Data capture: issues, best practices, options, lessons learned

Deborah Paul
iDigBio, Florida State University

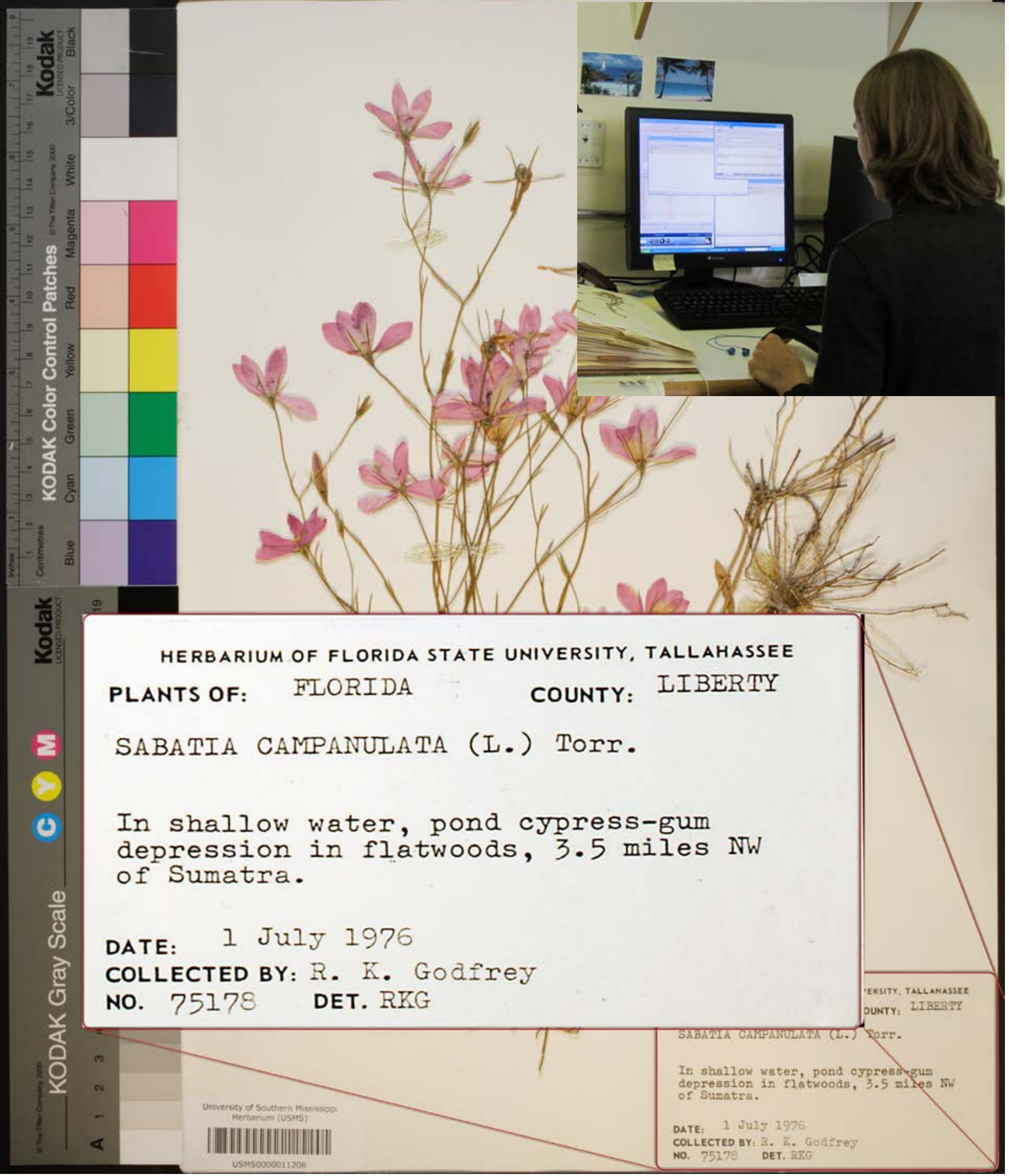
DigIn iDigBio Marine Invertebrate Digitisation Workshop
4-5 February 2019
@idbdeb @iDigBio

iDigBio is funded by grants from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program [DBI-1115210 (2011-2018) and DBI-1547229 (2016-2021)]. 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.

Goals of data capture

- Read and transcribe written materials
- Move accurate data into database

- Topics
  - parsing
  - label variation, ancillary data, derived data
  - wet collections digitization tasks and resources
    - data capture is one part
  - data quality – make a plan
  - data organization and doer happiness
  - future data collection



HERBARIUM OF FLORIDA STATE UNIVERSITY, TALLAHASSEE

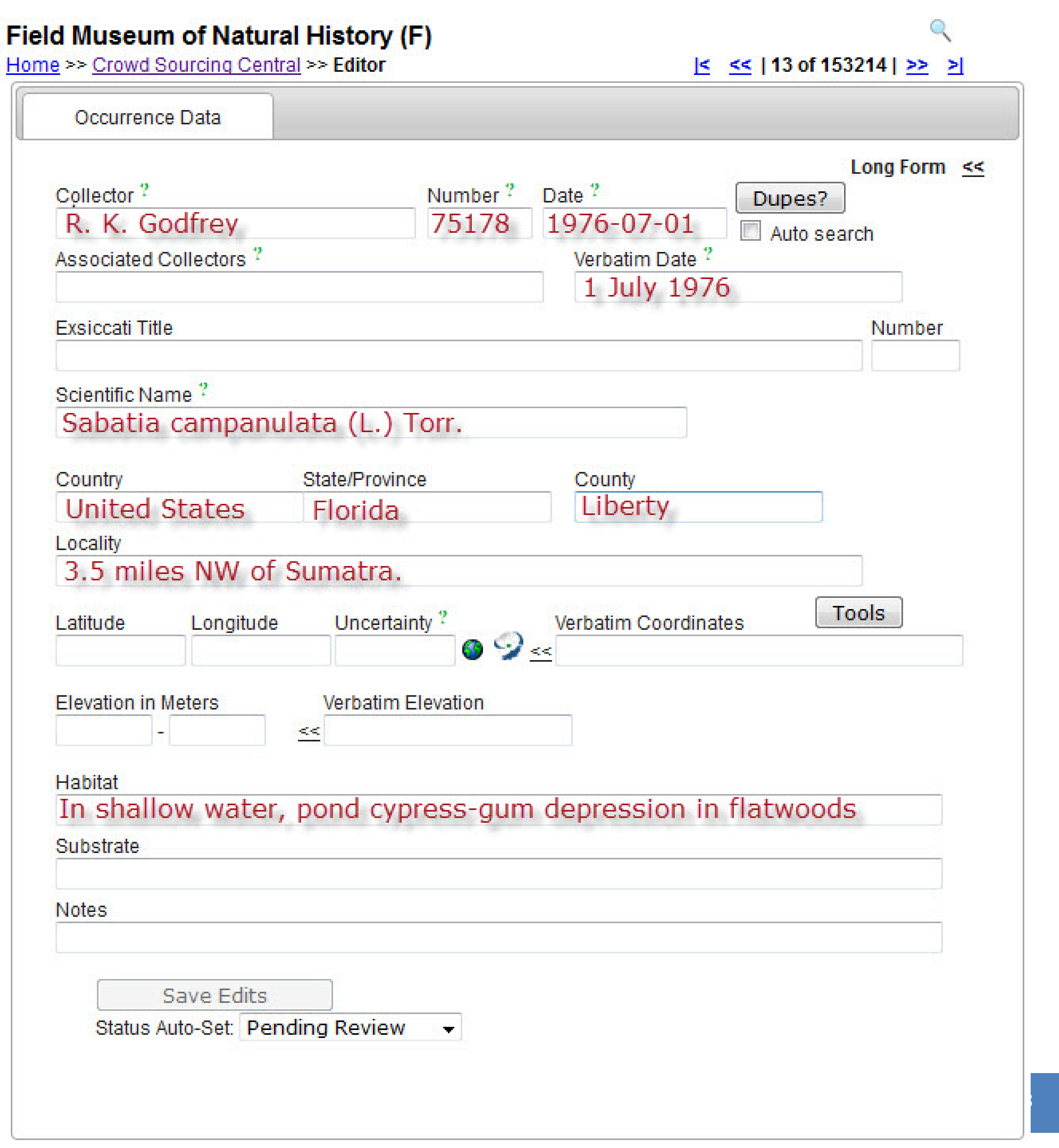
PLANTS OF: FLORIDA COUNTY: LIBERTY

SABATIA CAMPANULATA (L.) Torr.

In shallow water, pond cypress-gum depression in flatwoods, 3.5 miles NW of Sumatra.

DATE: 1 July 1976
COLLECTED BY: R. K. Godfrey
NO. 75178 DET. RKG

ERSITY, TALLAHASSEE
OUNTY: LIBERTY
SABATIA CAMPANULATA (L.) Torr.
In shallow water, pond cypress-gum depression in flatwoods, 3.5 miles NW of Sumatra.
DATE: 1 July 1976
COLLECTED BY: R. K. Godfrey
NO. 75178 DET. RKG



Field Museum of Natural History (F)

Home >> Crowd Sourcing Central >> Editor

|< << | 13 of 153214 | >> >|

Occurrence Data

Long Form <<

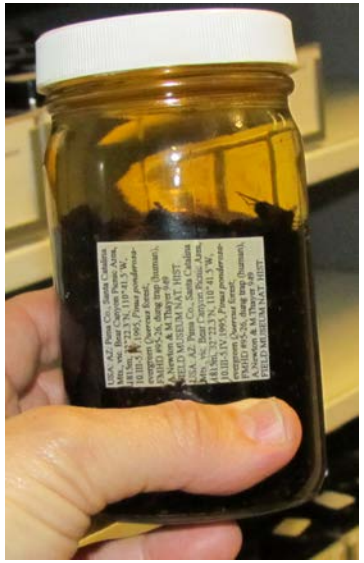
Collector: R. K. Godfrey
Number: 75178
Date: 1976-07-01
Dupes?
□ Auto search
Associated Collectors:
Verbatim Date: 1 July 1976
Exsiccati Title:
Number:
Scientific Name: Sabatia campanulata (L.) Torr.
Country: United States
State/Province: Florida
County: Liberty
Locality: 3.5 miles NW of Sumatra.
Latitude:
Longitude:
Uncertainty:
Verbatim Coordinates:
Tools
Elevation in Meters: -
Verbatim Elevation:
Habitat: In shallow water, pond cypress-gum depression in flatwoods
Substrate:
Notes:

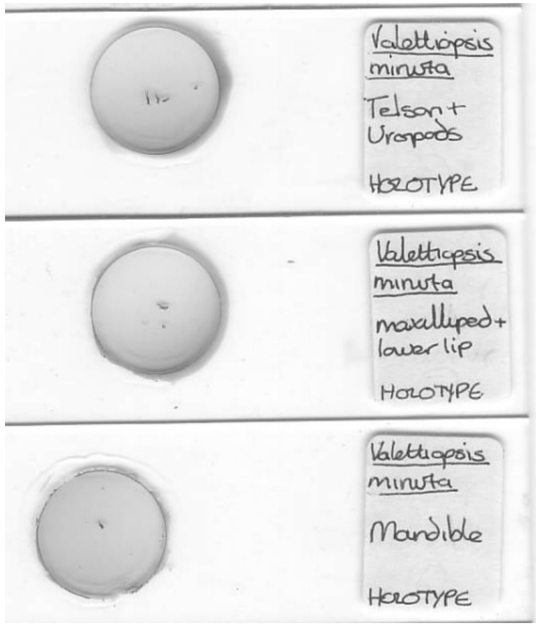
Save Edits

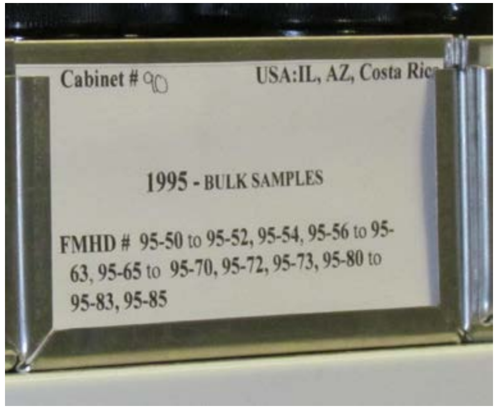
Status Auto-Set: Pending Review

Data Capture Challenges

- ink
- **wet**
- typed
- pencil
- fragile
- printed
- curved
- stacked
- obscured
- handwritten
- uneven lines
- colored paper
- non-planar surfaces
- non-standard terms
- non-standard formats


USA: AZ: Pima Co., Santa Catalina Mts., vic. Bear Canyon Picnic Area, 1815m, 32°22.3'N, 110°41.5'W, 10.III-5.IV.1995, Pinus ponderosa-evergreen Quercus forest, FMHD #95-26, dung trap (human), A. Newton & M. Thayer 949 FIELD MUSEUM NAT. HIST
USA: AZ: Pima Co., Santa Catalina Mts., vic. Bear Canyon Picnic Area, 1815m, 32°22.3'N, 110°41.5'W, 10.III-5.IV.1995, Pinus ponderosa-evergreen Quercus forest, FMHD #95-26, dung trap (human), A. Newton & M. Thayer 949 FIELD MUSEUM NAT. HIST

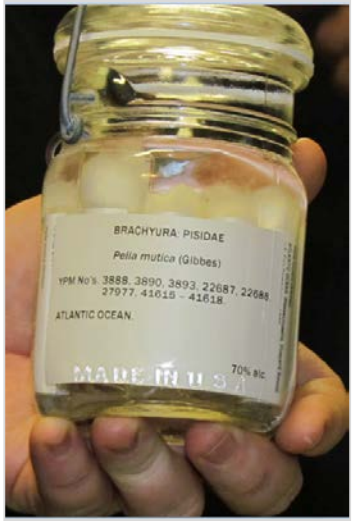


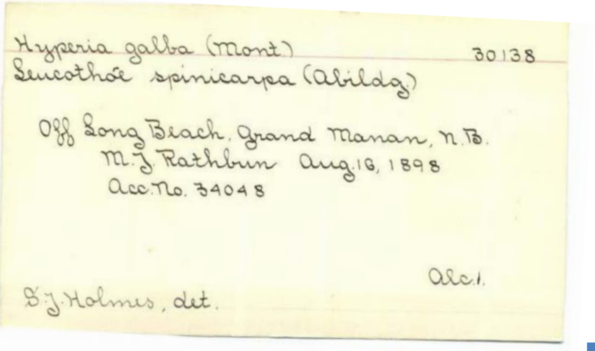


Cabinet # 90 USA:IL, AZ, Costa Ric

1995 - BULK SAMPLES

FMHD # 95-50 to 95-52, 95-54, 95-56 to 95-63, 95-65 to 95-70, 95-72, 95-73, 95-80 to 95-83, 95-85





Hyperia galba (Mont.) 30138
Leucothoe spinicarpa (Abildg.)

Off Long Beach, Grand Manan, N.B.
M. J. Rathbun Aug. 16, 1898
Acc. No. 34048

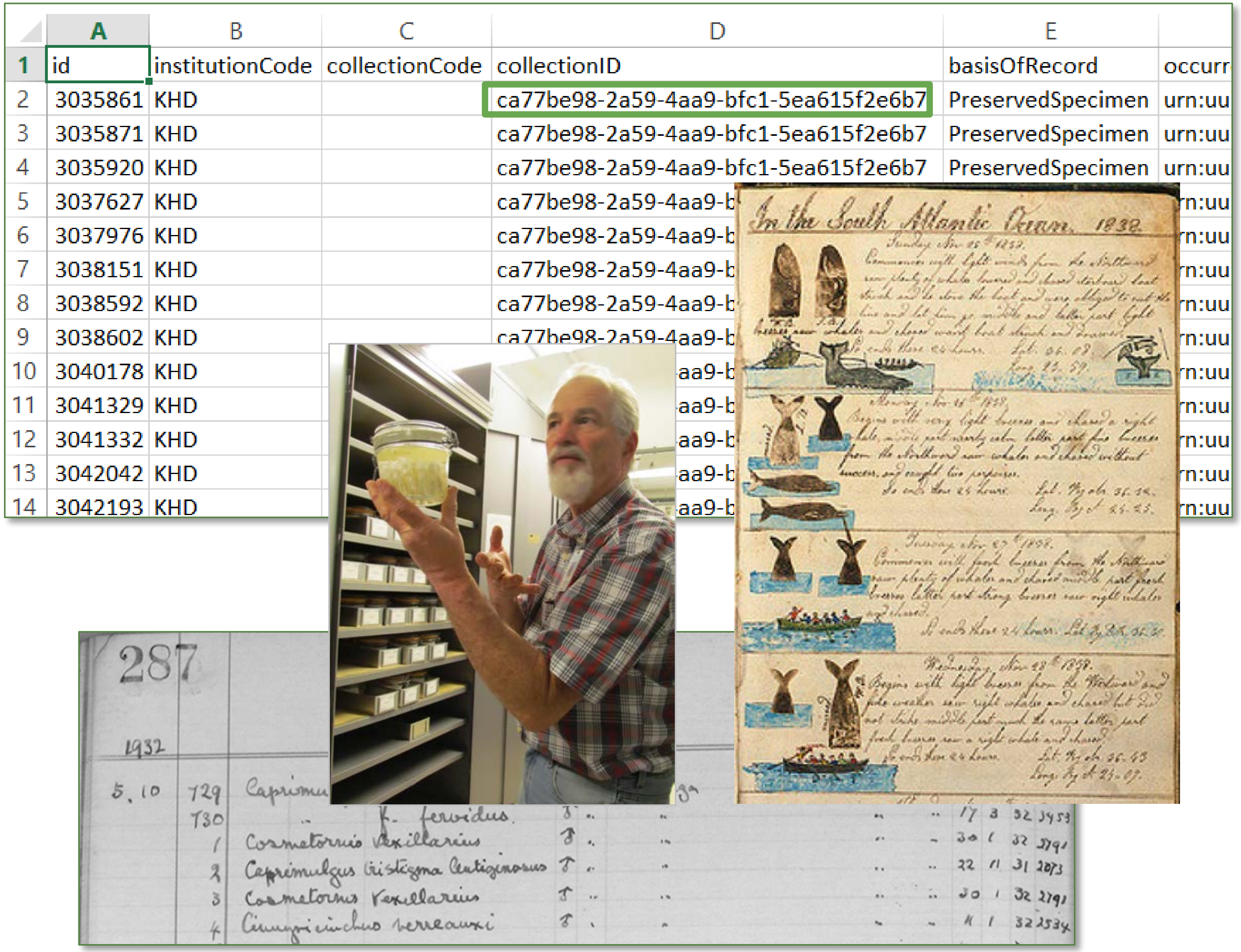
Alc.

S. J. Holmes, det.

Mobilization

other data formats needing **capture and *standardization*** **in order to share**

- spreadsheets
- log books
- field notes
- other derivative objects
- storage formats



	A	B	C	D	E	
1	id	institutionCode	collectionCode	collectionID	basisOfRecord	occurr
2	3035861	KHD		ca77be98-2a59-4aa9-bfc1-5ea615f2e6b7	PreservedSpecimen	urn:uu
3	3035871	KHD		ca77be98-2a59-4aa9-bfc1-5ea615f2e6b7	PreservedSpecimen	urn:uu
4	3035920	KHD		ca77be98-2a59-4aa9-bfc1-5ea615f2e6b7	PreservedSpecimen	urn:uu
5	3037627	KHD		ca77be98-2a59-4aa9-b		rn:uu
6	3037976	KHD		ca77be98-2a59-4aa9-b		rn:uu
7	3038151	KHD		ca77be98-2a59-4aa9-b		rn:uu
8	3038592	KHD		ca77be98-2a59-4aa9-b		rn:uu
9	3038602	KHD		ca77be98-2a59-4aa9-b		rn:uu
10	3040178	KHD		aa9-b		rn:uu
11	3041329	KHD		aa9-b		rn:uu
12	3041332	KHD		aa9-b		rn:uu
13	3042042	KHD		aa9-b		rn:uu
14	3042193	KHD		aa9-b		rn:uu

287

1932

5.10 729 Caprimu

730 K. fervidus ♂ 17 3 32 3453

1 Cosmetornis Vexillarius ♂ 30 1 32 3791

2 Caprimulgus Bistigma Lentiginosus ♂ 22 11 31 2073

3 Cosmetornis Vexillarius ♂ 30 1 32 2791

4 Cinnyricinclus verreauxi ♂ 11 1 32 2534

In the South Atlantic Ocean. 1858

Sunday Nov 25th 1858.
Commences with light winds from the Northward saw plenty of whales lowered and chased starboard boat struck and he stove the boat and were obliged to cut the line and let him go middle and latter part light breezes saw whales and chased waist boat struck and [illegible]
So ends these 24 hours. Lat. 36.08 Long. 23.59

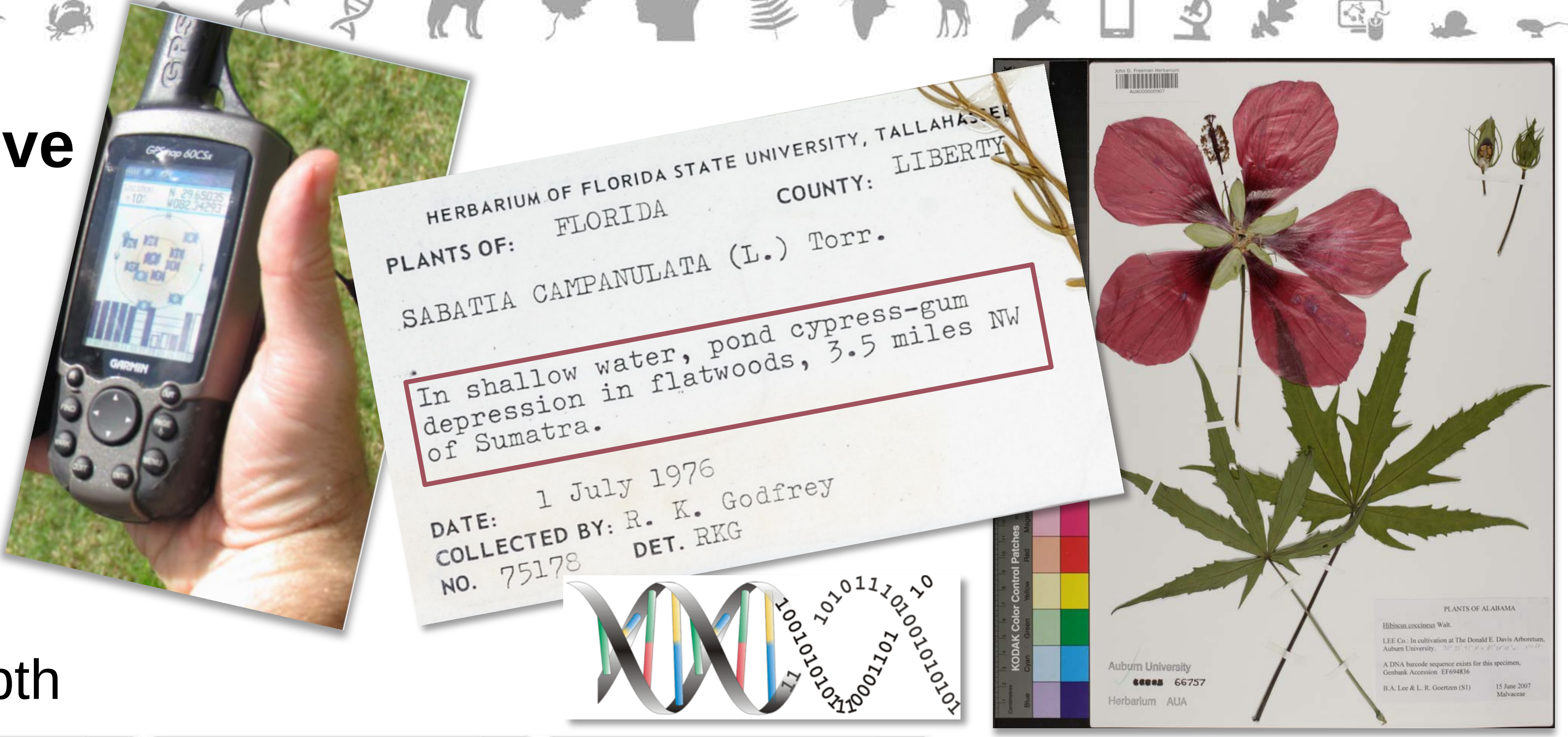
Monday Nov 26th 1858.
Begins with very light breezes and chased a right whale, middle part nearly calm latter part fine breezes from the Northward saw whales and chased without success, and caught two porpoises.
So ends these 24 hours. Lat. By obs. 36.34 Long. By Ch 24.25

Tuesday Nov 27th 1858.
Commences with fresh breezes from the Northward saw plenty of whales and chased middle part fresh breezes latter part strong breezes saw right whales and chased.
So ends these 24 hours. Lat By Obs. 36.3[illegible]

Wednesday Nov 28th 1858.
Begins with light breezes from the Westward and fine weather saw right whales and chased but did not strike middle part much the same latter part fresh breezes saw a right whale and chased.
So ends these 24 hours. Lat. By obs 36.43 Long. By Ch 23.07

Extract and Derive

- Geolocation
- Phenology
- Habitat
- Ecology
- Morphology
- Stratigraphy / Depth
- DNA…



HERBARIUM OF FLORIDA STATE UNIVERSITY, TALLAHASSEE

PLANTS OF: FLORIDA COUNTY: LIBERTY

SABATIA CAMPANULATA (L.) Torr.

In shallow water, pond cypress-gum depression in flatwoods, 3.5 miles NW of Sumatra.

DATE: 1 July 1976

COLLECTED BY: R. K. Godfrey DET. RKG

NO. 75178

PLANTS OF ALABAMA

Hibiscus coccineus Walt.

LEE Co.: In cultivation at The Donald E. Davis Arboretum, Auburn University.

A DNA barcode sequence exists for this specimen, Genbank Accession EF694836

B.A. Lee & L. R. Goertzen (S1) 15 June 2007 Malvaceae

Auburn University 66757 Herbarium AUA

1 of 3

1.00 mm

Epistenia coeruleata, Westwood, 1832, 1991-05-14

2 of 3

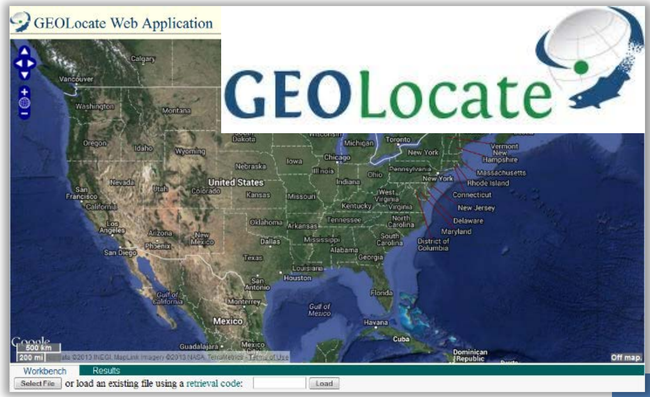
2.0 mm

Epistenia coeruleata, Westwood, 1832, 1991-05-14

3 of 3

1.00 mm

Epistenia coeruleata, Westwood, 1832, 1991-05-14

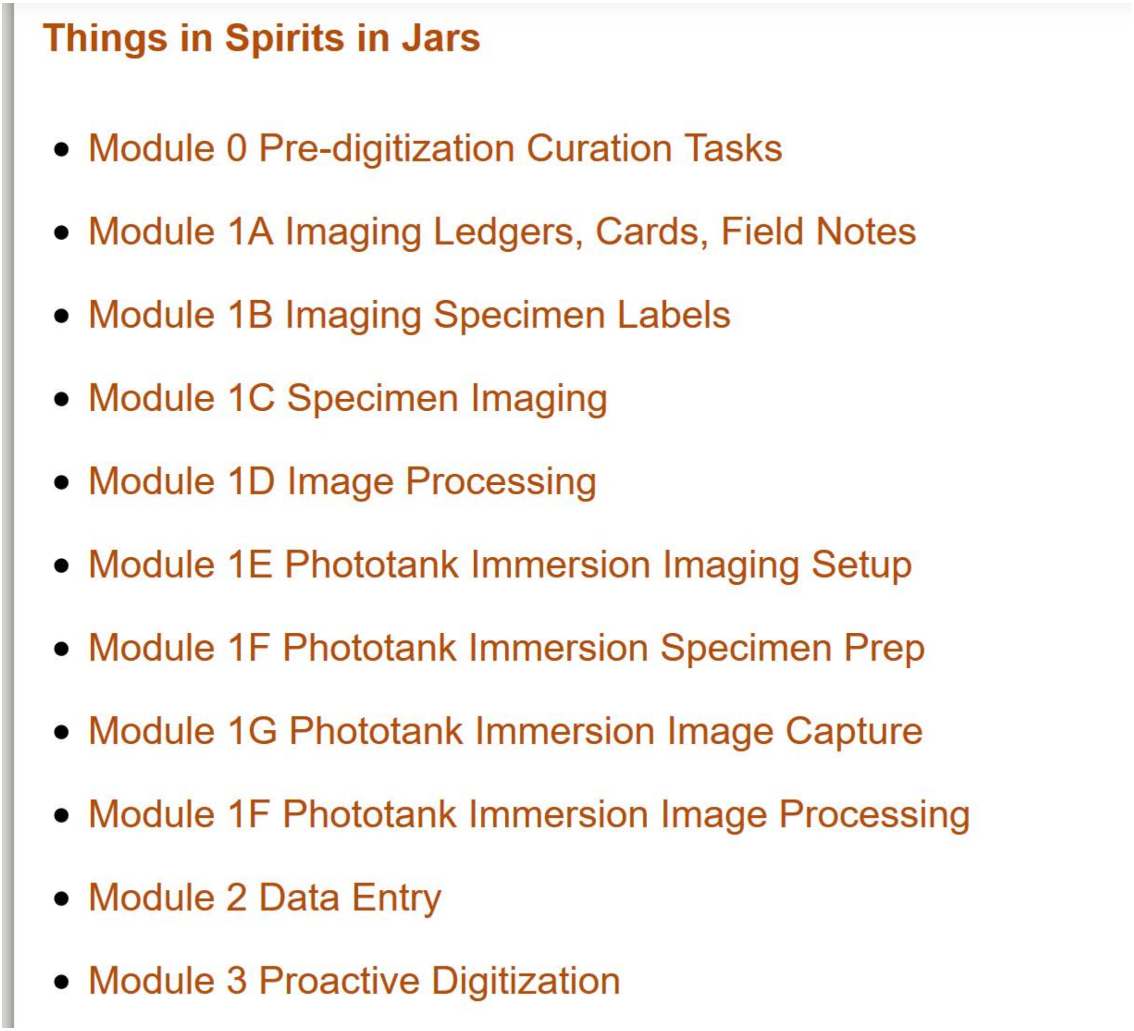


GEOLocate Web Application

GEOLocate

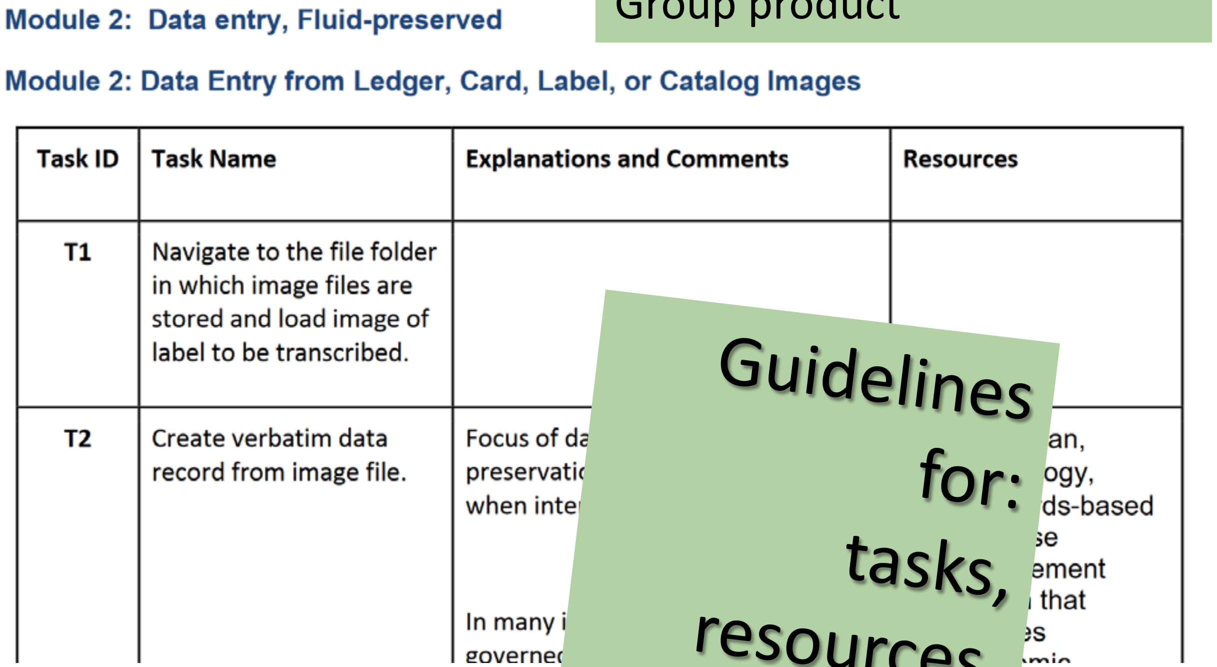
Wet Collections Tasks: what to consider?

iDigBio DROID3 Working Group product



Things in Spirits in Jars

- Module 0 Pre-digitization Curation Tasks
- Module 1A Imaging Ledgers, Cards, Field Notes
- Module 1B Imaging Specimen Labels
- Module 1C Specimen Imaging
- Module 1D Image Processing
- Module 1E Phototank Immersion Imaging Setup
- Module 1F Phototank Immersion Specimen Prep
- Module 1G Phototank Immersion Image Capture
- Module 1F Phototank Immersion Image Processing
- Module 2 Data Entry
- Module 3 Proactive Digitization



Module 2: Data entry, Fluid-preserved

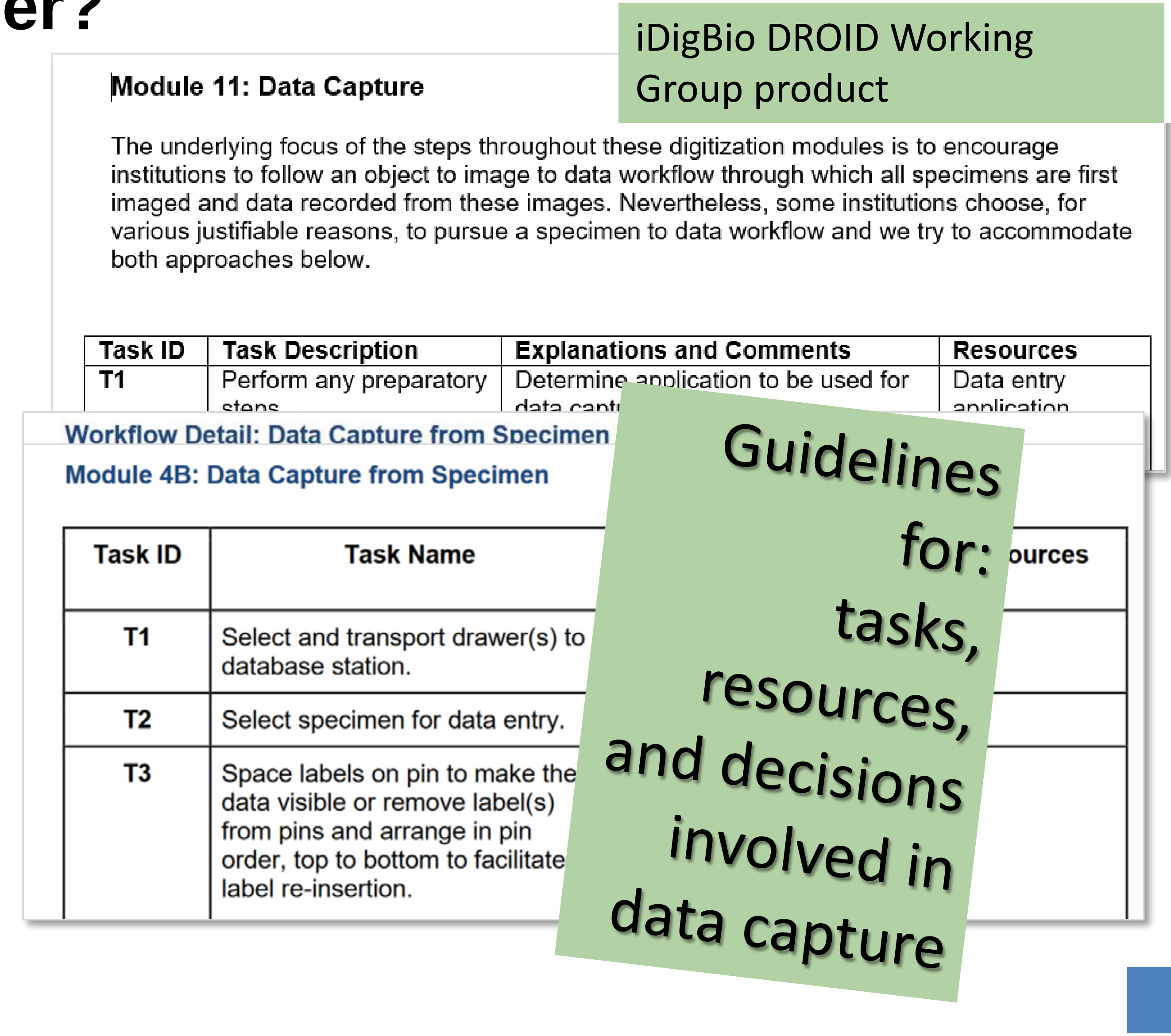
Module 2: Data Entry from Ledger, Card, Label, or Catalog Images

Task ID	Task Name	Explanations and Comments	Resources
T1	Navigate to the file folder in which image files are stored and load image of label to be transcribed.		
T2	Create verbatim data record from image file.	Focus of da preservatio when inte In many i governe	an, ogy, ds-based se ement that es mic

Guidelines for: tasks, resources, and decisions involved in digitizing wet collections

Data Capture: what to consider?

- data from image or data from label
- identifier for the object
  - local to global
  - never reuse
- how much data to capture?
  - all or some?
- is there useful existing digitized data?
  - taxonomy, geography, collector names
- do you have the database fields you need?
  - where to put the data
  - verbatim and / or interpreted
- do you have to move the specimens?
  - can you take the data capture to the shelves?



iDigBio DROID Working Group product

Module 11: Data Capture

The underlying focus of the steps throughout these digitization modules is to encourage institutions to follow an object to image to data workflow through which all specimens are first imaged and data recorded from these images. Nevertheless, some institutions choose, for various justifiable reasons, to pursue a specimen to data workflow and we try to accommodate both approaches below.

Task ID	Task Description	Explanations and Comments	Resources
T1	Perform any preparatory steps	Determine application to be used for data captu	Data entry application

Workflow Detail: Data Capture from Specimen

Module 4B: Data Capture from Specimen

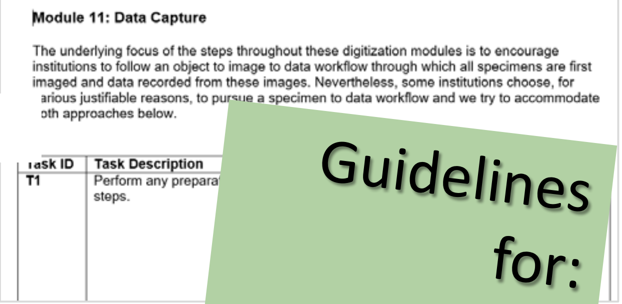
Task ID	Task Name	ources
T1	Select and transport drawer(s) to database station.	
T2	Select specimen for data entry.	
T3	Space labels on pin to make the data visible or remove label(s) from pins and arrange in pin order, top to bottom to facilitate label re-insertion.	

Guidelines for: tasks, resources, and decisions involved in data capture

Data Capture: what to consider? Part 2

- transcription issues
  - parsing (what goes in which field), implicit values
  - text missing from authority file
- data quality checks
  - transcription errors, erroneous information on labels
  - human and automated checks
- written protocols
  - iterative improvements, updates when equipment or software changes

iDigBio DROID Working Group product



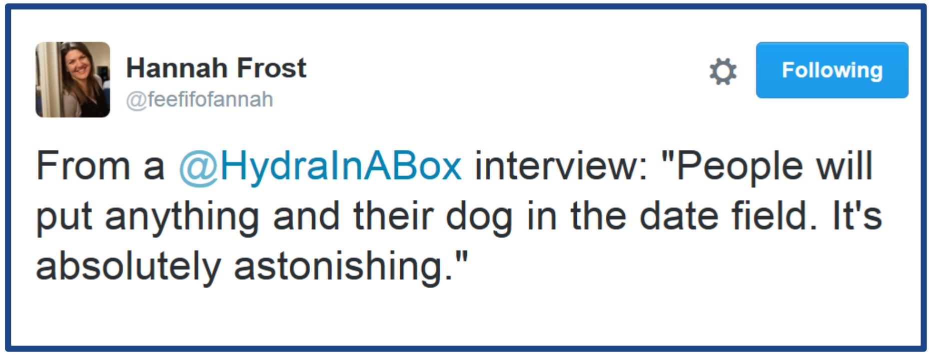
Module 11: Data Capture

The underlying focus of the steps throughout these digitization modules is to encourage institutions to follow an object to image to data workflow through which all specimens are first imaged and data recorded from these images. Nevertheless, some institutions choose, for arious justifiable reasons, to pursue a specimen to data workflow and we try to accommodate oth approaches below.

ask ID	Task Description
T1	Perform any prepara steps.

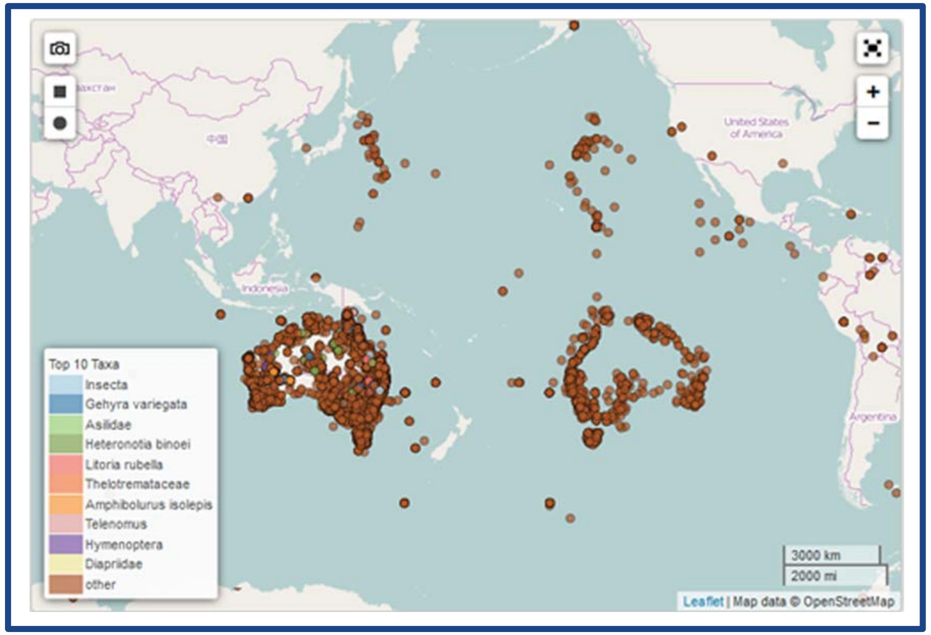
Guidelines for: tasks, resources, and decisions involved in data capture

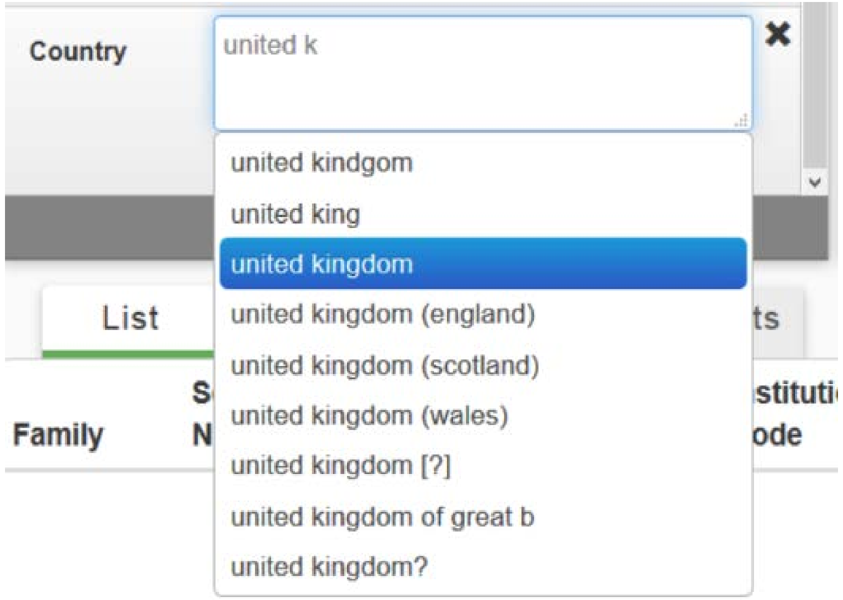
Data quality: an issue at many levels



Hannah Frost
@feefifofannah

From a @HydraInABox interview: "People will put anything and their dog in the date field. It's absolutely astonishing."





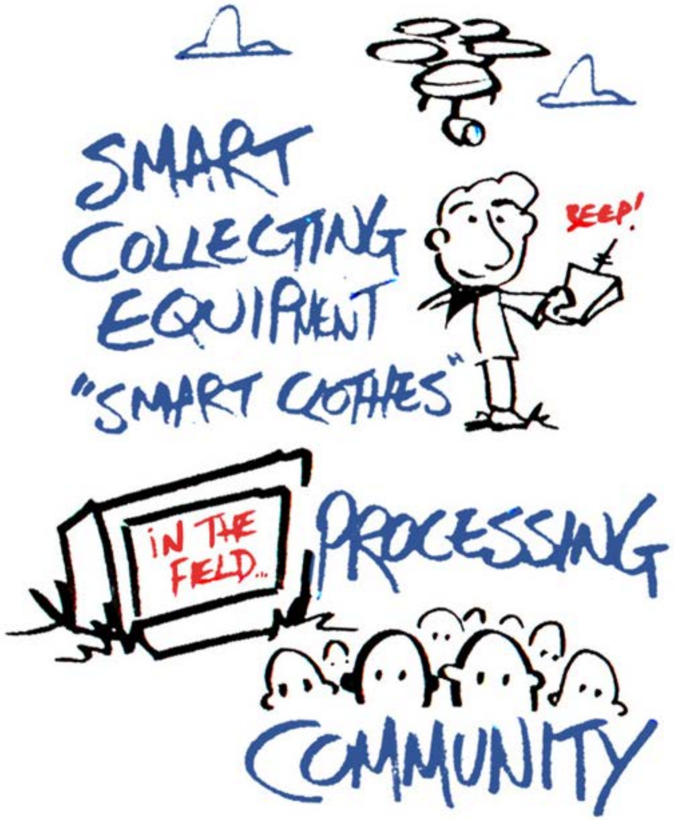
Country: united k

- united kindgom
- united king
- united kingdom
- united kingdom (england)
- united kingdom (scotland)
- united kingdom (wales)
- united kingdom [?]
- united kingdom of great b
- united kingdom?

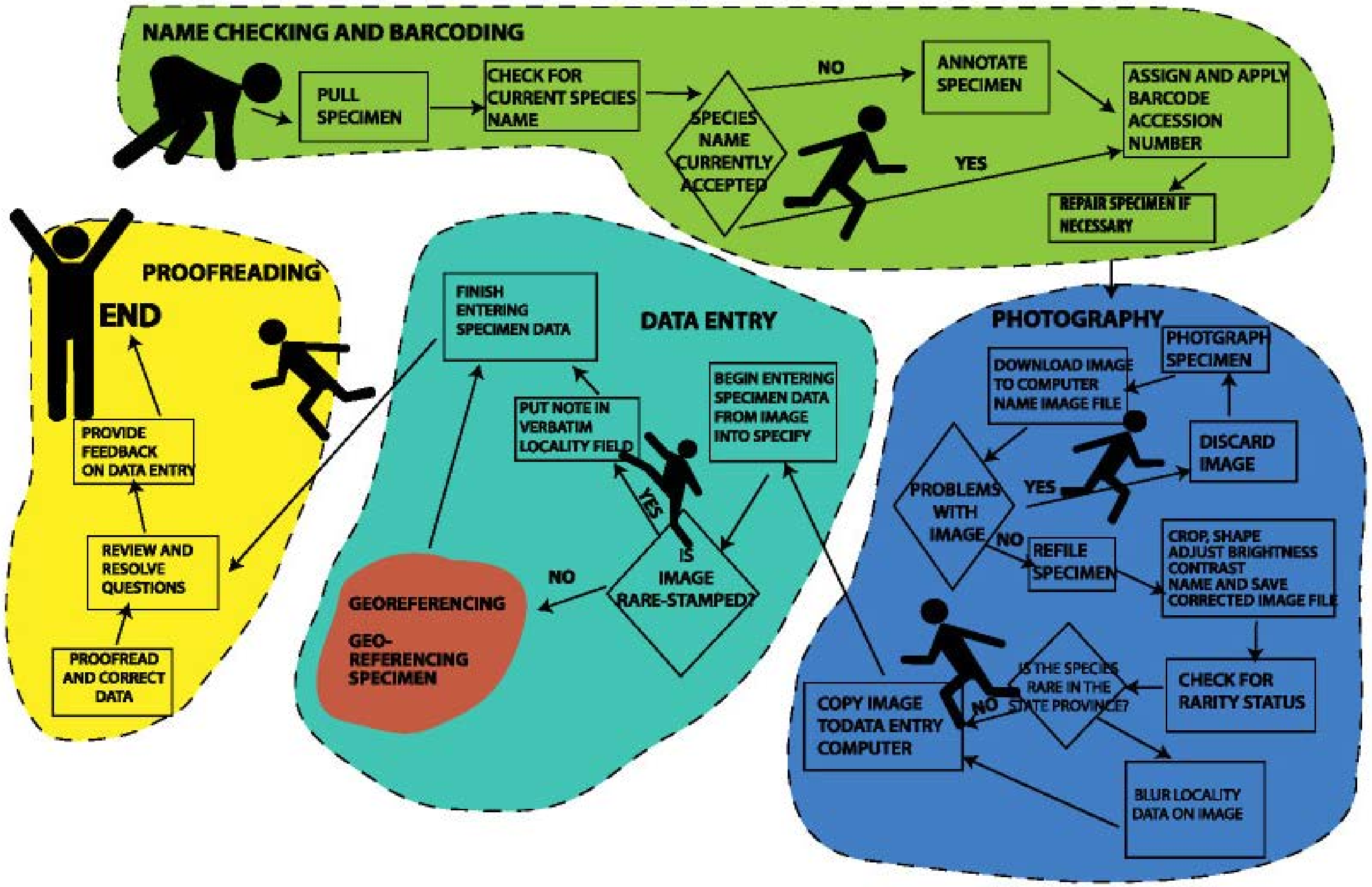
196 Countries in the world, but 1100 distinct values in the country field

Do-er happiness for *productivity and data quality*

From the field ➡ *born digital*



By @GoodStickLtd
Cisco Pitstop 2016


NAME CHECKING AND BARCODING
PULL SPECIMEN
CHECK FOR CURRENT SPECIES NAME
SPECIES NAME CURRENTLY ACCEPTED
NO
ANNOTATE SPECIMEN
YES
ASSIGN AND APPLY BARCODE ACCESSION NUMBER
REPAIR SPECIMEN IF NECESSARY
PHOTOGRAPHY
PHOTGRAPH SPECIMEN
DOWNLOAD IMAGE TO COMPUTER NAME IMAGE FILE
PROBLEMS WITH IMAGE
YES
DISCARD IMAGE
NO
REFILE SPECIMEN
CROP, SHAPE ADJUST BRIGHTNESS CONTRAST NAME AND SAVE CORRECTED IMAGE FILE
CHECK FOR RARITY STATUS
IS THE SPECIES RARE IN THE STATE PROVINCE?
NO
COPY IMAGE TODATA ENTRY COMPUTER
BLUR LOCALITY DATA ON IMAGE
DATA ENTRY
BEGIN ENTERING SPECIMEN DATA FROM IMAGE INTO SPECIFY
IS IMAGE RARE-STAMPED?
YES
PUT NOTE IN VERBATIM LOCALITY FIELD
NO
GEOREFERENCING
GEO-REFERENCING SPECIMEN
FINISH ENTERING SPECIMEN DATA
PROOFREADING
PROOFREAD AND CORRECT DATA
REVIEW AND RESOLVE QUESTIONS
PROVIDE FEEDBACK ON DATA ENTRY
END

By Dorothy Allard

Thanks – insights?

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 grants from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program [DBI-1115210 (2011-2018) and DBI-1547229 (2016-2021)]. 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.