Data capture: aka data digitization

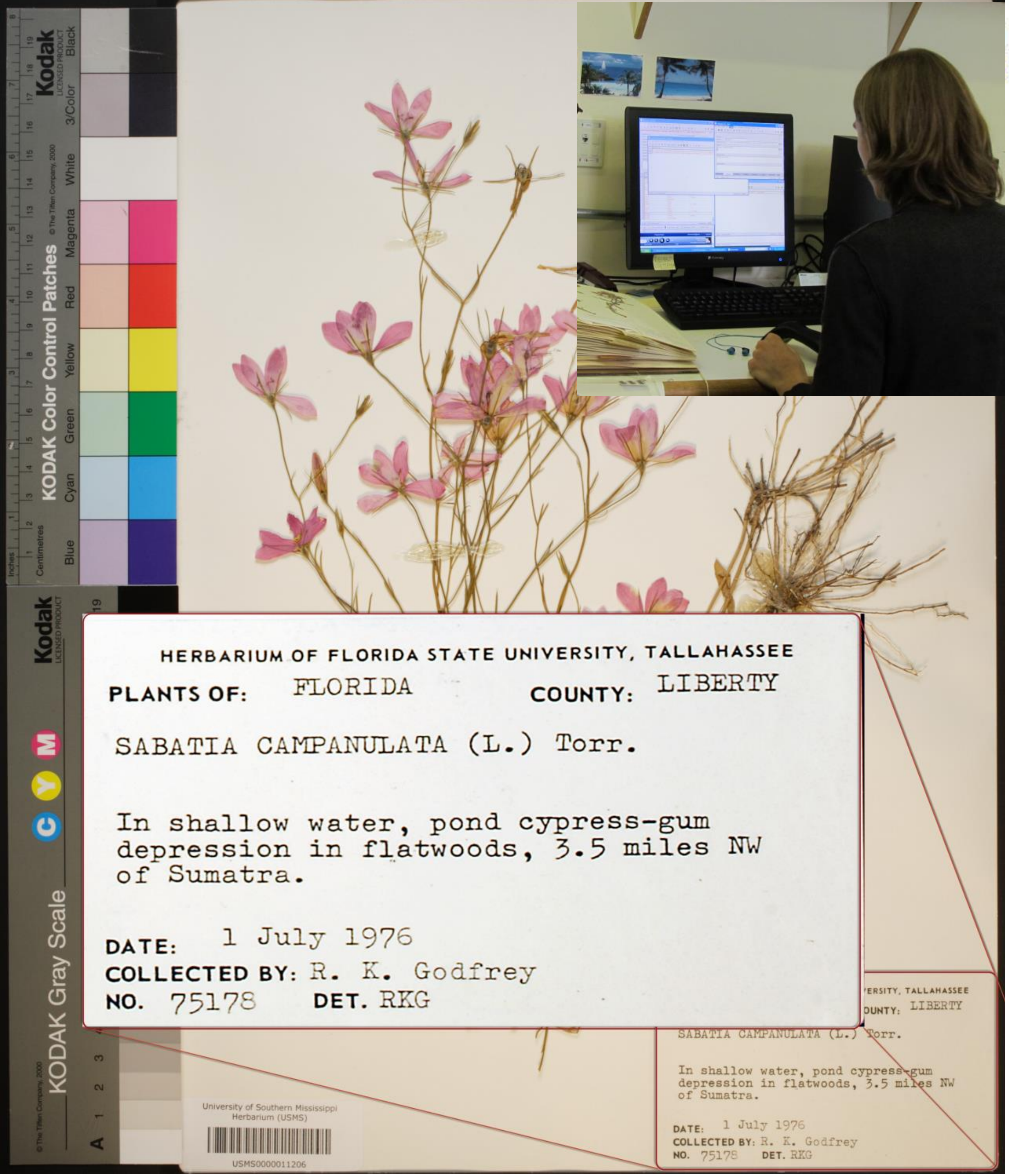
some best practices, options, lessons learned

Deborah Paul
iDigBio, Florida State University
Bristol UK Swans Practical Digitisation Workshop
9 March 2018
@idbdeb @iDigBio

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.

Goals of data capture

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



HERBARIUM OF FLORIDA STATE UNIVERSITY, TALLAHASSEE

PLANTS OF: FLORIDA COUNTY: LIBERTY

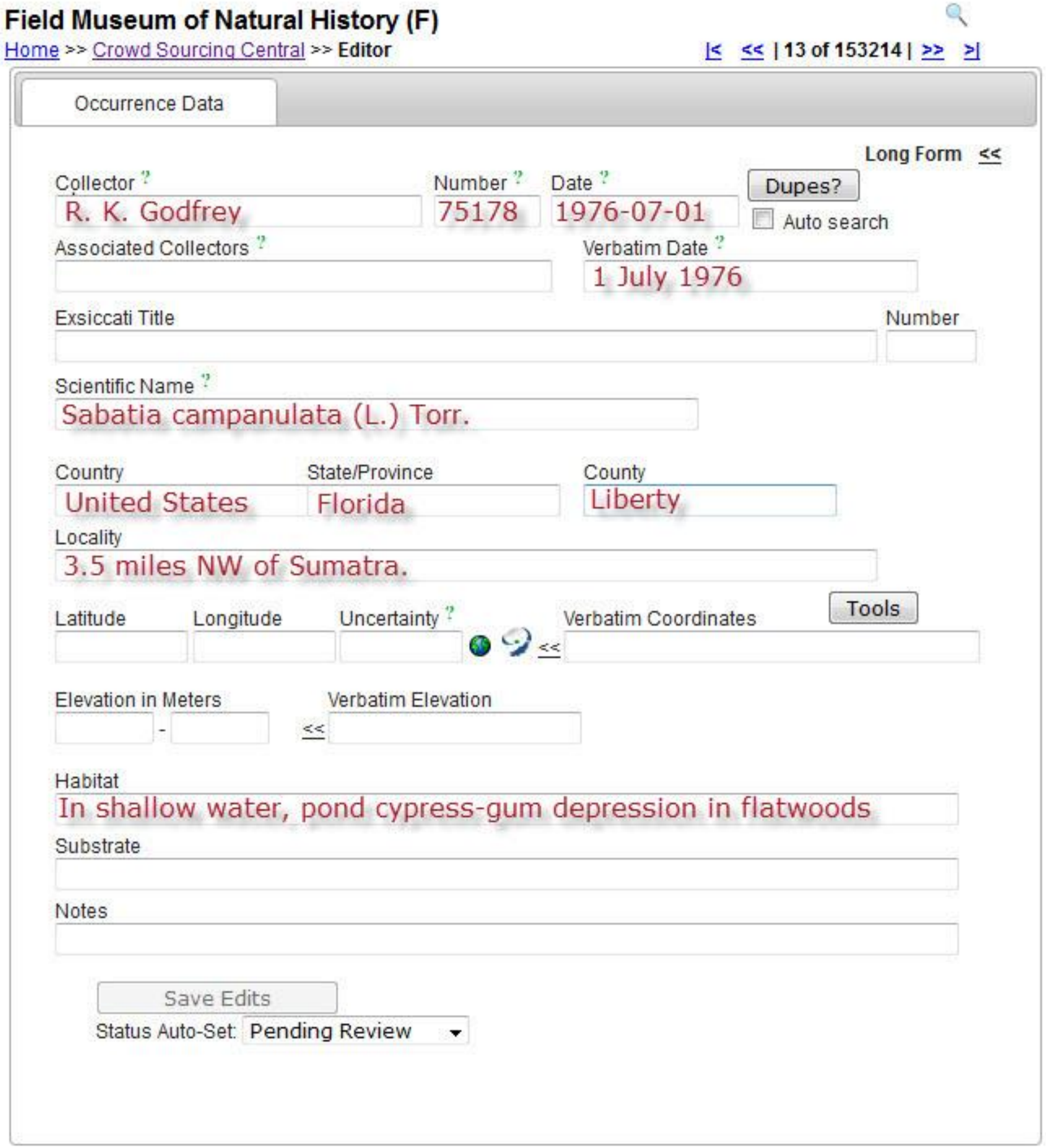
SABATIA CAMPANULATA (L.) Torr.

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

University of Southern Mississippi Herbarium (USMS)
USMS0000011206



Field Museum of Natural History (F)

Home >> Crowd Sourcing Central >> Editor

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

Occurrence Data

Long Form <<

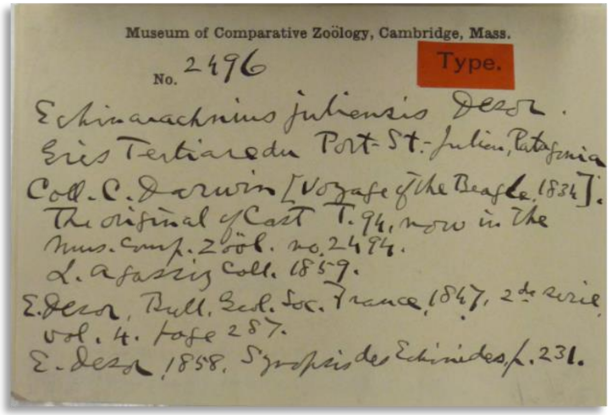
Collector: R. K. Godfrey
Number: 75178
Date: 1976-07-01
Dupes?
□ Auto search
Associated Collectors:
Verbatim Date: 1 July 1976
Exsiccati Title:
Number:
Scientific Name: Sabatia campanulata (L.) Torr.
Country: United States
State/Province: Florida
County: Liberty
Locality: 3.5 miles NW of Sumatra.
Latitude: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

Save Edits

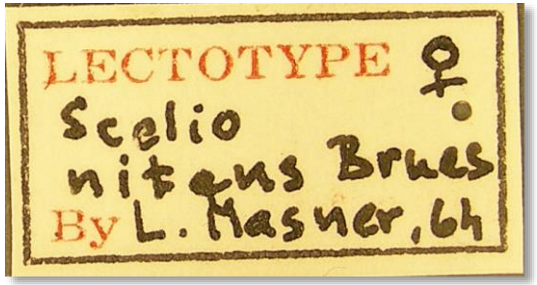
Status Auto-Set: Pending Review

Data Capture Challenges

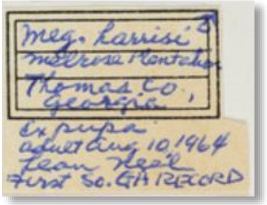
- ink
- typed
- pencil
- printed
- stacked
- handwritten
- uneven lines
- colored paper
- non-planar surfaces
- non-standard terms
- non-standard formats



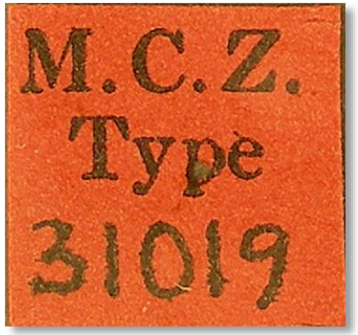
Museum of Comparative Zoölogy, Cambridge, Mass.
No. 2496 Type.
Echinarachnius juliensis Desor.
Grès Tertiare du Port-St.-Julien, Patagonia
Coll. C. Darwin [Voyage of the Beagle, 1834].
The original of Cast T. 94, now in the Mus. Comp. Zoöl. no. 2494.
L. Agassiz coll. 1859.
E. Desor, Bull. Geol. Soc. France, 1847, 2de serie, vol. 4. page 287.
E. Desor, 1858. Synopsis des Echinides, p. 231.



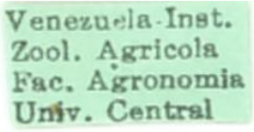
LECTOTYPE ♀
Scelio
nitens Brues
By L. Masner, 64



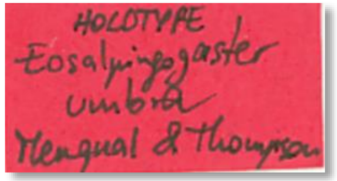
meg. harrisi
melkrose Montalu
Thomas Co., Georgia
ex pupa
adult Aug 10, 1964
Jean Neal
First So. GA RECORD



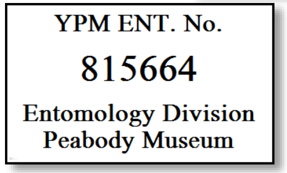
M.C.Z.
Type
31019



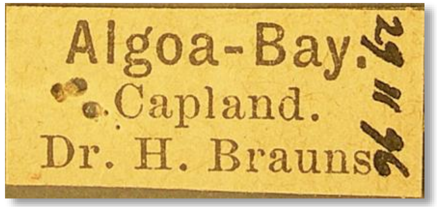
Venezuela-Inst.
Zool. Agricola
Fac. Agronomia
Univ. Central



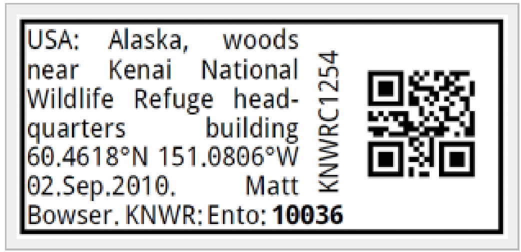
HOLOTYPE
Eosalpingogaster
umbra
Mengual & Thompson



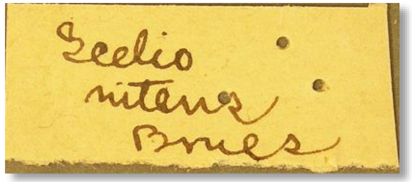
YPM ENT. No.
815664
Entomology Division
Peabody Museum



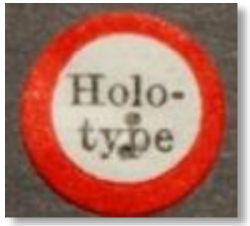
Algoa-Bay.
Capland.
Dr. H. Brauns
29 ॥ 96



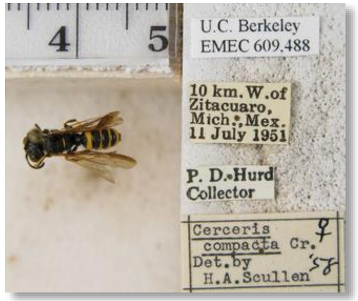
USA: Alaska, woods near Kenai National Wildlife Refuge headquarters building
60.4618°N 151.0806°W
02.Sep.2010. Matt Bowser. KNWR:Ento:10036
KNWRC1254



Scelio
nitens
Brues



Holo-
type

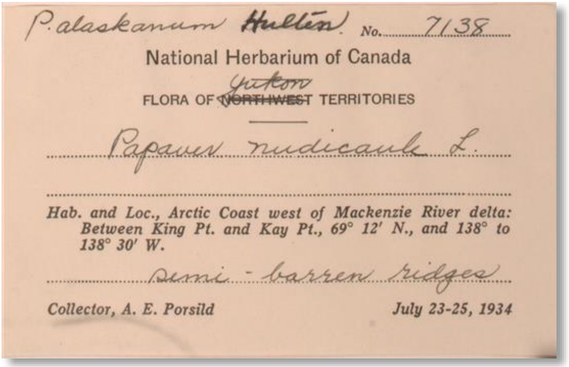


U.C. Berkeley
EMEC 609.488

10 km. W. of
Zitacuaro,
Mich., Mex.
11 July 1951

P. D. Hurd
Collector

Cerceris ♀
compacta Cr.
Det. by '58
H.A. Scullen



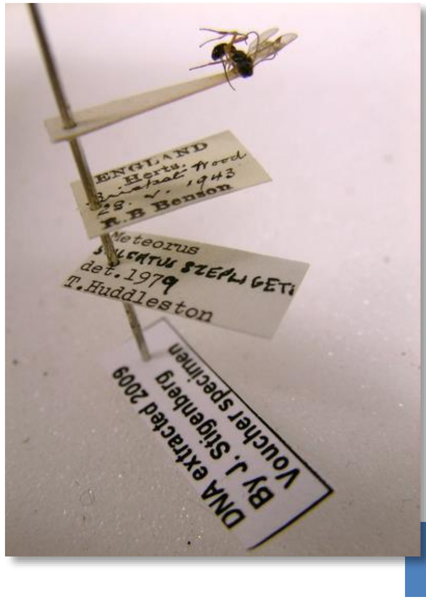
P. alaskanum Hultén No. 7138
National Herbarium of Canada
FLORA OF Yukon ~~NORTHWEST~~ TERRITORIES

Papaver nudicaule L.

Hab. and Loc., Arctic Coast west of Mackenzie River delta: Between King Pt. and Kay Pt., 69° 12′ N., and 138° to 138° 30′ W.

semi-barren ridges

Collector, A. E. Porsild July 23-25, 1934



ENGLAND
Herts: Wood
Bristol
28. V. 1943
R.B Benson

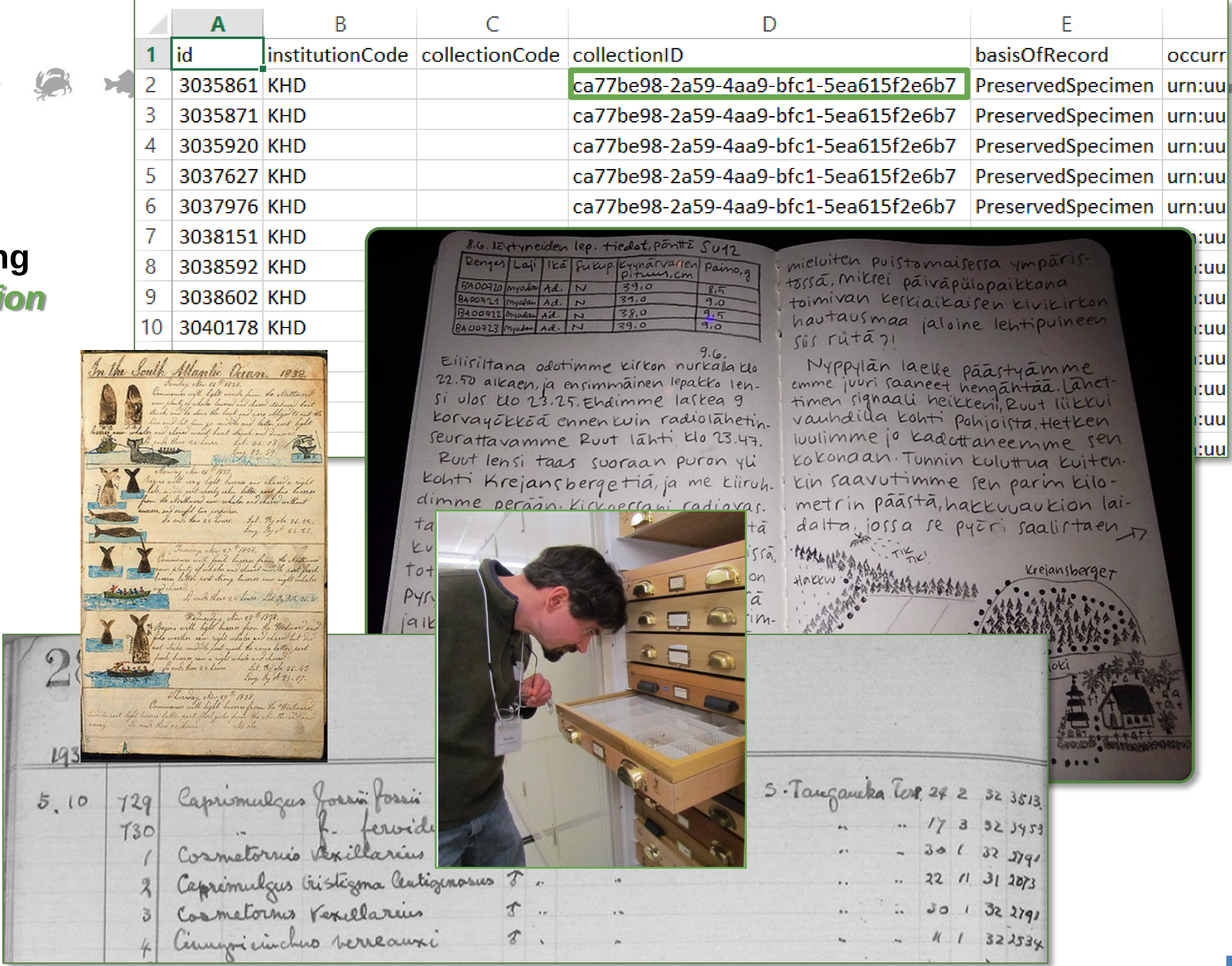
Meteorus
sulcatus Szepligeti
det. 1979
T.Huddleston

DNA extracted 2009
By J. Stigenberg
Voucher specimen

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-bfc1-5ea615f2e6b7	PreservedSpecimen	urn:uu
6	3037976	KHD		ca77be98-2a59-4aa9-bfc1-5ea615f2e6b7	PreservedSpecimen	urn:uu
7	3038151	KHD				:uu
8	3038592	KHD				:uu
9	3038602	KHD				:uu
10	3040178	KHD				:uu

8.6. löytyneiden lep. tiedot, pöntti SU12

Rengas	Laji	Ikä	Sukup.	Kyynärvarren pituus, cm	Paino, g
BA00720	myodau	Ad.	N	39,0	8,5
BA00721	myodau	Ad.	N	39,0	9,0
BA00722	myodau	Ad.	N	38,0	9,5
BA00723	myodau	Ad.	N	39,0	9,0

9.6.

Eilisiltana odotimme kirkon nurkalla klo 22.50 alkaen, ja ensimmäinen lepakko lensi ulos klo 23.25. Ehdimme laskea 9 korvayökköä ennenkuin radiolähetin-seurattavamme Ruut lähti klo 23.47. Ruut lensi taas suoraan puron yli kohti Krejansbergetiä, ja me kiiruhdimme perään. Kiskoessani radiovas-

ta ... tä
ku ... issä,
tot ... on
pyr ... ä
jalk ... im-

mieluiten puistomaisessa ympäristössä. Mikrei päiväpiilopaikkana toimivan keskiaikaisen kivikirkon hautausmaa jaloine lehtipuineen siis riitä?!

Nyppylän laelle päästyämme emme juuri saaneet hengähtää. Lähettimen signaali heikkeni, Ruut liikkui vauhdilla kohti pohjoista. Hetken luulimme jo kadottaneemme sen kokonaan. Tunnin kuluttua kuitenkin saavutimme sen parin kilometrin päästä, hakkuuaukion laidalta, jossa se pyöri saalistaen →

In the South Atlantic Ocean. 1838.

Sunday Nov 25th 1838.
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 drawed.
So ends these 24 hours. Lat. 36. 09 Long. 25. 59.

Monday Nov 26th 1838.
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. 32 Long. By [illegible] 24. 25.

Tuesday Nov 27th 1838.
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. [illegible]

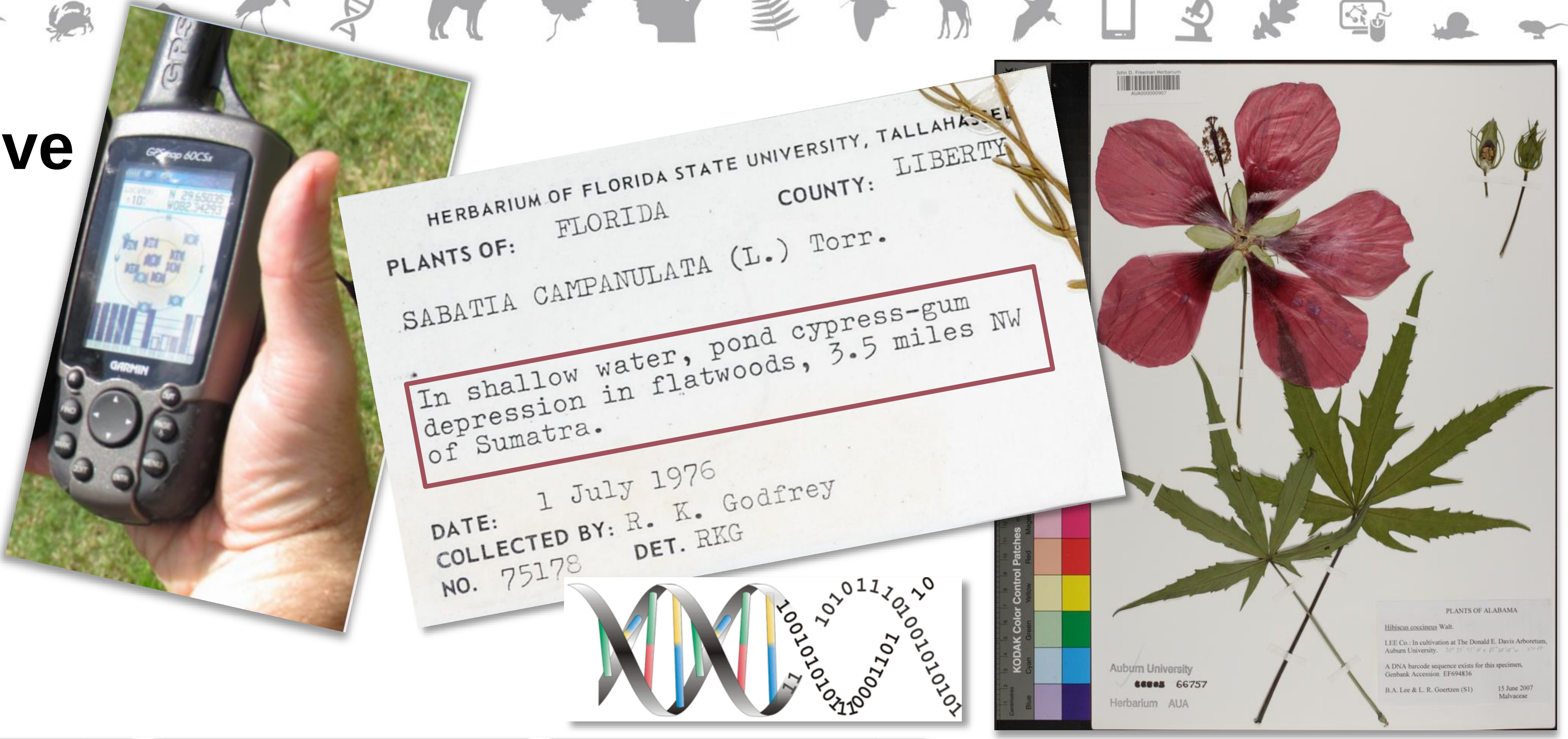
Wednesday Nov 28th 1838.
Begins with light breezes from the Westward and fine weather saw right whales and chased but did not strike middle part much the same latter part fresh breezes saw a right whale and chased.
So ends these 24 hours. Lat. By obs 35. 43 Long. By [illegible] 24. 07.

Thursday Nov 29th 1838.
Commences with light breezes from the Westward middle part light breezes latter part hard gales from the Northward and rainy. So ends these 24 hours. No obs.

28
193
5. 10 | 729 | Caprimulgus fossii fossii | S. Tanganyika Terr. | 24 | 2 | 32 | 3513
730 | " f. fervida | .. .. | 17 | 3 | 32 | 3453
1 | Cosmetornis vexillarius | .. .. | 30 | 1 | 32 | 2791
2 | Caprimulgus bistigma lentiginosus | ♂ | .. | .. .. | 22 | 11 | 31 | 2073
3 | Cosmetornis vexillarius | ♂ | .. | .. .. | 30 | 1 | 32 | 2791
4 | Cinnyricinclus verreauxi | ♂ | . | .. .. | 11 | 1 | 32 | 2534

Extract and Derive

- Geolocation
- Phenology
- Habitat
- Ecology
- Morphology
- Stratigraphy
- 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

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

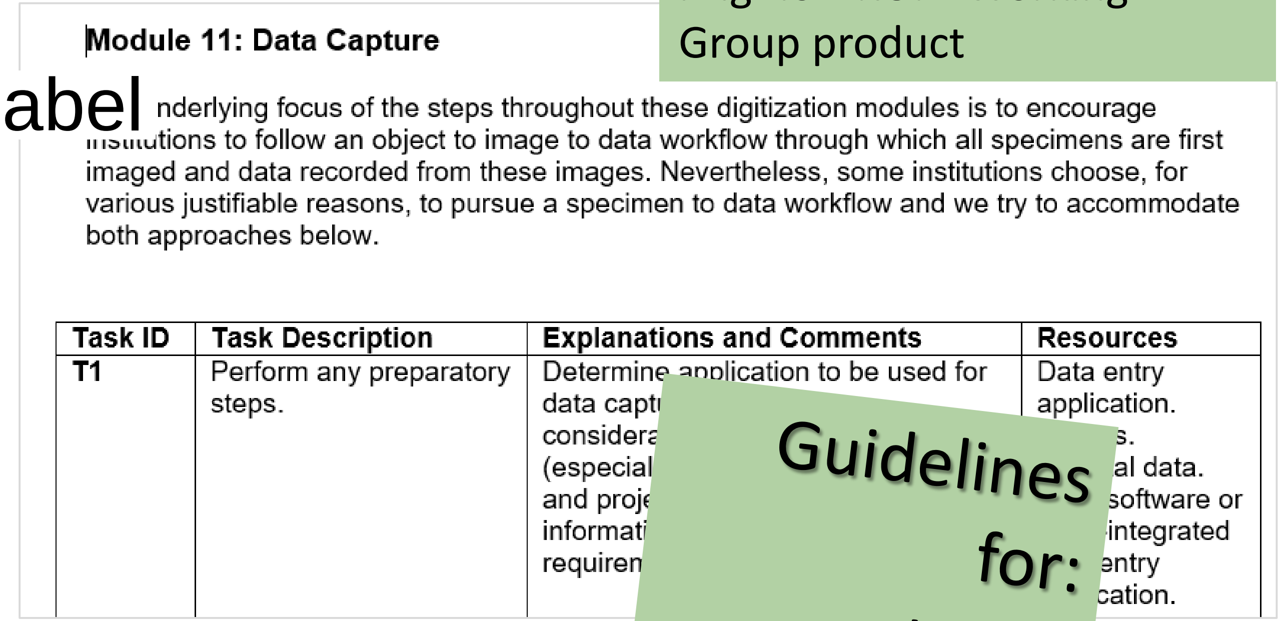
Epistenia coeruleata, Westwood, 1832, 1991-05-14

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

GEOLocate Web Application
GEOLocate
United States
Mexico
Workbench
Results
Select File
or load an existing file using a retrieval code:
Load

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



Module 11: Data Capture

nderlying 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 considera (especial and proje informati requirem	Data entry application. s. al data. software or integrated entry cation.

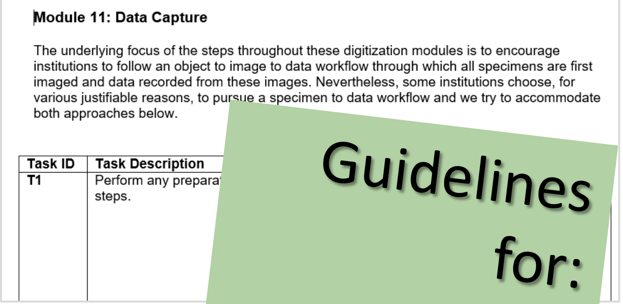
iDigBio DROID Working Group product

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



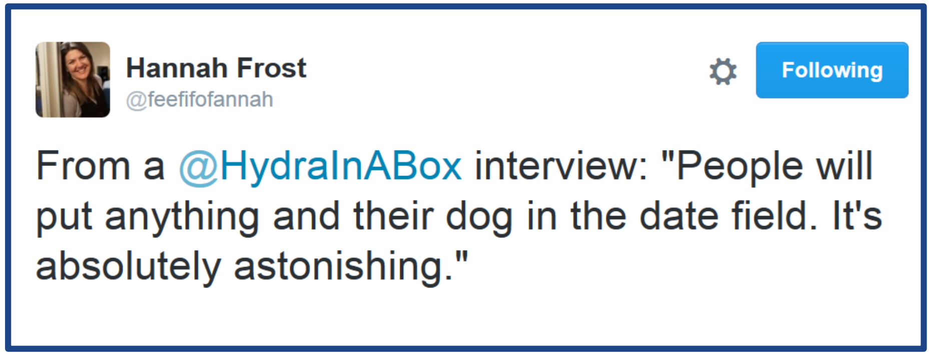
Module 11: Data Capture

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 various justifiable reasons, to pursue a specimen to data workflow and we try to accommodate both approaches below.

Task ID	Task Description
T1	Perform any prepara steps.

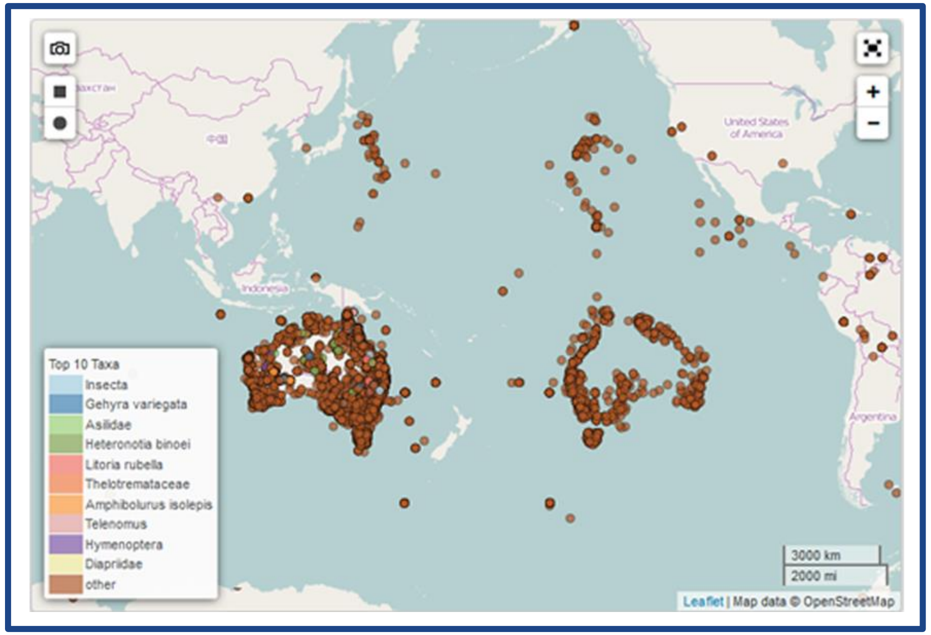
Guidelines for: tasks, resources, and decisions involved in data capture

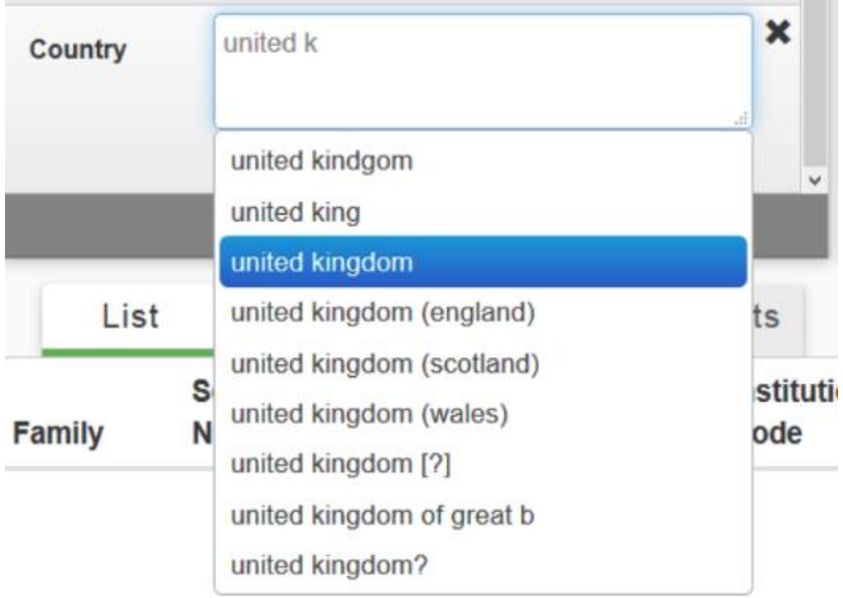
Data quality: an issue at many levels



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

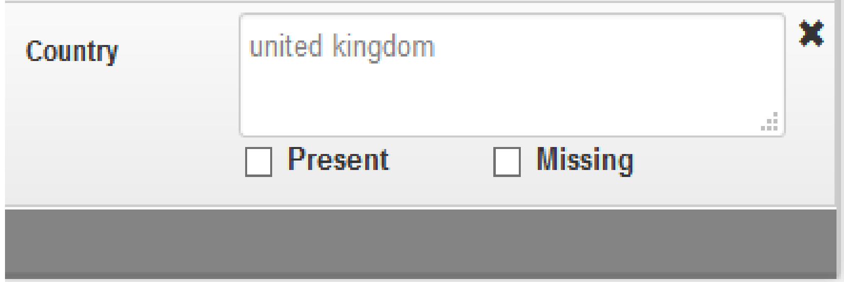
united kindgom
united king
united kingdom
united kingdom (england)
united kingdom (scotland)
united kingdom (wales)
united kingdom [?]
united kingdom of great b
united kingdom?

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

Data cleaning

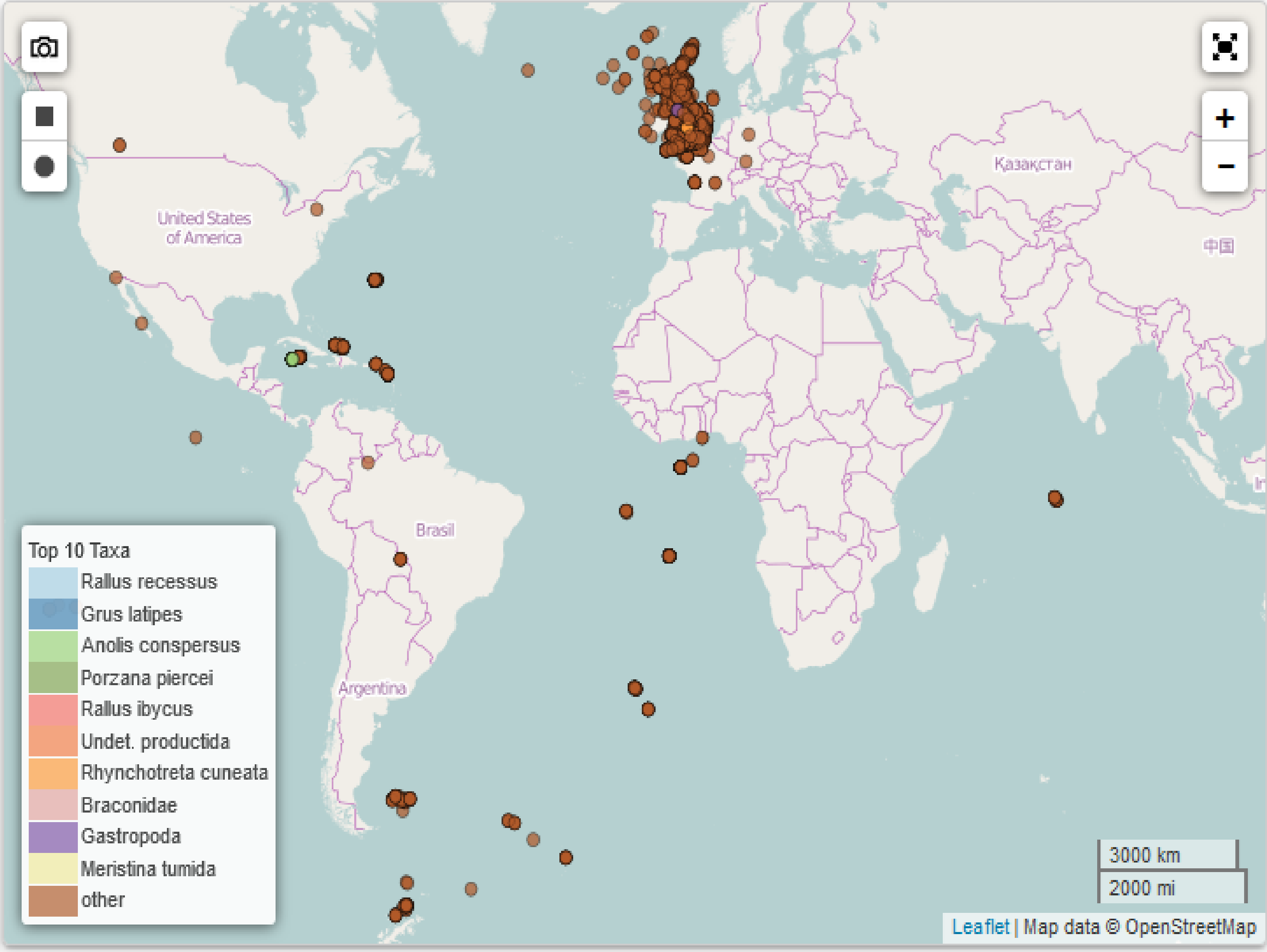
country=
“united kingdom”

Or does it?



Country: united kingdom

☐ Present ☐ Missing



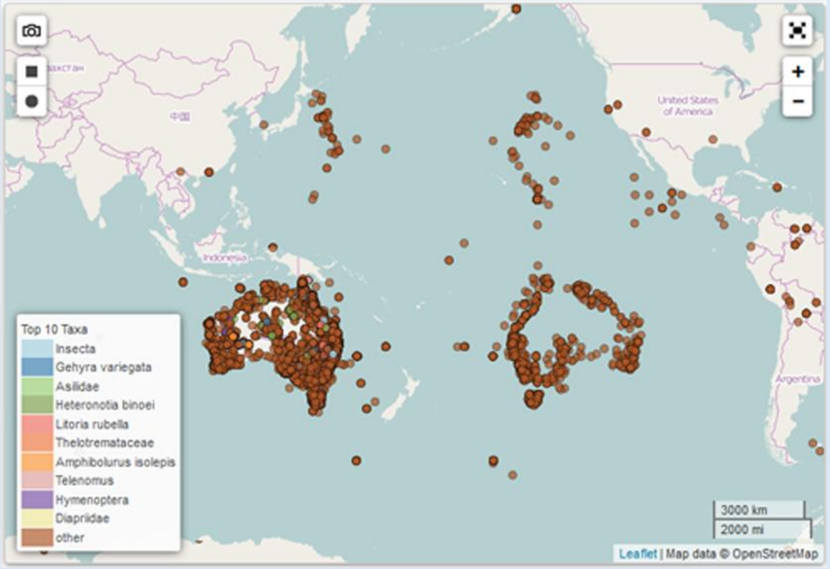
Data Quality:
Grooming and tics (adapted from Joanna McCaffrey at iDigBio)

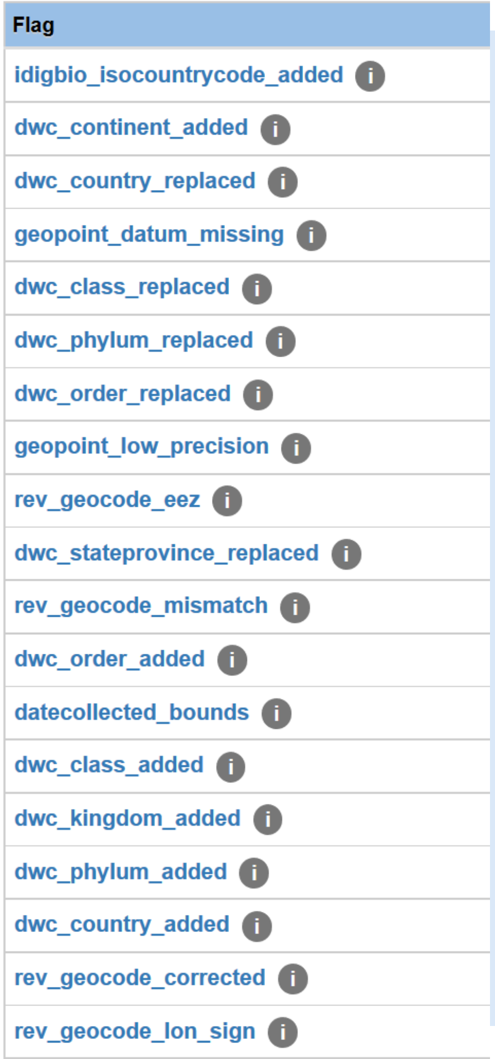
Your dataset **is no longer just for making labels**, there are other considerations for being digital, and out in the wild:

1. Put dates in ISO 8601 format, i.e., YYYY-MM-DD, e.g., 2015-09-17
2. Parse out scientific name
3. Conversely, put the piece parts into a scientific name
4. Provide as much higher taxonomy as your feel comfortable with, fill in tribe, sub+super family, kingdom, division, class, order) get out of ‘family’ land.
5. Make sure lat and lon coordinates are in degrees decimal, and no N, S, E, W
6. Do not export '0' in fields to represent no value, e.g., lat or lon
7. Put elevation in METERS units in the elevation field without the units (e.g., the fields dwc:minimumElevationInMeters and dwc:maximumElevationInMeters already assume the numeric values are in meters, so there no need to include the units with the data)
8. And not to get too esoteric, do not use un-escaped newline characters or embedded tabs
9. Watch out for diacritics (à á â ã ä å, save in UTF-8)

iDigBio Data Quality (DQ) Flags enhance Digitization Workflows

Example: spot and fix georeferencing issues.





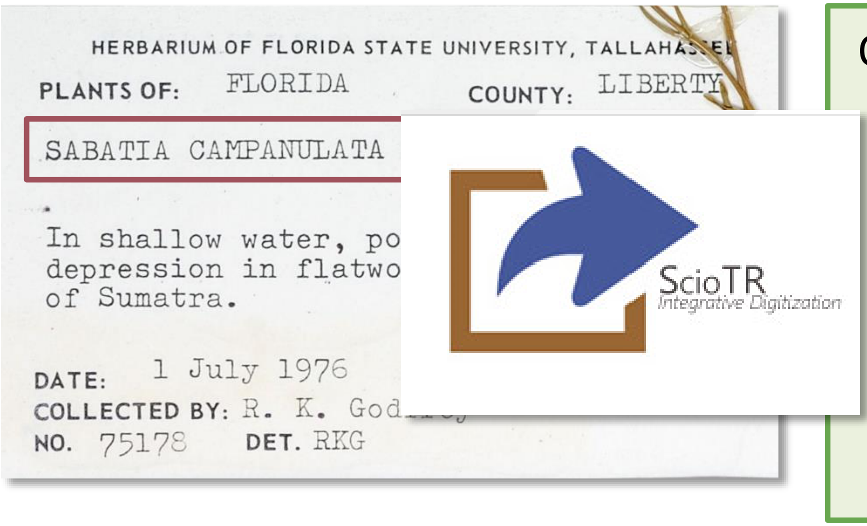
Flag
idigbio_isocountrycode_added ⓘ
dwc_continent_added ⓘ
dwc_country_replaced ⓘ
geopoint_datum_missing ⓘ
dwc_class_replaced ⓘ
dwc_phylum_replaced ⓘ
dwc_order_replaced ⓘ
geopoint_low_precision ⓘ
rev_geocode_eez ⓘ
dwc_stateprovince_replaced ⓘ
rev_geocode_mismatch ⓘ
dwc_order_added ⓘ
datecollected_bounds ⓘ
dwc_class_added ⓘ
dwc_kingdom_added ⓘ
dwc_phylum_added ⓘ
dwc_country_added ⓘ
rev_geocode_corrected ⓘ
rev_geocode_lon_sign ⓘ

- Value of data aggregation
- Planning for feedback integration

Some Newer Data Capture Options

- Voice technology

- Touch screen technology



HERBARIUM OF FLORIDA STATE UNIVERSITY, TALLAHASSEE

PLANTS OF: FLORIDA COUNTY: LIBERTY

SABATIA CAMPANULATA

In shallow water, po
depression in flatwo
of Sumatra.

DATE: 1 July 1976

COLLECTED BY: R. K. Godfrey

NO. 75178 DET. RKG

ScioTR
Integrative Digitization

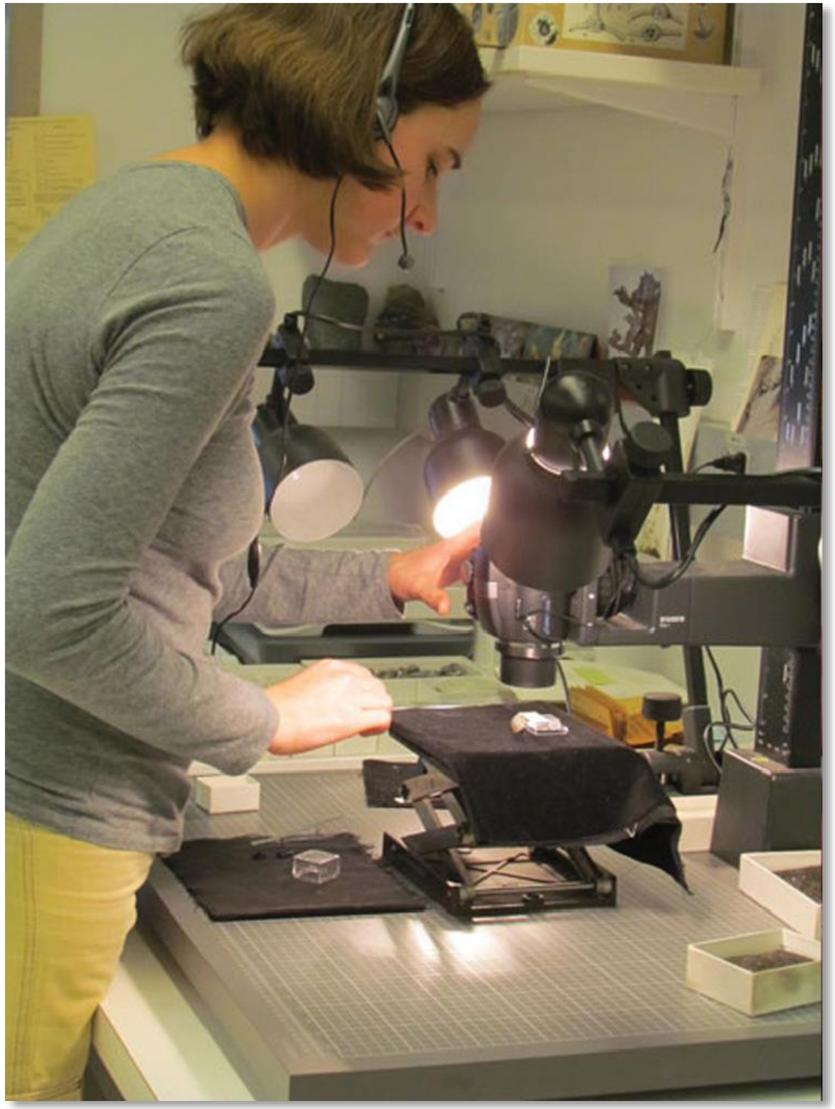
Occurrence (Verbatim)

Verbatim Institution
HERBARIUM OF FLORIDA STATE UNIVERSITY,

Catalog Number
75178

Verbatim Scientific Name
SABATIA CAMPANULATA (L.) Torr.

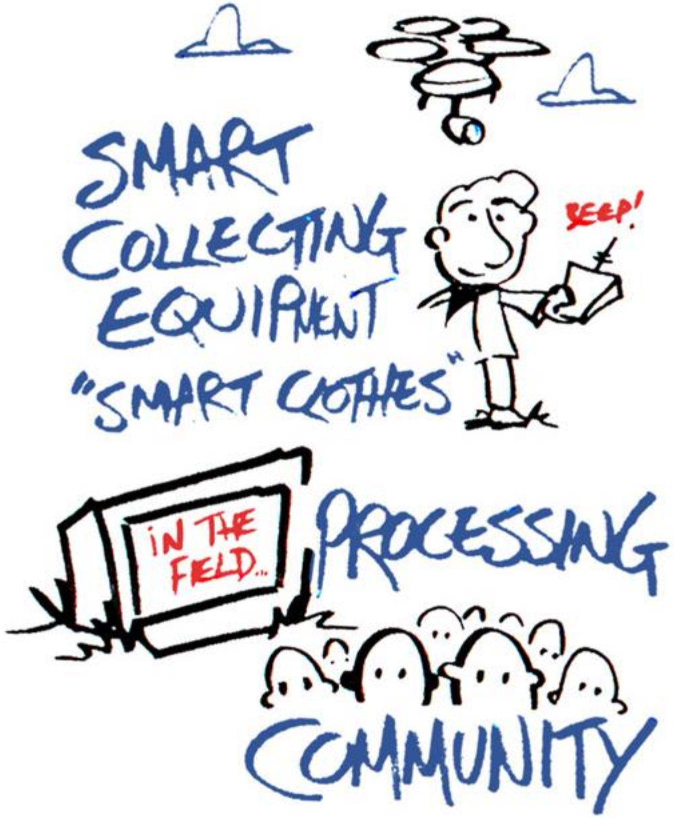
Recorded By…



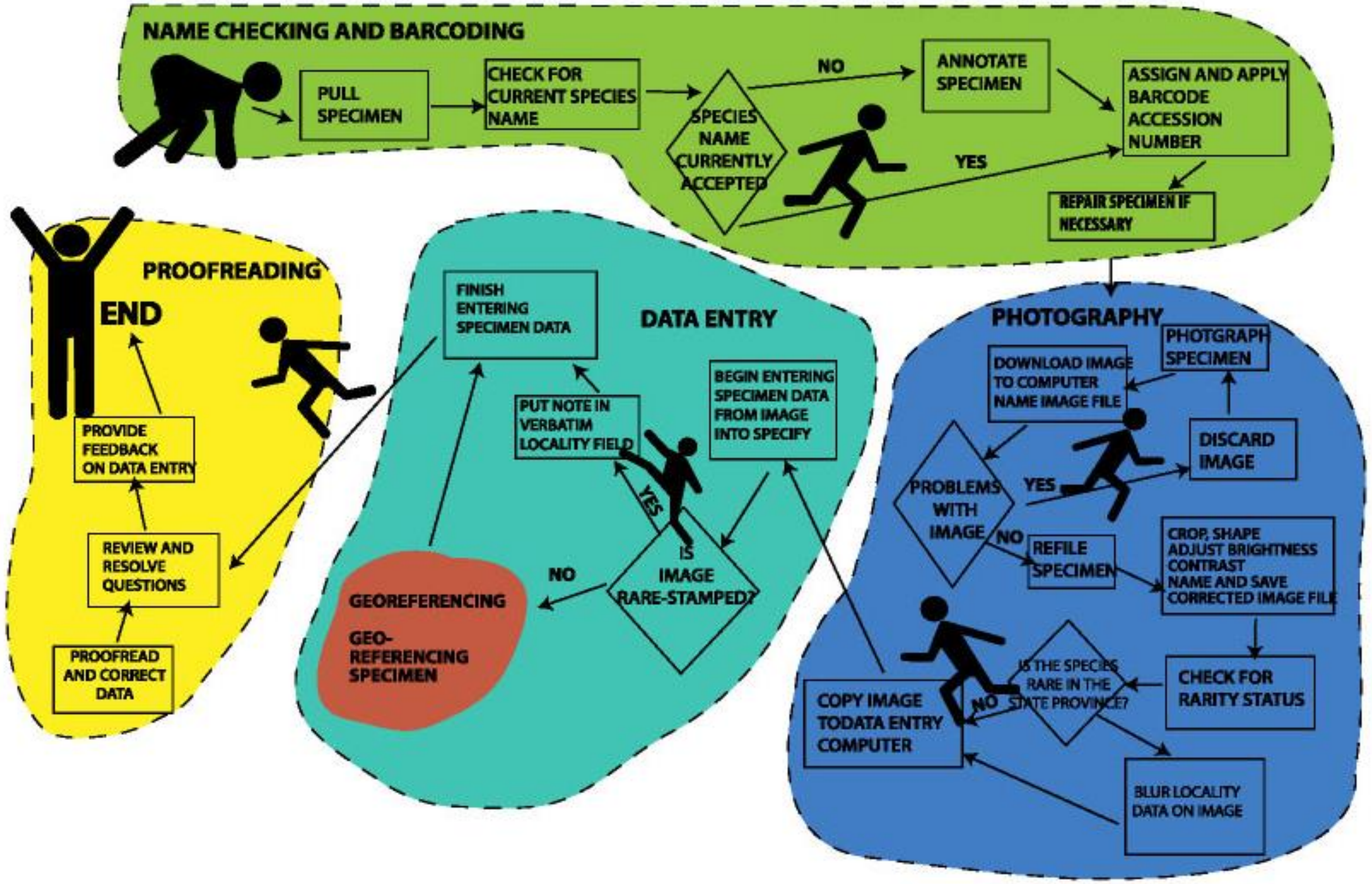
- OCR + ML + NLP

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

YES

ANNOTATE SPECIMEN

ASSIGN AND APPLY BARCODE ACCESSION NUMBER

REPAIR SPECIMEN IF NECESSARY

PHOTOGRAPHY

PHOTGRAPH SPECIMEN

DOWNLOAD IMAGE TO COMPUTER NAME IMAGE FILE

DISCARD IMAGE

PROBLEMS WITH IMAGE

YES

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

BLUR LOCALITY DATA ON IMAGE

COPY IMAGE TODATA ENTRY COMPUTER

DATA ENTRY

FINISH ENTERING SPECIMEN DATA

BEGIN ENTERING SPECIMEN DATA FROM IMAGE INTO SPECIFY

PUT NOTE IN VERBATIM LOCALITY FIELD

IS IMAGE RARE-STAMPED?

YES

NO

GEOREFERENCING

GEO-REFERENCING SPECIMEN

PROOFREADING

END

PROOFREAD AND CORRECT DATA

REVIEW AND RESOLVE QUESTIONS

PROVIDE FEEDBACK ON DATA ENTRY

By Dorothy Allard

Thanks – questions?

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. All images used with permission or are free from copyright.