Integrating Augmented OCR and Georeferencing in Natural History Collection (NHC) Digitization

Deborah L. Paul
dpaul@fsu.edu
Institute for Digital Information (FSU), iDigBio

SPNHC Symposium: Introduction to Digitization and Dissemination of Natural History Data: iDigBio, BISON and other initiatives

Thursday 20 June 2013

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.

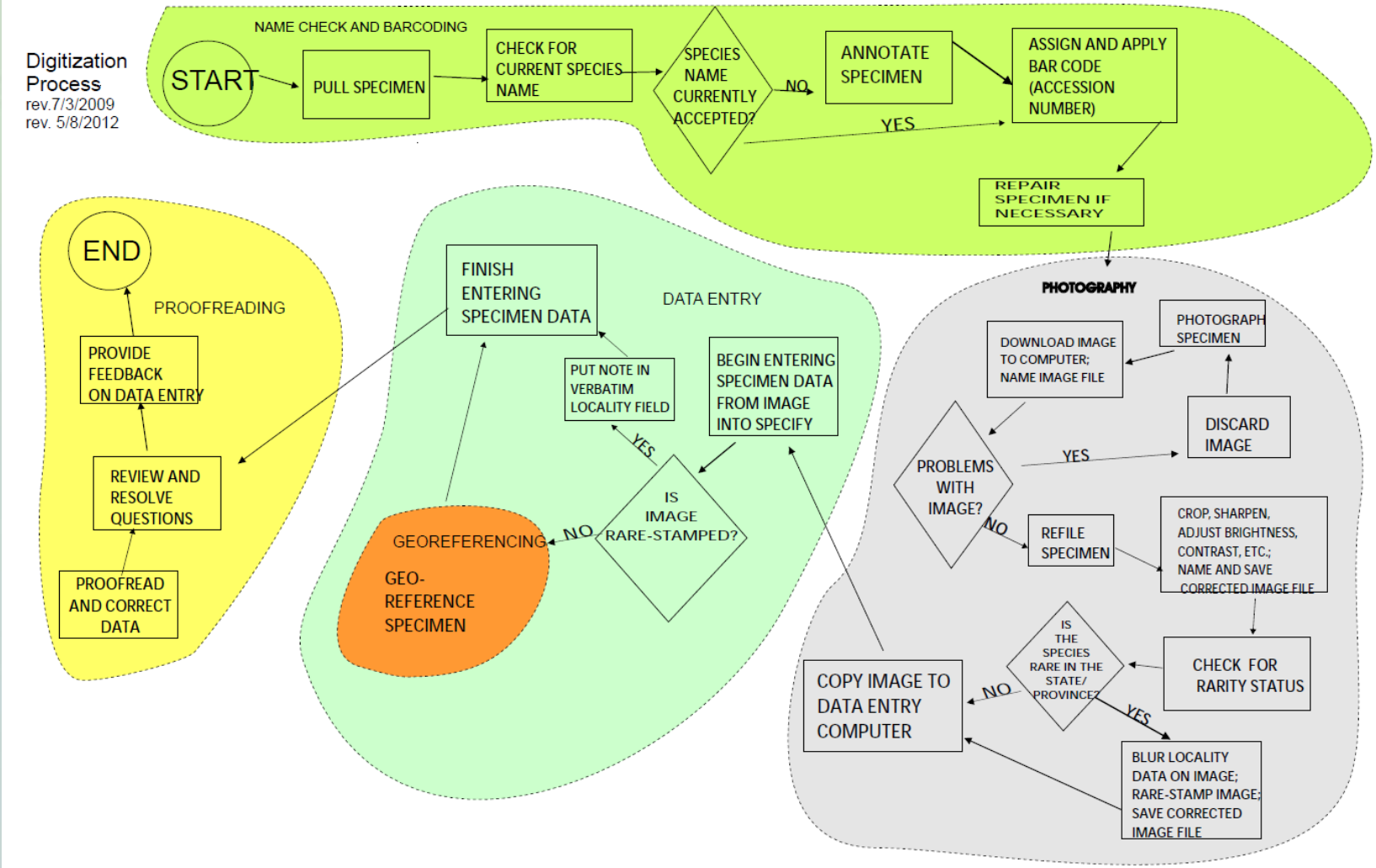
iDigBio
Integrated Digitized Biocollections

3,270,330 *Specimen Records*

391,312 *Media Records*

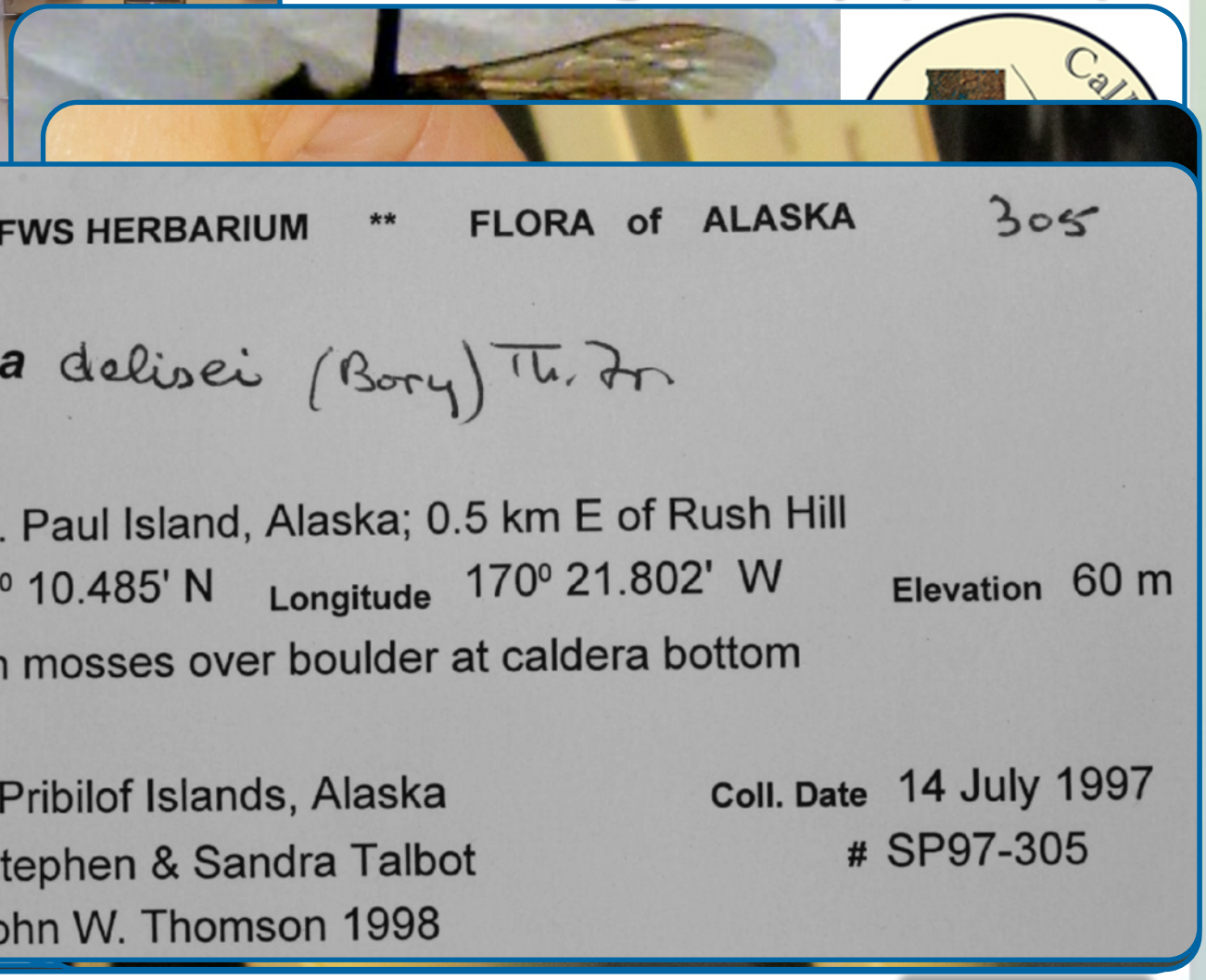
80 *Recordsets*

Making data and images of millions of biological specimens



iDigBio
Integrated Digitized Biocollections

Augmenting OCR Working Group (aOCR)



AKFWS HERBARIUM ** FLORA of ALASKA 305

Cetraria delisei (Bory) Th. Fr.

Location St. Paul Island, Alaska; 0.5 km E of Rush Hill

Latitude 57° 10.485' N Longitude 170° 21.802' W Elevation 60 m

Habitat On mosses over boulder at caldera bottom

Quad. Map Pribilof Islands, Alaska Coll. Date 14 July 1997

Collector Stephen & Sandra Talbot # SP97-305

Det. John W. Thomson 1998

iDigBio
Integrated Digitized Biocollections

APIARY PROJECT

Apiary Workflow

apiary.brit.org/drupal/modules/apiary_project/workflow/index.php?workflow_id

ANALYZE SPECIMEN | TRANSCRIBE TEXT | PARSE TEXT | GETTING STARTED | ?

PLANTS OF TEXAS

COLLECTED UNDER THE AUSPICES OF SOUTHERN METHODIST UNIVERSITY
AND THE BUREAU OF PLANT INDUSTRY, U. S. DEPARTMENT OF AGRICULTURE

Triodanis Holzingeri McVaugh

Roadside bank near post-oak woods
about 2 miles west of Decatur, Wise Co.

ROGERS MCVAUGH, No. 7197 June 19, 1945

ROI Type:Undefined Edit
Transcription:Undefined
Parsing:Undefined
Delete ROI

ROI Type:Annotation/Other Edit
Transcription:Undefined
Parsing:Undefined
Delete ROI

ROI Type:Annotation/Other Edit
Transcription:Undefined
Parsing:Undefined
Delete ROI

ROI Type:Primary Label Edit
Transcription:Undefined
Parsing:Undefined
Delete ROI

ROI Type:Barcode Edit
Transcription:Undefined
Parsing:Undefined
Delete ROI

My Queue Summary + ?
5 specimens
5 ROIs
open queue

ap-specimen:Specimen-794 - Analyzing ?
No roi currently selected
Status: not started
Add specimen note

ROI Legend ?
- Primary Label
- Annotation/Other
- Barcode
- Undefined

BRIT Plant to planet.

iDigBio Integrated Digitized Biocollections

Symbiota

Promoting Bio-Collaboration

Occurrence Data | Determination History | Images | Admin

Collector Info

Catalog Number | Occurrence ID | Collector | Number | Date

1281887

Dupes

Associated Collectors | Other Catalog Numbers

Latest Identification

Scientific Name: | Author:

ID Qualifier: | Family:

Identified By: | Date Identified:

Locality

Country | State/Province | County | Municipality

Locality:

Locality Security

Latitude | Longitude | Uncertainty | Datum | Elevation in Meters | Verbatim Elevation

Tools | ft.

Misc

Habitat:

Associated Taxa:

Description:

Notes:

Curation

Type Status: | Disposition:

Reproductive Condition: | Establishment Means: | Cultivated

Owner Code: | Basis of Record: PreservedSpecimen | Language:

Label Processing

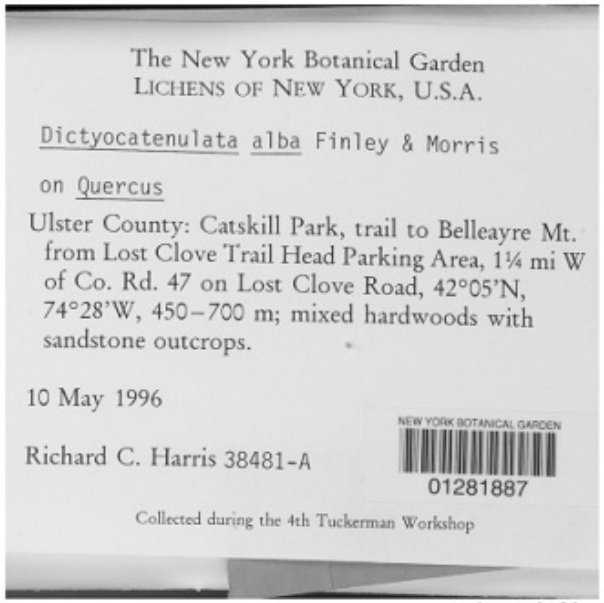


Image 1 of 1

```
The New York Botanical Garden
LICHENS of NEW YORK, U.S.A.
Dictyocatenulata alba Finley & Morris
On QUERCUS
Ulster County: Catskill Park, trail to Belleayre
Mt.
from Lost Clove Trail Head Parking Area, 1â€™/1
mi W
of Co. Rd. 47 on Lost Clove Road, 42Â°05â€™N,
74Â°28â€™X/, 450-700 rn; mixed hardwoods with
sandstone outcrops. -
1 O May 1 99 6
NEW VOHK BOTANICAL GAâ€¢IDEN
Richard C. Harris 38481-/-
01281887
Collected during the 4th Tuckerman Workshop
```

iDigBio Integrated Digitized Biocollections

Symbiota

Promoting Bio-Collaboration

Image 1 of 1

The New York Botanical Garden
LICHENS oF NEW YORK, U.S.A.
Dictyocatenulata alba Finley & Morris
On QUEPCUS
Ulster County: Catskill Park, trail to Belleayre Mt.
from Lost Clove Trail Head Parking Area, 1â€³/1 mi W
of Co. Rd. 47 on Lost Clove Road, 42Â°05â€³N, 74Â°28â€³X/, 450-700 rn; mixed hardwoods with sandstone outcrops. -
1 O May 1 99 6
NEW VOHK BOTANICAL GAâ€IDEN
Richard C. Harris 38481-/-
01281887
Collected during the 4th Tuckerman Workshop

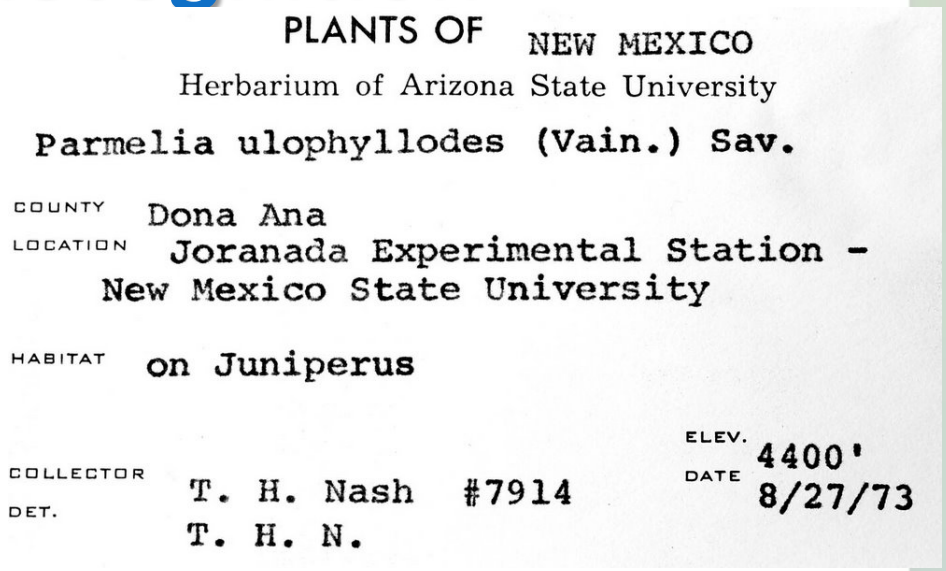
The herbaria at the Royal Botanic Garden Edinburgh (RBGE) and The New York Botanical Garden (NY) use a similar recently developed workflow:

1. Capture **minimal data** for each specimen
   - (e.g. barcode, geographic region, and the taxon name on the specimen folder).
2. Capture **high quality digital images** of each specimen
3. Process images with ABBYY® **OCR software**
4. Develop and/or use software tools to **search OCR text** output and **sort the images/data** based on principal data elements (e.g. by collector and country)
5. Sort image/data to enable **faster data capture** & **higher accuracy** by data transcriptionists, including **duplicate record matching**.

iDigBio
Integrated Digitized Biocollections

Optical Character Recognition

- Tesseract V3
- Dual cycle
  - Automatic
  - Manual review
- Expected hurdles
  - Handwritten labels
  - Old fonts
  - Faded labels
  - Form labels
- Adjustable image variables


PLANTS OF NEW MEXICO
Herbarium of Arizona State University

Parmelia ulophyllodes (Vain.) Sav.

COUNTY Dona Ana
LOCATION Joranada Experimental Station -
New Mexico State University

HABITAT on Juniperus

ELEV. 4400'
COLLECTOR T. H. Nash #7914 DATE 8/27/73
DET. T. H. N.

PLANTS OF NEW r~1Exlco
Herbarium of Arizona State University
Parmelia ulophyllodes (Vain.) Sav.
COUNTY â€œÂ°â€ â€œâ€œ
Joranada Experimental Station -
New Mexico State University
"â€œâ€œâ€œ' on Juniperus
ELEV. ‘ 4400
EEILLEETUR DATE
DU T. H. Nash #7914 8/27/73
T. H. N.

aOCR Working Group

- Bryan Heidorn, Director, School of Information Resources and Library Science, University of Arizona
- Robert Anglin, LBCC TCN Developer
- Reed Beaman, iDigBio Senior Personnel, FLMNH
- Jason Best, Director Bioinformatics, Botanical Research Institute of Texas
- Renato Figueiredo, iDigBio IT
- Edward Gilbert, Symbiota Developer, LBCC and SCAN TCNs
- Nathan Gnanasambandam, Xerox
- Stephen Gottschalk, Curatorial Assistant, NYBG
- Peter Lang, Pre Sales Engineer, ABBYY
- Deborah Paul, iDigBio User Services
- Elspeth Haston, Royal Botanic Garden, Edinburgh
- John Mignault, New York Botanical Garden
- Anna Saltmarsh, Royal Botanic Gardens, KEW
- Nahil Sobh, Developer, InvertNet
- Alex Thompson, iDigBio Developer
- William Ulate, Biodiversity Heritage Library (BHL), MOBOT
- Kimberly Watson, Curatorial Assistant, NYBG
- Qianjin Zhang, Graduate Student, University of Arizona

aOCR Highlights

- faster and more efficient data transcription, image analysis
- share what's possible: Wiki, Publications, Meetings, Workshops, Hackathons
  - Publications - see **http://idigbio.org/biblio**
  - The SALIX Method: A semi-automated workflow for herbarium specimen digitization. June 2013 Taxon 62(3):581-590 A Barber, D Lafferty, L Landrum
  - aOCR overlap with Public Participation Working Group and DROID working Groups
- POST-DOC! New post-doc position at iDigBio
  - Public Participation in Digitization of Biodiversity Collections
  - data standards, tool integration & enhancement planning
  - New programmer position
- Public Participation tools
  - Symbiota / LBCC TCN
  - Notes From Nature

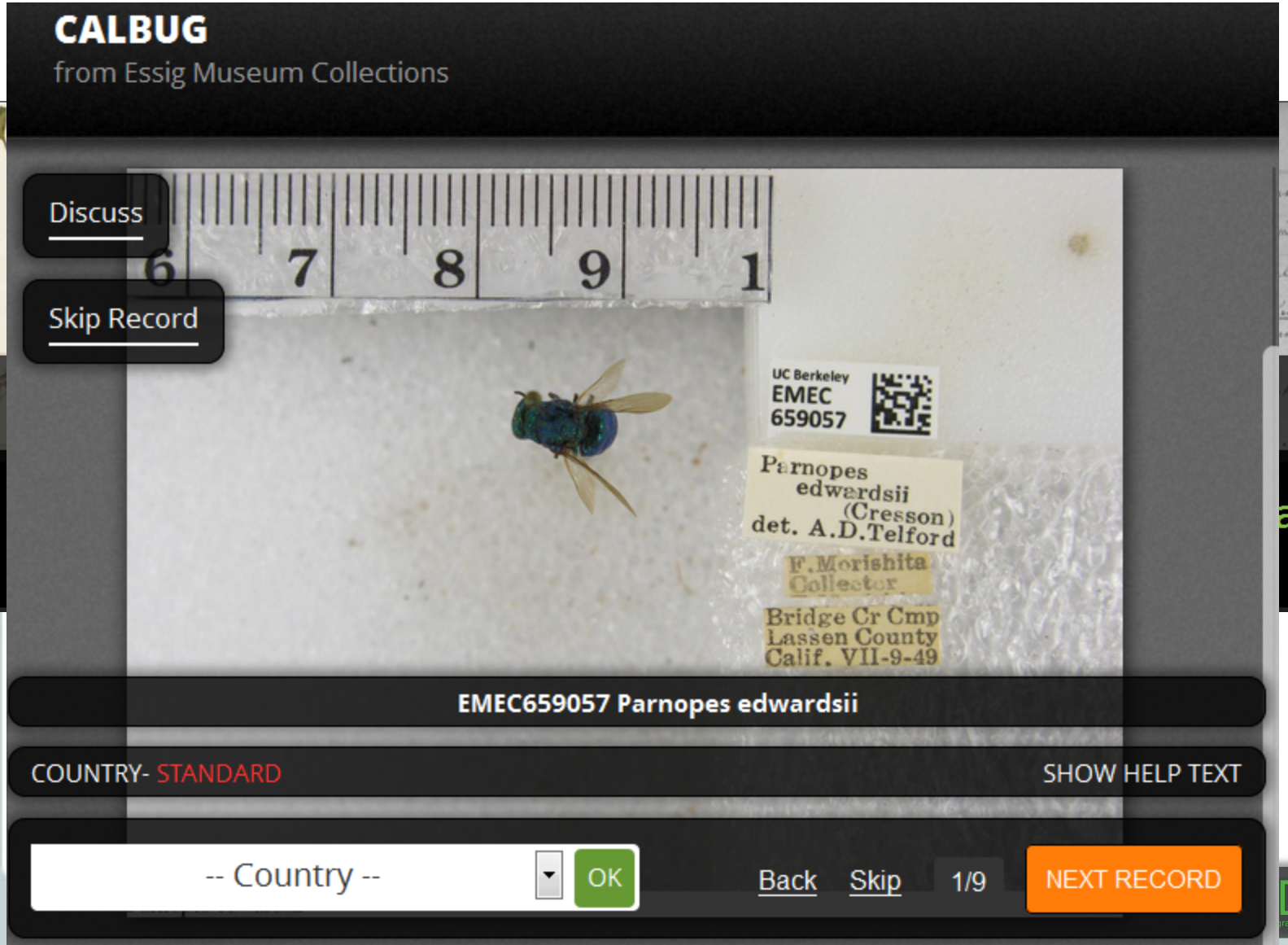
iDigBio
Integrated Digitized Biocollections

Notes from Nature .

COLLECTIONS ABOUT DISCUSS BLOG

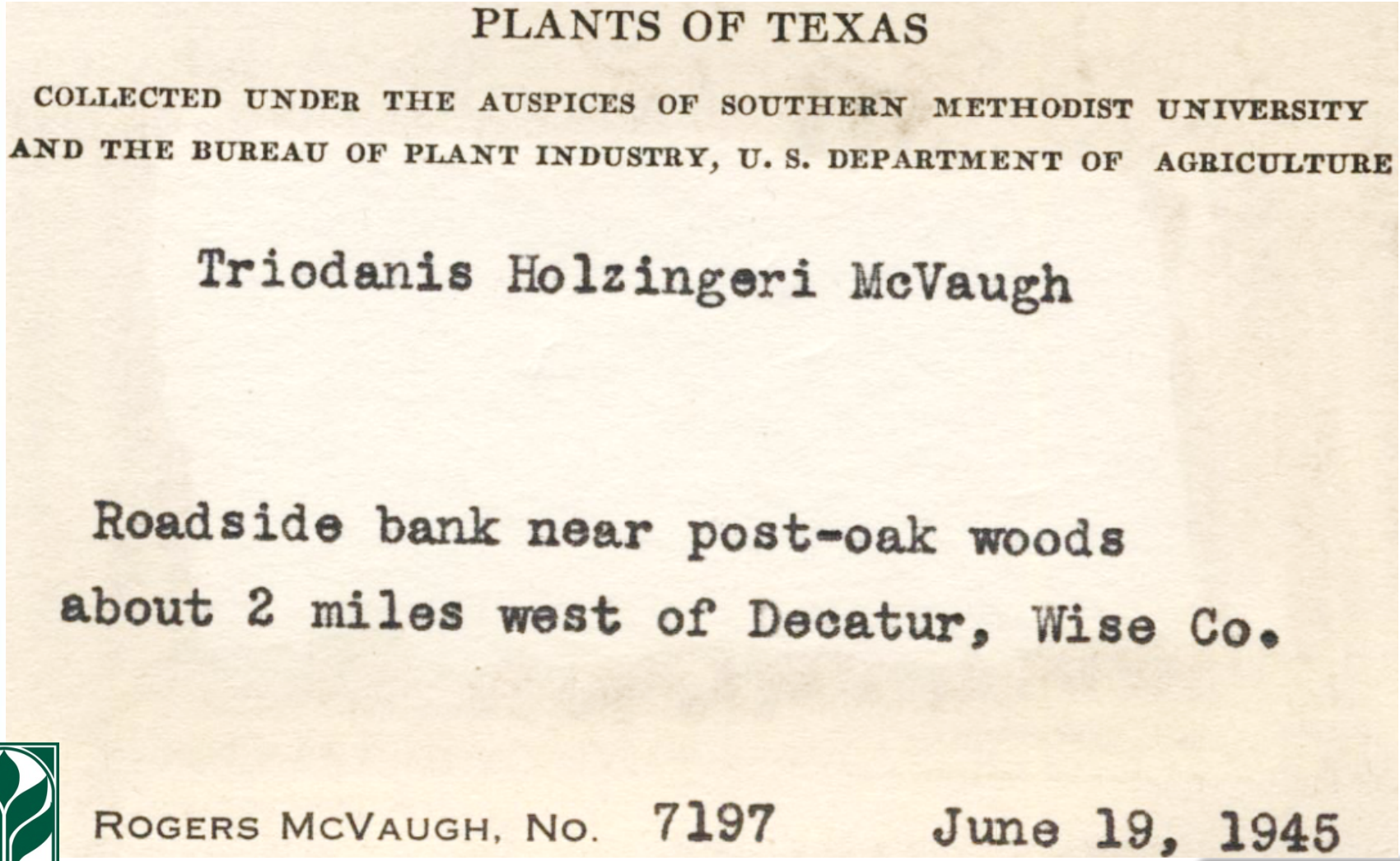
http://www.notesfromnature.org/

2 collections available




and now for something a bit different…

iDigBio Georeferencing Working Group (GWG)



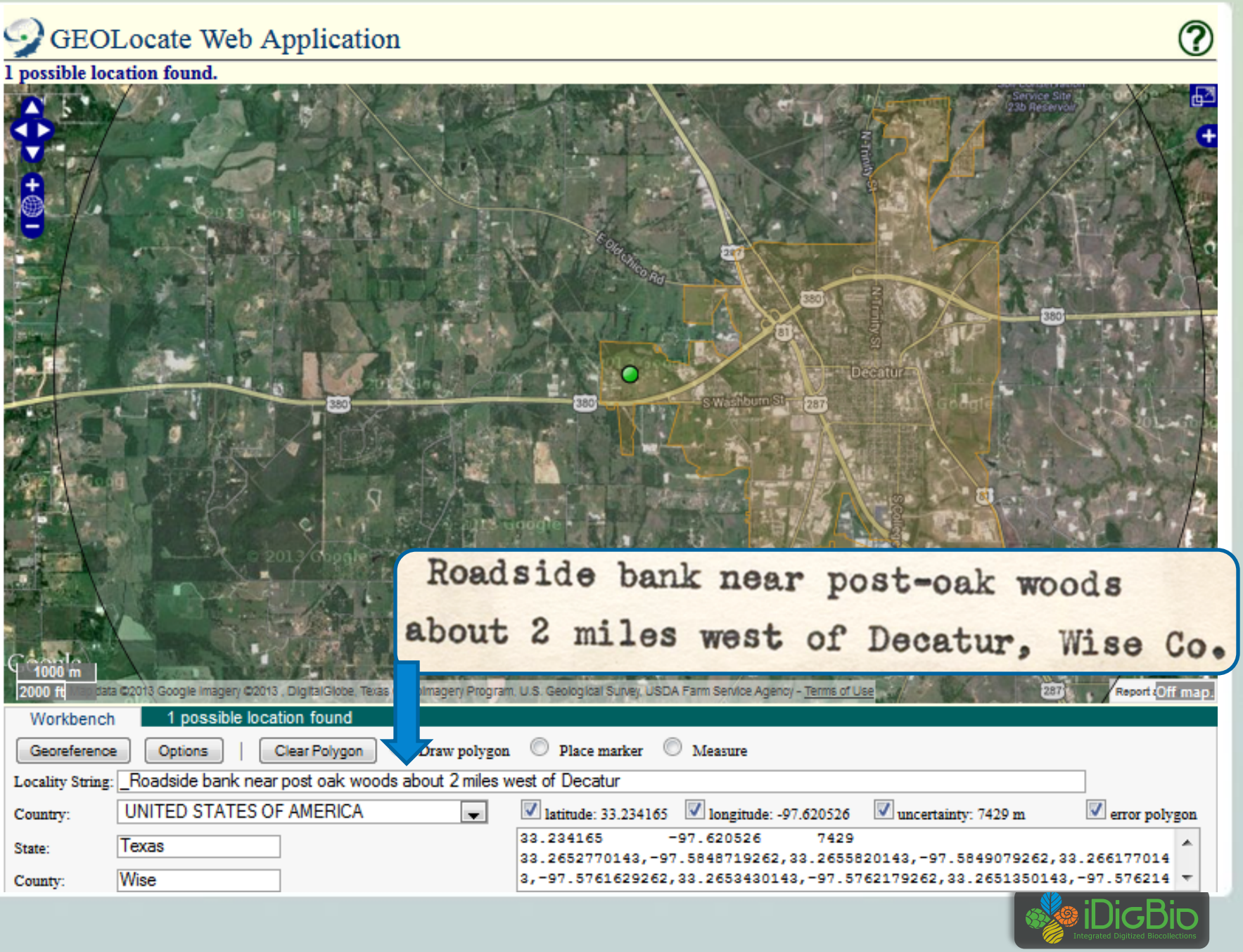
PLANTS OF TEXAS

COLLECTED UNDER THE AUSPICES OF SOUTHERN METHODIST UNIVERSITY AND THE BUREAU OF PLANT INDUSTRY, U. S. DEPARTMENT OF AGRICULTURE

Triodanis Holzingeri McVaugh

Roadside bank near post-oak woods about 2 miles west of Decatur, Wise Co.

ROGERS MCVAUGH, No. 7197 June 19, 1945


GEOLocate Web Application
1 possible location found.
Roadside bank near post-oak woods
about 2 miles west of Decatur, Wise Co.
Workbench
1 possible location found
Georeference
Options
Clear Polygon
Draw polygon
Place marker
Measure
Locality String: _Roadside bank near post oak woods about 2 miles west of Decatur
Country: UNITED STATES OF AMERICA
State: Texas
County: Wise
latitude: 33.234165
longitude: -97.620526
uncertainty: 7429 m
error polygon
33.234165 -97.620526 7429
33.2652770143,-97.5848719262,33.2655820143,-97.5849079262,33.266177014
3,-97.5761629262,33.2653430143,-97.5762179262,33.2651350143,-97.576214

Georeferencing in a workflow…

- where does georeferencing fit in my workflow?
  - as I enter data?
  - after all data is entered (usually, but not always)
  - some combination?
- what's the fastest / cheapest / most accurate method?
- do I need experts to do this?
  - who will do it?
  - how much will it cost?
  - what are the issues?

GEOLocate: Community Edition

Collaborative Georeferencing

Welcome debpaul, user since 1/29/2012. | LOGOUT

- Home
- About Us
- Contact Us
- My Communities
- Portal Tutorial

Community: FSU

Data Sources | Members | Settings

Data source management operations

Add new community data source via CSV files

Click on an item's header to expand/collapse its content.

community-wide		
nolat	date added: Thursday, March 15, 2012	delete
labeldata	date added: Thursday, March 15, 2012	delete
Leon1	date added: Tuesday, April 24, 2012	delete
Leon2	date added: Tuesday, April 24, 2012	delete

Top of page.

iDigBio Integrated Digitized Biocollections

GWG Resources at iDigBio

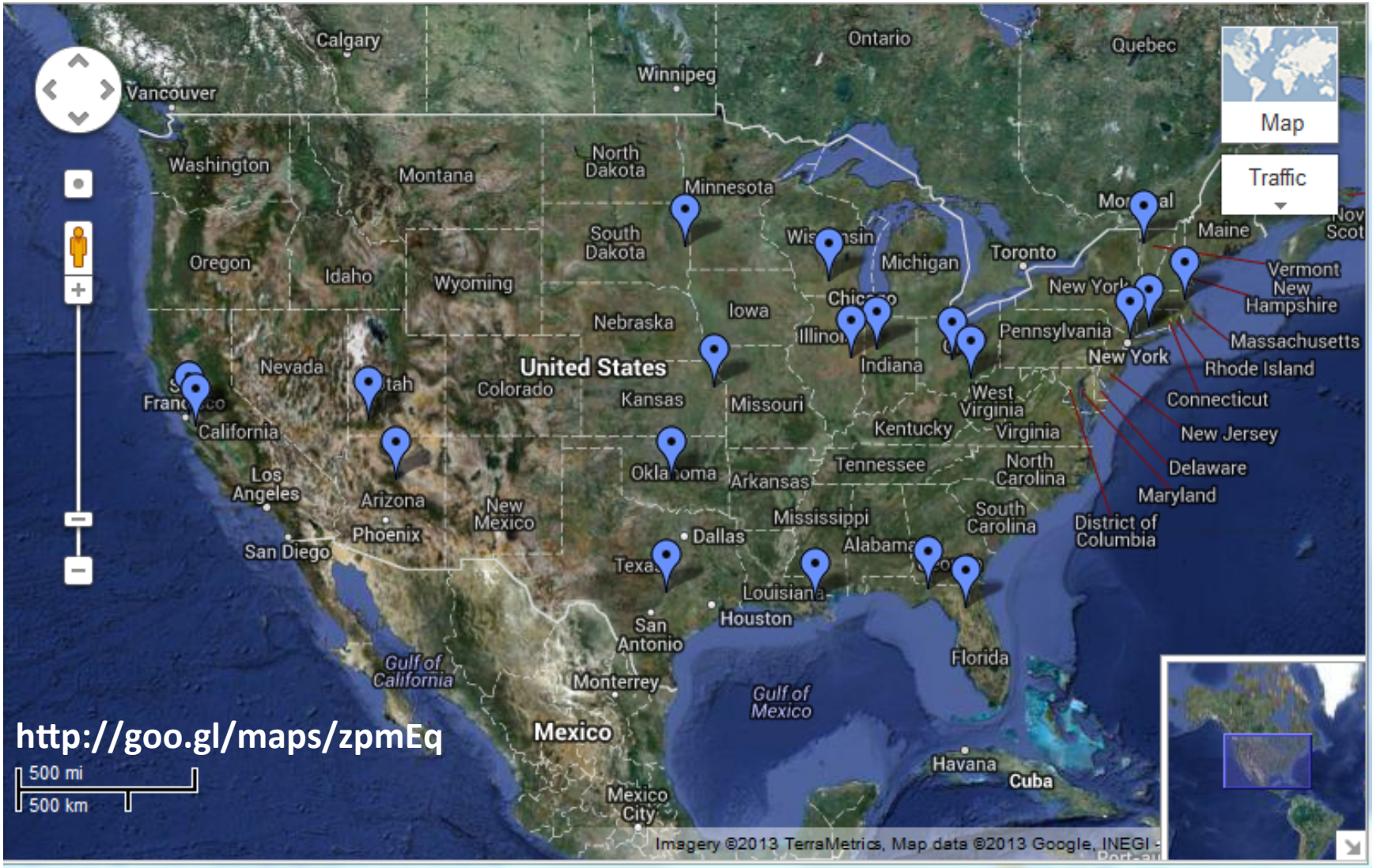
Remote Participation:

- Want to attend Train the Trainer's Workshop Remotely? Please log in the software works.

Reading Materials and Resources:

1. Georeferencing Quick Reference Guide
2. Guide to Best Practices for Georeferencing - Chapman, A.D. and J. W
3. iDigBio Georeferencing Wiki
4. HerpNET Georeferencing Resources
5. Pre Workshop Survey Questions iDigBio Asked
6. Group Notes - Take Workshop Notes Together Here
7. [Post - Workshop Survey Questions]
8. Got a Georeferencing Question? Post it at the iDigBio Georeferencing

A Georeferencing Trainer Near You…



http://goo.gl/maps/zpmEq

iDigBio
Integrated Digitized Biocollections

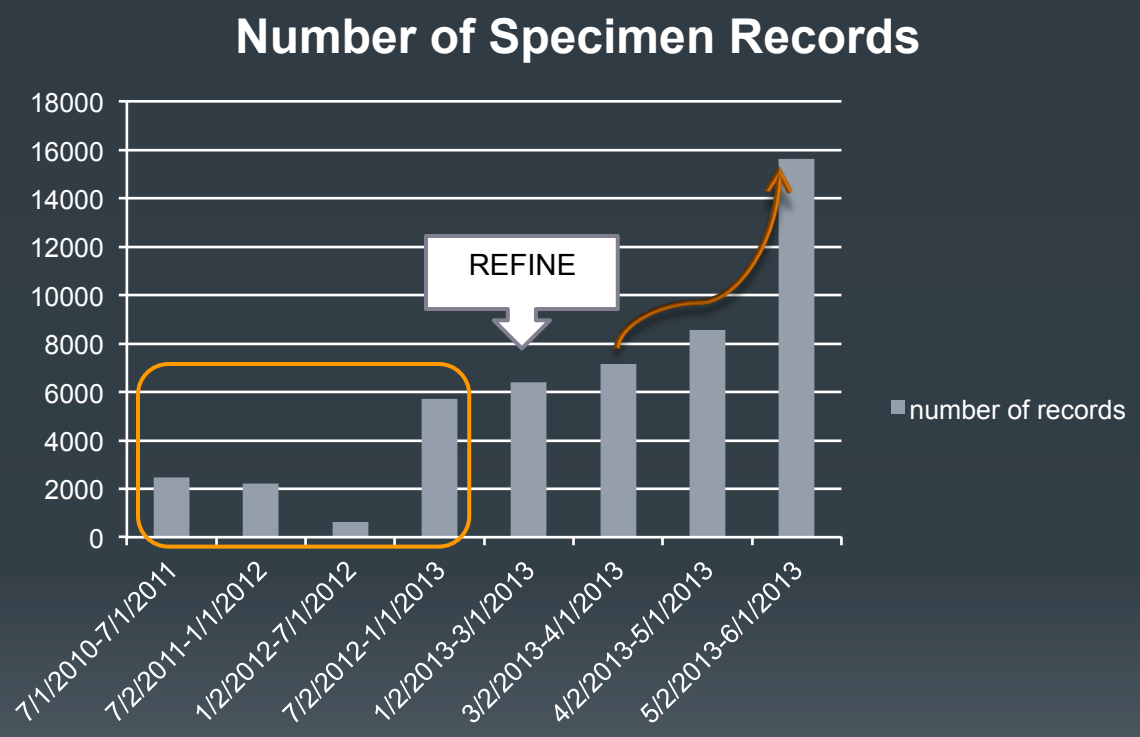
Videos at http://vimeo.com/idigbio

iDigBio
Integrated Digitized Biocollections

What's possible?

- **Digitization workflows: a modular approach to achieving an effective plumbing system**. Ann Molineux, Liath Appleton, Angella Thompson, Louis G. Zachos* SPNHC 2013
  - Non-vertebrate Paleontology Laboratory, Texas Natural Science Center, The University of Texas at Austin; *Department of Geology & Geological Engineering, University of Mississippi
- Common Goal: clean, enhanced, fit-for-use data
  - **openrefine.org**

Migration rate change



- iDigBio 'Train the trainers' georeferencing workshop 10/8/2012
- In-house transfer of training knowledge 11/2011
- Current Specify records 50589 [c. 161903 specimens]

Train-the-Trainers
…the model
…the numbers

- As of March 2013
  - 13 of 25 from TTT #1 (October 2012) responded to survey
  - 7 people have done one or more trainings
    - > 70 people trained
- Records georeferenced since the training (in < 5 months)
  - over 4400

iDigBio
Integrated Digitized Biocollections

GWG News

- iDigBio's Georeferencing Working Group (GWG)
  - Train-the-Trainers 1, 2
    - remote attendance – see you there? Aug 12 - 16
  - Repatriation
  - GBIF eLearning
    - September 2013
  - georeferencing.org
- Public Participation in Science
  - georeferencing
  - GEOLocate
  - Expert or Non Expert? What's the magic number?

iDigBio GWG +

- David Bloom, VertNet Coordinator, Berkeley
- Nelson Rios, GEOLocate Developer, Tulane
- John Wieczorek, Information Architect/VertNet, Berkeley
- Carol Spencer, Staff Curator/VertNet, Berkeley
- Reed Beaman, FLMNH
- Una Farrell, Co-PI, Paleoniches TCN, Kansas
- Paul Heinrich, Co-PI, SCAN TCN, NAU
- Mary Barkworth, Utah State University
- Gil Nelson, iDigBio Digitization
- Deborah Paul, iDigBio User Services

FLORIDA MUSEUM OF NATURAL HISTORY

GEOLocate

GBIF

VertNet

BioGeomancer

Paleoniches TCN

UtahState University

SCAN TCN

iDigBio Integrated Digitized Biocollections

Thank you, from the aOCR and GWG!

- NSCA
- SPNHC 2013
- South Dakota Schools of Mines
- …the entire community

iDigBio
Integrated Digitized Biocollections