



iDigBio

Integrated Digitized Biocollections



iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (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. All images used with permission or are free from copyright.

Gaps in Biodiversity Data: Challenges for Digitization

2-3 December 2015
Missouri Botanical Garden
St. Louis

Gil Nelson
gnelson@bio.fsu.edu
iDigBio/Florida State University



iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (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. All images used with permission or are free from copyright.

The Nature of Gaps in the Availability of Digitized Data

At the institutional level:

- **Strategically ignored data in otherwise digitized collections**
 - **Specimen selection (taxon, geography, endangered status, etc.)**
 - **Data selection**
- **Specimens not yet in the digitization pipeline**

At the community level:

- **Unfunded collections**
- **Small collections**
- **Institutions that choose not to aggregate their data**

The Nature of Gaps in the Availability of Digitized Data

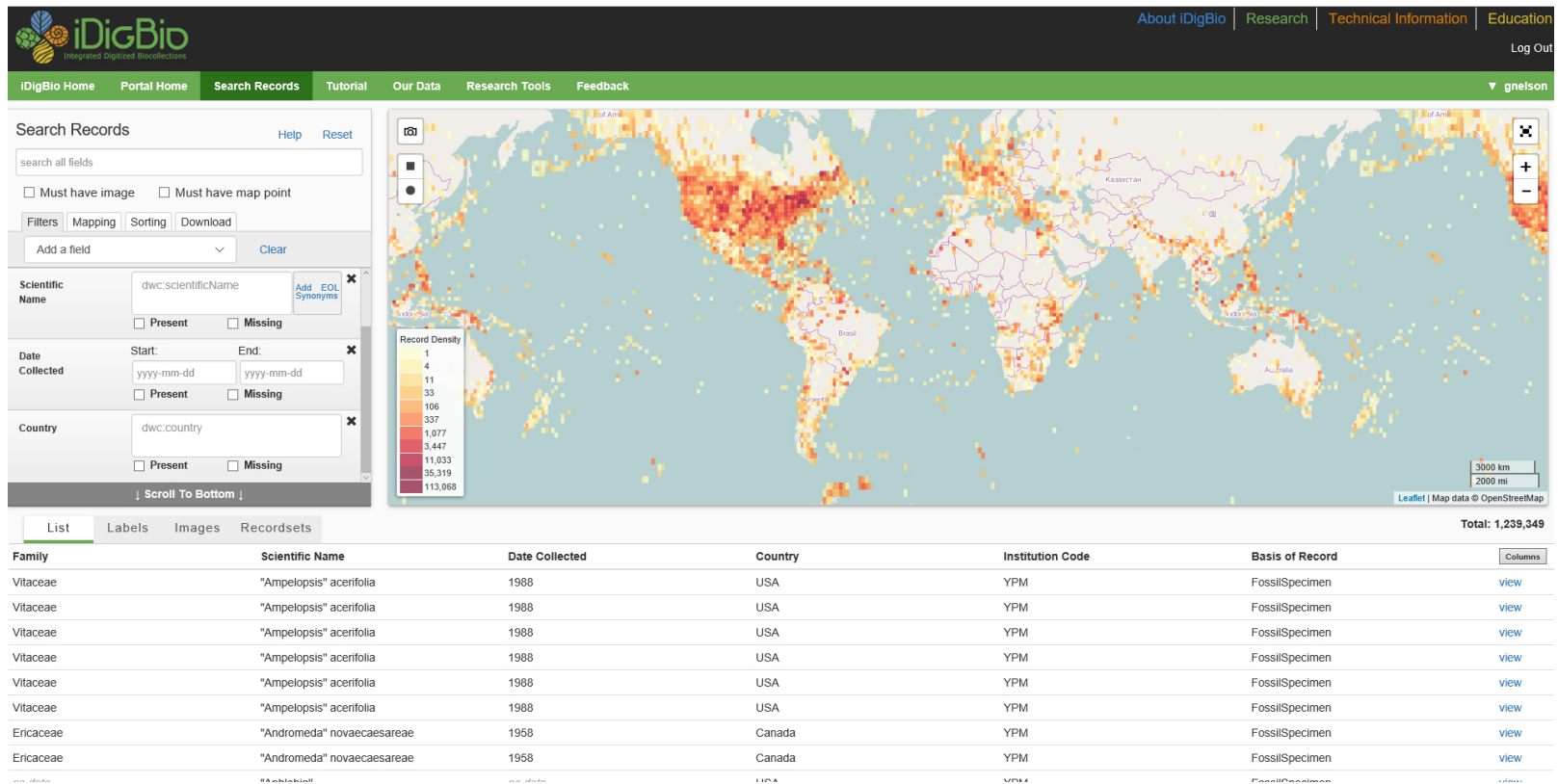
At the institutional level:

- **Strategically ignored data in otherwise digitized collections**
 - Specimen selection (taxon, geography, endangered, types, etc.)
 - Data selection
- **Specimens not yet in the digitization pipeline**

At the community level:

- **Unfunded collections**
- **Small collections**
- **Institutions that choose not to aggregate their data**

Unevenness in data in the iDigBio portal reflects our mission to accept and ingest all contributed biodiversity specimen data with few restrictions or requirements.



The screenshot displays the iDigBio portal interface. At the top, there is a navigation bar with links for 'About iDigBio', 'Research', 'Technical Information', and 'Education'. Below this is a green header with 'iDigBio Home', 'Portal Home', 'Search Records', 'Tutorial', 'Our Data', 'Research Tools', and 'Feedback'. The user 'gnelson' is logged in.

The main content area is divided into two sections. On the left is the 'Search Records' panel, which includes a search bar, filter options (Must have image, Must have map point), and search criteria for Scientific Name, Date Collected, and Country. On the right is a world map showing record density, with a legend indicating values from 1 to 113,068. The map shows high density in North America and Europe.

Below the map is a table of search results. The table has columns for Family, Scientific Name, Date Collected, Country, Institution Code, Basis of Record, and a 'Columns' button. The total number of records is 1,239,349.

Family	Scientific Name	Date Collected	Country	Institution Code	Basis of Record	Columns
Vitaceae	"Ampelopsis" acerifolia	1988	USA	YPM	FossilSpecimen	view
Vitaceae	"Ampelopsis" acerifolia	1988	USA	YPM	FossilSpecimen	view
Vitaceae	"Ampelopsis" acerifolia	1988	USA	YPM	FossilSpecimen	view
Vitaceae	"Ampelopsis" acerifolia	1988	USA	YPM	FossilSpecimen	view
Vitaceae	"Ampelopsis" acerifolia	1988	USA	YPM	FossilSpecimen	view
Vitaceae	"Ampelopsis" acerifolia	1988	USA	YPM	FossilSpecimen	view
Ericaceae	"Andromeda" novaecaesareae	1958	Canada	YPM	FossilSpecimen	view
Ericaceae	"Andromeda" novaecaesareae	1958	Canada	YPM	FossilSpecimen	view

The completeness of digitized biodiversity data that get to any aggregator depends on:

- **what gets digitized,**
- **how it gets digitized,**
- **and what is selected to be shared.**



All dependent on institutional decisions.

Four Basic Assumptions about Data Gaps Directly Attributable to Digitization Practices

The availability of comprehensive, robust, and complete digital datasets from biodiversity specimens is directly dependent on the assumptions made when designing and implementing digitization protocols.

Data that are missing in the record of any particular digitized specimen may often be traced to intentional decisions than to a lack of data to be digitized.

We are still very early in the biodiversity specimen digitization adventure.

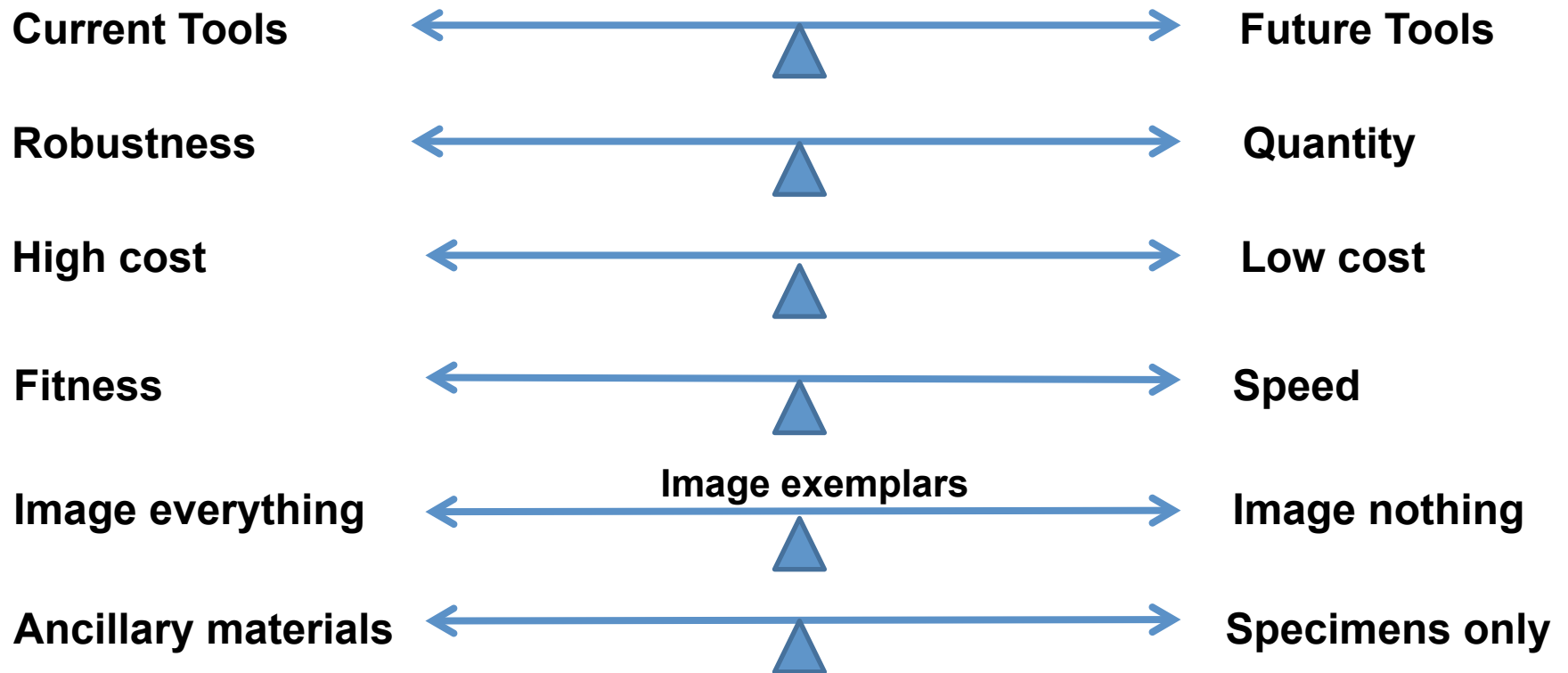
Small collections that may contain the darkest of dark data are underrepresented in digitization projects.



Gaps and biases are often designed into digitization protocols for several important reasons:

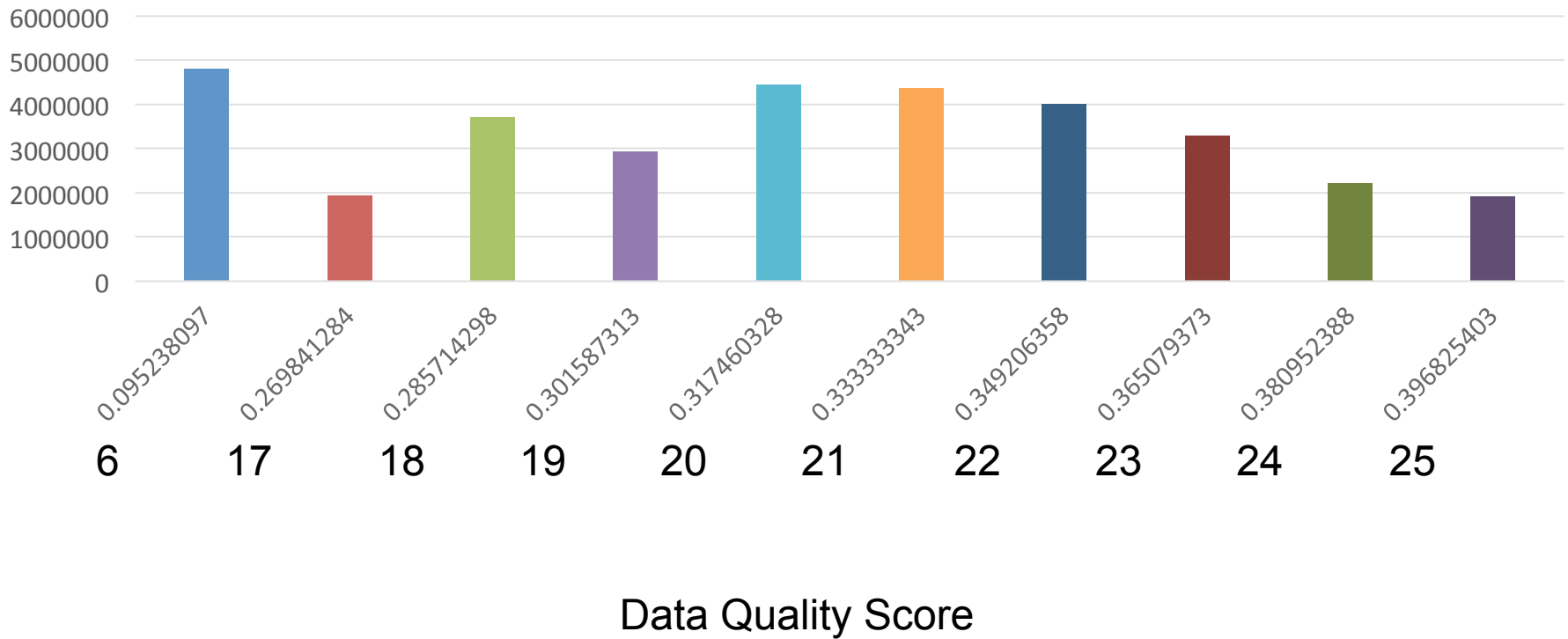
- **Funder expectations**
- **Tendency toward low cost per specimen**
- **What place imaging holds in the process**

Digitization Decision Continua that Influence or Result in Data Gaps





Records



Gaps Due to Differential Strategies for Data Enrichment

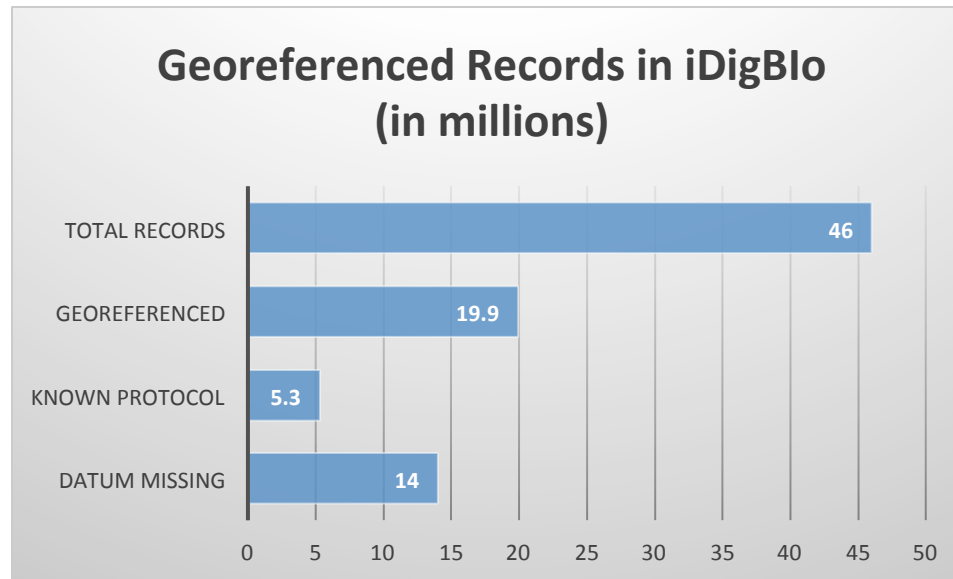
Coding and normalizing:

- Phenological data, anomalies & outliers
- Ecological/habitat descriptions/nomenclature/parsing
- Density/associated species/abundance/habitat health
- Morphological characteristics and variation

Georeferencing:

- Protocol documentation
- Resolution inconsistencies
 - Geographic centroids (county, park, state)
 - Label data
 - Datum
 - Method
- Documentation

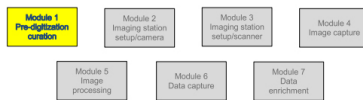
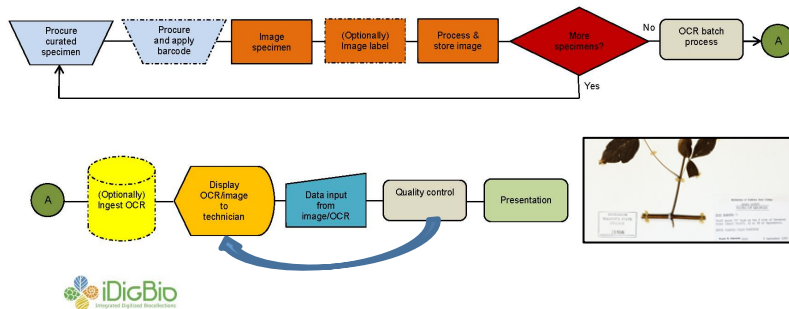
Georeferencing Consistency



Closing Digitization Gaps

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.



Module 1: Pre-digitization Curation Task List

Task ID	Task Description	Explanations and Comments	Resources
T1	Apply storage locator barcodes to storage locations (rooms, cabinets, shelves, folders, drawers, etc).	Most useful when systematically digitizing an entire collection. Otherwise potentially helpful with herbarium inventory. May be less helpful for collections that are digitizing in random order or only portions of the collection related to specific projects, or with significant separation between the pre-digitization curation, databasing, and image capture modules.	Barcodes, QRcode, DataMatrix.
T2	Select specimens to digitize.	For herbaria, this often includes all specimens. Where this is not the case, selection should follow the institution's pre-determined digitization policies or project management plan.	Digitization policy manual or project management plan.
T3	Associate/insert machine readable barcodes/documents with/into folders.	Some institutions create machine readable documents to gather data at the cabinet and/or folder level. Documents might contain such information as family, higher geography, and current identification ("file-as name"). These data will be read and associated with individual collection records in Module 4, T1 or Module 7.	QRcodes, DataMatrix, 1D barcode, or OCR-readable documents for insertion into specimen folders.

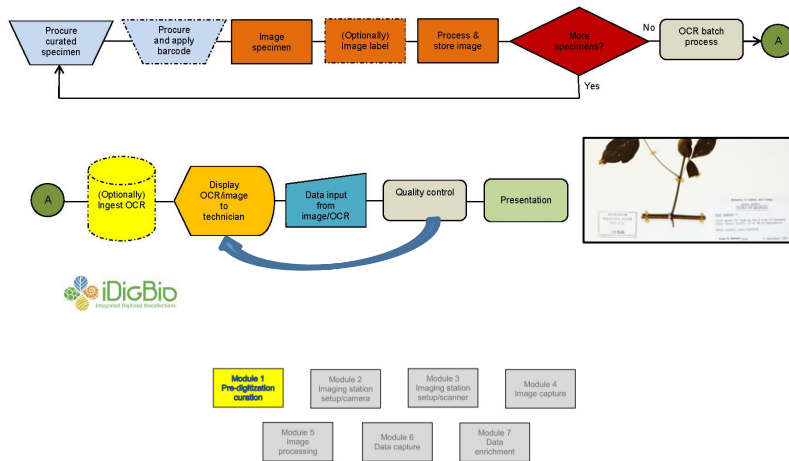
iDigBio has made gains in facilitating the development of digitization workflows in several communities.

The challenges of institutional variation has encouraged our working groups to provide maximum accommodation via the development of modular, more or less “plug and play” approaches that preserve institutional flexibility.

Closing Digitization Gaps

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.



Module 1: Pre-digitization Curation Task List

Task ID	Task Description	Explanations and Comments	Resources
T1	Apply storage locator barcodes to storage locations (rooms, cabinets, shelves, folders, drawers, etc).	Most useful when systematically digitizing an entire collection. Otherwise potentially helpful with herbarium inventory. May be less helpful for collections that are digitizing in random order or only portions of the collection related to specific projects, or with significant separation between the pre-digitization curation, databasing, and image capture modules.	Barcodes, QRcode, DataMatrix.
T2	Select specimens to digitize.	For herbaria, this often includes all specimens. Where this is not the case, selection should follow the institution's pre-determined digitization policies or project management plan.	Digitization policy manual or project management plan.
T3	Associate/insert machine readable barcodes/documents with/into folders.	Some institutions create machine readable documents to gather data at the cabinet and/or folder level. Documents might contain such information as family, higher geography, and current identification ("file-as name"). These data will be read and associated with individual collection records in Module 4, T1 or Module 7.	QRcodes, DataMatrix, 1D barcode, or OCR-readable documents for insertion into specimen folders.

Adoption of discipline consensus workflows based on research community.

Community agreement on the essential core data requirements that should drive digitization workflows and contribute to research.

Agree on sets of community-based priorities for addressing current data gaps.



iDigBio
Integrated Digitized Biocollections

Thank you!

