iDigBio
Integrated Digitized Biocollections

Digitization Tasks and Components
An Overview for Herbaria

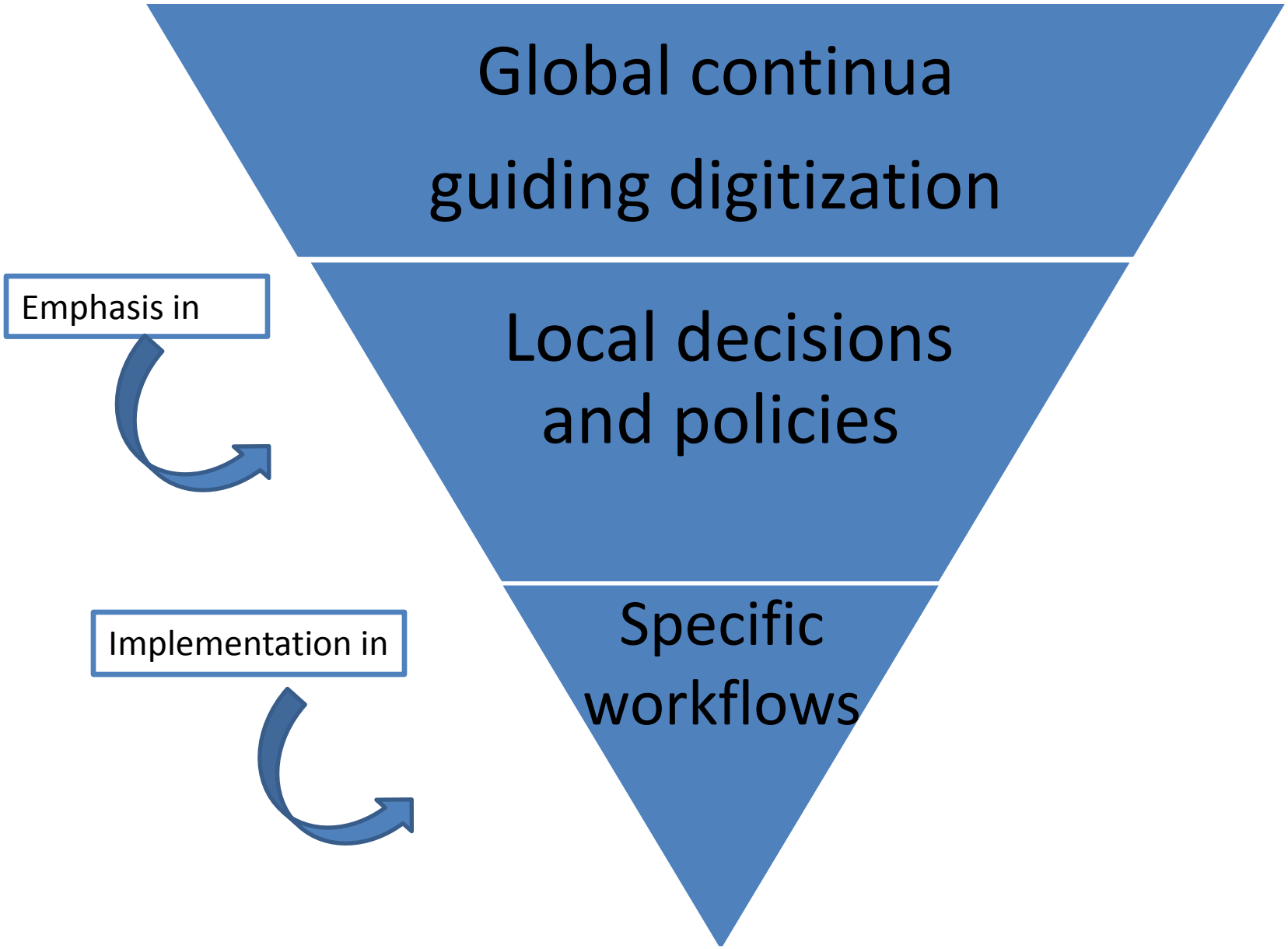
Gil Nelson
September 16-18, 2012
Valdosta State University

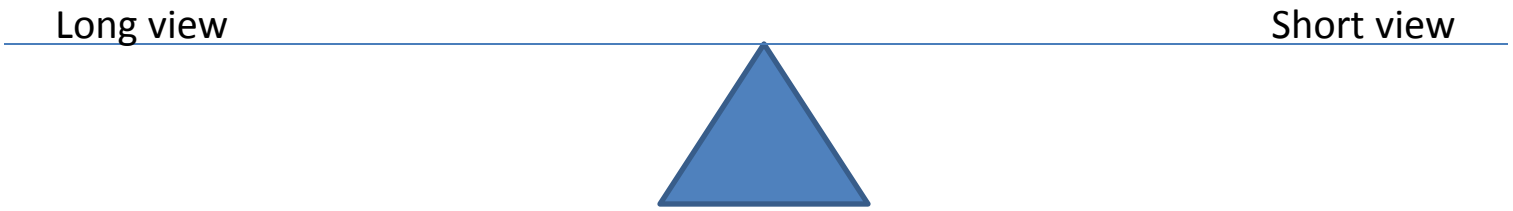
Ultimate Goals of Effective Workflows

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

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

Improvement level: Collaboration and workflow sharing across the herbarium 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 balancing long term goals with short term constraints, including a commitment to implementing future enhancements.

Pressures mitigating the long view

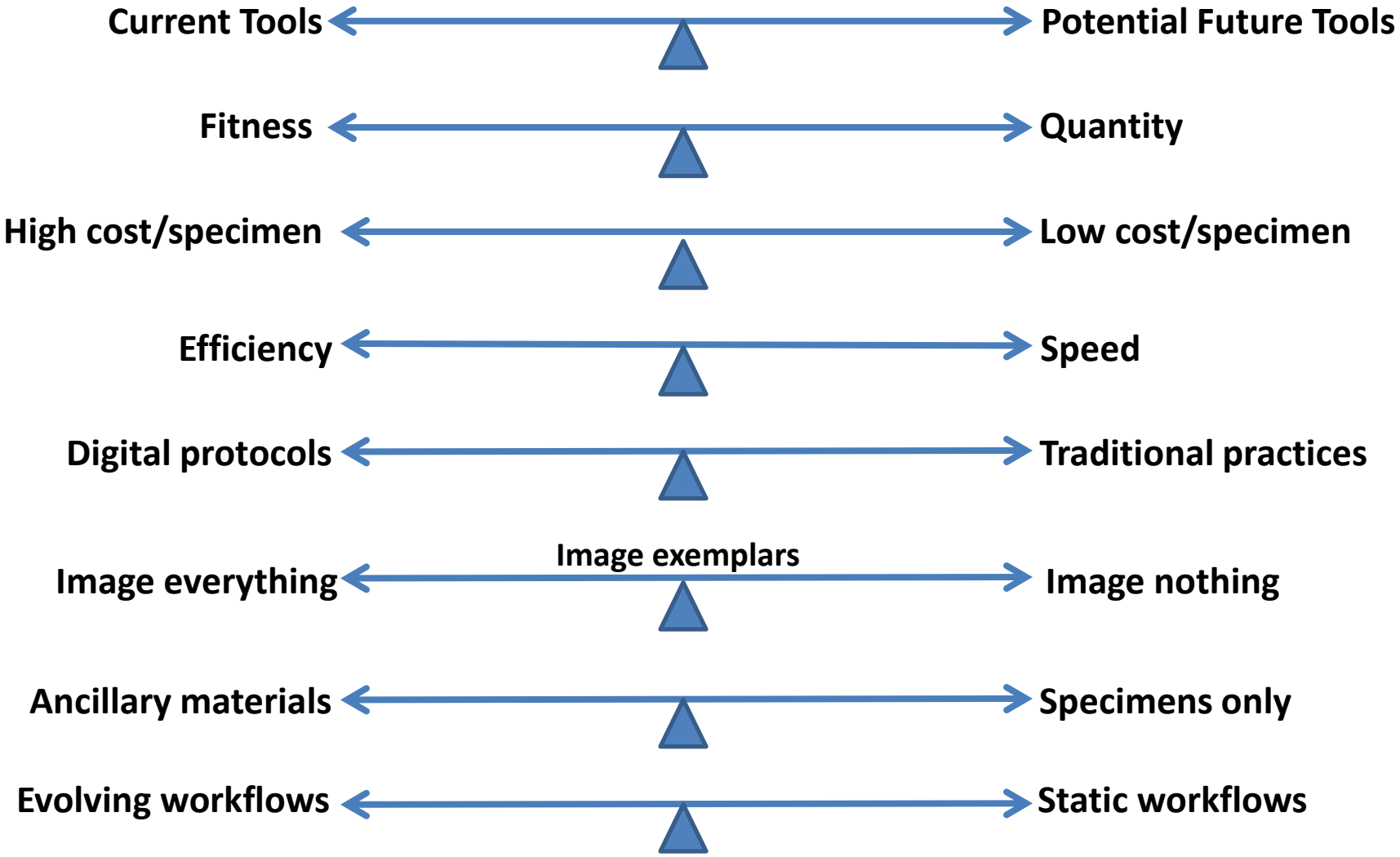
So much data, so little time.
Our collections are not getting smaller.
The funding agencies have high output expectations.
We only have 3 years to get this done.
All of our data and all of our specimens are important.
Let's just use the images!
We'll do the minimum now and enhance it later.
(while avoiding the Scarlet O'Hara syndrome).

Tracks to Digitization

• **Taking the inside track** 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** 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.

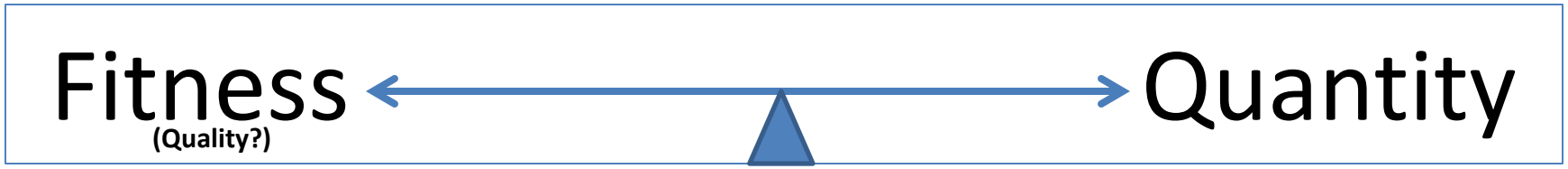
Scan and Deliver: Managing User-initiated Digitization in Special Collections and Archives, 2011
J. Schaffner, F. Snyder. S. Supple

Global Digitization Continua



Future Tools Favoring the Inside/Middle Tracks

- OCR, NLP, and ICR (handwriting analysis) improvements
- Automated image analysis/computer vision for data extraction
- Data mining labels
- Robotic technologies, conveyor belts, etc. (Paris, Northeast Herbaria TCN)
- Improvements in discovery/capture/use of duplicates (SGR, Symbiota)
- Improvements in voice recognition and other data entry technologies
- Post-digitization tools for curation and quality control
- Field-based data capture



Facilitators

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

Facilitators

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

Efficiency vs. Speed

False dichotomy?

Is increased speed the inevitable outcome of improved efficiency?

Is increased speed always and necessarily the desired outcome of improved efficiency?

Efficiency vs. Speed

Improving Efficiency

Reduce or eliminate redundancy (e.g., label data entry)
Reduce or eliminate unnecessary steps in a workflow
Maintain an evidently logical, easy-to-follow workflow
Mitigate monotony for technicians
Reduce or eliminate travel time
Reduce technician fatigue
Ensure sustained output
Increase output over the long term

Efficiency vs. Speed

A rested, happy, satisfied tortoise is usually better than a harried hare!

Productivity vs. Cost

Issues to Resolve in Assessing Productivity

Comparability: Just 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 cost:

- Is this a competitive event?
- Output per hour at given fitness?
- $$ per specimen at given fitness?
- Accounting for variances in prep type, regional pay rates, data robustness, etc.?

Continuous Workflow Improvement

Develop written workflows

Continuous evaluation of written and production workflows by:

- Technicians
- Workflow managers
- Collections mangers

With particular attention to:

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

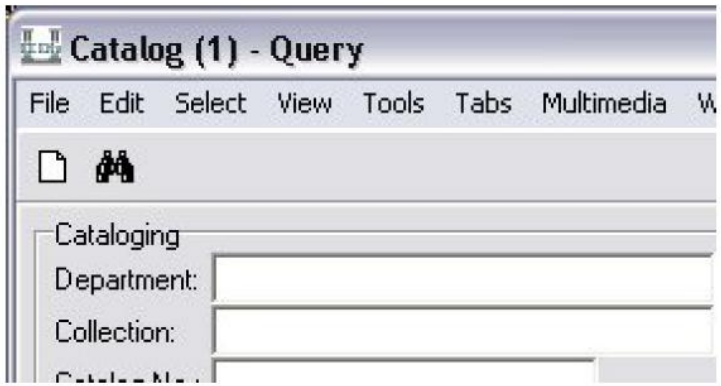
Written Protocols

How to catalog specimens:

Open KE EMu from the link on the desktop using the login you are provided with. Refer to the attached screenshot as you enter data (until you become a pro).

A menu, at right, will pop up with various options (technically "modules") that you can choose from. All volunteers will only need to work in the "Catalog" module in the future, your account will show "Catalog" as your only option.

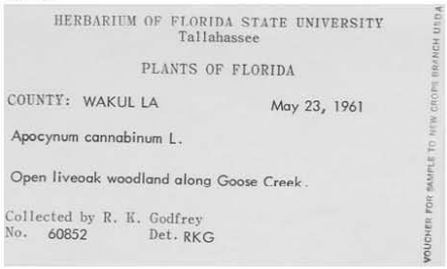
When the catalog opens, chose either the paper icon (next to the binoculars) or click on File->New.



This will give you the data entry form for YPM specimens (it should look similar to the attached form).

Written Protocols

Ex.4:


HERBARIUM OF FLORIDA STATE UNIVERSITY
Tallahassee

PLANTS OF FLORIDA

COUNTY: WAKULLA May 23, 1961

Apocynum cannabinum L.

Open liveoak woodland along Goose Creek.

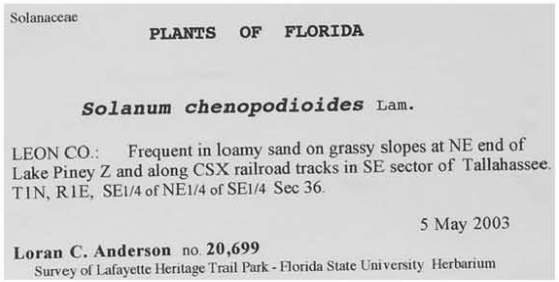
Collected by R. K. Godfrey
No. 60852 Det. RKG

VOUCHER FOR SAMPLE TO NEW CROPS BRANCH USDA

Voucher Flag: Unknown.
Notes: Voucher for sample to New Crops Branch USDA.

When a specimen has been examined for a flora, you do not enter anything in **Notes**. The information relating to the use of the specimen for a flora is dealt with in **Verification Info**. See **Verification Info** directions for further information.

Ex. 5: Bar Code 5765


Solanaceae

PLANTS OF FLORIDA

Solanum chenopodioides Lam.

LEON CO.: Frequent in loamy sand on grassy slopes at NE end of Lake Piney Z and along CSX railroad tracks in SE sector of Tallahassee. T1N, R1E, SE1/4 of NE1/4 of SE1/4 Sec 36.

5 May 2003

Loran C. Anderson no. **20,699**
Survey of Lafayette Heritage Trail Park - Florida State University Herbarium

Voucher Flag: Flora
Notes:
Verification Level Flag: 3
Further Identification Comments: Survey of Lafayette Heritage Trail Park.

Unique Species ID:
Must be selected from dropdown list. If the species does not exist in the dropdown list, check to see whether an accepted synonym is in the list. Go to The University of South Florida Atlas of Florida Vascular Plants website for this (http://www.plantatlas.usf.edu/). If you are certain the species is nowhere in the list, ask the supervisor to confirm this and to add the species in question to the list.

CAUTION! When entering a new taxon for the first time, be aware the system remembers the last taxon you entered. If you do not choose the new taxon from the **Unique Species ID** drop-down list, the specimen you are data-entering will be in the system with the incorrect (prior) taxon name.

10

Written Protocols

The following pages will now take you through the data entry windows.

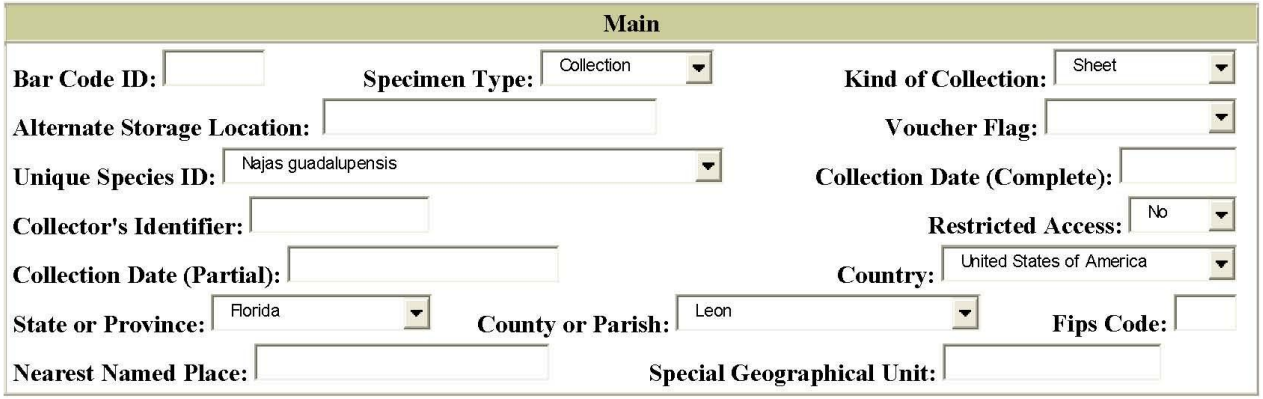
SIDEBAR:

These sidebar links will be present in every window you enter. You will need to use the Add Specimen, List Specimens, Search Specimens, Search People, Add Person and List People links at various points during your data entry.

Add Specimen: This is the entry point into the data entry windows used to enter new records.
List Specimens: If you need to return to a specimen after you have completed entering the data you can click on "List Specimens", locate the barcode of the specimen you are looking for, and click on that.
Search Specimens: Search by barcode or other fields to find a particular specimen or specimens as needed. The Public version of this search window is also useful for checking some data entry issues.
Search People: Use this to verify if a person in is the People List or not. If not, they will need to be added with Add Person.
Add Person: If when you are selecting a collector from the dropdown list, their name does not appear on that list, you must add them into the database. Click on Add Person and enter the **Last Name**, and **First Name** fields. All other fields in this window can be left blank.
List People: This can be selected in order to confirm the presence of a person in the database.

LOGGING ON/ DATA ENTRY:

1. Login with Username and password you have been given.
2. Click on Add Specimen in sidebar. You will now enter the main data entry window.



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

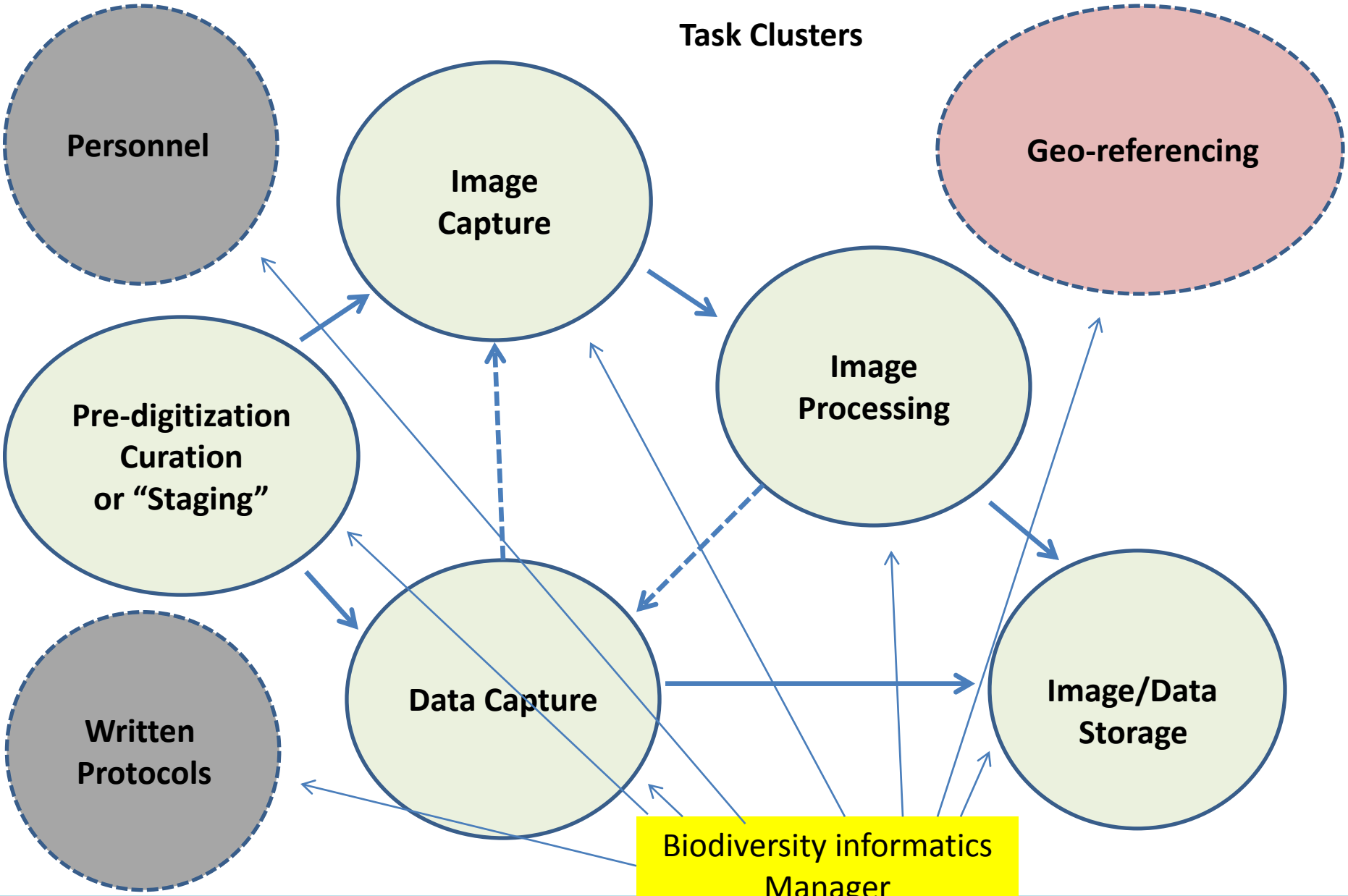
Observing Digitization Practices in Biological and Paleontological Collections

28 Collections
10 Museums
Spanning biological and paleontological collections
Insects and other invertebrates, plants, birds, mammals
Wet, dry

Acknowledgments

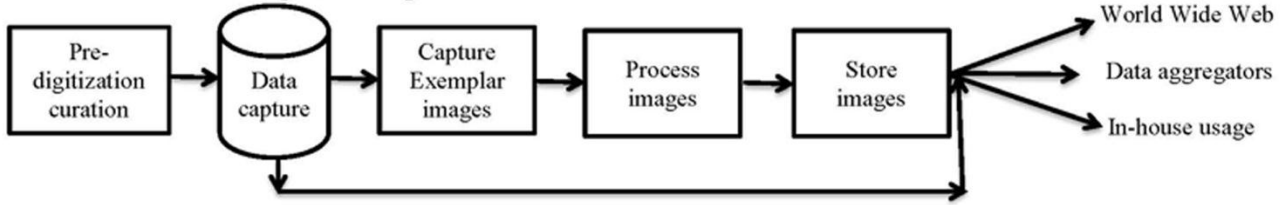
American Museum of Natural History
Botanical Research Institute of Texas
Florida Museum of Natural History
Florida State University
Harvard Herbarium
Museum of Comparative Zoology (Harvard)
New York Botanical Garden
SERNEC
Specify Software Project (University of Kansas)
Symbiota Software Project (Arizona State University)
Tall Timbers Research Station and Land Conservancy
Tulane University Museum of Natural History
University of Kansas Insect Museum
Valdosta State University
Yale Peabody Museum

Herbarium Digitization Workshop

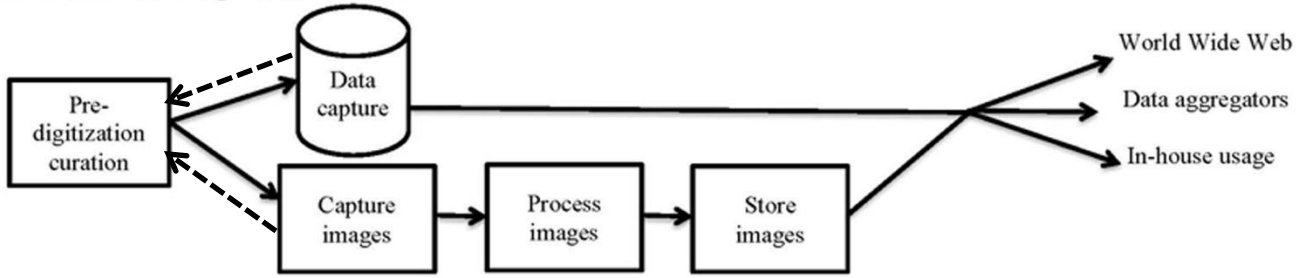


Dominant Digitization Patterns Observed

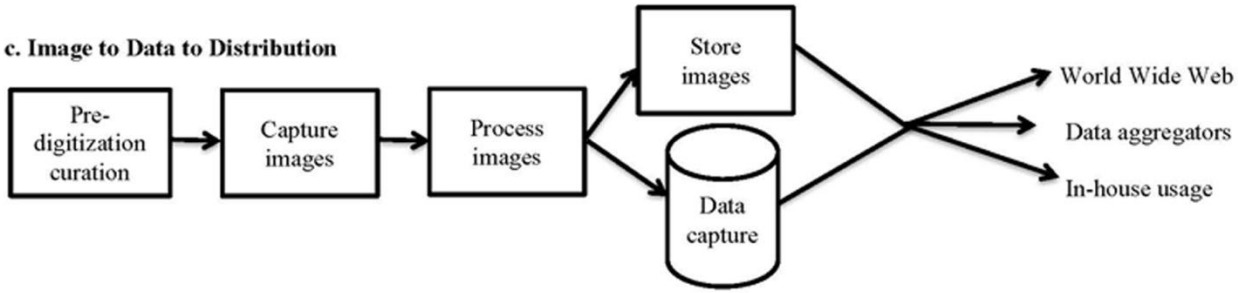
a. Data to Occasional or Optional Image to Distribution



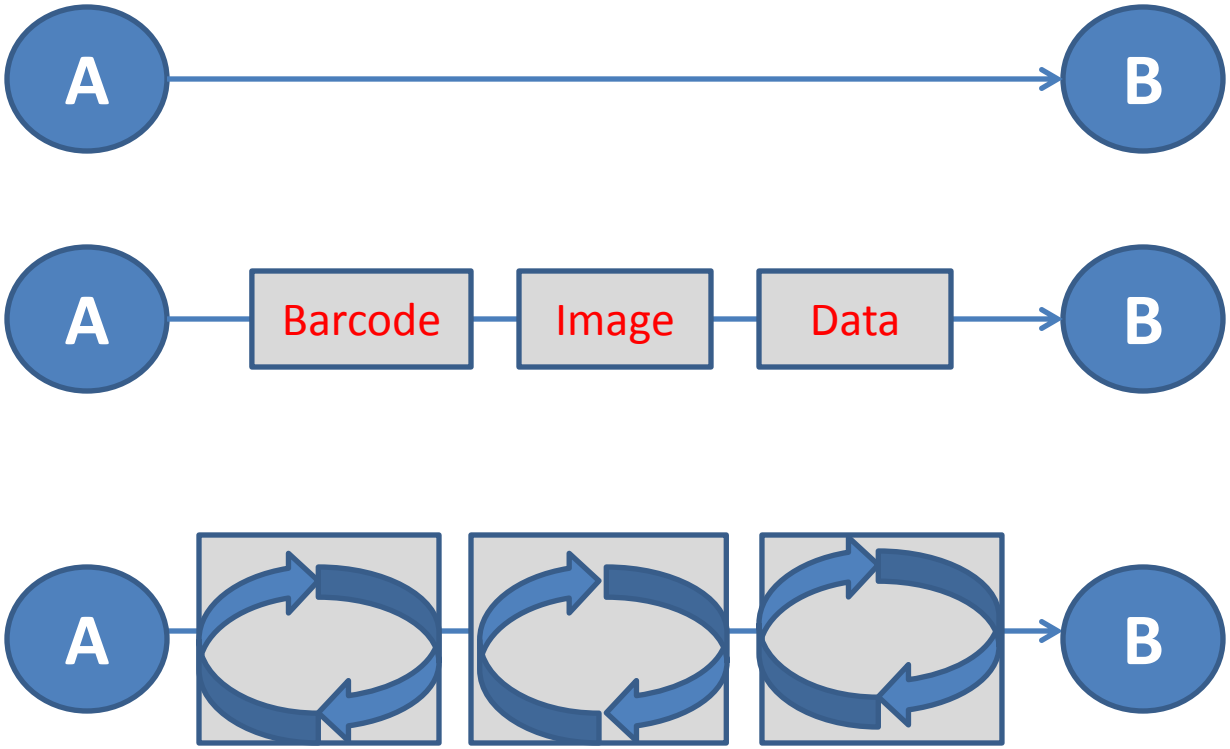
b. Parallel Data/Image to Distribution



c. Image to Data to Distribution



Linear vs. Iterative

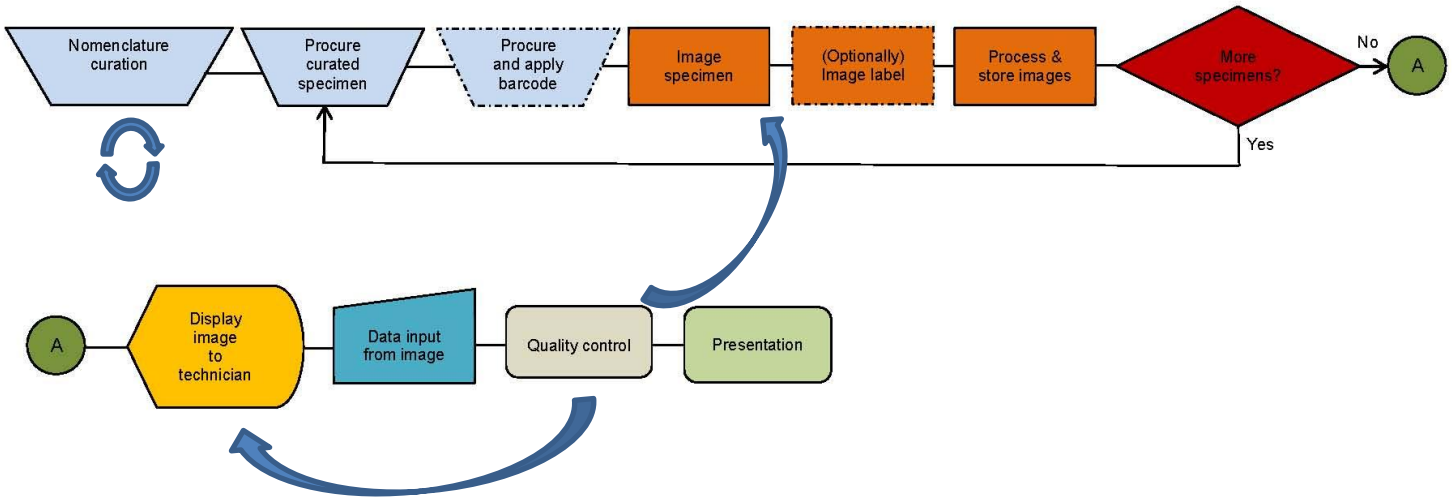


Personnel specialization & availability | Reduces bottlenecks | Technician preference

Object to Image to Data Workflow

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.



Identifiers

Identifiers should be persistent and unique, meaning that they are:*

- **assigned once and only once, and**
- **forever associated with a single object.**

Does not mean that a record must have one and only one identifier.

Two important reasons for creating and maintaining identifiers:

- Provide a handle that can be used for keeping track of all characteristics of an object, including its primary properties, commentary on those properties, and relationships with other object.
- Provide a handle that can be used to provide services for the object, including versioning, those that deliver the data and metadata of the object, and those that link the object to services for other objects.

2.1. Benefits of URIs*

The choice of syntax for global identifiers is somewhat arbitrary; it is their global scope that is important. The ***Uniform Resource Identifier***, [URI], has been successfully deployed since the creation of the Web. There are substantial benefits to participating in the existing network of URIs, including linking, bookmarking, caching, and indexing by search engines, and there are substantial costs to creating a new identification system that has the same properties as URIs.

Good practice: Identify with URIs

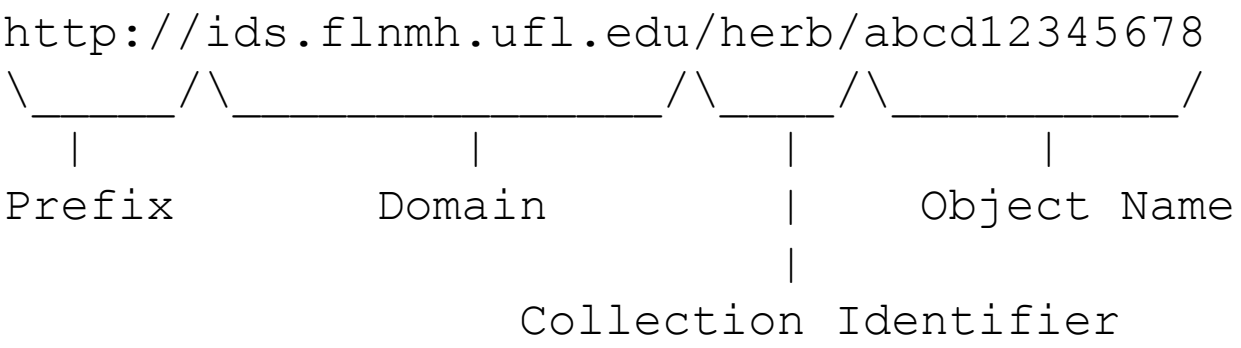
To benefit from and increase the value of the World Wide Web, agents should provide URIs as identifiers for resources.

A resource should have an associated URI if another party might reasonably want to create a hypertext link to it, make or refute assertions about it, retrieve or cache a representation of it, include all or part of it by reference into another representation, annotate it, or perform other operations on it. Software developers should expect that sharing URIs across applications will be useful, even if that utility is not initially evident.

***Architecture of the World Wide Web* (http://www.w3.org/TR/webarch/#identification**

In its GUID document, iDigBio recommends that providers adopt the http URI (Universal Resource Identifier) scheme for all identifiers. Though this scheme results in a pattern that resembles a URL (Universal Resource Locator), URI's do not have to be actionable or resolvable through a web browser. Identifier patterns should be registered with iDigBio.

Pattern:



Sample schemes for VSU:

- for specimen records (collection objects):
  - **valdosta.edu/vsc/co/<barcode value>**
- for taxon records:
  - **valdosta.edu/vsc/tx/<TaxonID>**

Thank You!