Digitizing Biological Collections

Integrated Digitized Biocollections

Getting Started with Digitization

Gil Nelson
Wet Collections Digitization Workshop
4-6 May 2013

Figure 1. Dorsal (left) and ventral views (right) of the male holotype of Rhabdopus pseudacuifolius sp. n. (NMW 16301:5).
Scale bar = 5 mm.
Ultimate Goals of Biological Collections Digitization

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.
Digitizing Biological Collections

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 robust digitization.
• Taking the short view means 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!
  Doing the minimum now and enhance it later.
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.
Future Tools Favoring the Inside/Middle Tracks

- 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.
Digitizing Biological Collections

Digitization Decision Continua

- Current Tools
- Potential Future Tools
- Maximum Fitness
- Quantity
- 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
Digitizing Biological Collections

Robust

Facilitators

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

Spartan

Facilitators

- Emphasize output
- Skeletal datasets
- Defer validation/cleaning
- Deferred quality control
- Deferred georeferencing
- Deferred digital curation
- Record current determination
- Staff generalization
- Large collection
- Emphasize data
- Low quality images
Focusing on Efficiency vs. Speed

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
Five task clusters that enable efficient and effective digitization of biological collections
ZooKeys 209: 19–45, doi: 10.3897/zookeys.209.3135
Thank You!