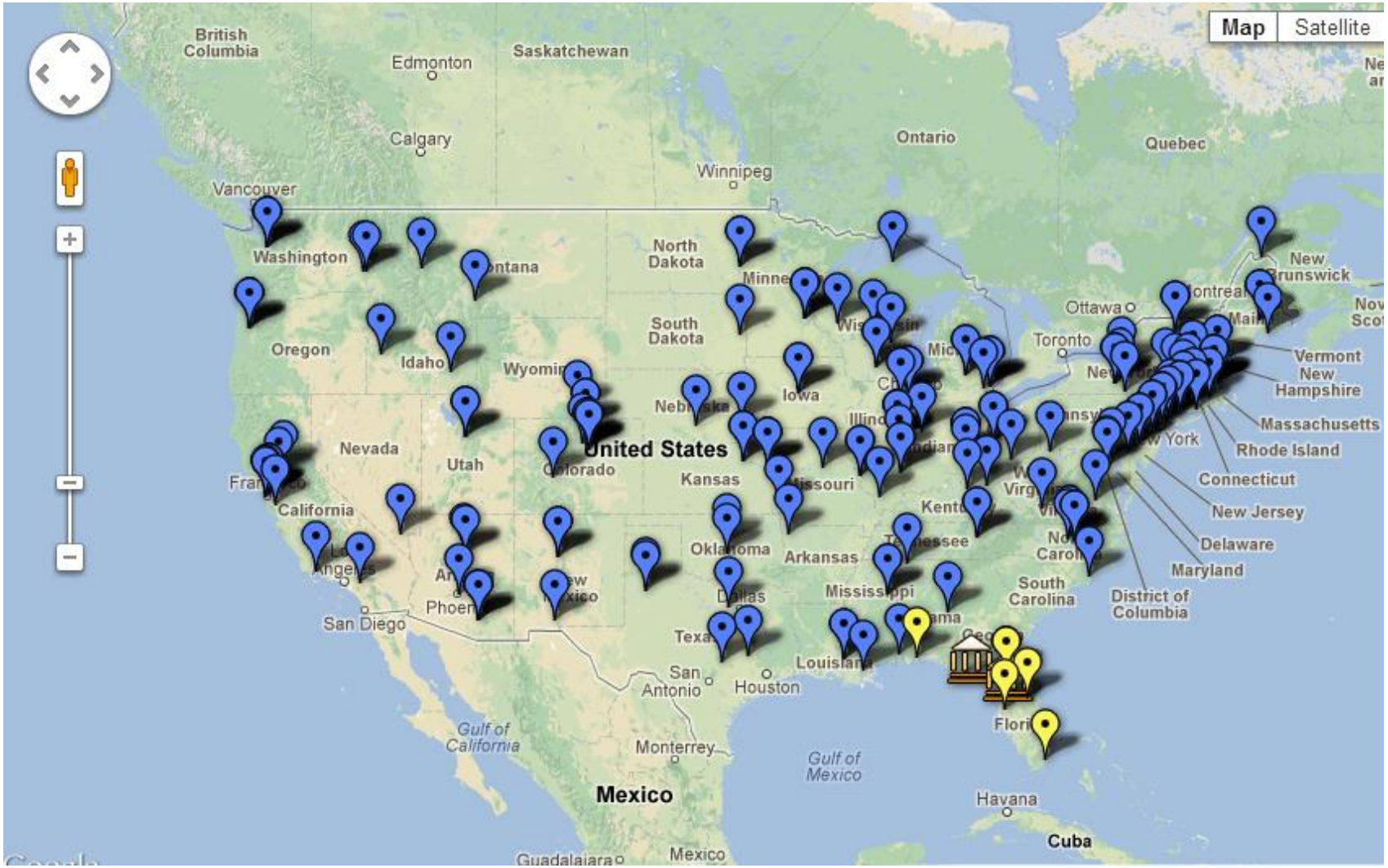
- **Provide portal access to biodiversity data in a cloud-computing environment**

- **Plan for long-term sustainability of the national digitization effort**
  - Expand participation: partners and data sources

Seven Thematic Collections Networks (TCNs)

- InvertNet: An Integrative Platform for Research on Environmental Change, Species Discovery and Identification (*Illinois Natural History Survey, University of Illinois*) http://invertnet.org
- Plants, Herbivores, and Parasitoids: A Model System for the Study of Tri-Trophic Associations (*American Museum of Natural History*) http://tcn.amnh.org
- North American Lichens and Bryophytes: Sensitive Indicators of Environmental Quality and Change (*University of Wisconsin – Madison*) http://symbiota.org/nalichens/index.php http://symbiota.org/bryophytes/index.php
- Digitizing Fossils to Enable New Syntheses in Biogeography-Creating a PALEONICHES-TCN (*University of Kansas*)
- The Macrofungi Collection Consortium: Unlocking a Biodiversity Resource for Understanding Biotic Interactions, Nutrient Cycling and Human Affairs (*New York Botanical Garden*)
- Mobilizing New England Vascular Plant Specimen Data to Track Environmental Change (*Yale University*)
- Southwest Collections of Anthropods Network (SCAN): A Model for Collections Digitization to Promote Taxonomic and Ecological Research (*Northern Arizona University*) http://hasbrouck.asu.edu/symbiota/portal/index.php

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

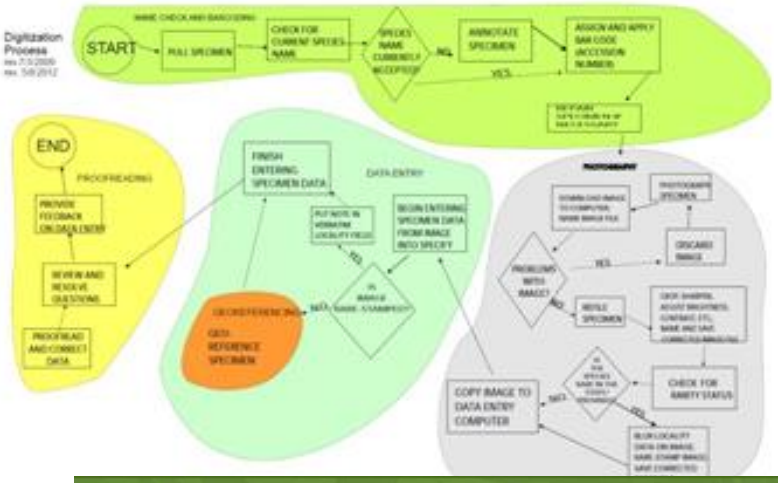


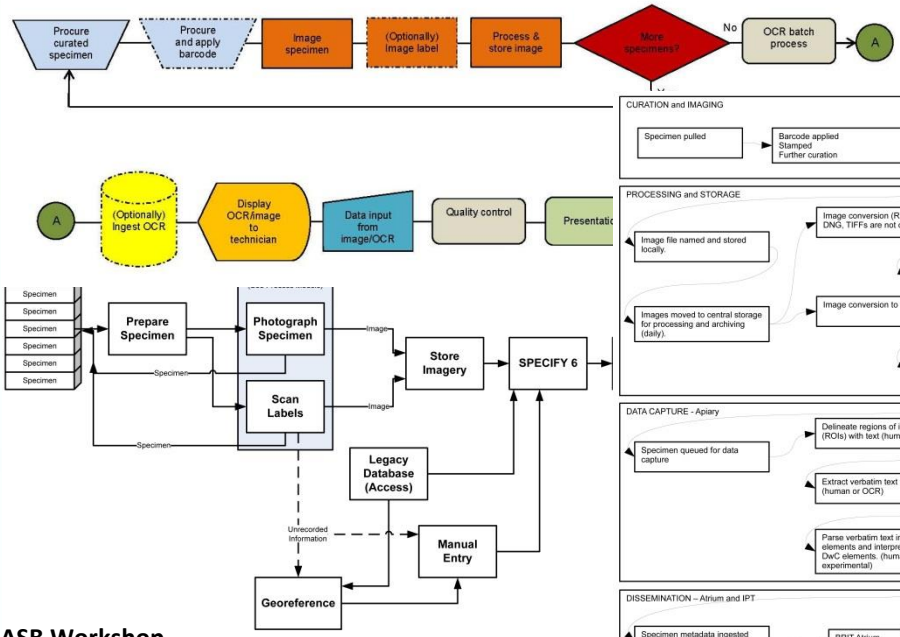
7 TCNs, 130+ participating institutions, 49 states

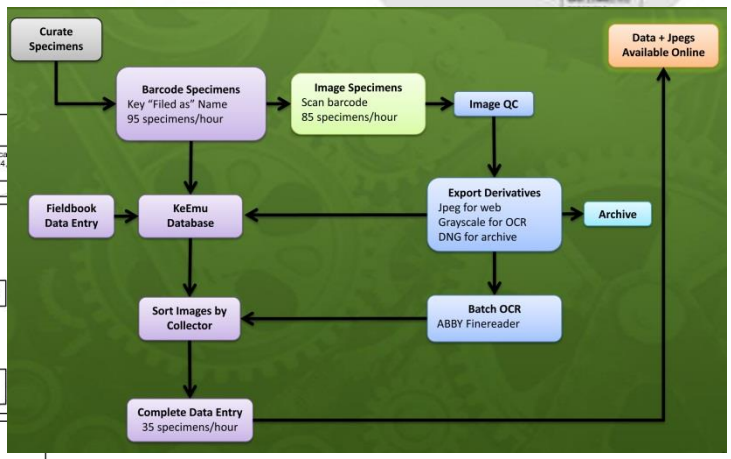
iDigBio

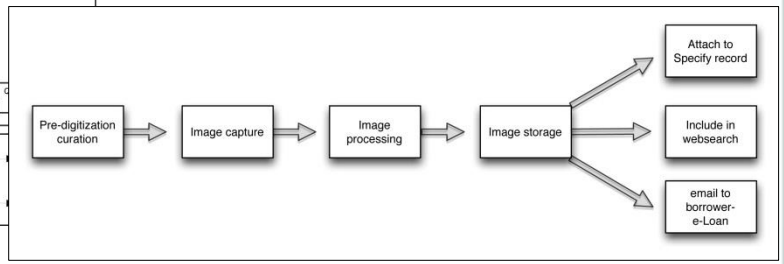
Integrated Digitized Biocollections

Digitization Workflows









ASB Workshop
Gil Nelson/Deb Paul
13 April 2013
Charleston, WV

iDigBio Integrated Digitized Biocollections

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

ZooKeys 209: 19–45 (2012)
doi: 10.3897/zookeys.209.3135
www.zookeys.org

RESEARCH ARTICLE

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

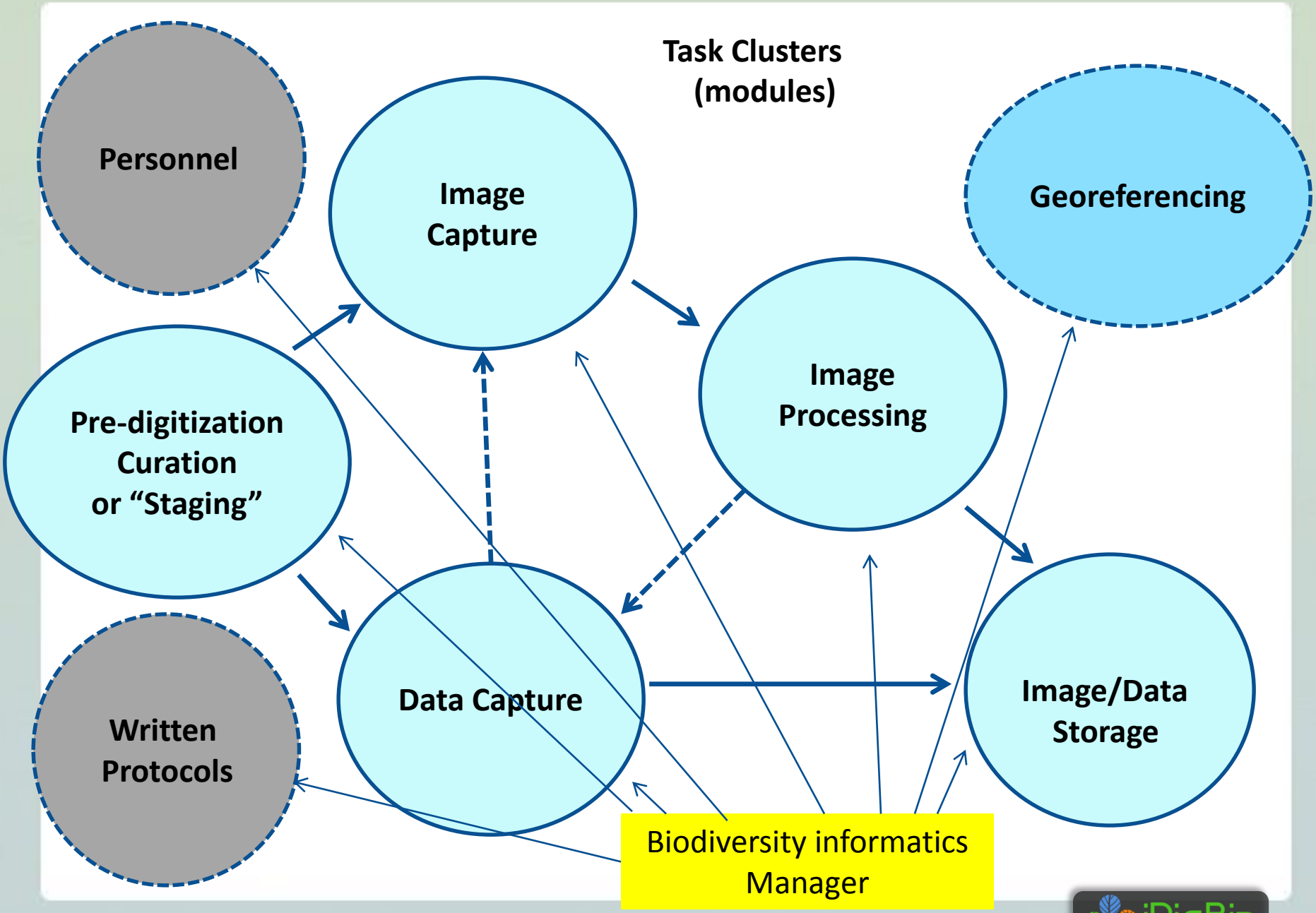
Five task clusters that enable efficient and effective digitization of biological collections

Gil Nelson[1], Deborah Paul[1], Gregory Riccardi[1], Austin R. Mast[2]

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)

iDigBio
Integrated Digitized Biocollections

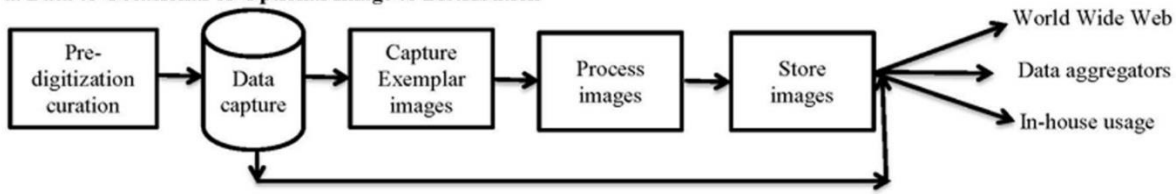

Task Clusters
(modules)
Personnel
Image
Capture
Georeferencing
Pre-digitization
Curation
or "Staging"
Image
Processing
Written
Protocols
Data Capture
Image/Data
Storage
Biodiversity informatics
Manager

Workflows Patterns Observed

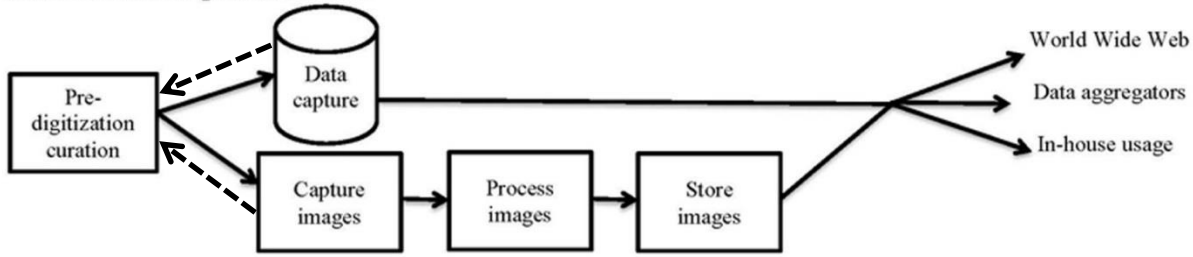
Dominant Digitization Patterns Observed

Figure 1: Dominant Digitization Workflows

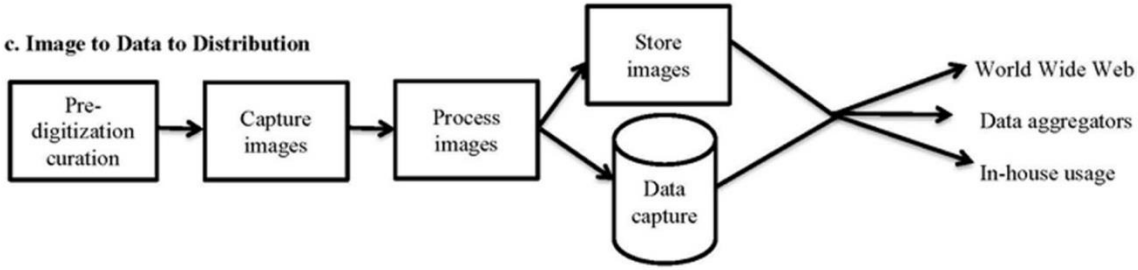
a. Data to Occasional or Optional Image to Distribution



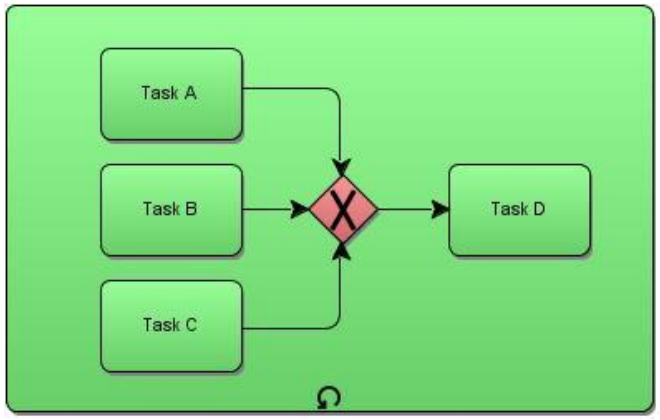
b. Parallel Data/Image to Distribution



c. Image to Data to Distribution



Values of defined workflows



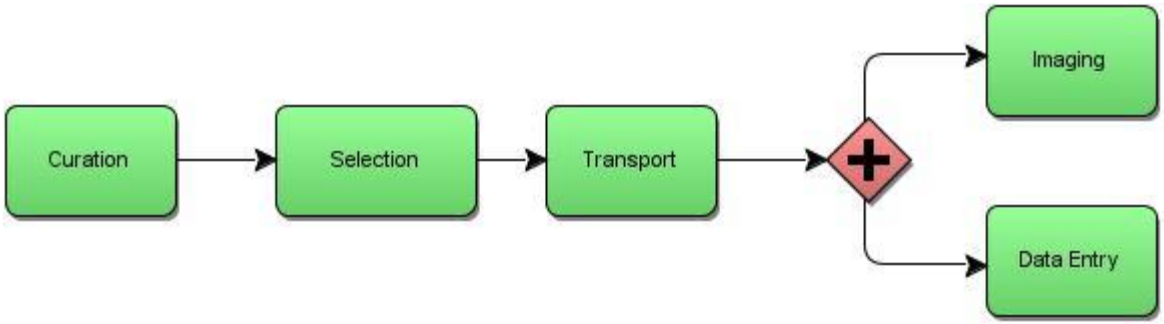
- Promote efficiency and automation of processes
- Facilitate routing and scheduling of activities
- Provide for balancing workloads
- Ensure that processes are visible and predictable
- Allow for escalations and notifications
- Enhance tracking of tasks
- Foster collaboration of all parties involved
- Stimulate the convergence of process and information
- Promote continuous evaluation and redesign

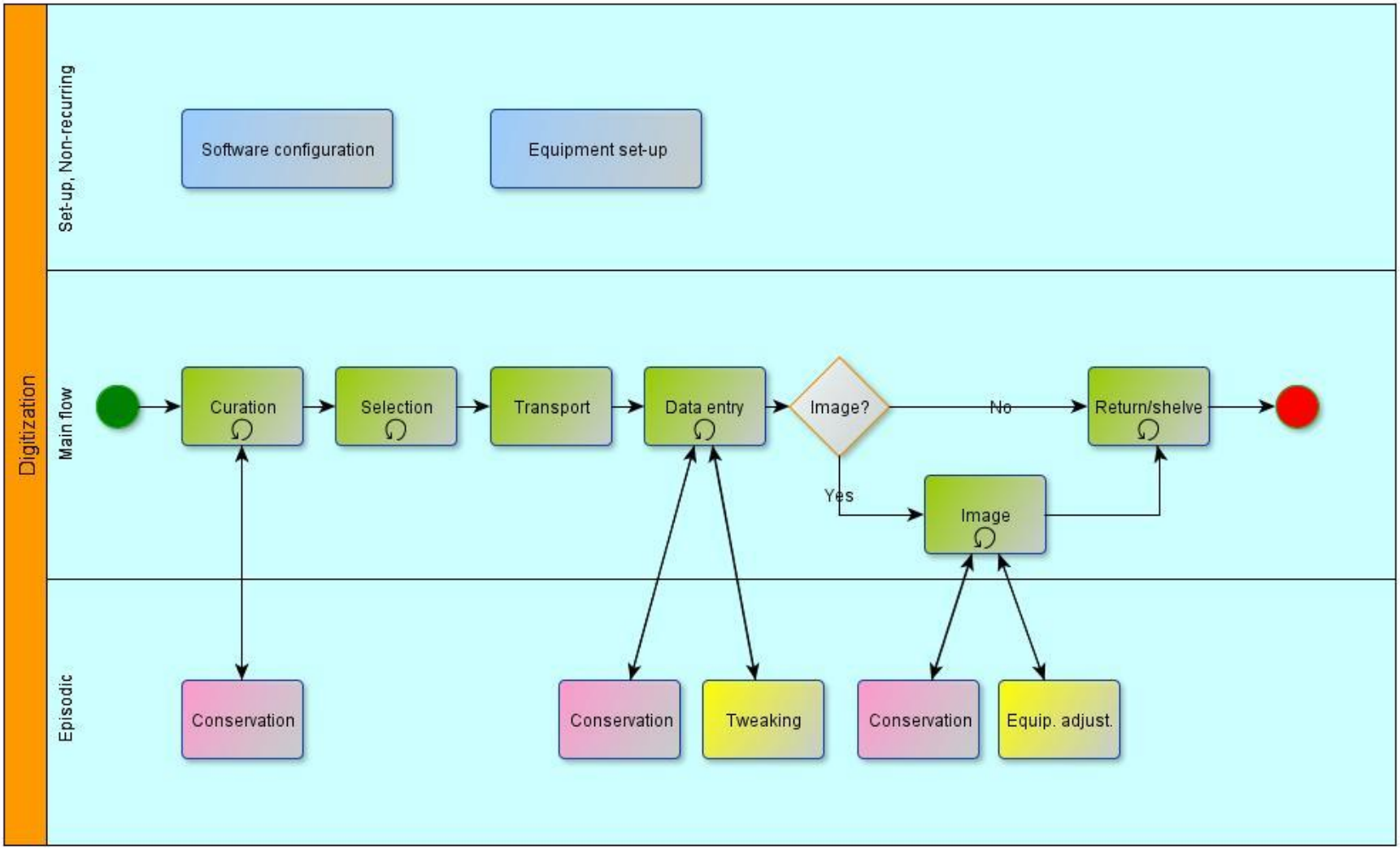
Pre-planning a Workflow Process

- Identify the database management system and imaging equipment to be used.
- Identify the process/module for which the workflow will be designed.
- Identify (*in excruciating detail*!) the tasks (or task clusters) that constitute the process/module.
- Identify the specific actions to be taken and the attributes (if any) associated with these actions.
- Identify roles (and only secondarily the people who will fill them).
- Identify points/processes/parameters for notifications and escalations.
- Identify dependencies, transitions, and iterations.
- Determine minimal data requirements for defining a complete record.
- Determine how records and objects will be uniquely identified in a global environment.
- Determine how identifiers will be assigned.
- Determine if/how identifiers will be affixed to the specimen/lot/collection object.
- Determine a consistent file naming strategy for images, attachments, other related materials.
- Determine file storage needs and location for data, images, and ancillary materials.
- Define and diagram flow.

Example Processes (Modules), their Cycles and Dependencies

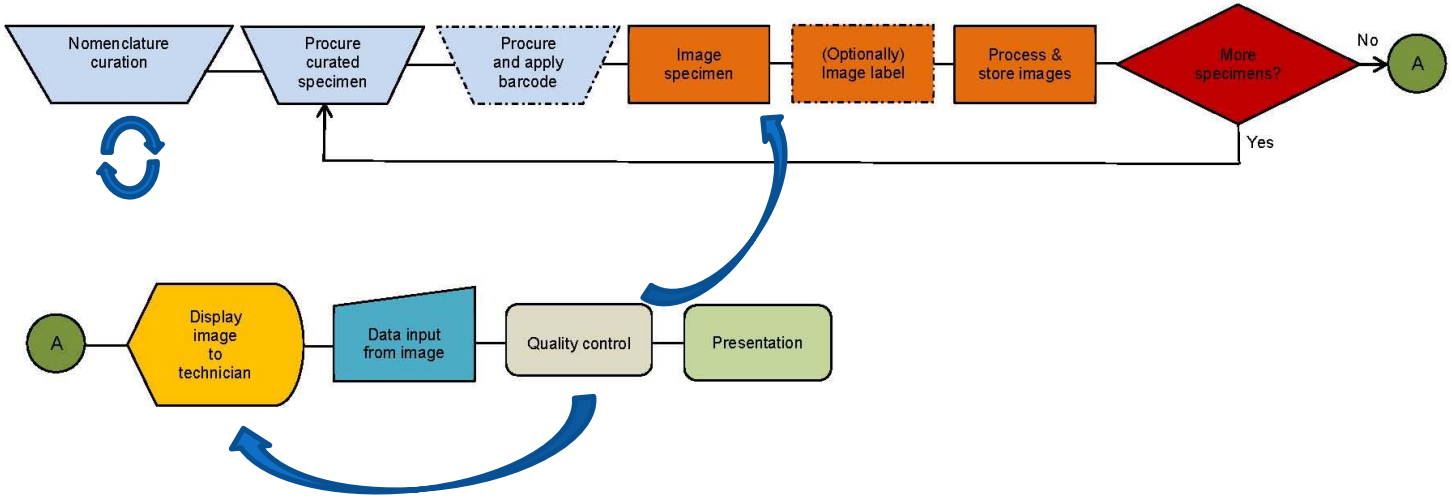
Process	Cycle	Dependency
Software configuration	Once/non-recurring	
Equipment set-up	Once/non-recurring	
Specimen curation	Recurring	
Specimen selection	Recurring	Pre-digitization curation
Specimen transport	Recurring	Specimen selection, imaging, data entry
Conservation	Episodic	Curatorial processes, imaging, data entry
Data entry	Recurring/tasks iterative	Specimen transport
Imaging	Recurring/tasks iterative	Specimen transport
Equipment adjustment	Episodic	Data entry/imaging
Software update/tweaking	Episodic	
Specimen return/shelving	Recurring	Imaging or data entry




Digitization
Set-up, Non-recurring
Software configuration
Equipment set-up
Main flow
Curation
Selection
Transport
Data entry
Image?
No
Return/shelve
Yes
Image
Episodic
Conservation
Conservation
Tweaking
Conservation
Equip. adjust.

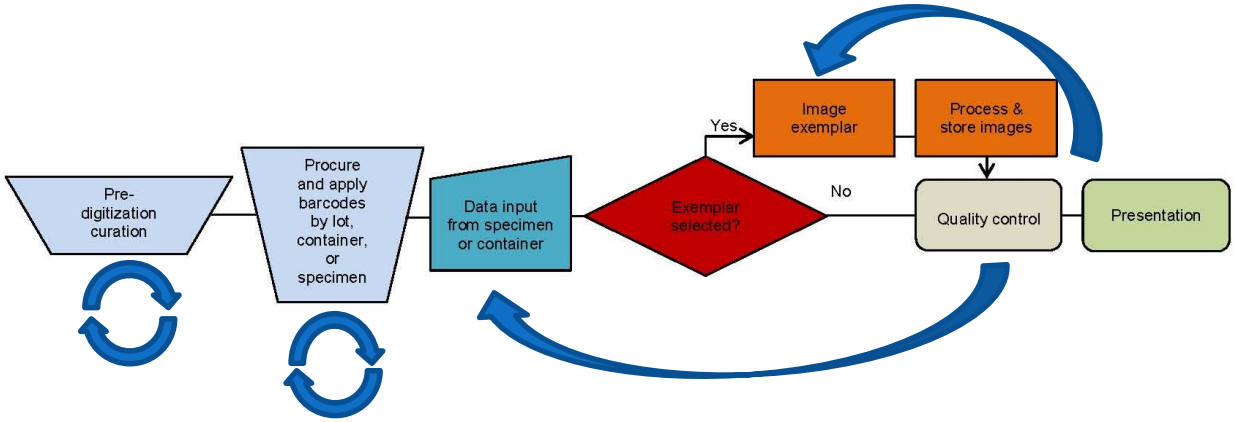
O2I2D(2)—Existing Specimen Workflow: Object to Image to Data

This workflow is designed for capturing images of existing specimens and using these images as the basis for data capture. Depending upon preparation type, barcodes are sometimes applied inline as the step immediately previous to imaging (shown optionally below) and other times en masse within an independent step during which several dozen or several hundred barcodes are applied in preparation for imaging. Pre-digitization curation and annotation is particularly important in this workflow to ensure that the current nomenclature to be used in data entry is obvious and clearly visible in the image.



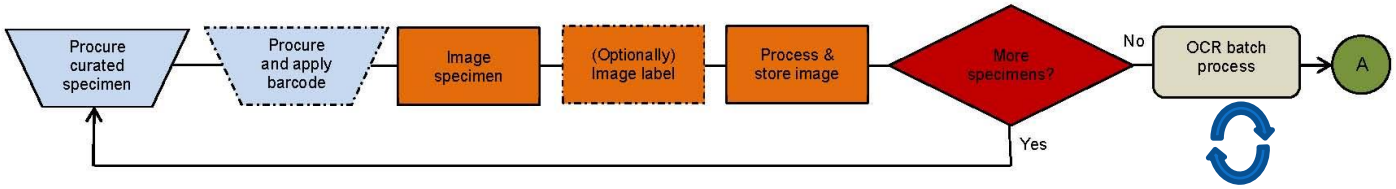
O2D2EI—Existing Specimen Workflow: Object to Data to Exemplar Images

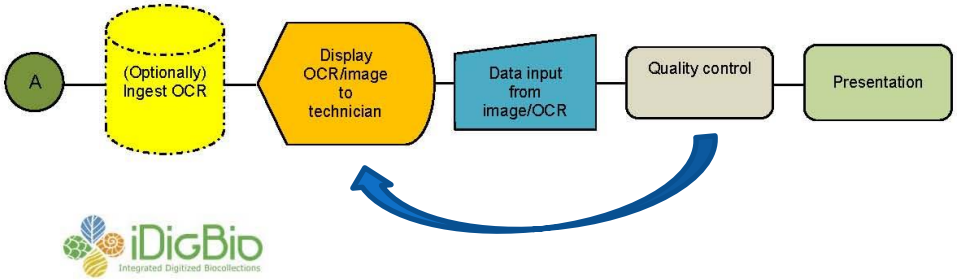
This workflow is in use for collections that capture data in specimen lots, collecting events, taxon container, or other aggregates, but capture images only for exemplar specimens. Data capture is effected from specimen labels. Depending upon preparation type, barcodes are usually applied inline—often to the containing tray or container—as the step immediately preceding data entry. Hence, barcodes may designate a single specimen or an aggregate of specimens, such as a unit tray within an insect drawer or ethanol-filled container in a wet collection. Barcode application is executed prior data entry and image capture usually follows data entry. Pre-digitization curation, including nomenclatural annotations and specimen organization, is usually important in this workflow.



O2I2D(1)—Existing Specimen Workflow Using Optical Character Recognition: Object to Image to Data

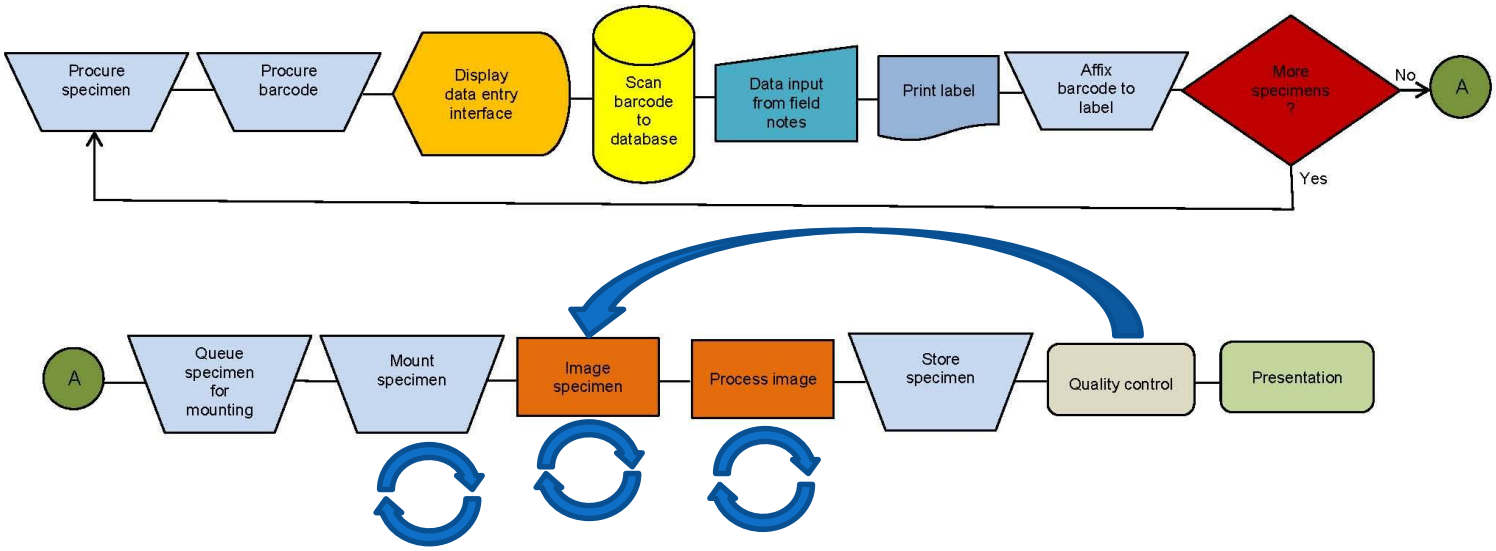
This workflow is designed to capture images of existing specimens, pass the images through optical character recognition (OCR) software, and use the combination of image and OCR output to capture data. There are variations on this workflow. For example, depending on preparation type, barcodes are sometimes applied inline as the step immediately previous to imaging (shown optionally below) and other times en masse within an independent step during which several dozen or several hundred barcodes are applied in preparation for imaging. OCR may also occur in various ways: 1) in batch (as shown below), with numerous images being processed following the close of one or more imaging sessions, 2) "on the fly" as a record and its associated image are loaded for data entry, or 3) one image at a time as a step immediately following the imaging of each specimen. OCR output may be ingested into a field in the database (shown optionally below), stored as individual text files within the computer's file system, or virtually processed at the time the image is presented to the data entry technician. The presentation of images and OCR to data entry technicians occurs in a single interface in which database fields, OCR output, and specimen image are simultaneously visible. Pre-digitization curation and annotation is particularly important in this workflow to ensure that the current nomenclature to be used in data entry is obvious and clearly visible in the image and/or OCR output.





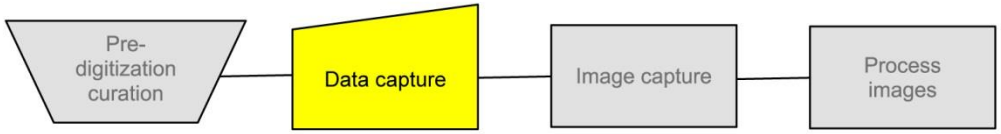
FN2D2I—New Specimen Workflow: Field notes to data to image

This workflow is designed for actively growing collections in which new specimens are regularly added. Collectors, especially in herbaria, typically keystroke label data from field notes, store the label with the specimen, and queue the specimen for mounting. Following mounting, the specimen is treated as an existing specimen with the data entered into the database by a technician, who re-keys the data previously keyed by the collector. The workflow proposed here eliminates the second keying of label data by capturing label data into the database as the label is prepared, allowing the label to be printed from the database immediately following data entry. The workflow assumes a database management system with functionality for printing labels, as well as a strategy that includes the application of bar codes to the newly printed label rather than to the specimen sheet.



A sample, detailed task list.

1. Open Capture NX2 and View NX2.
2. Open Camera Control Pro 2.
3. Open default.ncc as settings file:
   Settings->Load Control Settings
   ->My Documents->CameraSettings->default.ncc.
4. Create a folder in X:\SpecimenImages\NEF, using the current date as the folder name, as 2013-04-14.
5. Retrieve next specimens to image from cabinet.
6. Insert Image “From Here” tag to proper place in cabinet.
7. Set image number in Camera Control 2 to next bar code:
   tools->download options
   Edit
   Start numbering at: <Enter next bar code number; no leading zeros>.
8. In Download Options, set the default folder to the one you created in step 4.
9. Position specimen in frame, ensuring complete specimen is visible.
10. Open Live View, position the focus square on specimen.
11. Click AF to test.
12. Click AF and Shoot.
13. Once the first image loads, navigate to it in Capture NX2 or View NX2.
14. Open the image, zoom in and check margins to ensure all of the specimen is visible.
15. Repeat 8-11 until satisfied, resetting image number each time.
16. Close Live View.
17. Load next specimen in frame.
18. Use remote release on camera and record the images.
19. As you shoot, check each image bar code to ensure it is in sequence with the one preceding it and matches the next one in the series.
20. For out-of-sequence bar codes, change the number in the download options.
21. Repeat 17-20 until all specimens are imaged.



Guiding Principles

Follow a modular approach

- "Plug and play" modules are preferred.
- Simple modules involving a limited number of tasks are easier to troubleshoot and maintain.
- Divide large modules into sub-modules.
- Modules are generally self-contained but tangential.
- There is no consensus workflow, virtually all workflows are customized.

Assign roles deliberately

- Adjust to strengths of each technician--using students and volunteers requires flexibility in role assigned to personnel rather personnel assigned to role.

Create task lists

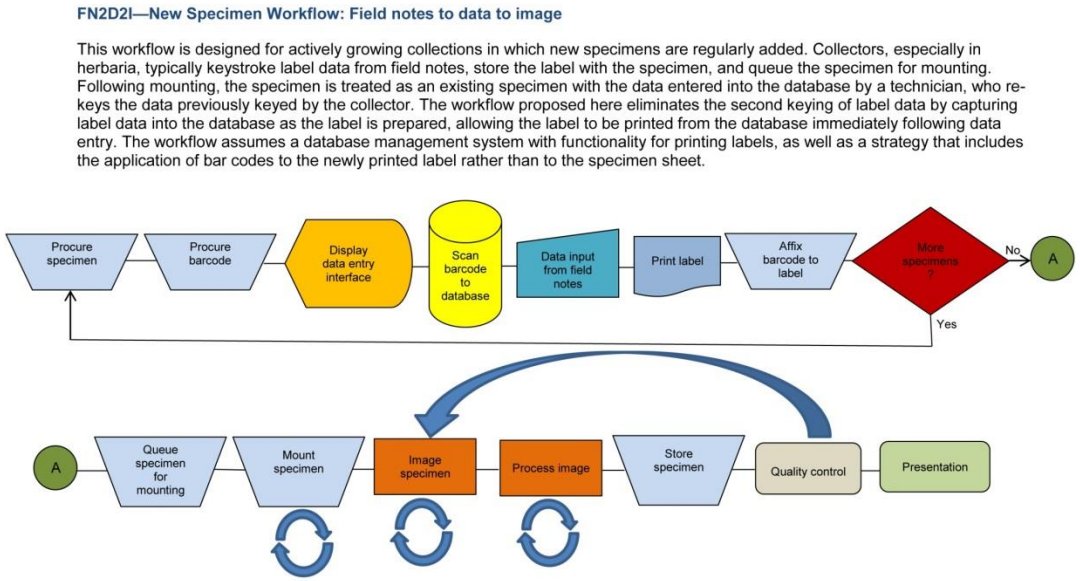
- Complete.
- Clear.
- Succinct.
- Ordered.
- Reusable.

Guiding Principles

- **Segmenting clusters and subroutines**
  - Standalone repetitive processes.
    - Barcoding.
    - Imaging.
    - Image processing.
    - Re-shelving.
  - Conservation and repair.
  - Georeferencing.
  - OCR.

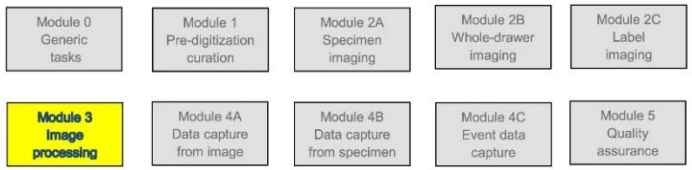
Workflows Working Groups

- The Flat Sheets and Packets Working Group has completed modules and associated tasks for herbarium and related collections (October 2012).
- The Pinned Things in Trays and Drawers has finished and posted its work for entomology (January 2013).
- 3D Objects in Spirits in Jars and Vials is nearing completion of its workflows for fluid-preserved specimens (April 2013).
- 3D Objects in Drawers and Trays workflows group to start work in April (June 2013).
- Preparation-independent workflows to follow (2013).



Modular Approach

Workflow Detail: Specimen Image Processing (Pinned Things)



Module 3: Specimen Image Processing

Task ID	Task Name	Explanations and Comments	Resources
T1	Transfer images from camera to immediate image processing storage.	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.	Ample storage space with backup procedures (also see T8-T9).
T2	Adjust orientation and crop images, as necessary.	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).

University of Florida • Florida Museum of Natural History • Dickinson Hall (Museum Rd. & Newell Dr.) • Gainesville, FL 32611 • 352-273-1906
iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (#EF1115210)

Workflow Modules and Task Lists

One outgrowth of the DROID (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, Pinned Specimens in Trays and Drawers, invested its time developing modules to support effective entomological digitization workflows. Other preservation 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 orderly, 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.

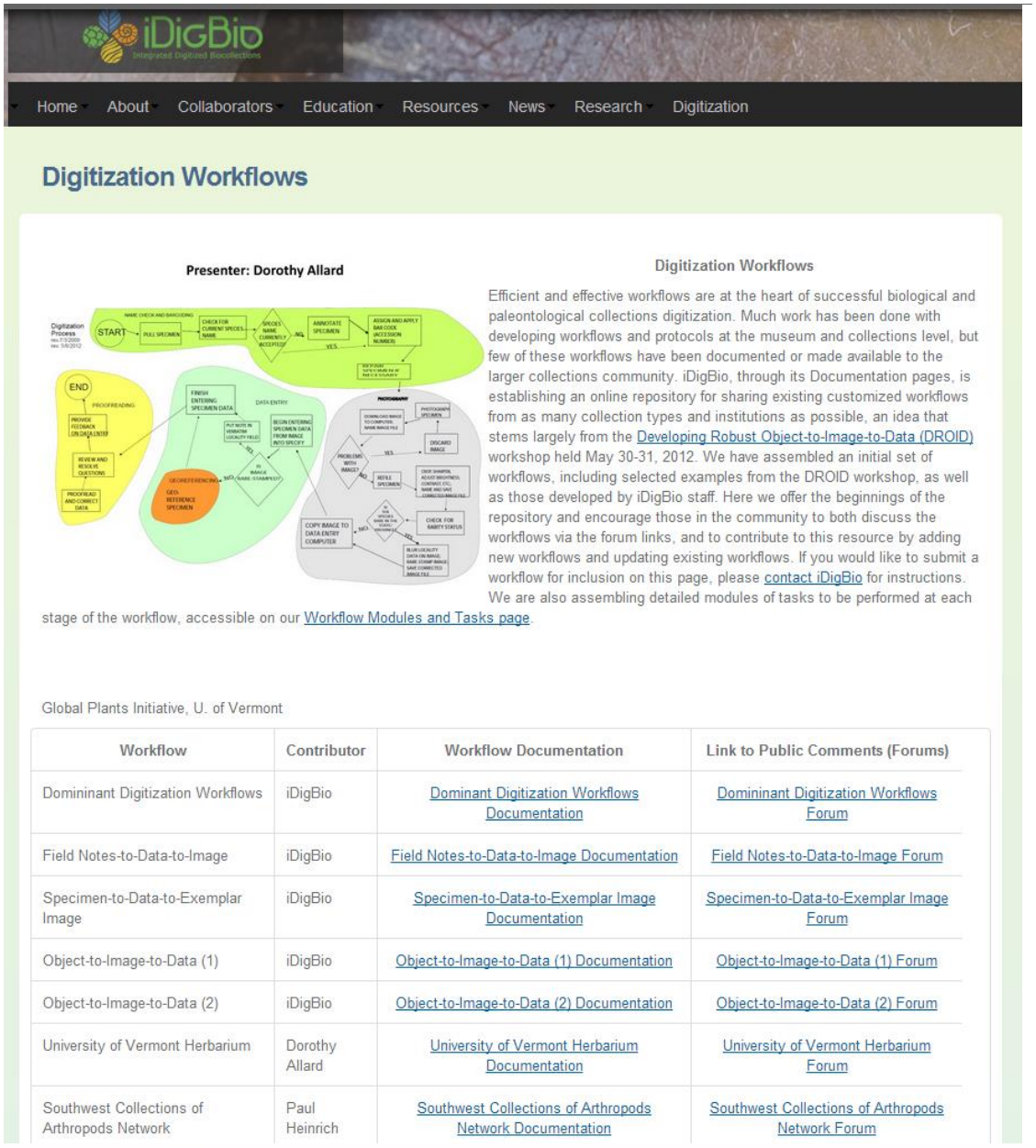
Links to published modules as they are completed are provided below:

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

Pinned Things in Trays and Drawers Working Group - Dried Insects

- Module 0 Generic Tasks Applicable to Two or More Modules
- Module 1 Pre-digitization Curation Tasks
- Module 2A Specimen Imaging Tasks
- Module 2B Whole-drawer Imaging Tasks
- Module 2C Label Imaging Tasks
- Module 3 Image Processing Tasks
- Module 4A Data Capture From Image Tasks
- Module 4B Data Capture From Specimen Tasks
- Module 4C Event Data Capture Tasks
- Module 5 Quality Assurance Tasks

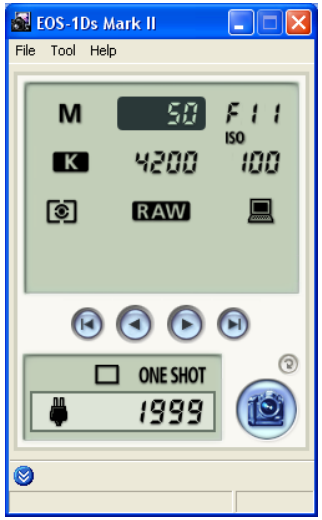


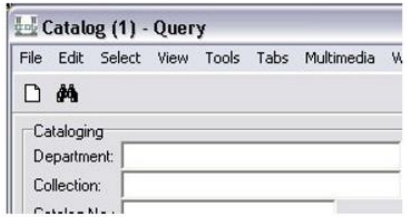
Posted To Collaborative Workflows Page Linked to the Digitization Resources Wiki

Documentation and Instructions

- **Written Protocols**
  - Essential!
  - Include illustrations/screen shots.
  - 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.





Continuous Workflow Improvement

Develop written workflows that reflect actual practice.

Continuous evaluation of written and actual workflows by:

- Technicians
- Workflow managers
- Collections managers

With particular attention to:

- Bottlenecks
- Redundancy
- Handling time
- Varying rates of productivity

Imaging Decisions

- Image purpose
  - Taxonomic judgments
  - Identification/annotation
  - Morphological examination
  - Documentation-of-occurrence voucher
  - DNA voucher
  - Exemplar of species, genus, family, order, other
- Prioritization of specimens to capture
- Related literature, vocalizations, live images, video, etc.
- Equipment/software considerations

Thank you!