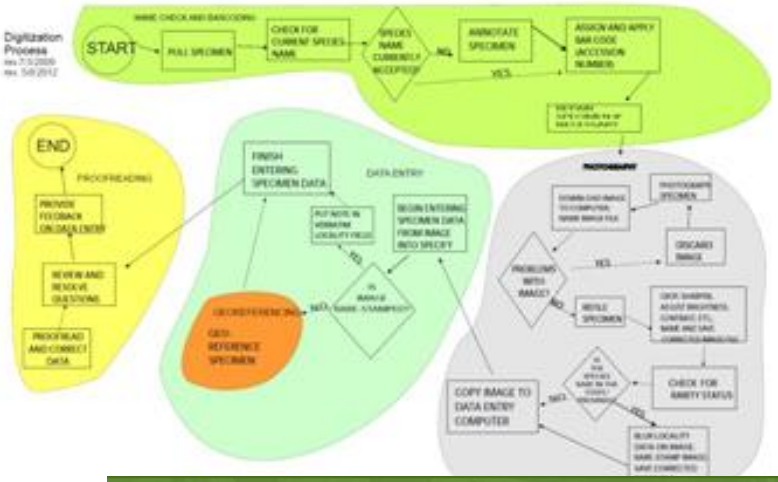
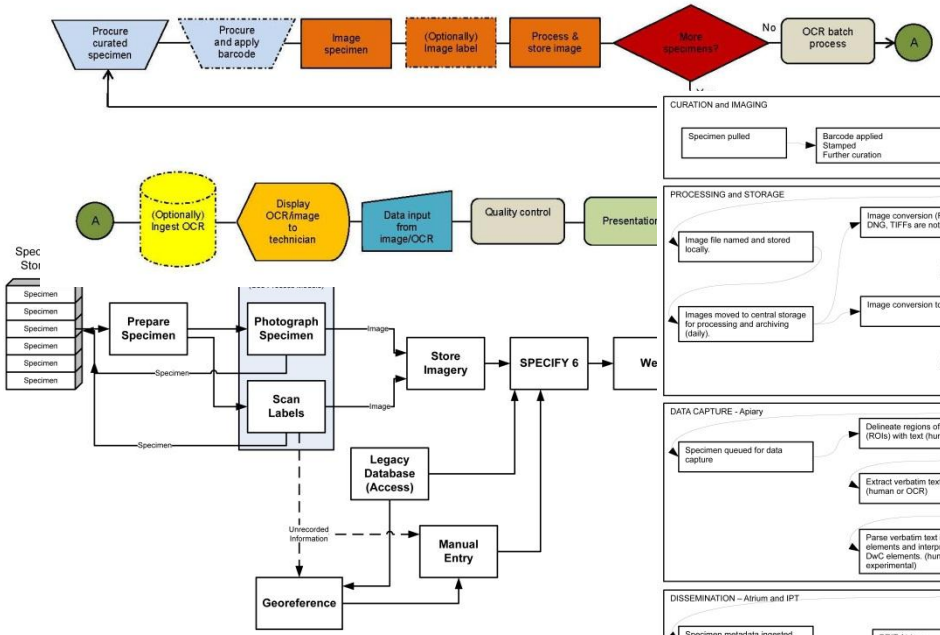
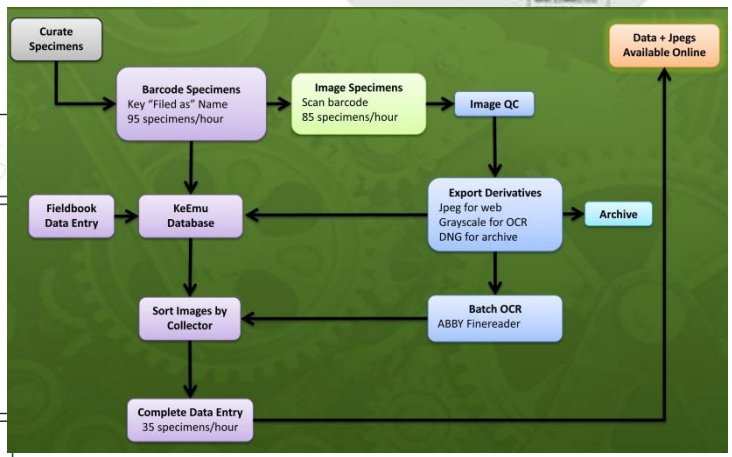
Herbarium Digitization Workshop

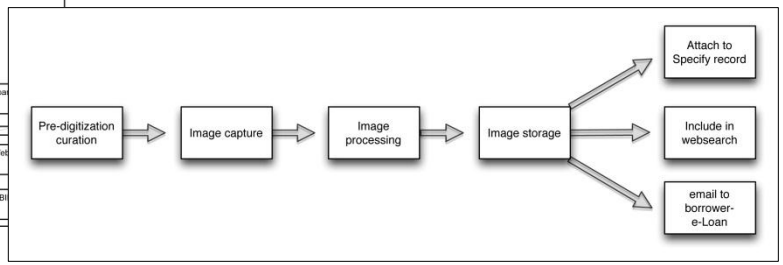
iDigBio
Integrated Digitized Biocollections

Digitization Workflows

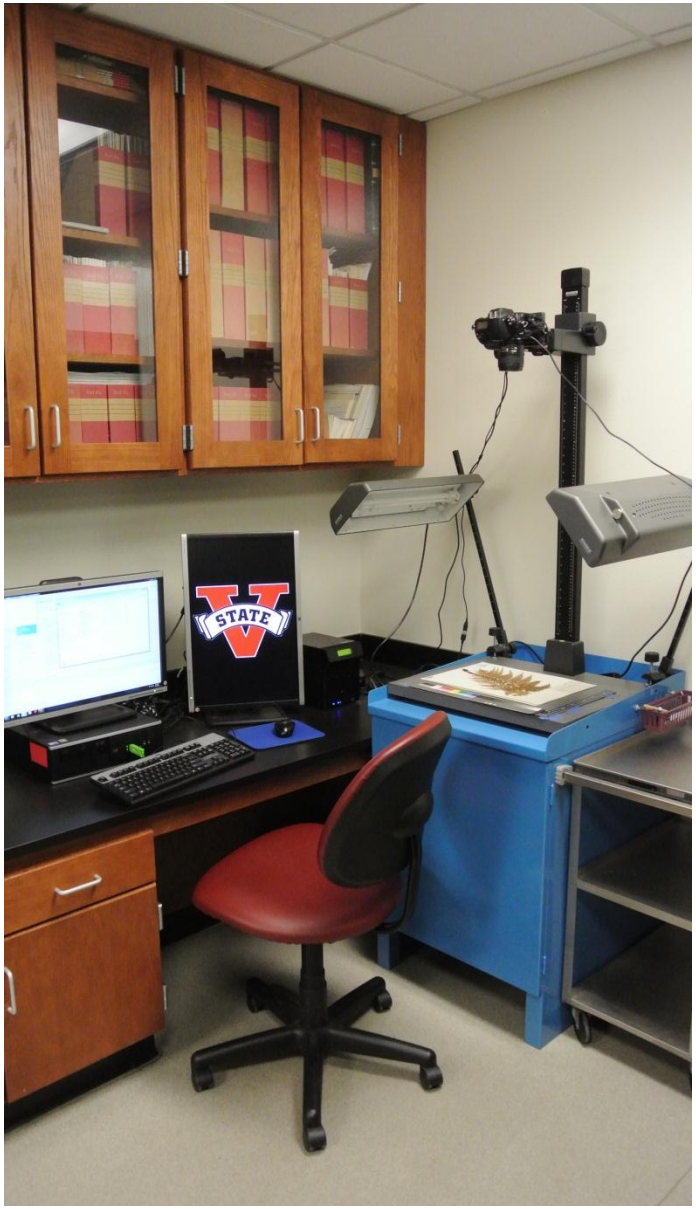








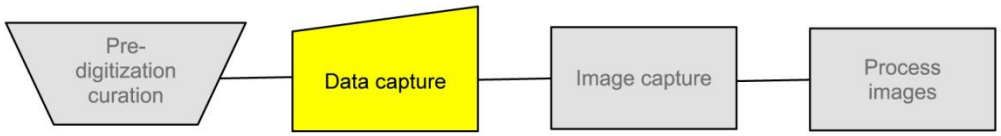
Gil Nelson
September 16-18, 2012
Valdosta State University



Dimensions of Workflow Design

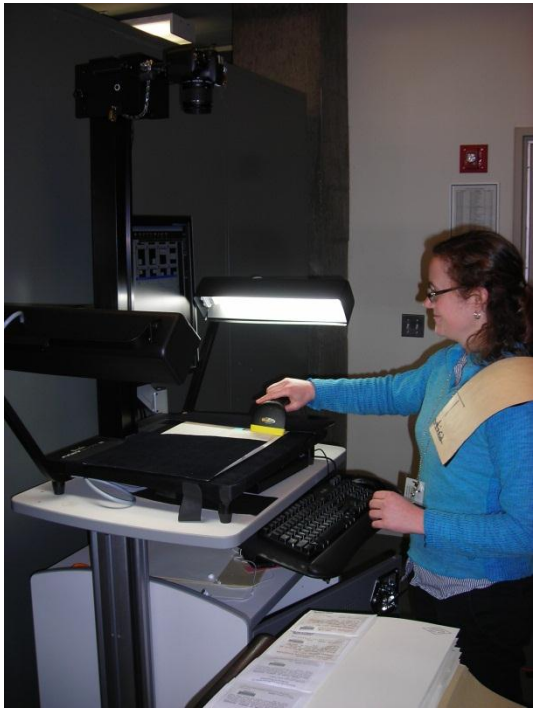
- **Efficiency**
  - Contrasted with speed
  - Reduce technician fatigue
  - Maintain technician focus
  - Optimize task execution times (time/motion)
    - Record statistics
    - Make adjustments

- **Conservation of movement**
  - Positioning and compactness of work station components
  - Left to right vs. right to left
  - Starting and ending locations
  - Proximity of equipment (including mouse)



Elements of Workflow Design

- **Follow a modular approach**
  - Plug and play modules
  - Modules are often self-contained and independent
  - There is no consensus workflow, virtually all workflows are customized
  - Adjust to strengths of each technician (using students requires flexibility in duties assigned to personnel rather personnel assigned to position)
  - Task Lists for each module
    - Plug and play tasks
    - Clear
    - Succinct
    - Ordered



Dimensions of Workflow Design

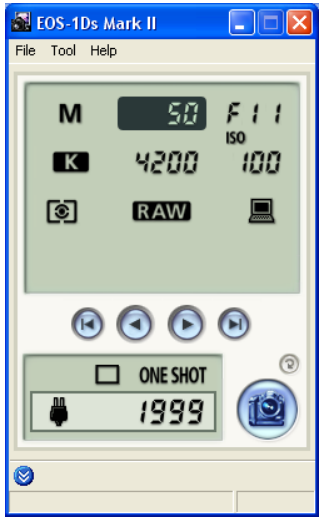
- **Multi-tasking**
  - Making the most of down time (regardless how long)
  - Nesting shorter tasks (start one task, start another)
    - Overcoming distractedness
- **Workflow simulation and modeling**
  - Analyzing temporal juxtaposition of workflow task clusters
  - Analyzing spatial juxtaposition of workflow task clusters
- **Task list simulation and modeling**

Dimensions of Workflow Design

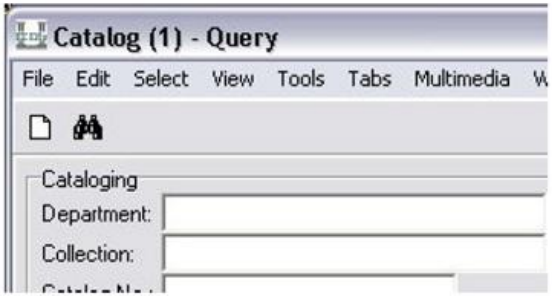
- **Segmenting clusters and subroutines**
  - Standalone repetitive processes
    - Barcoding
    - Imaging
    - Image processing
    - Re-filing
  - Conservation and repair
  - Georeferencing
  - OCR

Documentation and Instructions

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

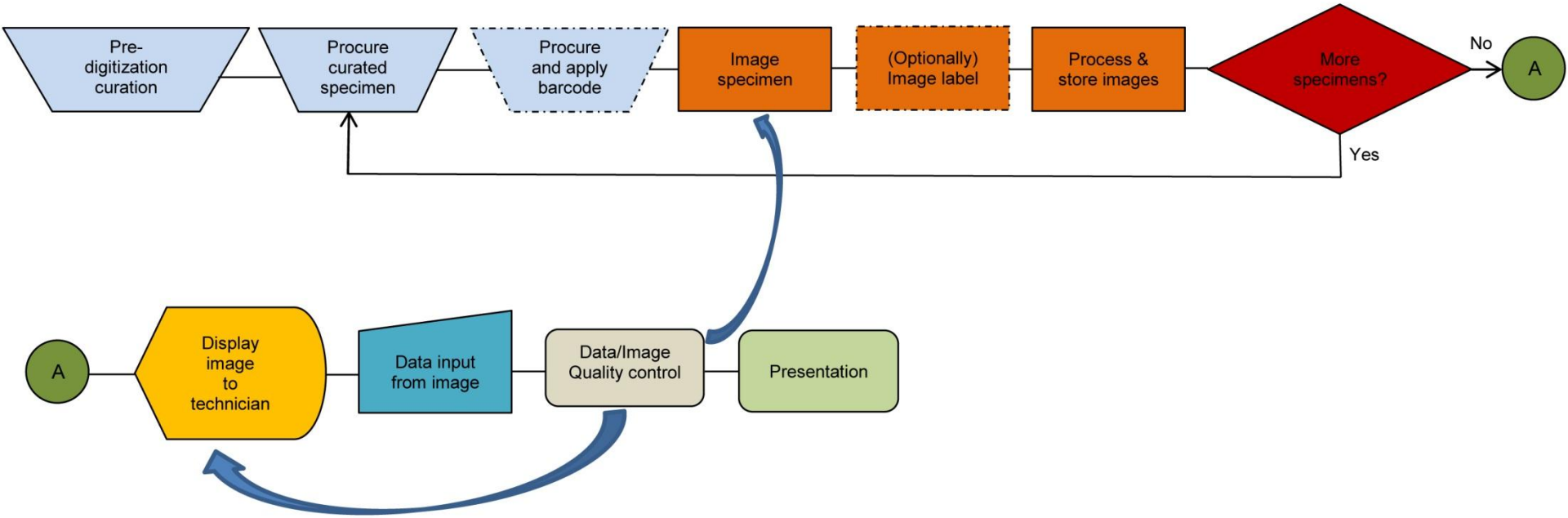


- **Feedback Loops**
  - Technicians: best source of efficiency adaptations, either by show or tell
  - Easy methods for receiving feedback
  - Personal copies of the protocol
  - Master copy available via Google docs for updates and suggestions



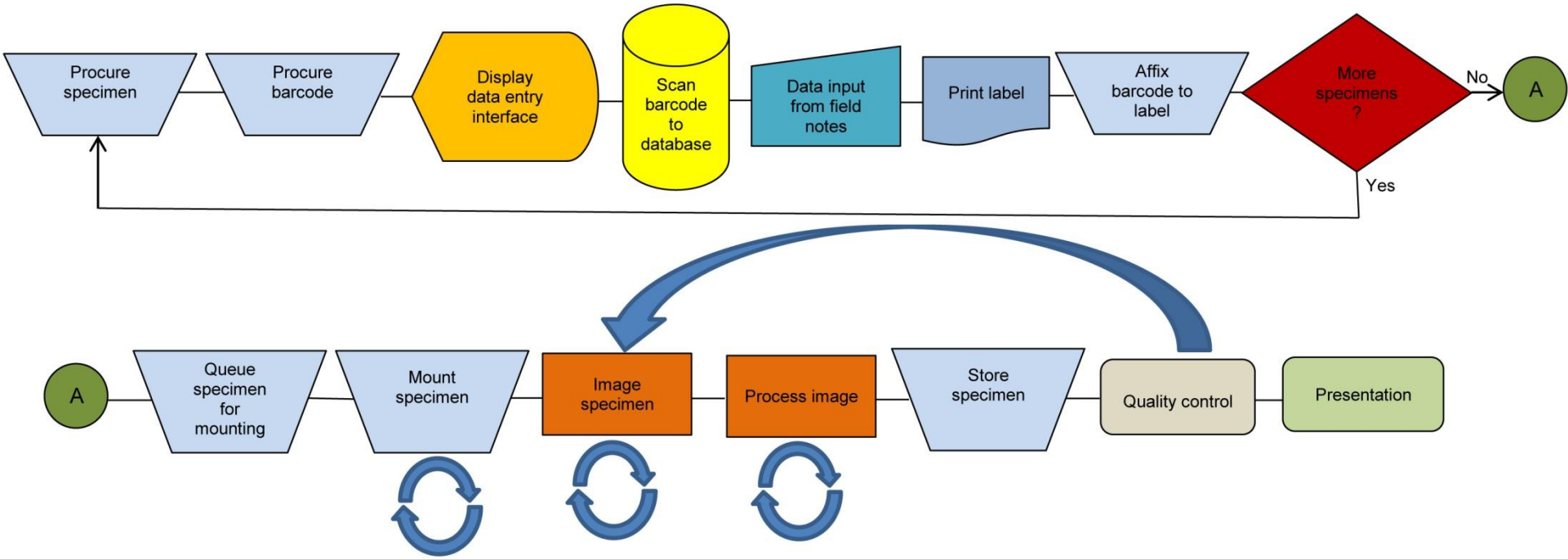
O2I2D(2)—Existing Specimen Workflow: Object to Image to Data

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.



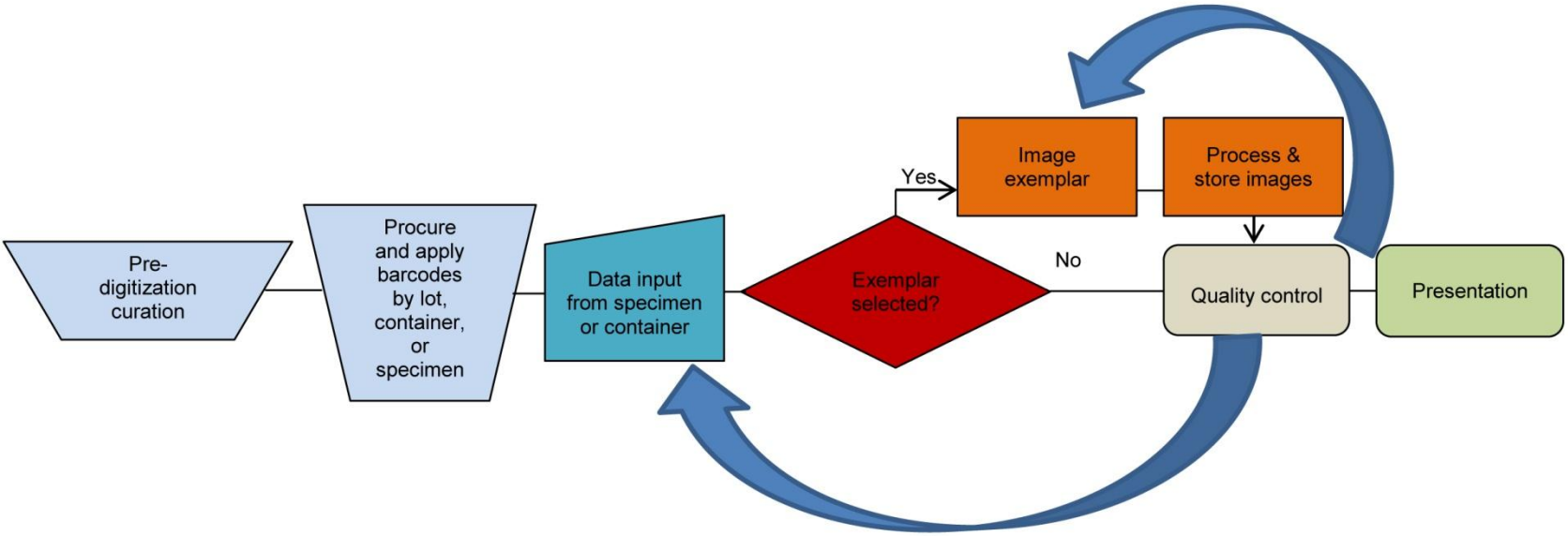
FN2D2I—New Specimen Workflow: Field notes to data to image

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.



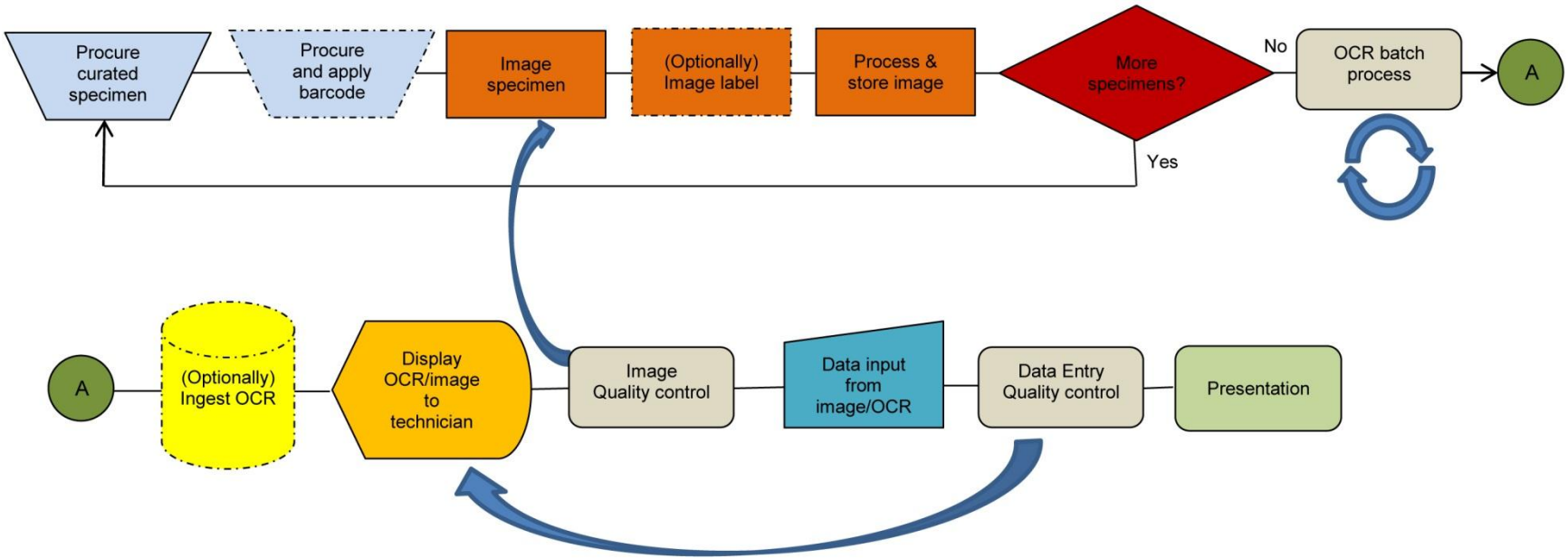
O2D2EI—Existing Specimen Workflow: Object to Data to Exemplar Images

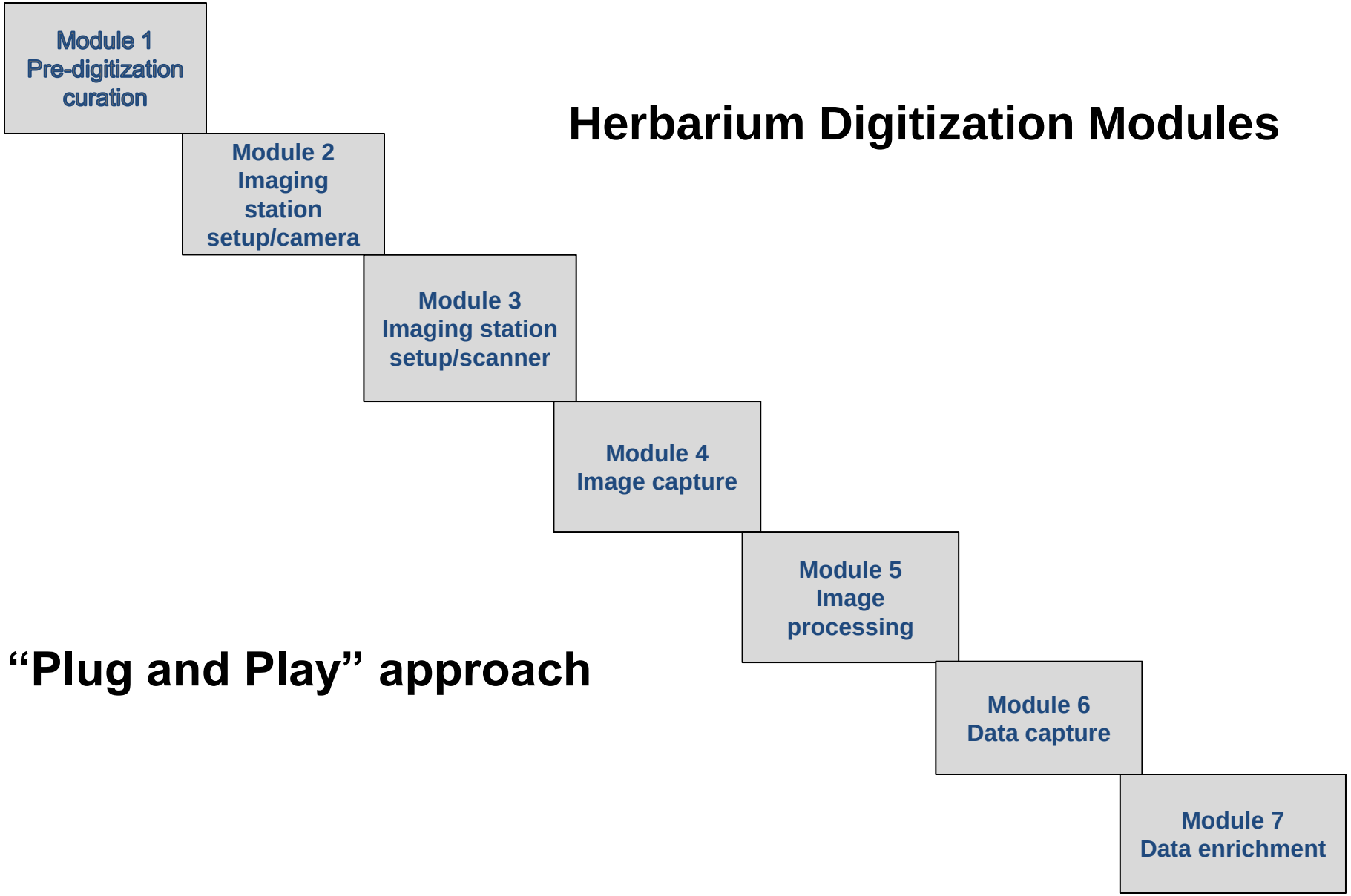
This workflow is in use for collections that capture data in specimen lots, collecting events, taxon container, or other aggregates, but capture images only for exemplar specimens. Data capture is effected from specimen labels. Depending upon preparation type, barcodes are usually applied inline—often 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.

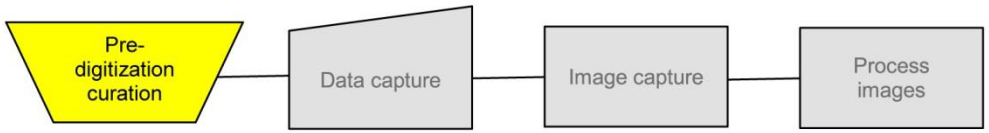




Herbarium Digitization Modules

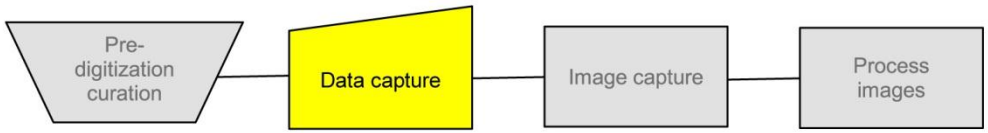
“Plug and Play” approach

Workflow Detail: Pre-digitization Curation



- select specimens to digitize (project based, family based, loan requests, all specimens),
- locate specimens,
- pull specimens,
- apply barcode to specimens (1D or 2D),
- relocate specimens to digitization area,
- document or flag location for return of removed specimens,
- annotate for preferred nomenclature and taxonomic interpretation,
- validate label data,
- attach a unique identifier (most often a 1- or 2-D barcode) to a specimen, container, tray, or drawer
- inspect for and re-route for repair specimen damage (may require a separate workflow),
- treat specimens for pests (freeze or otherwise treat),
- organize or reorganize the contents of cabinets and folders to facilitate digitization
- vet type specimens,
- select exemplars for digitization, when that approach is appropriate,
- potentially create a skeleton data record.

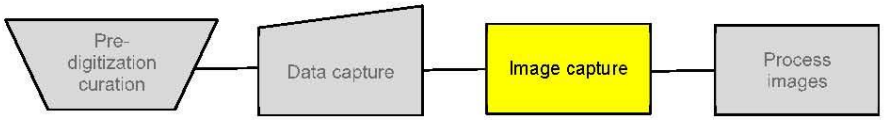
Workflow Detail: Data Capture



Data capture entails:

- reviewing the written data entry protocol,
- opening database software (e.g. KE EMu, Specify, Arctos, Apiary, Symbiota, custom interface, etc.),
- opening associated image files (if using an O2I2D protocol),
- entering barcode or unique specimen identifier
- selectingf taxon from look-up or dropdown dialog,
- selectingf taxon from look-up or dropdown dialog,
- selecting taxon name from pick list
- entering collector name, collector number, collection date
- attempting electronic search for duplicates (Scatter Gather Reconcile, Symbiota)
- attemping or using existing Optical Character Recognition (if using OCR),
- attempting Natural Language Processing (NLP)
- entering other label data,
- quality control by inspection or electronically.

Workflow Detail: Image Capture



The image capture process typically involves five groups of related tasks, including:

- original installation and setup of imaging station components (usually a one time activity)
- immediate pre-imaging equipment configuration and initialization,
- procuring/organizing the next batch of specimens for imaging,
- acquiring the image,
- checking quality.

Imaging station components vary by institution, organism being imaged, and intended use of the resulting images. Most common is a single-lens reflex digital camera fitted with a standard or macro lens and connected to manufacturer or third-party camera control software. A typical station includes components from the following list:

- camera and lens, camera and microscope, or flatbed scanner (e.g. HerbScan)
- panoramic robot (e.g. GigaPan or SatScan),
- cable connecting camera to computer for auto-transfer of images,
- camera control software (third party or camera manufacturer produced),
- image processing software (most common are Canon Digital Photo Professional, Nikon Capture NX2, Photoshop, and Lightroom),
- image stacking equipment and software, (e.g. Helicon Focus, Automontage),
- remote shutter release (wireless or tethered),
- copy stand and/or specimen holder,
- studio lighting, flash units, or light/diffuser box (e.g., Photo Box MK),
- scale bar,
- color standard,
- stamp to mark sheets, jars, trays, or folders that have been imaged, and
- associated instruments and tools (pinning blocks, forceps, latex gloves, etc.).

Immediate pre-imaging equipment configuration entails:

- reviewing image acquisition protocol,
- connecting or ensuring the connection of computer to camera,
- starting external studio lighting or lightbox, or checking, adjusting, and testing flash units and power supplies,
- starting camera control and image acquisition software,
- starting the camera,
- setting camera aperture, shutter speed, and focus point (or loading these attributes from a previously configured settings file),
- adjusting camera height,
- changing or attaching lenses,
- loading ancillary image management/processing software.

Acquiring the image entails:

- Placing specimen on copy stand or holder,
- Adjusting focus when manual focusing is required,
- Releasing the shutter via remote shutter release or mouse, as required,
- Readjusting focus and reshooting a single specimen multiple times when stacking software is in use,
- Not touching the camera, copy stand, or holder during image acquisition to minimize camera movement,
- Removing specimen from copy stand or holder and replacing.

Checking quality, every few images to ensure that:

- lighting, exposure, and focus remains constant,
- file naming is progressing according to plan,
- exposure remains correct,
- focus remains sharp,
- images lack imperfections such as blemishes or streaking,
- files are not corrupted,
- barcodes or identifiers are in place and readable.

Workflow Detail: Image Processing

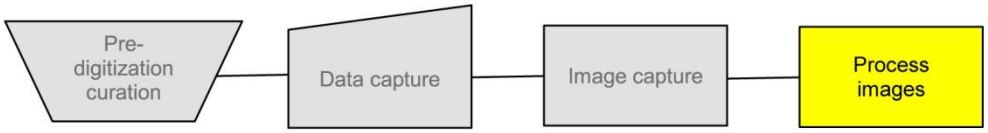


Image processing involves all tasks performed on an image or group of images following image capture. At least 11 tasks are typically addressed in this task cluster, not all of which are universally applied:

- preserve image metadata (EXIF, IPTC)
- quality control (second or third level quality control following that performed during image capture),
- barcode value extraction (usually via OCR software where barcodes are not scanned or recorded at image capture),
- file conversion (to TIFF, JPEG, PNG, etc.)
- image cropping,
- color balance, white balance, light level adjustments,
- image stacking (using Helicon Focus, AutoMontage, or others)
- redaction of sensitive data,
- file transfer to archive and create derivative for internet accessible storage,
- optical character recognition (OCR).

Thank you!