

Digitization Tools

Optical Character Recognition, Parsing, Consensus in Crowdsourcing

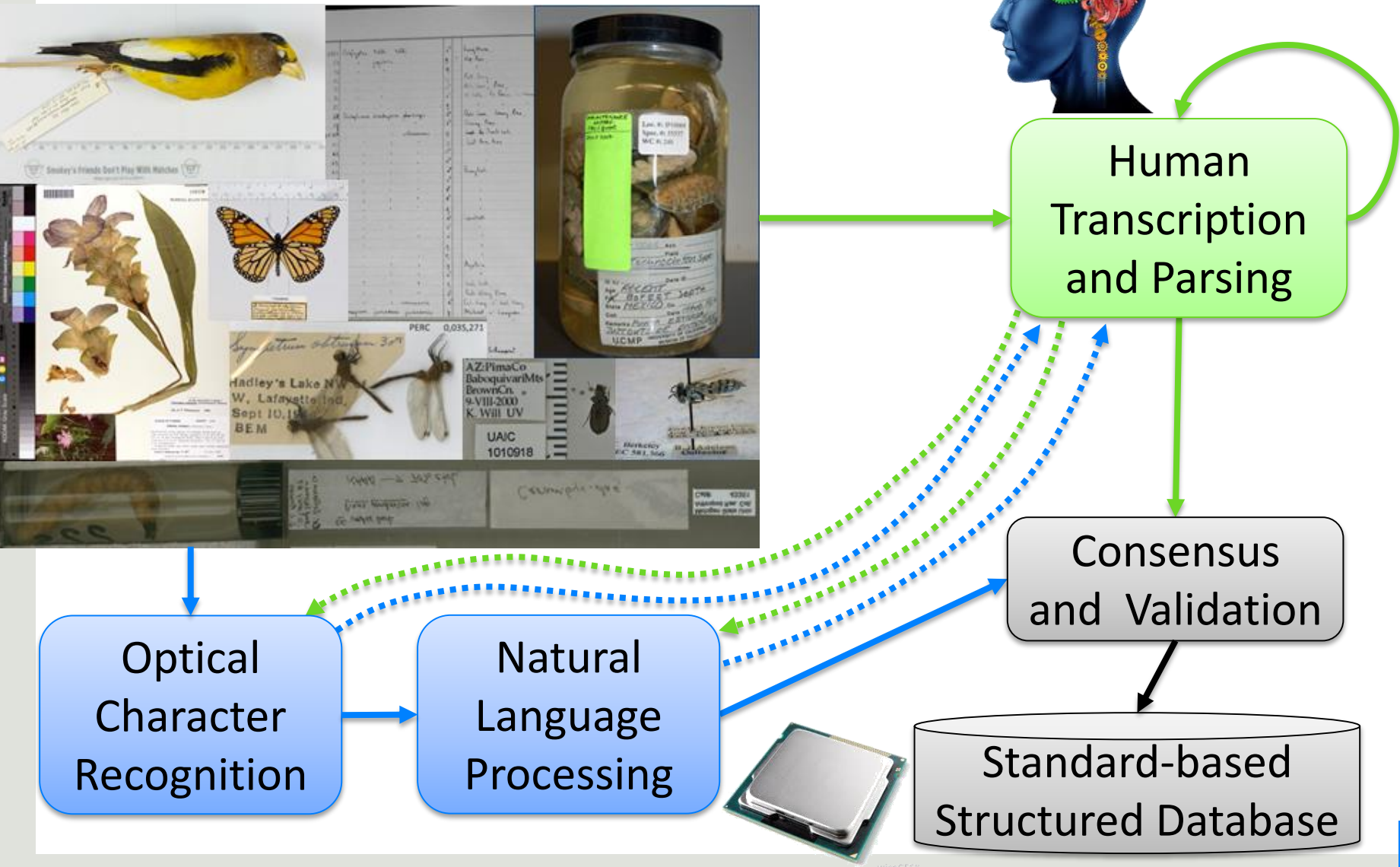
Andréa Matsunaga (ammatsun@ufl.edu)

Brazilian Biodiversity Information System (SiBBr) Launching Event
November 25, 2014
Brasília, 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.

Digitization Workflow Tools



Label

OCR

No.2L31.

National Herbarium of Canada
FLORA OF T TERRITORIES

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.

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

P. alaskanum *Hulten* No. 7138
National Herbarium of Canada
FLORA OF ~~NORTHWEST~~ *Yukon* 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

ABBYY

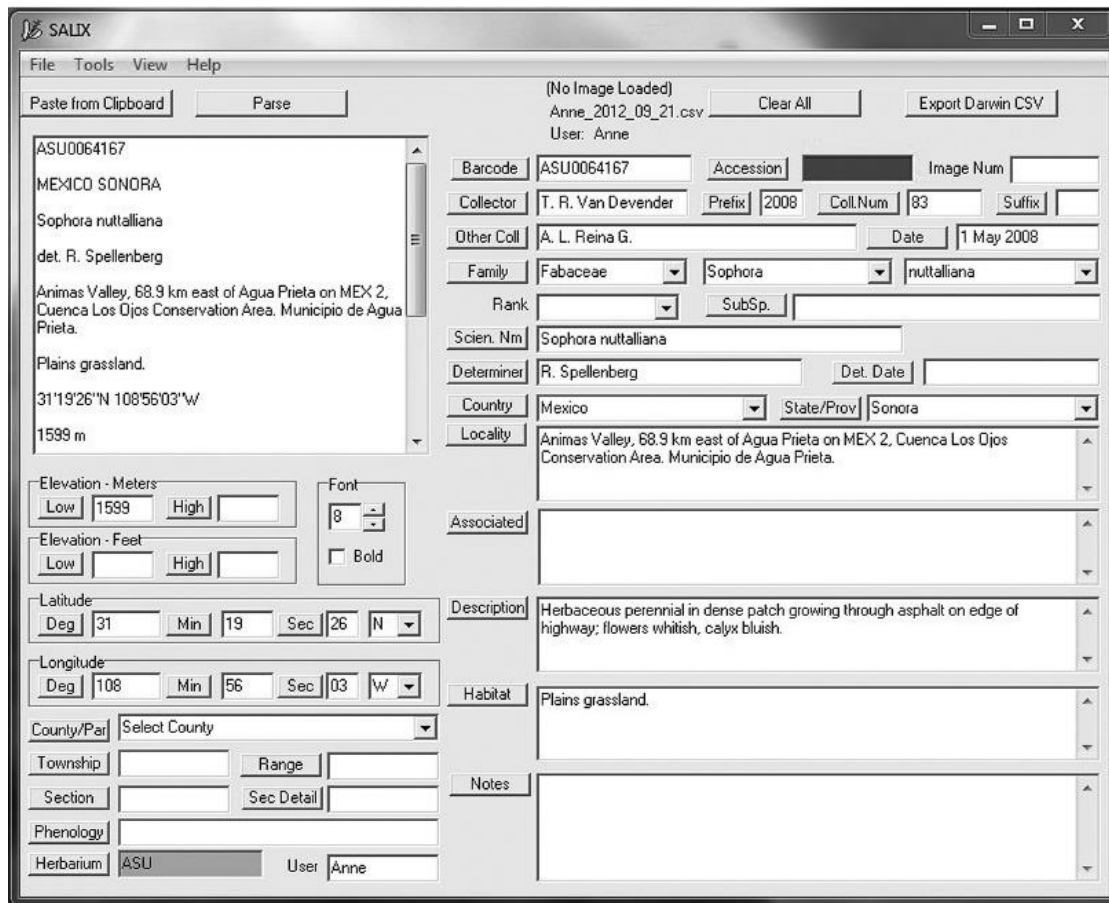
GOCR

OCRAD

OCRopus

Tesseract

SALIX (Semi-Automatic Label Information eXtraction)



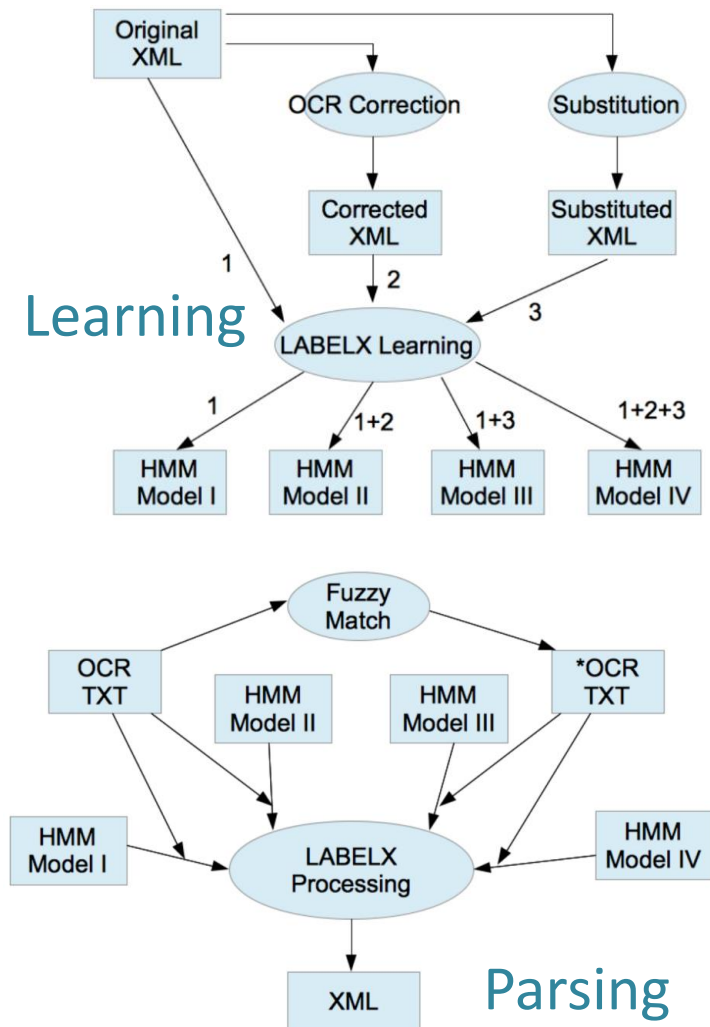
The screenshot shows the SALIX software interface with the following data:

- Barcode:** ASU0064167
- Accession:** [Empty]
- Image Num:** [Empty]
- Collector:** T. R. Van Devender
- Prefix:** 2008
- Coll Num:** 83
- Suffix:** [Empty]
- Other Coll:** A. L. Reina G.
- Date:** 1 May 2008
- Family:** Fabaceae
- Sophora:** [Empty]
- nuttalliana:** [Empty]
- Rank:** [Empty]
- SubSp:** [Empty]
- Scien. Nm:** Sophora nuttalliana
- Determiner:** R. Spellberg
- Det. Date:** [Empty]
- Country:** Mexico
- State/Prov:** Sonora
- Locality:** Animas Valley, 68.9 km east of Agua Prieta on MEX 2, Cuenca Los Ojos Conservation Area. Municipio de Agua Prieta.
- Associated:** [Empty]
- Description:** Herbaceous perennial in dense patch growing through asphalt on edge of highway; flowers whitish, calyx bluish.
- Habitat:** Plains grassland.
- Notes:** [Empty]
- Elevation - Meters:** Low 1599 High [Empty]
- Elevation - Feet:** Low [Empty] High [Empty]
- Latitude:** Deg 31 Min 19 Sec 26 N
- Longitude:** Deg 108 Min 56 Sec 03 W
- County/Par:** Select County
- Township:** [Empty] Range [Empty]
- Section:** [Empty] Sec Detail [Empty]
- Phenology:** [Empty]
- Herbarium:** ASU
- User:** Anne

- Parsing algorithm:
 - Built on compiled word statistics
 - Improve with use

- <http://taxonbytes.org/asu-herbarium-videos-on-seinet-and-ocr-based-digitization/>
- <http://daryllafferty.com/salix/>
- Barber, A., D. Lafferty & L.R. Landrum. 2013. The SALIX Method: A semi-automated workflow for herbarium specimen digitization. *Taxon* 62: 581-590.

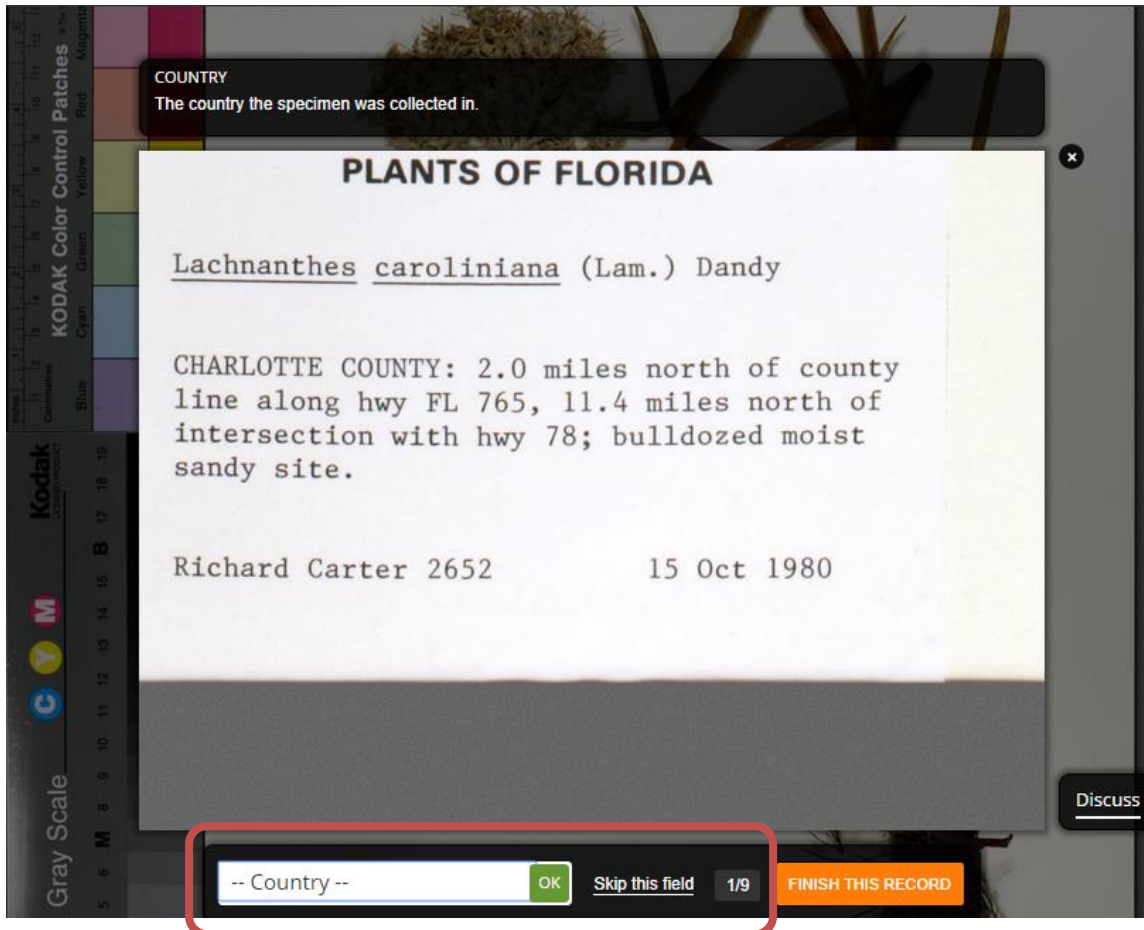
LABELX (Label Annotation through Biodiversity Enhanced Learning)



- Deals with:
 - Variability of label layouts
 - High-rate of OCR errors
- Machine learning:
 - Naive Bayes, Hidden Markov Models and N-Gramming
- Combined with human supervision
- Corrects some OCR errors

Crowdsourcing Transcription Projects

- NotesFromNature (<http://www.notesfromnature.org/>)
 - Zooniverse platform



COUNTRY
The country the specimen was collected in.

PLANTS OF FLORIDA

Lachnanthes caroliniana (Lam.) Dandy

CHARLOTTE COUNTY: 2.0 miles north of county line along hwy FL 765, 11.4 miles north of intersection with hwy 78; bulldozed moist sandy site.

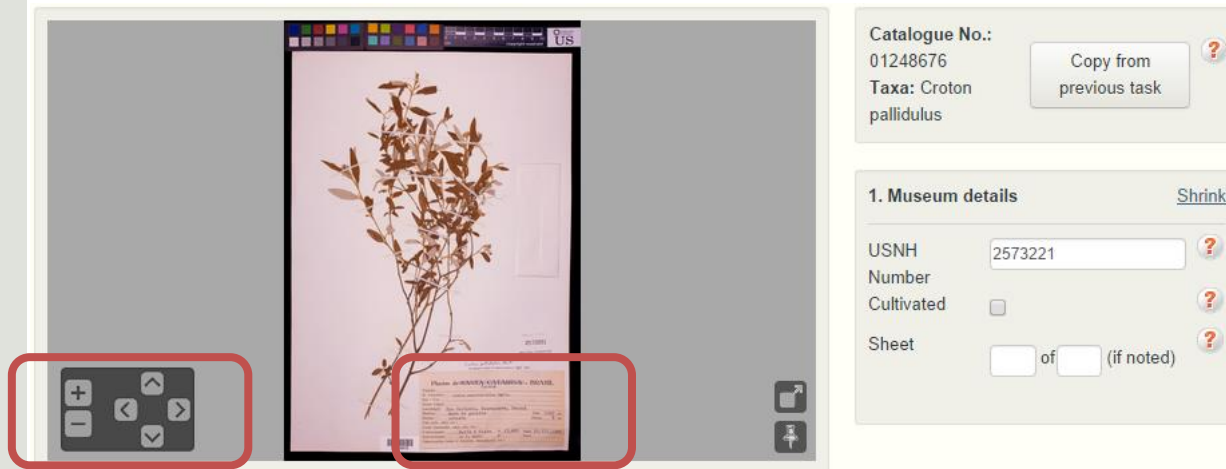
Richard Carter 2652 15 Oct 1980

-- Country -- OK Skip this field 1/9 FINISH THIS RECORD

- Once the user selects the region with the label, s/he can start transcribing and parsing information to a number of pre-defined fields
- For a requester, a pre-defined number of transcriptions are returned

Crowdsourcing Transcription Projects

- ALA (<http://volunteer.ala.org.au/>)
– Platform: Grails



- User zooms in to read the label and parse to the custom pre-defined terms
- Single worker followed by expert approval

2. Collection details

Collector(s)

Collector number

Collection Date (from) (to)

3. Location details

Verbatim Locality

State/Province/Territory

Country

Elevation - To

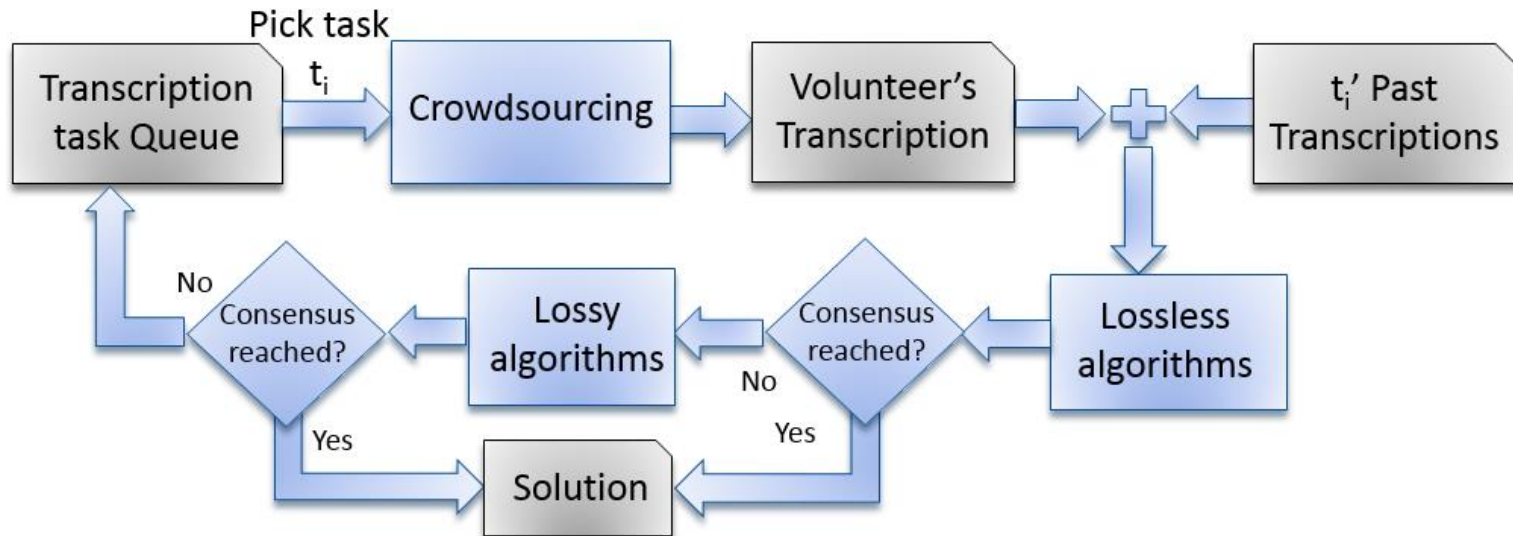
Verbatim Latitude

Verbatim Longitude

symbols:

symbols:

Consensus with worker controller



- Full consensus improvement from 1.8% to 84.2%
- Lossless algorithms have small impact except for *scientific author* and *collected by*
- Workforce savings can be as high as 55.8% when compared to a statically configured number of workers

★ Lichens Silver ★ Herb Silver ★ Herb All Silver

* | _____

Download 1000 results Cluster with Lingua

Title field name url
Summary field name conte
URL field name url
ID field name id

Read Solr clusters if present ☐
Use highlighter output if present ☒

Hide

Folders Circles FoamTree

Legend - Level of confidence that token is an accurately-transcribed word

extremely low	very low	low	undetermined	medium	high	very high
---------------	----------	-----	--------------	--------	------	-----------



Complete workflow

- Symbiota Software Project
 - <http://symbiota.org/docs/symbiota-workshops/>
 - <http://symbiota.org/docs/symbiota-introduction/symbiota-help-pages/>

New York Botanical Garden (NY)

Home >> Crowd Sourcing Central >> Editor

< << | 1 of 390 | >> >

Occurrence Data

Long Form <<

Collector ? Number ? Date ? Duplicates ?
 1 899 1988 Auto search

Associated Collectors ? Verbatim Date ?

Exsiccata Title Number

Scientific Name ?
 Nephroma antarcticum

Note: Full editing permissions are needed to edit an identification

Country State/Province County

Locality

Latitude Longitude Uncertainty ? Verbatim Coordinates Tools

Elevation in Meters Verbatim Elevation

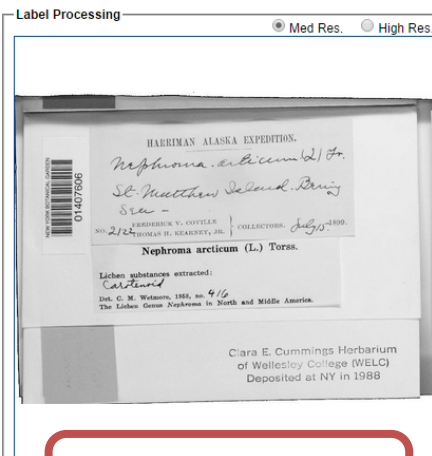
Habitat

Substrate

Notes

Save Edits

Status Auto-Set: Pending Review



OCR Image Options

☐ OCR whole image

☐ OCR w/ analysis

Image 1 of 1

HARRIMAN ALASKA EXPEDITION.
 be) (/*
 FREDERICK V. COVILLE I r ??
 1/2-1r. " " f
 COLLECTORS. /1 899. - f
 .W <- 'THOALAS H. KEARNEY, JR. J U^J/O
 Nephroma arcticum (L.) Torss.
 Lichen substances extracted:
 Det. G. M. Wetmore, 1958, no. H/Q,
 The Lichen Genus Nephroma in North and Middle
 America.
 Clara E. Cummings Herbarium

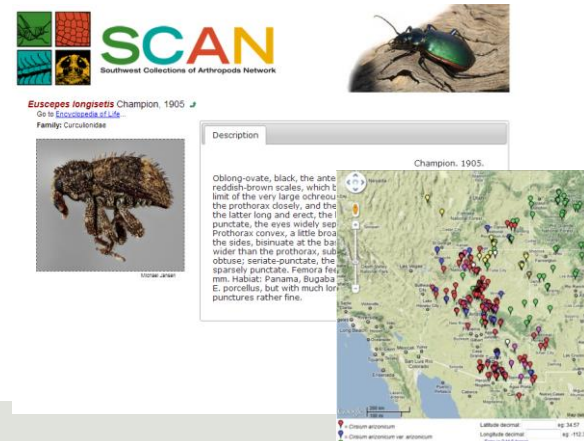
Notes:

Source:
 ABBY:2013-02-09

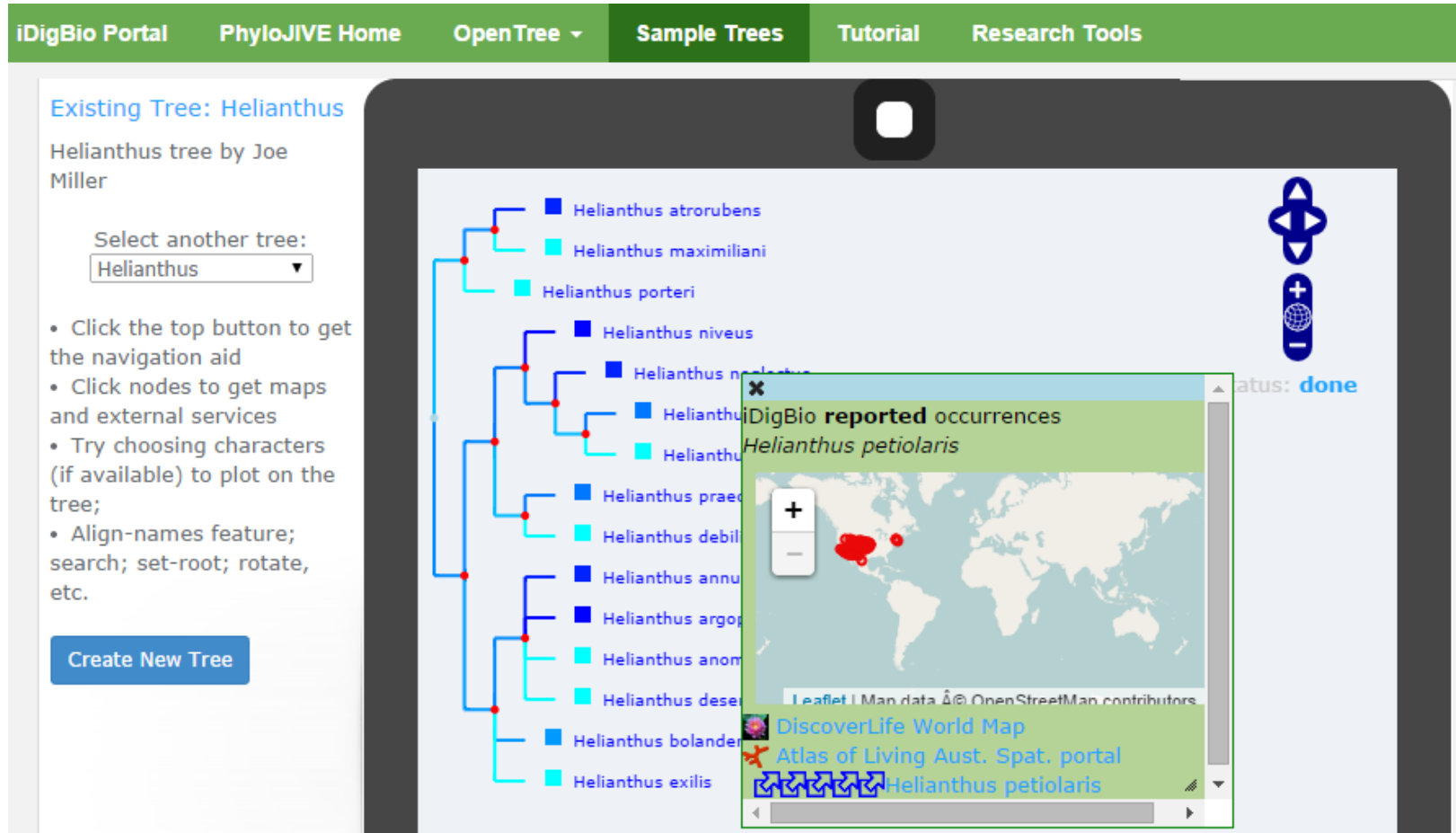
Save OCR Edits LBOC Parser 1 of 1

Delete OCR

- Ability to OCR and parse data
- Single worker followed by expert approval
- Open source
- Modular framework
- Community-based biodiversity portals



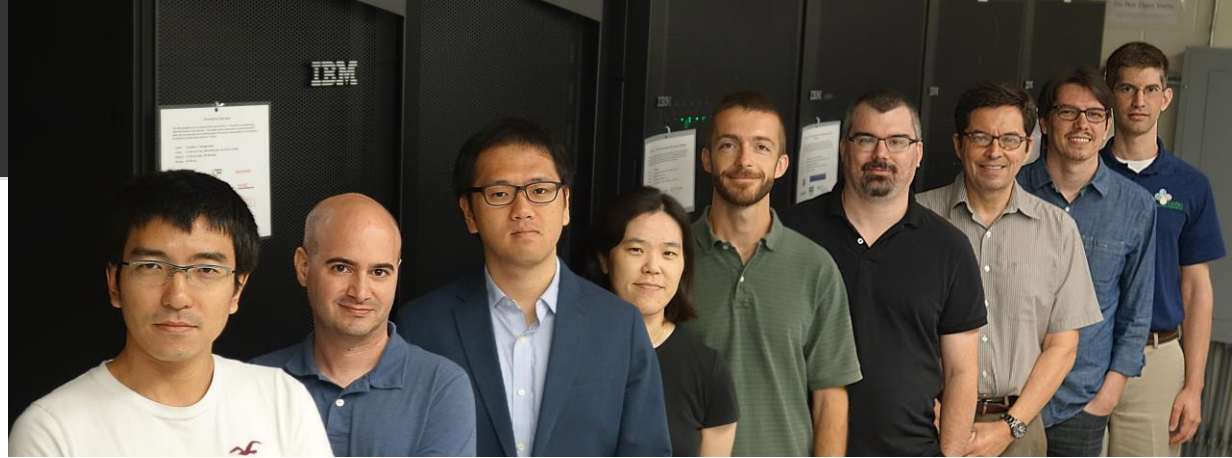
PhyloJIVE instance in iDigBio



The screenshot displays the iDigBio Portal interface for the PhyloJIVE Home. The top navigation bar includes links to iDigBio Portal, PhyloJIVE Home, OpenTree, Sample Trees, Tutorial, and Research Tools. The main content area is titled "Existing Tree: Helianthus" and shows a phylogenetic tree of Helianthus species. The tree is rooted and color-coded by species. A sidebar on the left provides instructions on how to interact with the tree, including clicking nodes to get maps and external services, and a "Create New Tree" button. A map window is open, showing the distribution of *Helianthus petiolaris* in North America. The map includes a legend for "DigBio reported occurrences" and a list of links for further information, such as "DiscoverLife World Map" and "Atlas of Living Aust. Spat. portal".

- Developed by Garry Jolley-Rogers, Joe Miller, and Temi Varghese
- Integrates biodiversity data with phylogeny.
- <http://phylojive.acis.ufl.edu/>

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

Thank you!