Herbarium Digitization

Overview and Guide to Resources

24 May 2014

TORCH VIII + iDigBio Digitization Workshop

Deborah Paul, on Twitter @idbdeb

Sul Ross State University, Alpine, Texas

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.

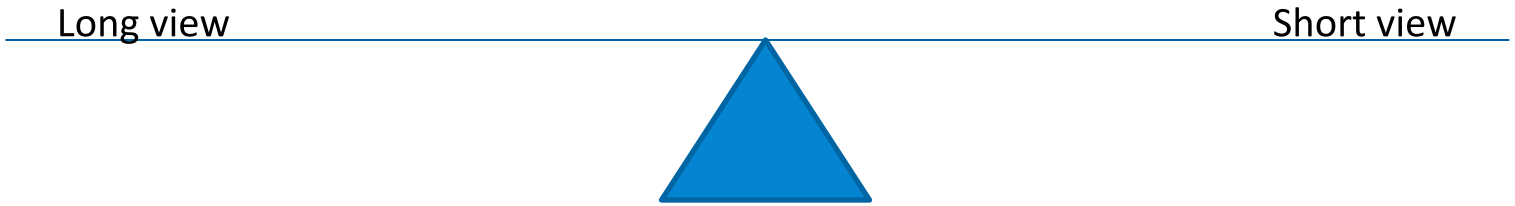
Biodiversity Digitization: Ultimate Goals

- 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 (worldwide).

Digitization Decision Making

- Global Needs and Policies
  - Local Policies and Decisions
    - Specific Workflows
- What to digitize?
- Can we digitize every bit of data associated with each object?
  - Skeleton records?
- *How to decide*?

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, including a commitment to implementing future enhancements?

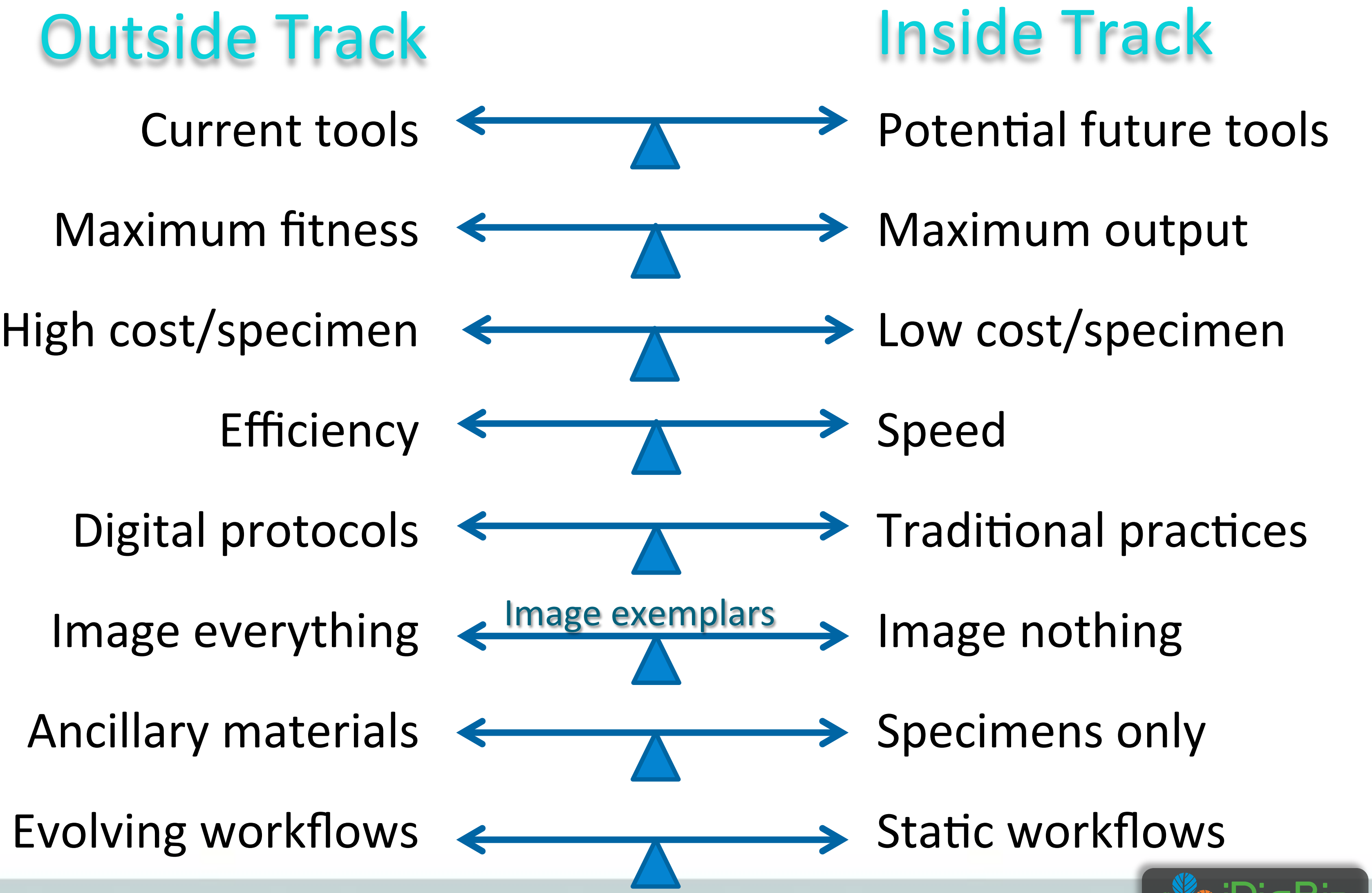
Pressures mitigating the long view

So much data, so little time.
Collections are not getting smaller (proactive vs. legacy).
Funding agencies have high output/low cost expectations.
We only have 3 years to get this done (sustainable models?).
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 (inside track).

Future tools favoring the short view

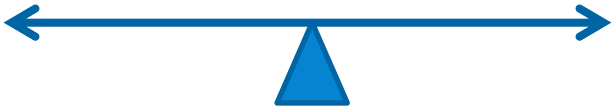
- OCR, NLP, and ICR (handwriting analysis) improvements.
- Automated image analysis for data extraction.
- Data mining of labels.
- Robotic technologies, conveyor belts, etc.
- Improvements in discovery/capture/use of duplicates.
- Improvements in voice recognition and other data entry technologies.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

Digitization Continua/Decision Points



Outside Track		Inside Track
Current tools		Potential future tools
Maximum fitness		Maximum output
High cost/specimen		Low cost/specimen
Efficiency		Speed
Digital protocols		Traditional practices
Image everything	Image exemplars	Image nothing
Ancillary materials		Specimens only
Evolving workflows		Static workflows

Long view



Short View

Facilitators	Facilitators
Emphasize fitness for use	Emphasize output
Robust datasets	Spartan datasets
Data validation/cleaning	Defer validation/cleaning
Integrated quality control	Deferred/minimum quality control
Integrated georeferencing	Deferred georeferencing
Intensive curation	Deferred or cursory curation
Record historical annotations	Record current determination
Staff specialization	Staff generalization
Small collection	Large collection
Emphasize images	Emphasize data
High quality images	Low quality images

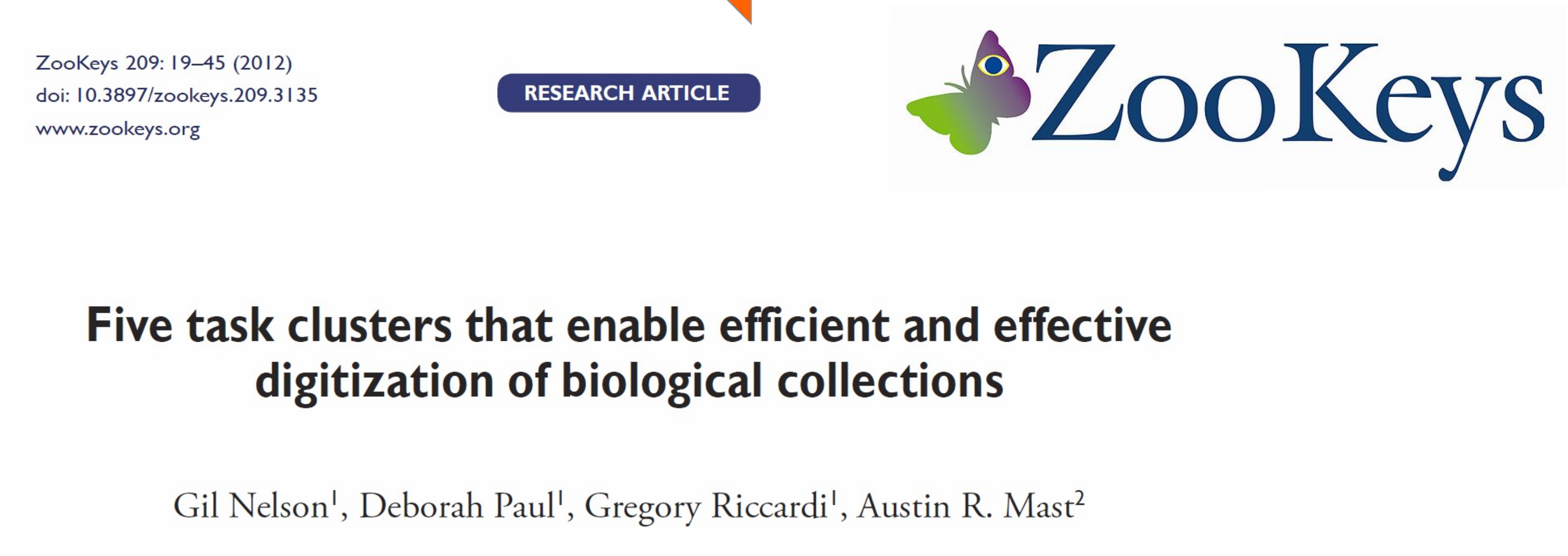
Robust

Spartan

Establishing a Baseline

- Find out what is going on in digitization
  - Benchmarking
  - Grounded Theory

Survey



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

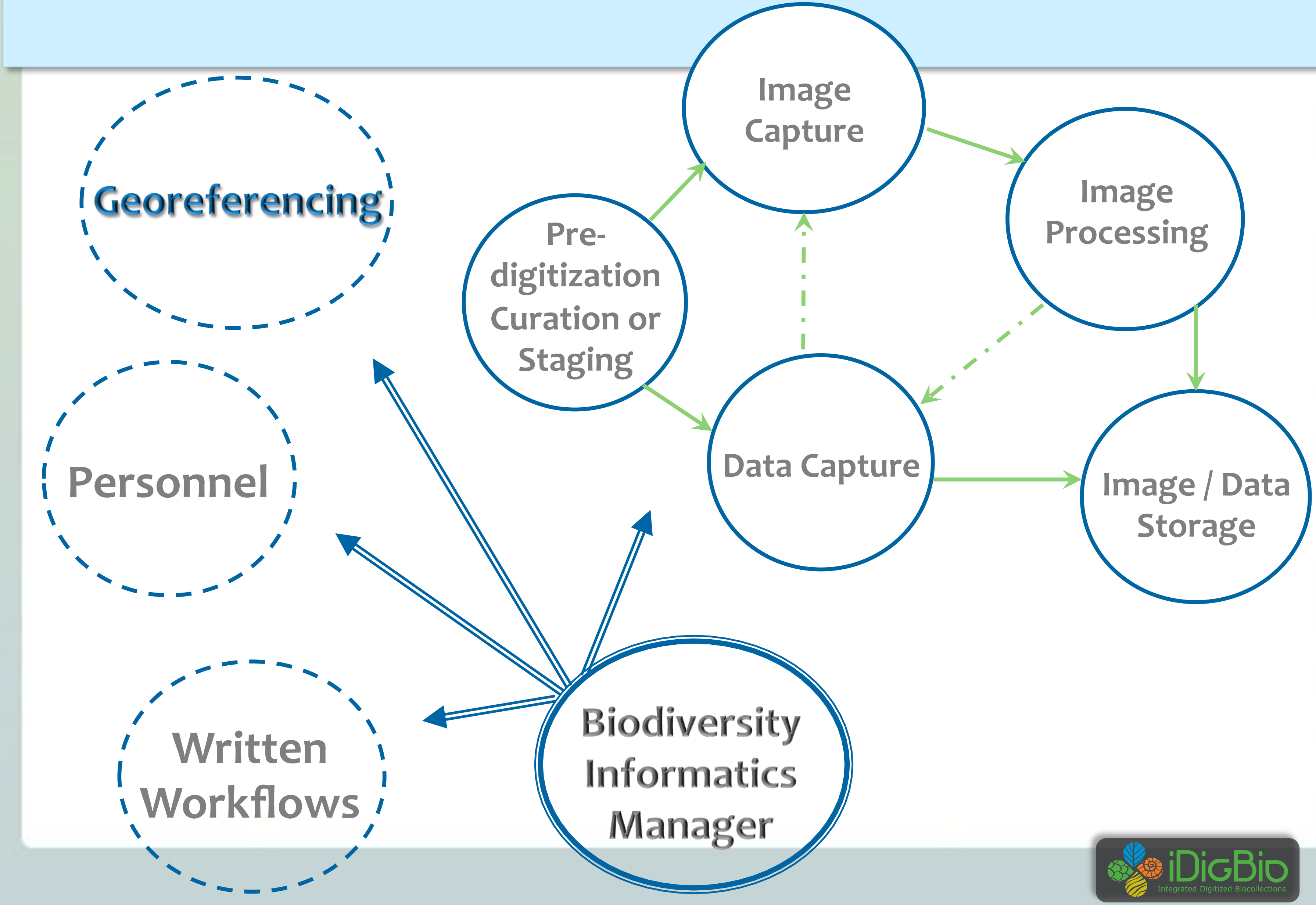
RESEARCH ARTICLE

ZooKeys

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]

KEY CLUSTERS



Georeferencing

Personnel

Written Workflows

Pre-digitization Curation or Staging

Image Capture

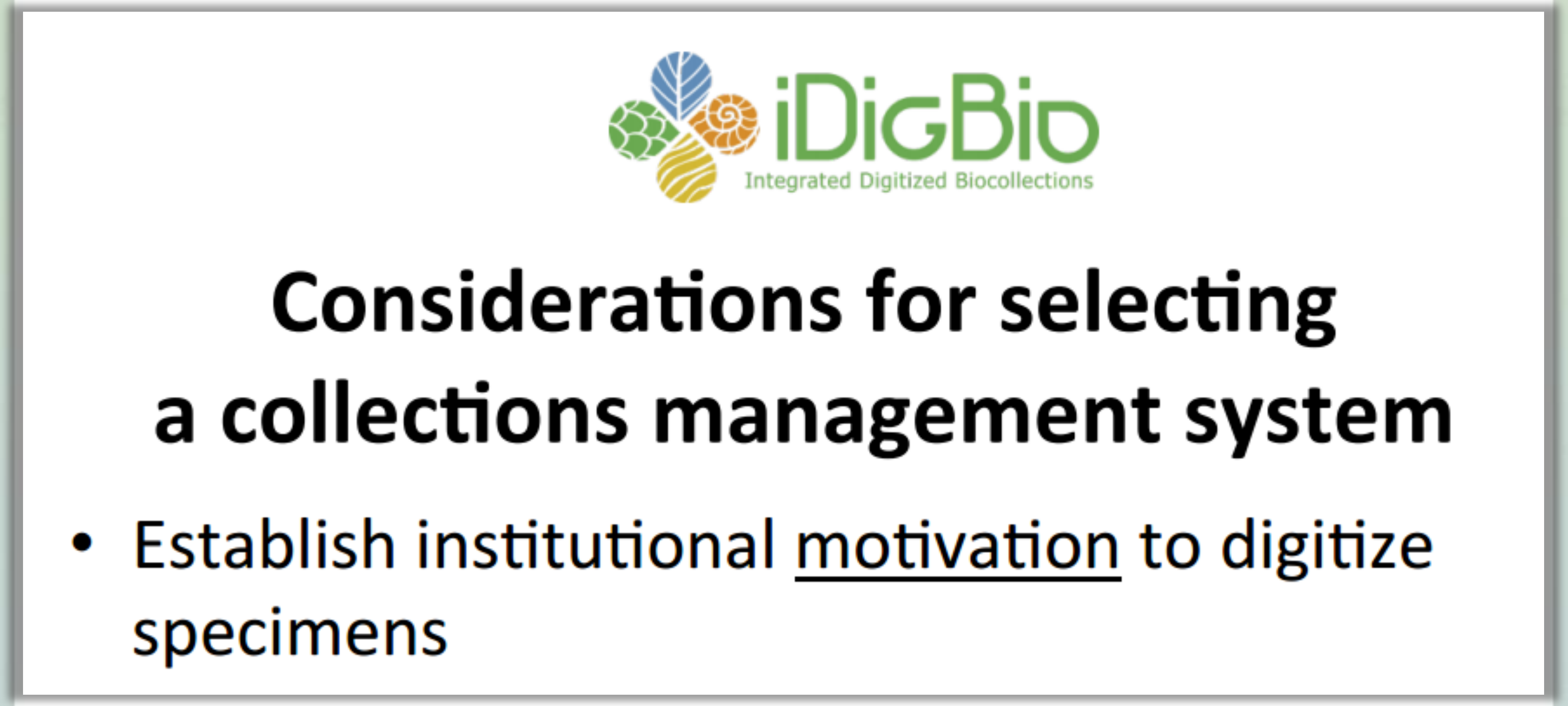
Image Processing

Data Capture

Image / Data Storage

Biodiversity Informatics Manager

What database? What suits best?



iDigBio
Integrated Digitized Biocollections

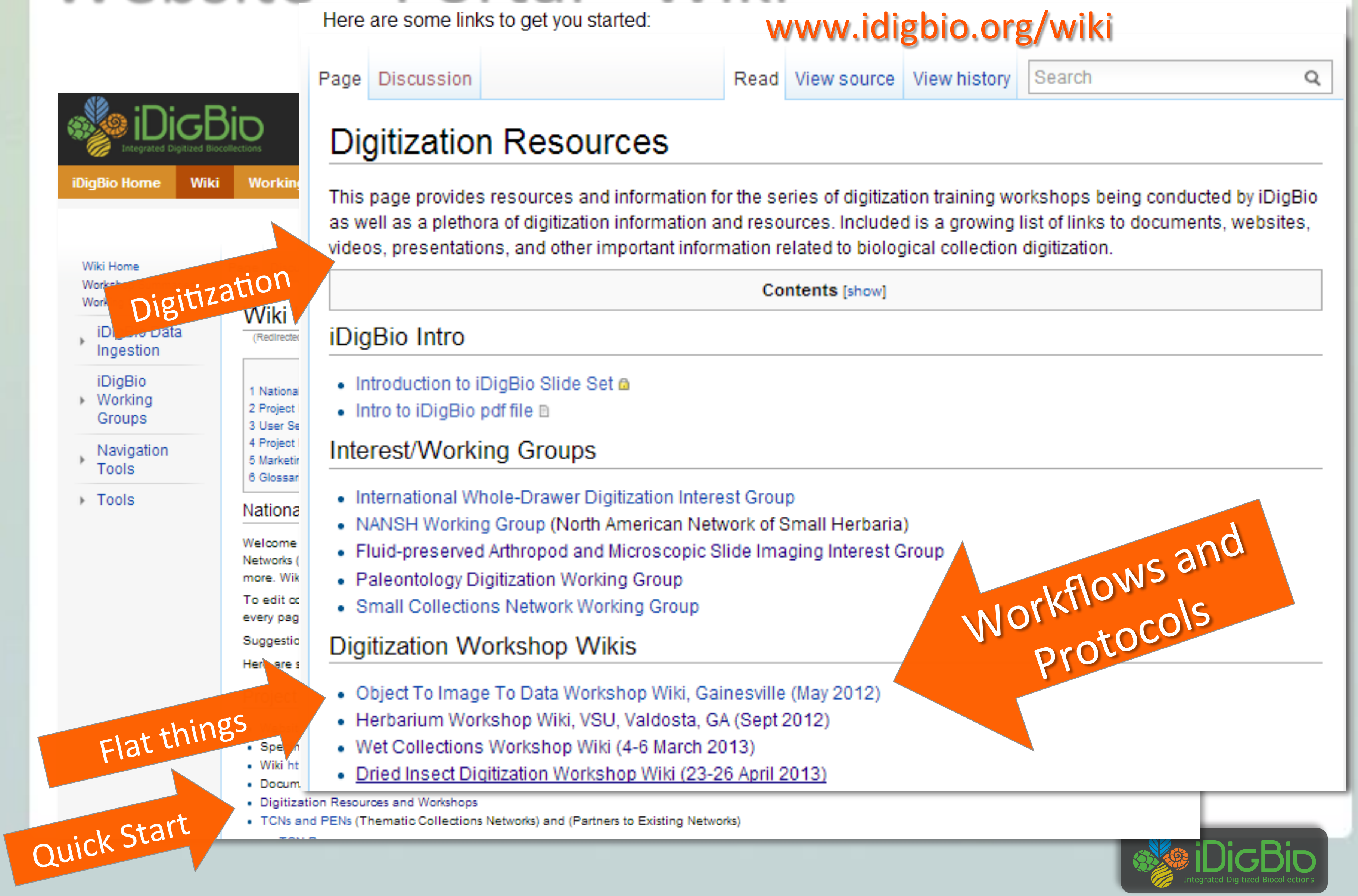
Considerations for selecting a collections management system

- Establish institutional motivation to digitize specimens

by Joanna McCaffrey, Digitizing Plant Specimens Workshop, 2012

Website – Portal - Wiki

www.idigbio.org/wiki



Here are some links to get you started:

Page | Discussion | Read | View source | View history | Search

Digitization Resources

This page provides resources and information for the series of digitization training workshops being conducted by iDigBio as well as a plethora of digitization information and resources. Included is a growing list of links to documents, websites, videos, presentations, and other important information related to biological collection digitization.

Contents [show]

iDigBio Intro

- Introduction to iDigBio Slide Set
- Intro to iDigBio pdf file

Interest/Working Groups

- International Whole-Drawer Digitization Interest Group
- NANSH Working Group (North American Network of Small Herbaria)
- Fluid-preserved Arthropod and Microscopic Slide Imaging Interest Group
- Paleontology Digitization Working Group
- Small Collections Network Working Group

Digitization Workshop Wikis

- Object To Image To Data Workshop Wiki, Gainesville (May 2012)
- Herbarium Workshop Wiki, VSU, Valdosta, GA (Sept 2012)
- Wet Collections Workshop Wiki (4-6 March 2013)
- Dried Insect Digitization Workshop Wiki (23-26 April 2013)

- Digitization Resources and Workshops
- TCNs and PENs (Thematic Collections Networks) and (Partners to Existing Networks)

Digitization

Workflows and Protocols

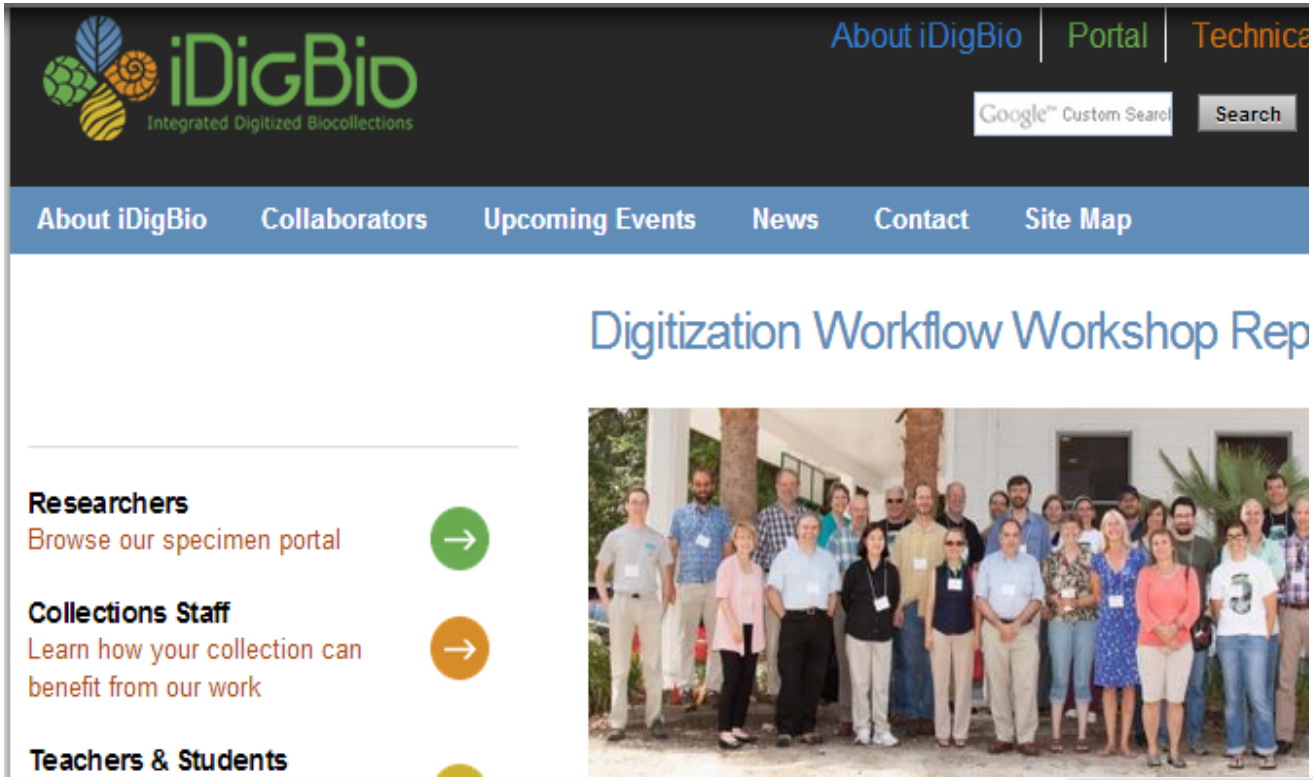
Flat things

Quick Start

Developing Robust Object to Image to Data Workflows (DROID)

Workshop design collaborative

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



Digitization Workflow Workshop Rep

Researchers
Browse our specimen portal

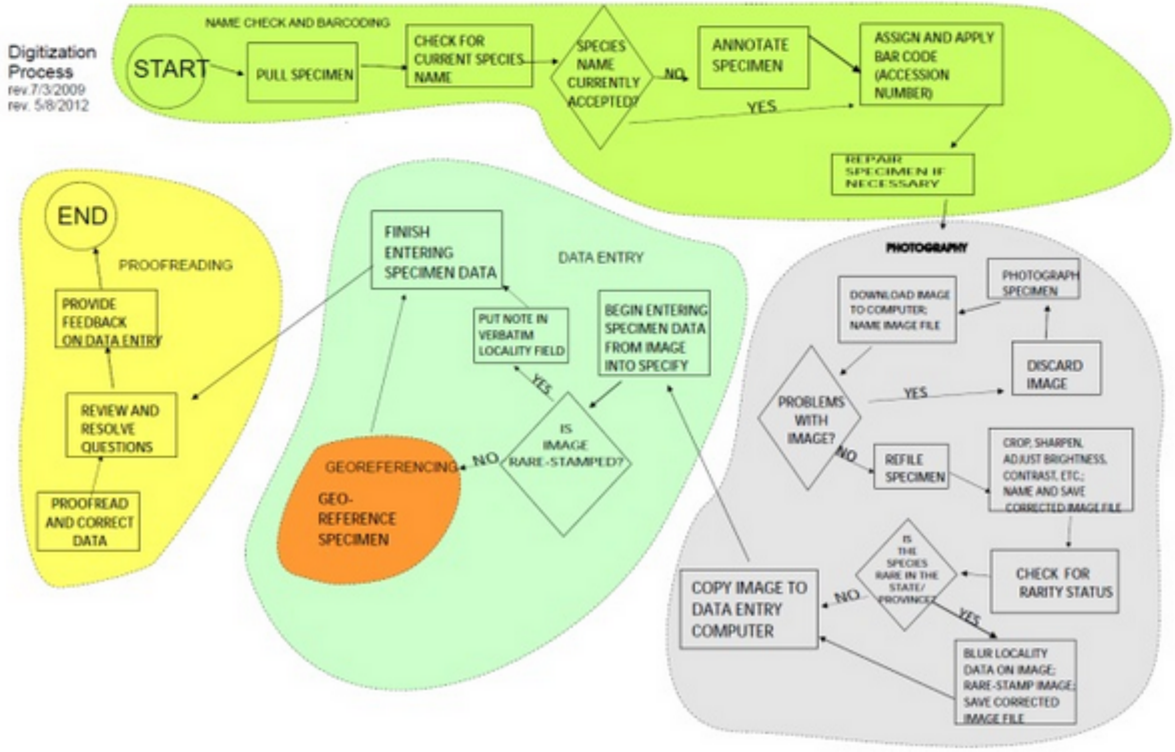
Collections Staff
Learn how your collection can benefit from our work

Teachers & Students

DROID 1: Flat Things

Your Input

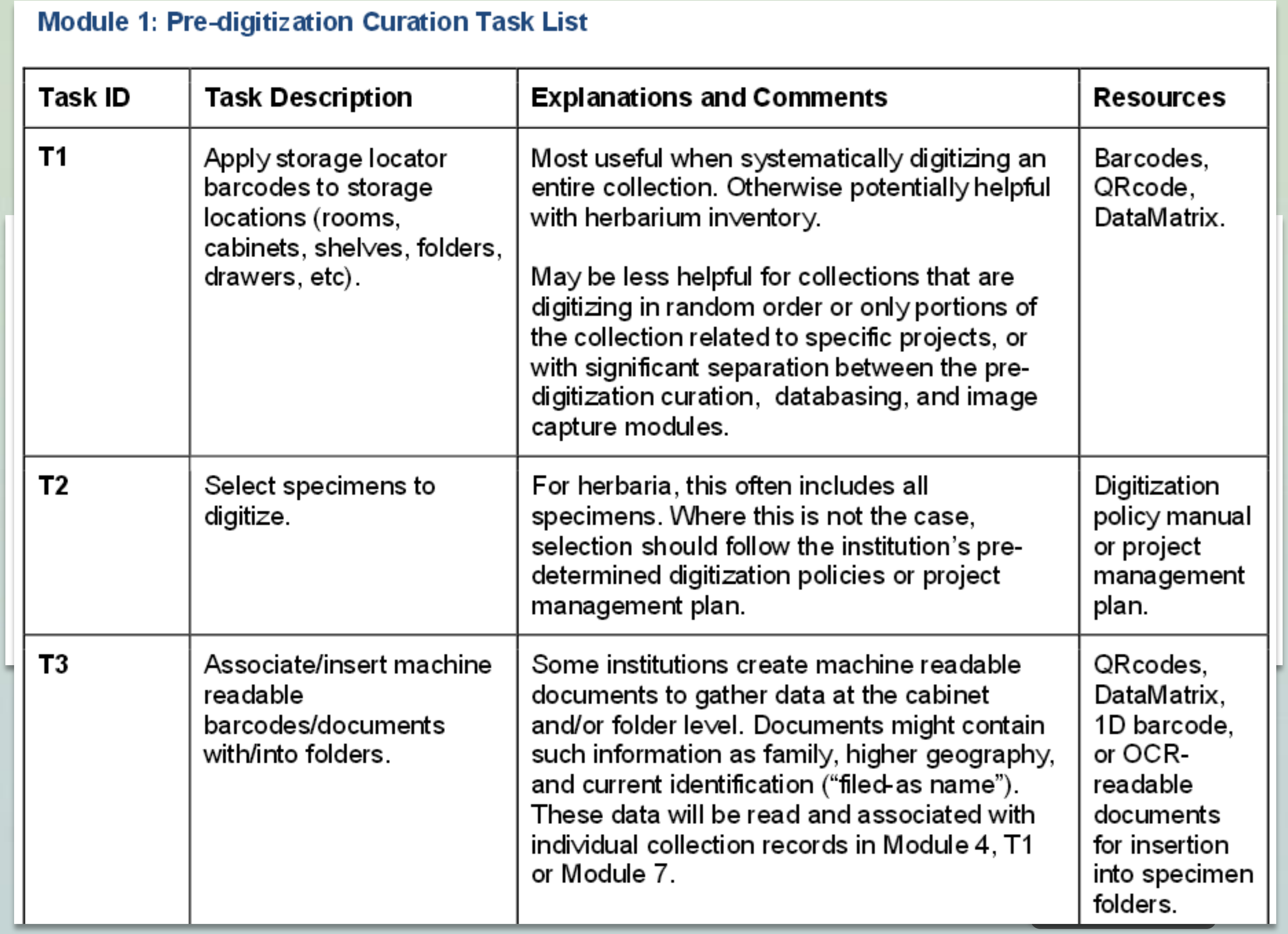
Presenter: Dorothy Allard


Digitization Process
rev.7/3/2009
rev. 5/8/2012
NAME CHECK AND BARCODING
START
PULL SPECIMEN
CHECK FOR CURRENT SPECIES NAME
SPECIES NAME CURRENTLY ACCEPTED?
NO
YES
ANNOTATE SPECIMEN
ASSIGN AND APPLY BAR CODE (ACCESSION NUMBER)
REPAIR SPECIMEN IF NECESSARY
PHOTOGRAPHY
PHOTOGRAPH SPECIMEN
DOWNLOAD IMAGE TO COMPUTER; NAME IMAGE FILE
PROBLEMS WITH IMAGE?
YES
DISCARD IMAGE
NO
REFILE SPECIMEN
CROP, SHARPEN, ADJUST BRIGHTNESS, CONTRAST, ETC.; NAME AND SAVE CORRECTED IMAGE FILE
CHECK FOR RARITY STATUS
IS THE SPECIES RARE IN THE STATE/ PROVINCE?
NO
YES
BLUR LOCALITY DATA ON IMAGE; RARE-STAMP IMAGE; SAVE CORRECTED IMAGE FILE
COPY IMAGE TO DATA ENTRY COMPUTER
DATA ENTRY
BEGIN ENTERING SPECIMEN DATA FROM IMAGE INTO SPECIFY
IS IMAGE RARE-STAMPED?
YES
PUT NOTE IN VERBATIM LOCALITY FIELD
NO
GEOREFERENCING
GEO-REFERENCE SPECIMEN
FINISH ENTERING SPECIMEN DATA
PROOFREADING
PROOFREAD AND CORRECT DATA
REVIEW AND RESOLVE QUESTIONS
PROVIDE FEEDBACK ON DATA ENTRY
END

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 these workflows have been documented or made available to the larger collections community.

iDigBio
Integrated Digitized Biocollections



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 ("filed-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.

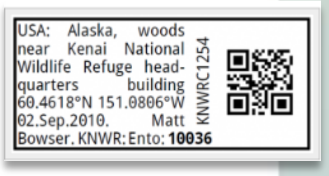
Predigitization Curation, AKA Staging

Personnel

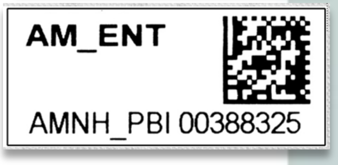
- taxonomic judgment
- personnel management
- uses standard references
- keen observational skills
- specimen-handling skills
- select specimens to image

Activities

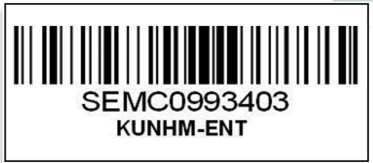
- Determination/annotation
  - (a professional)
- Data verification
  - (a professional)
- Drawer/cabinet organization
  - (trained techs)
- Re-pinning
  - (trained techs)
- Barcode application
  - (trained techs)


USA: Alaska, woods near Kenai National Wildlife Refuge head-quarters building 60.4618°N 151.0806°W 02.Sep.2010. Matt Bowser. KNWR:Ento:10036

KNWRC1254


AM_ENT

AMNH_PBI 00388325


SEMC0993403

KUNHM-ENT

Not all steps require a professional

Curation is a potential bottleneck

iDigBio
Integrated Digitized Biocollections

Predigitization (unanticipated) Benefits

- inspect /repair / specimen damage (ipm)
- collection health,
- inventory collection,
- re-pin / remount specimens
- replenish / replace preservatives
- attach a unique identifier
  - (most often a 1- or 2-D barcode)
  - to a specimen, container, or cabinet,
- discover important but
  - unknown, lost, or dislocated holdings
  - (e.g. those owned by other institutions or the federal government),
- update nomenclature and taxonomic interpretation,
- reorganize the cabinets, cases, trays, and containers,
- vet type specimens, and
- select exemplars for digitization / imaging

iDigBio
Integrated Digitized Biocollections

Data Capture

- Discipline-specific training or interest is often not particularly important

Data Capture is a potential bottleneck

Personnel

- Accurate
- Efficient
- Focused
- Tolerant of tedium
- Productive
- Speedy
- Oriented toward improving process

Process

- Keyboarding
- Voice capture
- OCR
  - QC
  - Data extraction
  - Barcode value extraction
- Data import

Source Documents

- Specimens/labels
- Images
- Ledgers/catalogs
- Field notebooks
- Monographs

Data Import Issues

- Source
  - Internal (legacy)
  - External
- Data quality/trust
- Data format
- Transformation/field mapping
- Post-import cleanup and quality control

Importance of written protocols

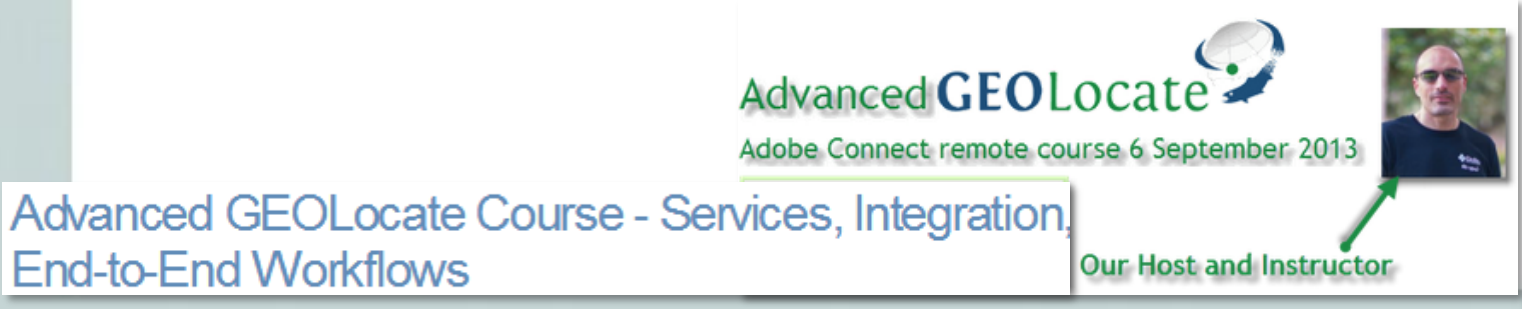
Georeferencing Working Group (GWG)

Current Resources

- Train-the-Trainers (TTT) I and II
- Online Workshop Resource
- Human Resources
  - Workforce Training
- listserve
- http://vimeo.com/idigbio
  - http://vimeo.com/idigbio/albums

Ongoing Work

- online training materials,
- Webinars
- Georef Workflows Help
- Georef Workshop Protocols
- Facilitating Georef Workshops
- http://www.georeferencing.org



Advanced GEOLocate

Adobe Connect remote course 6 September 2013

Advanced GEOLocate Course - Services, Integration, End-to-End Workflows

Our Host and Instructor

Review I

- Global decisions
  - Deciding to digitize (Digitization Maturity)
  - Funding
    - Information, Library Science, Museum Studies
    - Funding (IMLS, CLIR)
    - Expertise
    - Partners
  - Deciding what to digitize
  - Choosing collection management software
- Benchmarking – best practices discovery, group by Digitization task clusters
  - Predigitization curation
  - Data Capture
  - (Imaging)
  - Personnel
  - Georeferencing

Review II

- Developing robust object-to-image-to-data workflows (DROID)
- Benefits of digitization
- Importance of written protocols
  - Creating, managing workflows and protocols
  - Feedback
  - Sharing yours (DROID)
- Data from specimens or data from images?
- Workflow and data entry efficiency (bottlenecks)
- (Barcodes)
- The planet, needs the data.

To TORCH and our Sul Ross State hosts,

Thank you!

UF UNIVERSITY of FLORIDA

FLORIDA STATE UNIVERSITY 1851 VIRES ARTES MORES

FLORIDA MUSEUM OF NATURAL HISTORY

NSF

iDigBio Integrated Digitized Biocollections