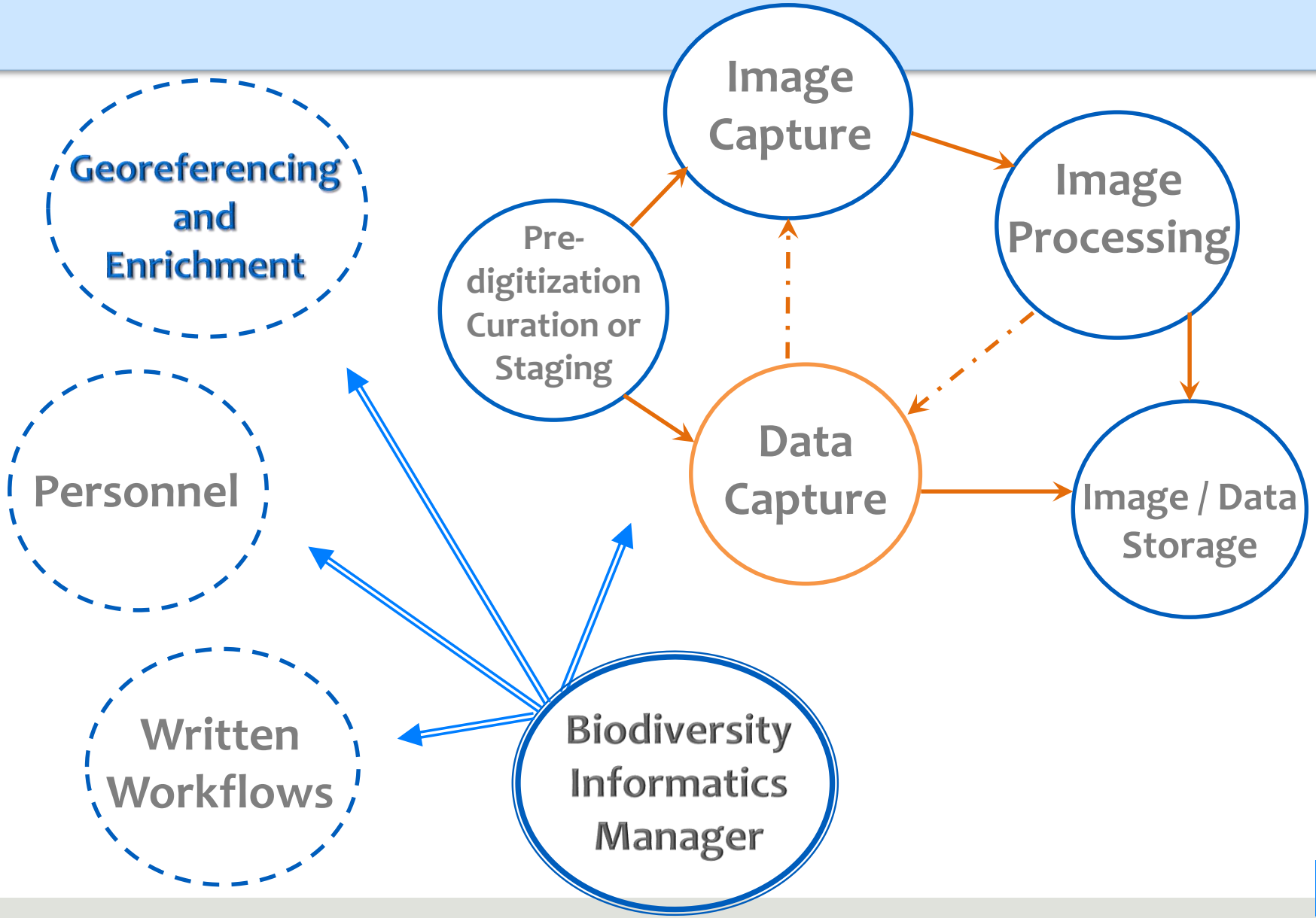
Task Cluster 4 - Electronic Data Capture: collection data management and data capture

Presenter: Deborah Paul
Florida State University
Integrated Digitized Biocollections (iDigBio)

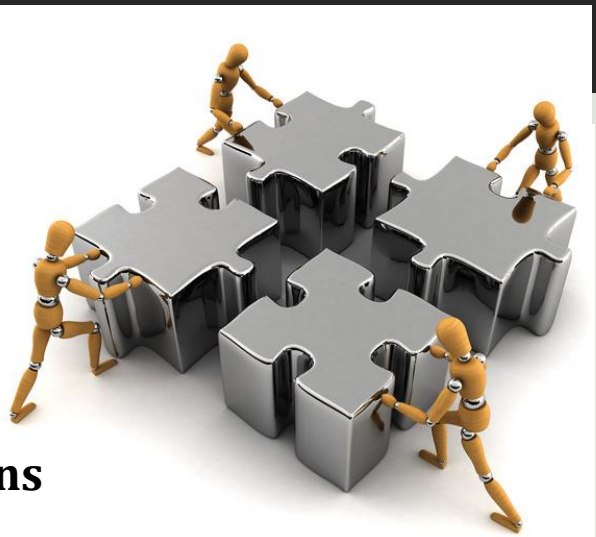
Brazilian Biodiversity Information System (SiBBr) Launching Event
25 November 2014
Brasilia, DF, Brazil

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.

DATA CAPTURE



Staffing

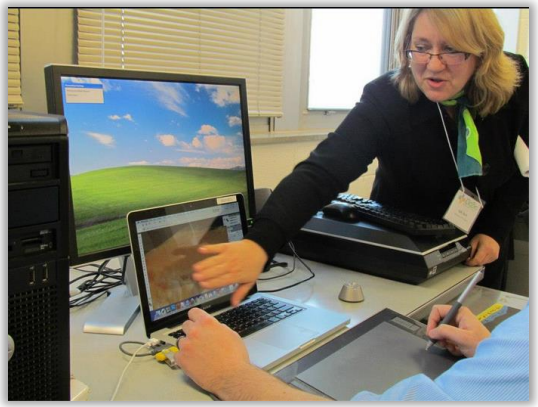


Data Entry, Imaging, Data Validation

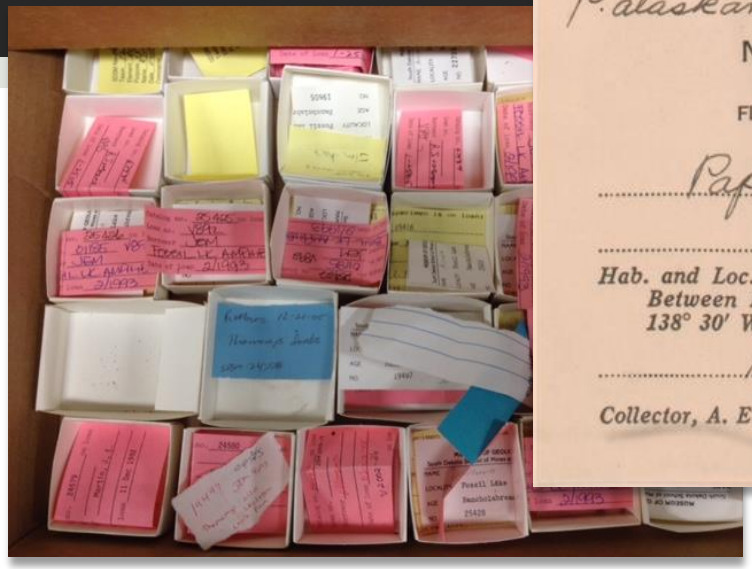
- In-house volunteers
- Paid staff
- Citizen scientists

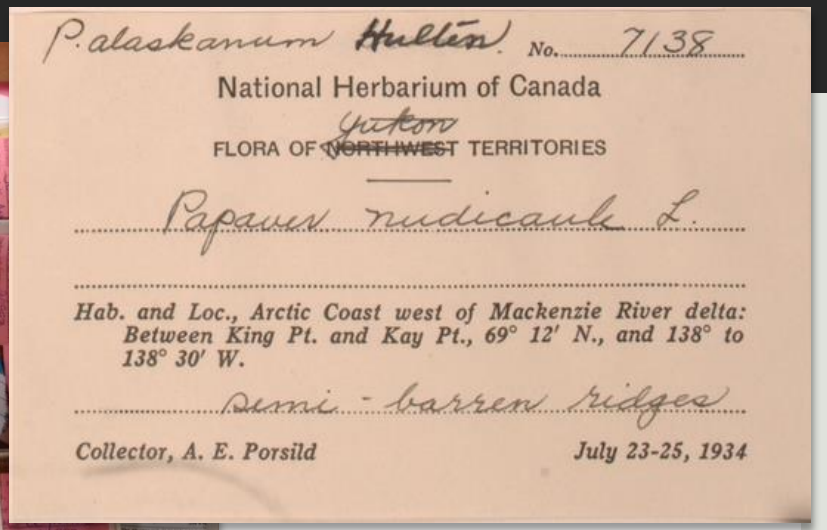
Considerations

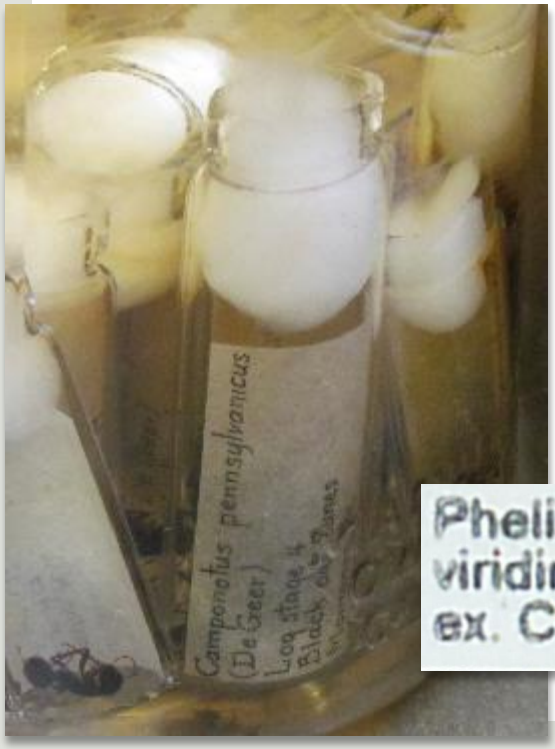
- Training
- Roles
- Autonomy
- Managing Staff and Data

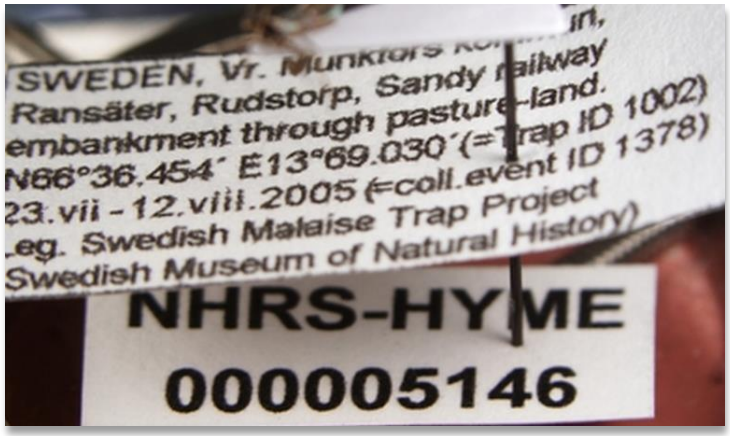


Labels
Notebooks
Card Files
Vials, …









Trend

Minimal Data Capture

- "filed as" name
- higher geography
- barcode
- image

- all sheets in folder get the same initial data
- only the barcode differs

How do we get these minimal records *completed*, or more complete?

Note darwin core / georeferencing standards

Your database

My database

map to a standard!

http://www.britishmuseum.org/images/rosettawriting384.jpg

Inside the 1899 Harriman Expedition

New York Botanical Garden (NY)

Occurrence Data

Long Form

Collector ? Wm. Trelease

Number ?

Date ?

Dupes?

Auto search

Associated Collectors ? De Alton Saunders

Verbatim Date ? 1899

Exsiccati Title

Number

Scientific Name ? Wahlenbergiella mucosa

Note: Full editing permissions are needed to edit an identification

Country: United States

State/Province: Alaska

County: Juneau

Locality: Taku Inlet

Latitude: 58.255678

Longitude: -134.078102

Uncertainty ? 17128

Verbatim Coordinates

Tools

Elevation in Meters

Verbatim Elevation

Habitat

Substrate

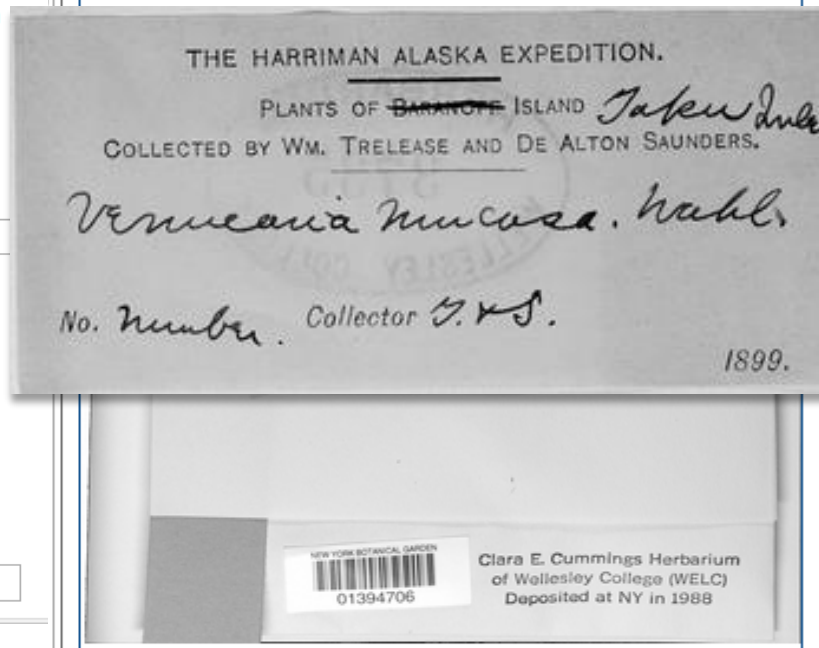
Notes

Save Edits

Status Auto-Set: Pending Review

Symbiota

Label Processing



OCR Image

Options

OCR whole image

OCR w/ analysis

Notes:

Source: ABBYY:2013-02-09

Save OCR Edits

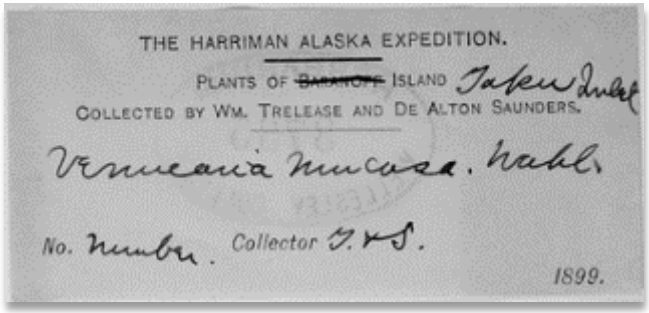
Parse OCR (LBCC)

1 of 1

Delete OCR

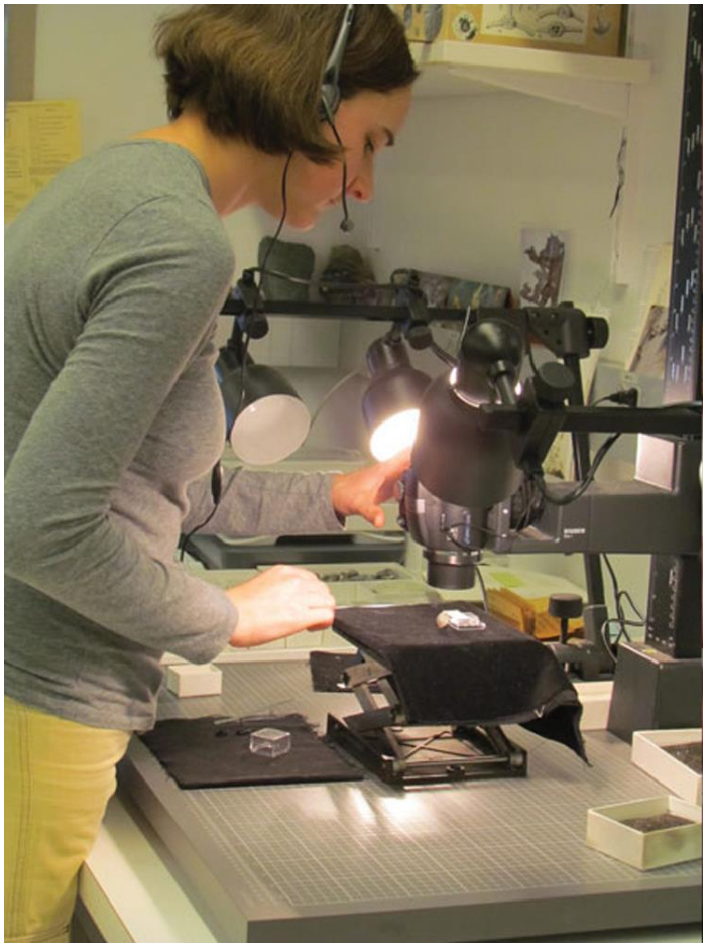
Specimen Data Capture

- Extracting label data
  - Before, during, after imaging (a choice)
- Entering data from label images
  - reduces specimen handling
  - can facilitate ability to read labels
  - creates a voucher for the label
- Database interface often customized
  - speeds data transcription and enhances accuracy
- Data often imported from spreadsheets (Specify)
- Online data entry (in-the-cloud)



Data Capture Options

- Data capture with voice* (shhhhhh)
- Using OCR software and OCR output parsed into database
  - vetted by a person
- OCR output is searchable!
- Records multi-keyed
- Crowd-sourcing



A few more key thoughts about data capture …

- database software and data-entry issues
  - ditto,
  - drop-downs,
  - automated scripts for validation
- error catching / data validation strategies
- tracking what has / has not been entered / imaged
- protocols / workflows continuously evaluated
- data quality / integrity

About iDigBio | Research | Technical Information | Education

iDigBio
Integrated Digitized Biocollections

Google™ Custom Search Search My account Log out

Making data and images of millions of biological specimens available on the web

24,705,794
Specimen Records

4,046,837
Media Records

357
Recordsets

Search the Portal

Why digitization matters
More about what we do and why

Digitization
Learn, share and develop best practices

Sharing Collections
Documentation on data ingestion

Working Groups
Join in, contribute, be part of the community

Proposals
New tool and workshop ideas

Citizen Scientists
How can you help biological collections?

Obrigada SiBBr! Find out more at …

https://www.idigbio.org/content/workflow-modules-and-task-lists

www.idigbio.org

facebook.com/iDigBio

twitter.com/iDigBio

vimeo.com/idigbio

idigbio.org/rss-feed.xml

webcal://www.idigbio.org/events-calendar/export.ics

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.