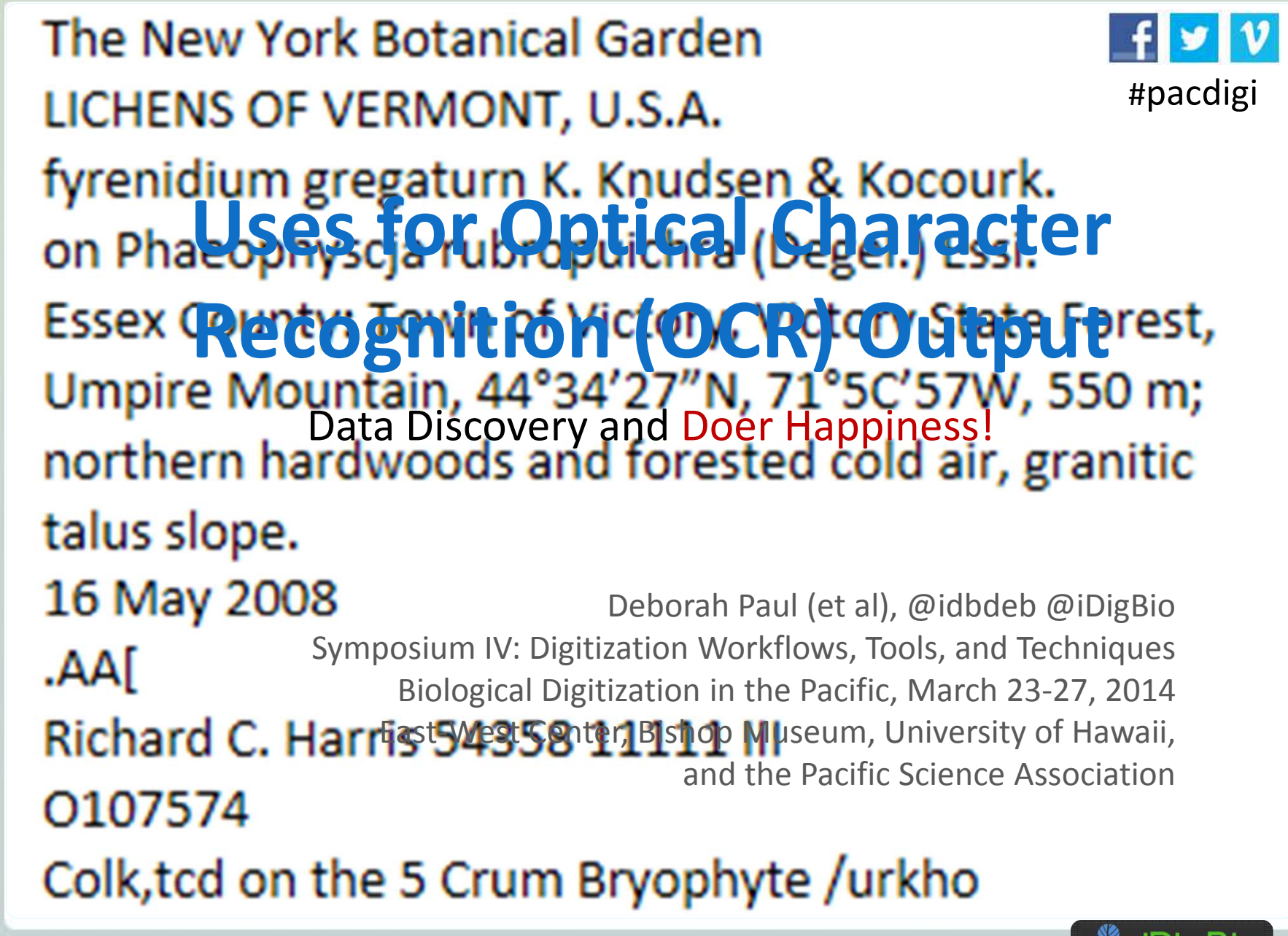


#pacdigi

Uses for Optical Character Recognition (OCR) Output

Data Discovery and Doer Happiness!

Deborah Paul (et al), @idbdeb @iDigBio

Symposium IV: Digitization Workflows, Tools, and Techniques
Biological Digitization in the Pacific, March 23-27, 2014
East West Center, Bishop Museum, University of Hawaii,
and the Pacific Science Association

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?

[Biological collection data capture: a rapid approach using curatorial data]

iDigBio
Integrated Digitized Biocollections

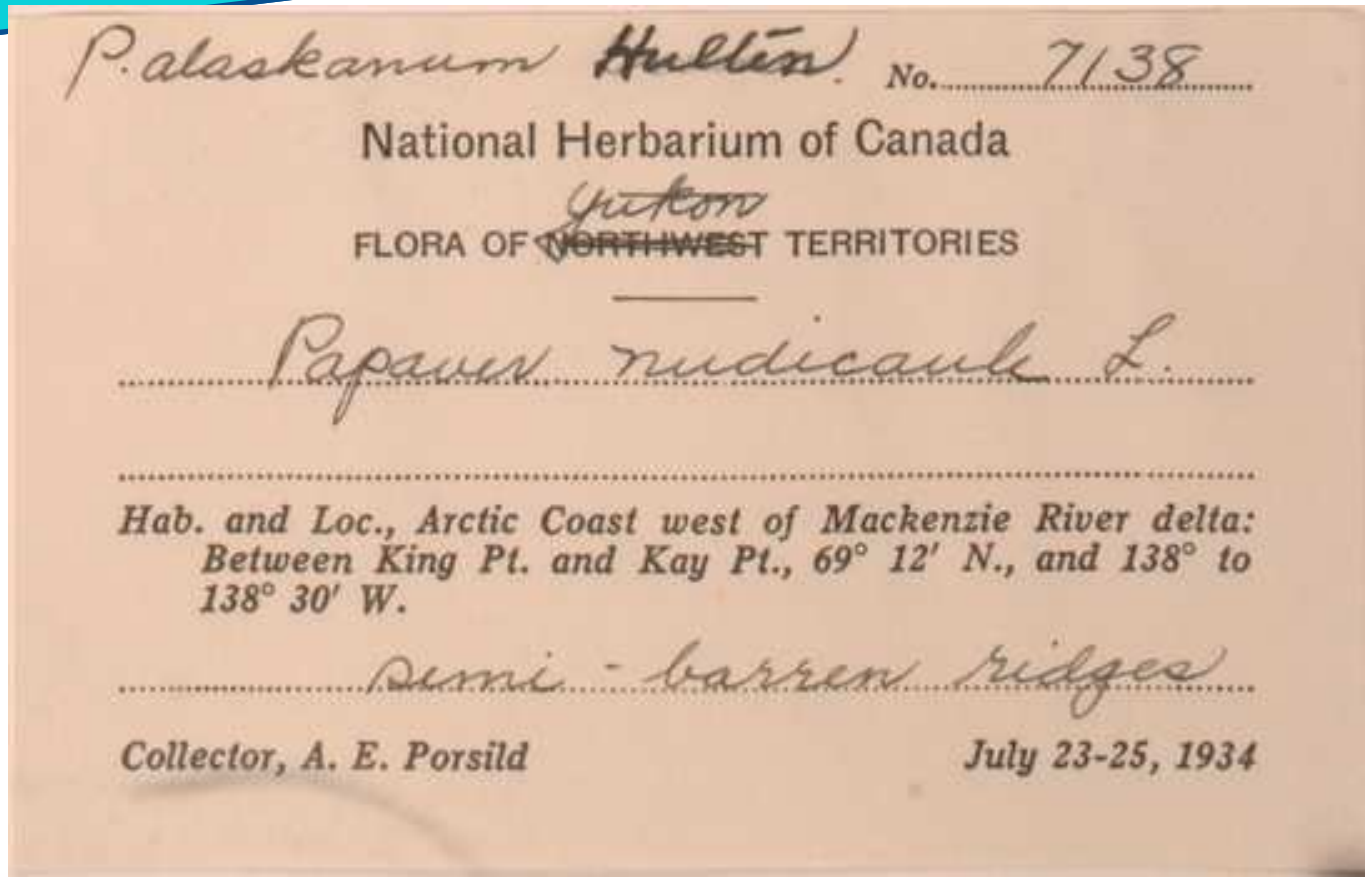
Would you like to…?

- enter records faster?
- use the ditto feature often?
- find duplicates quickly?
- find the labels
- find the labels with lots of handwriting?
- create your own record sets to transcribe?
  - by collector
  - by country or county
  - by your Great Aunt Susan
  - by taxon
  - by language
- create cogent sets to speed up validation and database updates?
- increase happiness of transcribers (paid and volunteer)?

Meaningful datasets make transcription
* faster
* less error prone
* *more fun!*

iDigBio
Integrated Digitized Biocollections

Got Text?
Got Handwriting?



P. alaskanum Hultén No. 7138

National Herbarium of Canada

FLORA OF ~~NORTHWEST~~ Yukon TERRITORIES

Papaver nudicaule L.

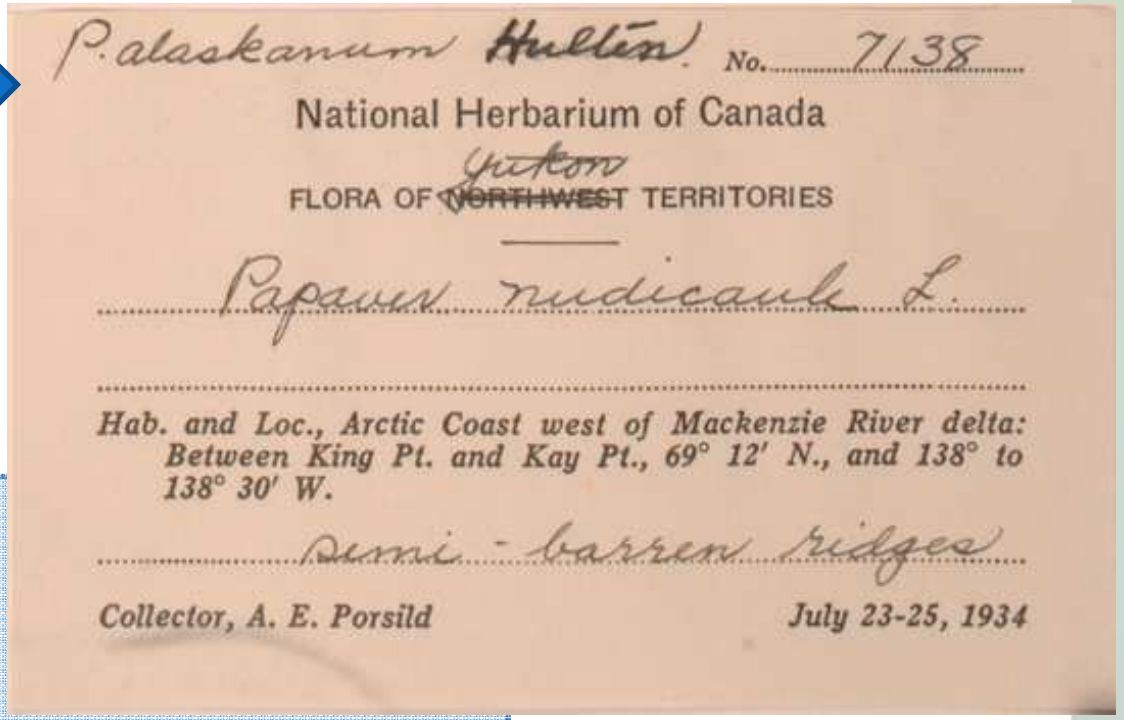
Hab. and Loc., Arctic Coast west of Mackenzie River delta: Between King Pt. and Kay Pt., 69° 12′ N., and 138° to 138° 30′ W.

Dunii - barren ridges

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

iDigBio
Integrated Digitized Biocollections

Label


P. alaskanum Hultén 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.

Dunes - barren ridges

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

OCR

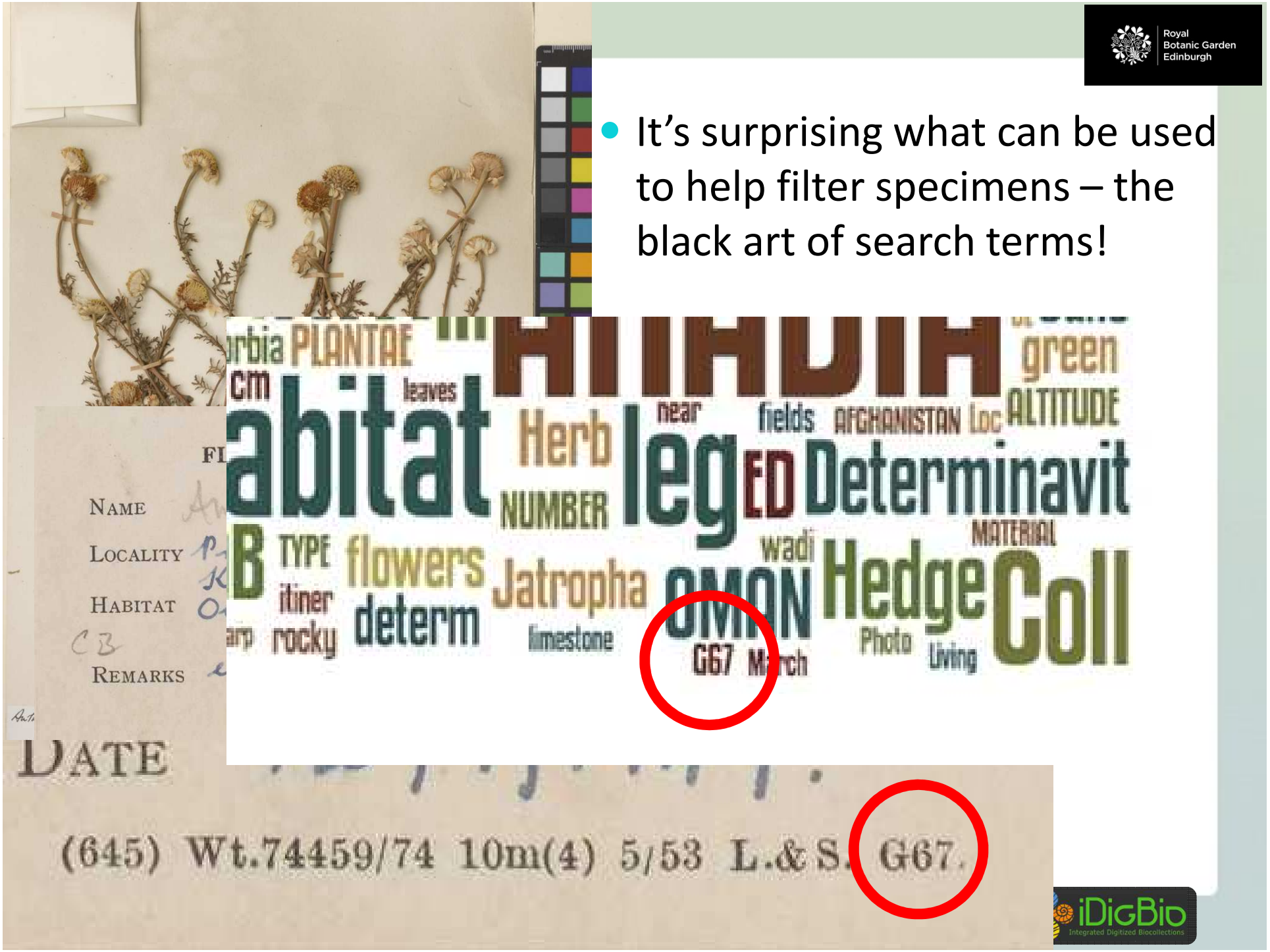
No. ....2L31.
National Herbarium of Canada
FLORA OF'T TERRITORIES
.
Hab. and Loc., Arctic Coast west of Mackenzie River delta:
Between King Pt. and Kay Pt., 69° 12' N., and 138° to 138° 30' W.
. .
Collector, A. E. Porsild July 23-25, 1934

Next imagine output from 1000s of labels!

iDigBio
Integrated Digitized Biocollections

Seeing the dark data…

Robyn E Drinkwater, Robert Cubey, Elspeth Haston at TDWG 2013.



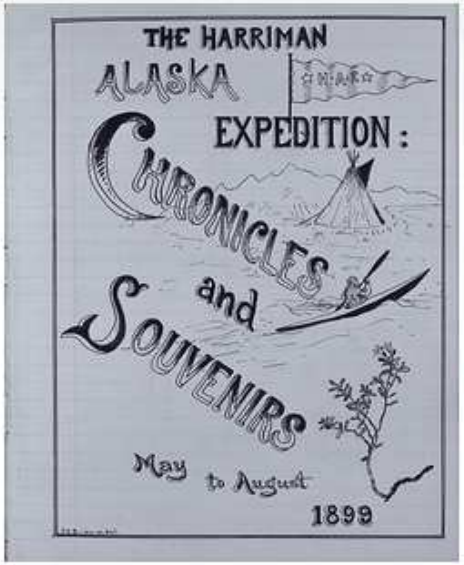
- It's surprising what can be used to help filter specimens – the black art of search terms!

Lichens, Bryophytes and Climate Change

Home | Project | Digitization | Data | Information | Volunteer | Calendar

Lichen Record Sets

Harriman Alaska Expedition



Front page of the expedition album created by its members.

In 1899 an interdisciplinary team of scientists and artists traveled from Seattle to Siberia and back exploring the Alaska coastline. Railroad magnate Edward Harriman commissioned the voyage as a hunting trip but it turned into a scientific expedition. The team he brought along to document the trip included some of the best and brightest minds in the country at the time.

Botanists on the trip included:
William Trelease - director of the Missouri Botanical Garden
De Alton Saunders - phycologist (algae)
Frederick Coville - USDA Chief Botanist and Founder of the United States National Arboretum
Thomas Kearney - research associate in botany at the California Academy of Sciences

Other notable members of the expedition:
John Muir - naturalist
Clinton Hart Merriam - zoologist
John Burroughs - author

Expedition party on a dock in Alaska. Photograph by Edward Curtis. Courtesy of the Missouri Botanical Garden.

Create Your Own Expedition or Record Set Link

Country: All Countries | State/Province: All States

Family: All Families | Genus: All Genera

Collection: All Collections

OCR Fragment:

Click To Go To Records | Reset

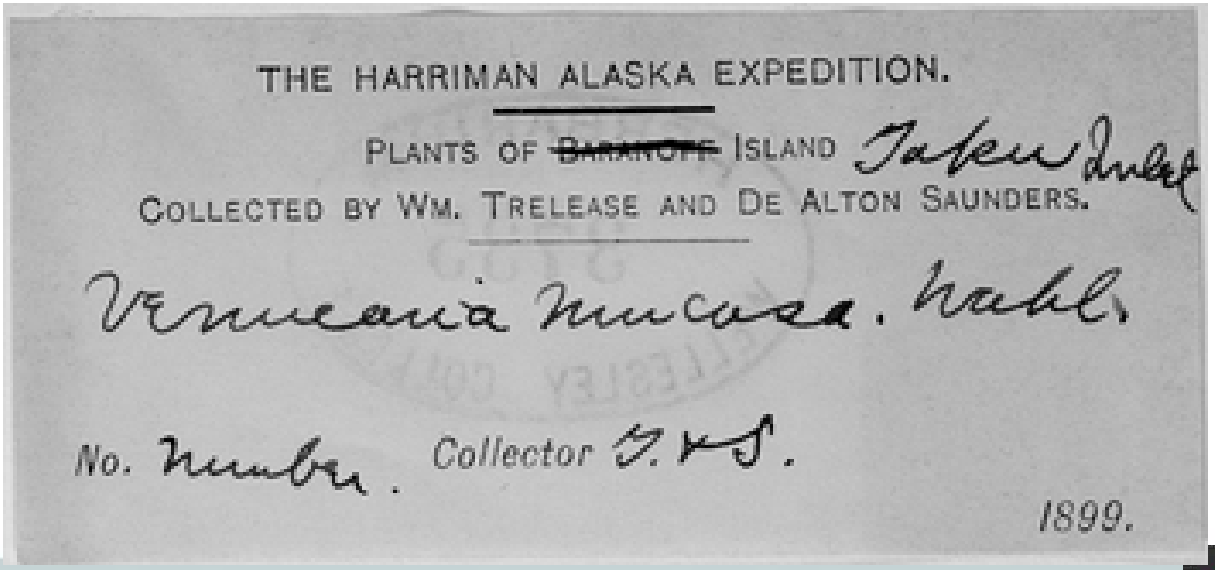
http://tinyurl.com/LichenRecords

iDigBio Integrated Digitized Biocollections

Inside the 1899 Harriman Expedition

Home >> Crowd Sourcing Central >> **Occurrence Record Table View** << | 1-416 of 416 records | >>

Symbiota ID	Catalog Number	Family	Scientific name	Author	Date Last Modified
964616	01394706	Verrucariaceae	*Wahlenbergiella mucosa*	(Wahlenb.) Gueidan & Thüs	2012-03-22 19:08:49
964634	01407608	Nephromataceae	*Nephroma antarcticum*	(Jacq.) Nyl.	2012-03-22 19:08:49
964643	01407614	Nephromataceae	*Nephroma antarcticum*	(Jacq.) Nyl.	2012-03-22 19:08:49
964645	01407669	Nephromataceae	*Nephroma bellum*	(Sprengel) Tuck.	2012-03-22 19:08:49
964649	01407673	Nephromataceae	*Nephroma bellum*	(Sprengel) Tuck.	2012-03-22 19:08:49
964654	01407671	Nephromataceae	*Nephroma bellum*	(Sprengel) Tuck.	2012-03-22 19:08:49
964678	01407616	Nephromataceae	*Nephroma antarcticum*	(Jacq.) Nyl.	2012-03-22 19:08:49



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

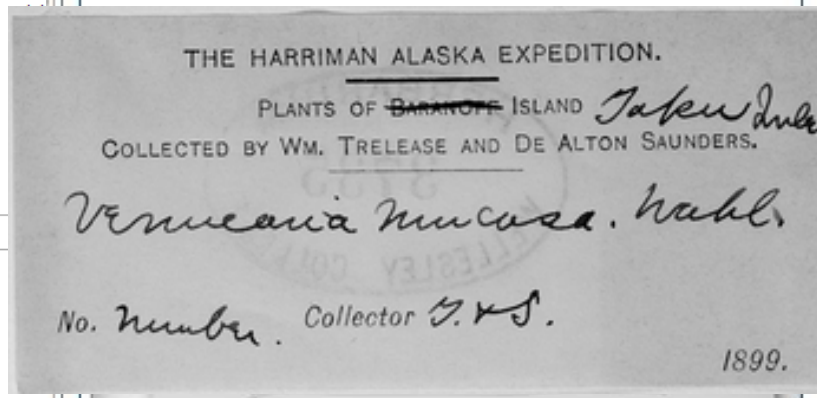
Substrate

Notes

Save Edits

Status Auto-Set: Pending Review

Label Processing



OCR Image

Options

Notes:

Source: ABBYY:2013-02-09

Save OCR Edits

Parse OCR (LBCC)

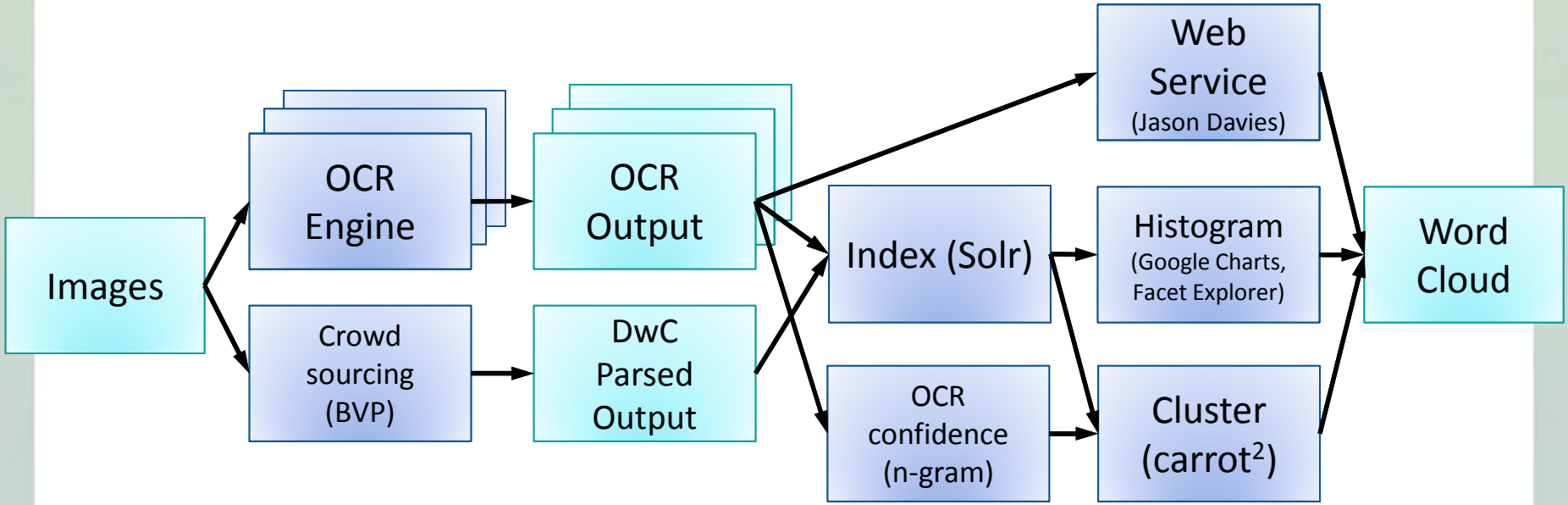
1 of 1

Delete OCR

iDigBio Integrated Digitized Biocollections

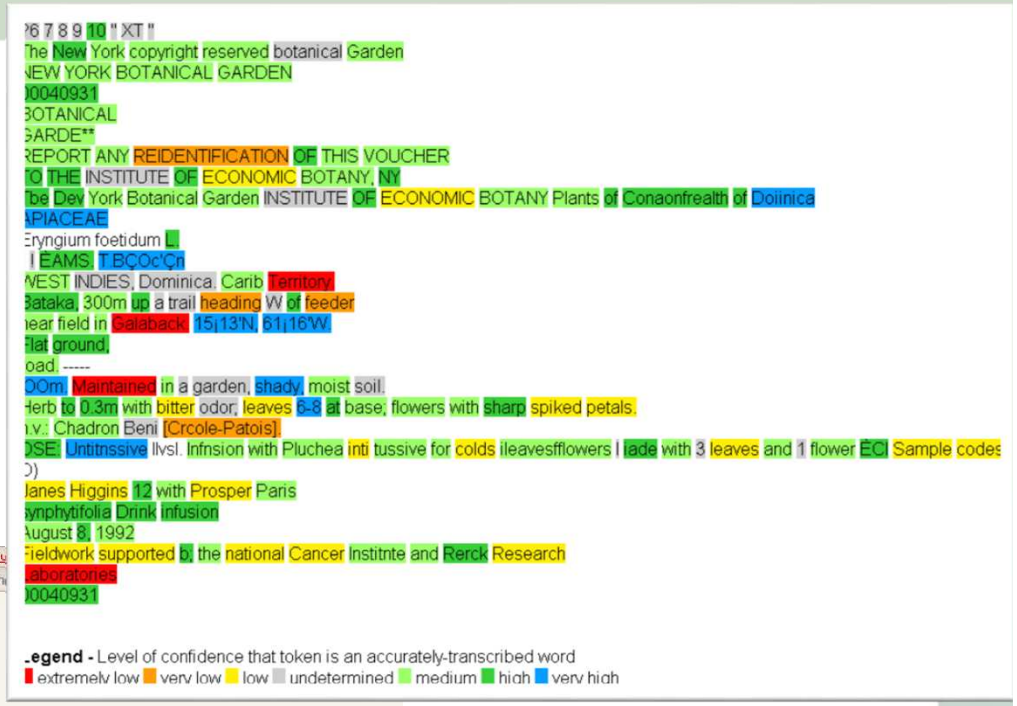
Some work from the recent iDigBio CITSCribe Hackathon

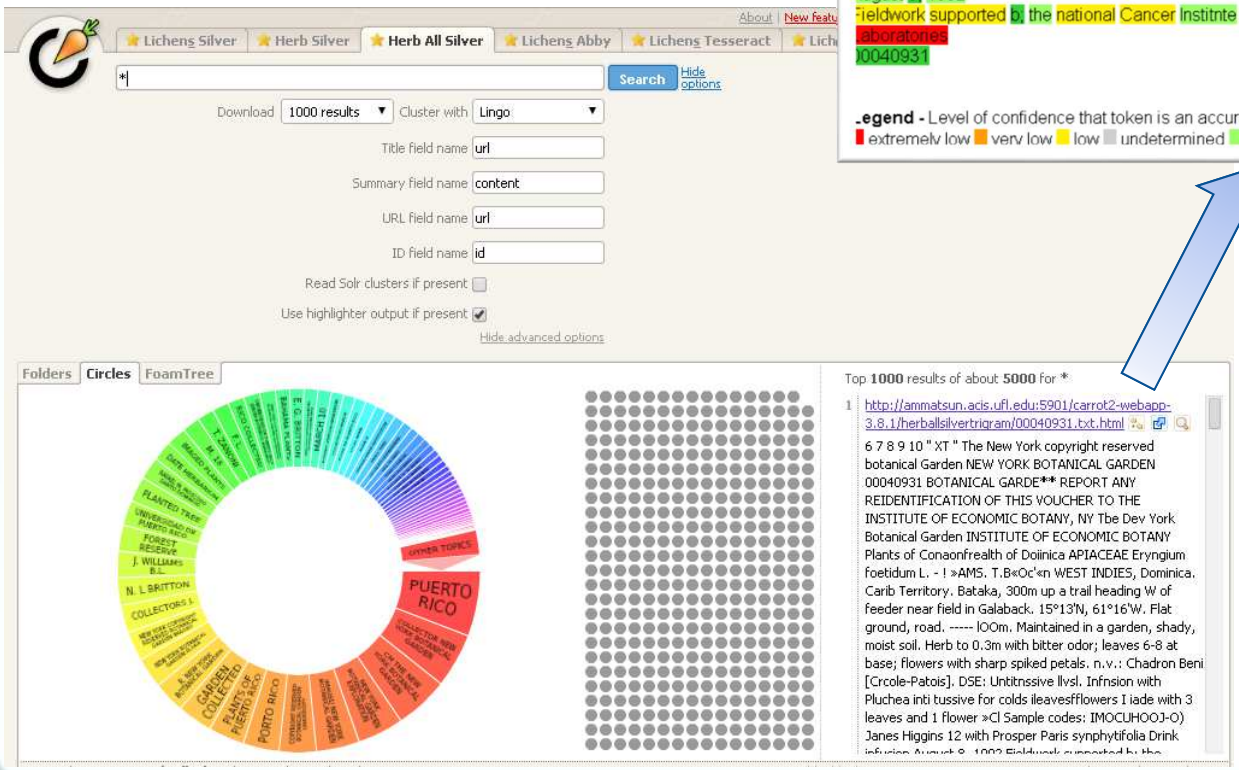
Overall Word Cloud Workflow



Google Charts: http://developers.google.com/chart/interactive/docs/gallery
N-gram: http://github.com/idigbio-citsci-hackathon/OCR-Error-Estimation
Facet explorer: http://github.com/idigbio-citsci-hackathon/facet-explorer
Jason Davies WC: http://www.jasondavies.com/wordcloud/
Apache Solr: http://lucene.apache.org/solr/
carrot²: http://project.carrot2.org/

Word Clouds using N-gram Scoring, Faceting, Solr + Carrot²





Imagine Integration with current software

GOT IT! LET ME TRANSCRIBE

GOT IT! LET ME CHOOSE MY TRANSCRIPTION GROUP

Choose your group:

Country

- Use for initial sort or validation

iDigBio
Integrated Digitized Biocollections

Working Group Collaboration
aOCR + DROID1 OCR workflow, …

Task ID	Task Name	Explanations and Comments	Resources (human, physical tool, software)
		Get help from folks already doing OCR (iDigBio aOCR wg!)	
	Select OCR software	Factors to use in making selection might include: • Cost • Batch processing capability • Output types (XML, Text, etc.) • Language Support • SDK availability • OS platform (Windows, Linux, Mac etc) • APIs • Authority File Support • Watched Folders - allowing use across multiple users. It will be helpful to review the next task to ensure adequate hardware/OS support.	Available OCR software and tools include: Proprietary • ABBYY Finereader • ABBYY Finereader Pro Edition • ABBYY Finereader Corporate Edition • ABBYY Recognition Server • OmniPage • Prime Recognition Open Source • Tesseract • J(G)OCR • OCRopus
	Install OCR software	**Hardware/OS issues** • Ensure that hardware meets requirements of the OCR software • Ensure sufficient memory capacity	

- Setting up OCR
- Running OCR
- Machine Learning
- Natural Language Processing

iDigBio
Integrated Digitized Biocollections

OCR use, a bit more…

- aOCR WG, JRA Synthesys3, …
- Who is using OCR output? …some examples
  - New York Botanical Garden
    - Caribbean Field Notebook Project
  - Royal Botanic Garden Edinburgh
  - Kew
  - Lichen, Bryophyte , Climate Change (LBCC) TCN
  - Biodiversity Heritage Library
- N-gramming (confidence scores)
- parsing algorithm optimization
- user-interface group
- exemplar ML and NLP workflows

iDigBio
Integrated Digitized Biocollections

Work presented here made possible by many and especially…

Mahalo!

- Andrea Matsunaga, Researcher, iDigBio
- Miao Chen, Indiana University, Data to Insight Center
- Jason Best, Botanical Research Institute of Texas
- Sylvia Orli, IT Head, Smithsonian Botany Department
- William Ulate, Technical Director, BHL
- Reed Beaman, Informatics Specialist, iDigBio
- iDigBio Augmenting Optical Character Recognition (aOCR) Working Group

NSF UF UNIVERSITY of FLORIDA FLORIDA STATE UNIVERSITY 1851 FLORIDA MUSEUM OF NATURAL HISTORY

iDigBio Integrated Digitized Biocollections