The New York Botanical Garden LICHENS OF VERMONT, U.S.A.



fyrenidium gregaturn K. Knudsen & Kocourk.
on Phaeophyscjarubropuichea (Segel.) 2531. er

Umpire Mountain, 44°34'27"N, 71°5C'57W, 550 m; Data Discovery and Doer Happiness! northern hardwoods and forested cold air, granitic talus slope.

16 May 2008

Deborah Paul (et al), @idbdeb @iDigBio

.AA[

Symposium IV: Digitization Workflows, Tools, and Techniques Biological Digitization in the Pacific, March 23-27, 2014

Richard C. Harrisst5/4e3t5/8nter, Bishop Museum, University of Hawaii, and the Pacific Science Association

0107574

Colk,tcd on the 5 Crum Bryophyte /urkho











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?



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

* move fun!



Got Text? Got Handwriting?

Palaskanum Hullin) No. 7/38
National Herbarium o	
FLORA OF CORTHWEST TE	ERRITORIES
Papaver mudi	caule L.
Hab. and Loc., Arctic Coast west of Between King Pt. and Kay Pt., 6 138° 30' W.	Mackenzie River delta: 69° 12′ N., and 138° to
semi - barr	en ridges
Collector, A. E. Porsild	July 23-25, 1934





Palaskanum Hullin No. 7138

National Herbarium of Canada

FLORA OF CORTHWEST TERRITORIES

Papaver mudicaule 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.

Demi - barren ridges

Collector, A. E. Porsild

July 23-25, 1934

OCP Y

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!



Seeing the dark data...









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

The leaves Herb leaves of fields afghanistan Loc ALTITUDE of the flowers of the f

DATE

Anta

LOCALITY

HABITAT

REMARKS

(645) Wt.74459/74 10m(4) 5/53 L.&S. G67.



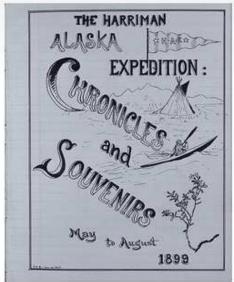


LBCC Lichens, Bryophytes and Climate Change

Project * Home Digitization * Data V 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



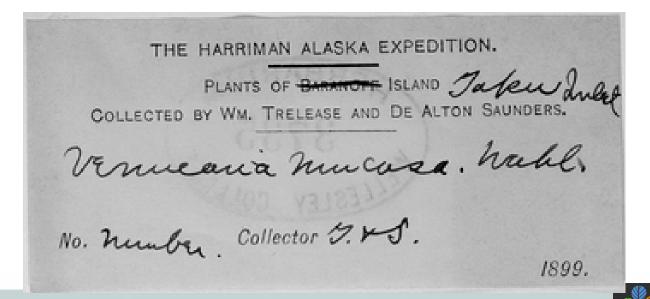
Expedition party on a dock in Alaska. Photograph by Edward Curtis. Courtesy of the Missouri Botanical Garden.





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
<u>964634</u>	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

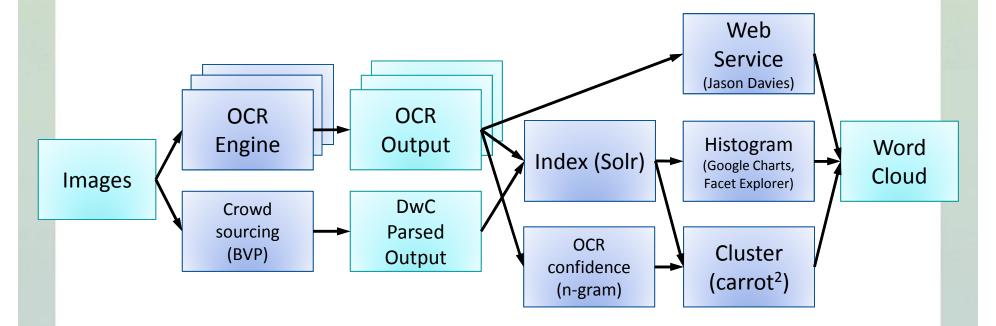


De Alton Saunders Exsiccati Title Number Scientific Name? Wahlenbergiella mucosa Note: Full editing permissions are needed to edit an identification Country State/Province Country Taku Inlet Latitude Longitude Uncertainty? Verbatim Coordinates Elevation in Meters Verbatim Elevation Substrate Notes Notes Notes Substrate Notes Save Edits Notes Collector S. V.S. Notes Notes Notes Notes Notes Notes Save Edits Notes Save Edits Notes Source: ABBYY:2013-02-09 Save OCR Edits Parse OCR (LBCC) 1 of	Collector? Wm. Trelease Associated Collectors? De Alton Saunders Exsiccati Title Scientific Name? Number? Date? Dupes? Auto search Verbatim Date? Number	PLANTS OF BARANOFF ISLAND Jaken In Collected by WM. Trelease and De Alton Saunders.
Collector? Number Date ? Dupes? Wm. Trelease Auto search Plants of Barance Island Jacker	Collector? Wm. Trelease Associated Collectors? De Alton Saunders Exsiccati Title Scientific Name? Number? Date? Dupes? Auto search Verbatim Date? Number	PLANTS OF BANANOFF ISLAND Jaken In Collected by WM. Trelease and De Alton Saunders.
Wm. Trelease Associated Collectors? De Alton Saunders Exsiccati Title Number Scientific Name? Wahlenbergiella mucosa Note Full edding permissions are needed to edit an identification Country State/Province County United States Alaska Juneau Locality Taku Inlet Laitude Longitude Uncertainty? Servetatim Coordinates Verbatim Coordinates Tools Elevation in Meters Verbatim Elevation Substrate Notes Notes Notes Notes Save Edits Notes Save Edits Plants of Banknor Island Jufeur Collector 9: Verbatim Saunders Collected by WM. Trelease and De Alton Saunders Verucearia Mucoca. Nation Country Collector 9: V.f. Notes No. Numbr. Collector 9: V.f. Notes Notes: Source: ABBYY:2013-02-09 Save OCR Edits Parse OCR (LBCC) 1 of	Wm. Trelease Associated Collectors? De Alton Saunders Exsiccati Title Scientific Name? Auto search Verbatim Date? Number	PLANTS OF BANANOFF ISLAND Jaken In Collected by WM. TRELEASE AND DE ALTON SAUNDERS.
Associated Collectors? Verbatim Date? 1899 COLLECTED BY WM. TRELEASE AND DE ALTON SAUNDERS Exsiccati Title Number Scientific Name? Wahlenbergiella mucosa Note Full editing permissions are needed to edit an identification Country United States Alaska Juneau Locality Taku Inlet Latitude Longitude Uncertainty? Verbatim Coordinates Tools Says 255678 134.078102 17128 Verbatim Coordinates Tools Substrate Notes: Notes: Notes: Source: ABBYY:2013-02-09 Save Edits PLANTS OF PLANTON SAUNDERS Collector 9: MS. No. Number No. Number No. Number Collector 9: MS. No. Number Collector 9: MS. No. Number No. Number Collector 9: MS. No. Number No. Number No. Number Collector 9: MS. No. Number Collector 9: MS. No. Number No. Numbe	Associated Collectors? De Alton Saunders Exsiccati Title Scientific Name?	Venuearia Mucasa. habl.
De Alton Saunders Exsiccati Title Number Number No. Number Collector O. F.J. No. Number Collector O	De Alton Saunders Exsiccati Title Number Scientific Name ?	Vermeana mucasa. habl.
Scientific Name? Wahlenbergiella mucosa Note. Full editing permissions are needed to edit an identification Country State/Province County United States Alaska Juneau Locality Taku Inlet Latitude Longitude Uncertainty? Verbatim Coordinates Tools Elevation in Meters Verbatim Elevation Substrate Notes: Notes Notes Save Edits No. Numbr. Collector 9: 15. No. Numbr. Collector 9: 15. No. Numbr. Collector 9: 15. Notes Notes Save Edits Notes: Source: ABBYY:2013-02-09 Save OCR Edits Parse OCR (LBCC) 1 of	Scientific Name ?	Vermeania mucasa. habl.
Scientific Name? Wahlenbergiella mucosa Note: Full editing permissions are needed to edit an identification Country State/Province County United States Locality Taku Inlet Latitude Longitude Uncertainty? Verbatim Coordinates Elevation in Meters Verbatim Elevation Substrate Notes Notes Notes Save Edits No. Numbr. Collector 9: 15. No. Numbr. Collector 9: 15. Notes Notes Notes Source: ABBYY:2013-02-09 Save OCR Edits Parse OCR (LBCC) 1 of	Scientific Name ?	No 2. La Collector T. V.S.
Country State/Province Country United States Alaska Juneau Locality Taku Inlet Latitude Longitude Uncertainty? Verbatim Coordinates 58.255678 -134.078102 17128		No 2 Collector 7 . J.
Country State/Province Country United States Alaska Juneau Locality Taku Inlet Latitude Longitude Uncertainty? Verbatim Coordinates 58.255678 -134.078102 17128		No 2 Collector 5 MJ.
Country State/Province County United States Alaska Uncality Taku Inlet Latitude Longitude Uncertainty? Verbatim Coordinates Verbatim Coordinates Verbatim Elevation Substrate Notes Notes Save Edits Save Edits Note State Province County Verbatim Coordinates OCR Image Options Notes: Source: ABBYY:2013-02-09 Save OCR Edits Parse OCR (LBCC) 1 of	wanienperdiella mucosa	
Country United States Alaska Uneau Locality Taku Inlet Latitude Longitude Uncertainty? Verbatim Coordinates Says 255678 Verbatim Elevation Substrate Notes Notes Save Edits Save Edits Alaska Juneau Verbatim Coordinates Tools OCR Image Options Notes: Save OCR Edits Parse OCR (LBCC) 1 of		
Locality Taku Inlet Latitude Longitude Uncertainty? Verbatim Coordinates 58.255678 -134.078102 17128		1899.
Taku Inlet Latitude Longitude Uncertainty? Verbatim Coordinates 58.255678 -134.078102 17128	United States Alaska Juneau	
Latitude Longitude Uncertainty? Verbatim Coordinates Tools Save Edits Verbatim Elevation in Meters Verbatim Elevation Verbatim Elevation OCR Image Options Clara E. Cummings Herbarium of Wolfestby/ College (WELC) Deposited at NY in 1988 Notes: Source: ABBYY:2013-02-09 Save OCR Edits Parse OCR (LBCC) 1 of	Locality	
Elevation in Meters Verbatim Elevation Habitat Notes: Save Edits Clara E Cummings Herbarium of Wolfesley College (WELC) Deposited at NY in 1988 OCR Image Options Notes: Save Edits Save Edits Parse OCR (LBCC) 1 of	·	
Save Edits Clara E Cummings Herbarium of Wolfesley College (WELC) Deposited at NY in 1988 OCR Image Options Notes: Save Edits Save Edits Save Edits ABBYY:2013-02-09 Save OCR Edits Parse OCR (LBCC) 1 of	Tools	
Elevation in Meters Verbatim Elevation -	Latitude Longitude Oncertainty Verbatiri Coordinates	
Notes: Source: ABBYY:2013-02-09 Save Edits Parse OCR (LBCC) 1 of	58.255678 -134.078102 17128	
Habitat OCR Image Options Notes: Substrate Source: ABBYY:2013-02-09 Save Edits Parse OCR (LBCC) 1 of	Flavation in Maters Verhatim Flavation	
Notes: Source: ABBYY:2013-02-09 Save Edits Parse OCR (LBCC) 1 of		
Notes: Source: ABBYY:2013-02-09 Save Edits Parse OCR (LBCC) 1 of		OOD Image Ontions
Source: ABBYY:2013-02-09 Save Edits Parse OCR (LBCC) 1 of	Habitat	OCR Image
Source: ABBYY:2013-02-09 Save OCR Edits Parse OCR (LBCC) 1 of	Not	tes:
ABBYY:2013-02-09 Save OCR Edits Parse OCR (LBCC) 1 of	Substrate	
ABBYY:2013-02-09 Save OCR Edits Parse OCR (LBCC) 1 of	Soi	urce:
Save OCR Edits Parse OCR (LBCC)	Notes	
Save OCR Edits Parse OCR (LBCC)		4-54
	Save Edits	Save OCR Edits Parse OCR (LBCC)
	Status Auto-Set: Pending Review ▼	



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

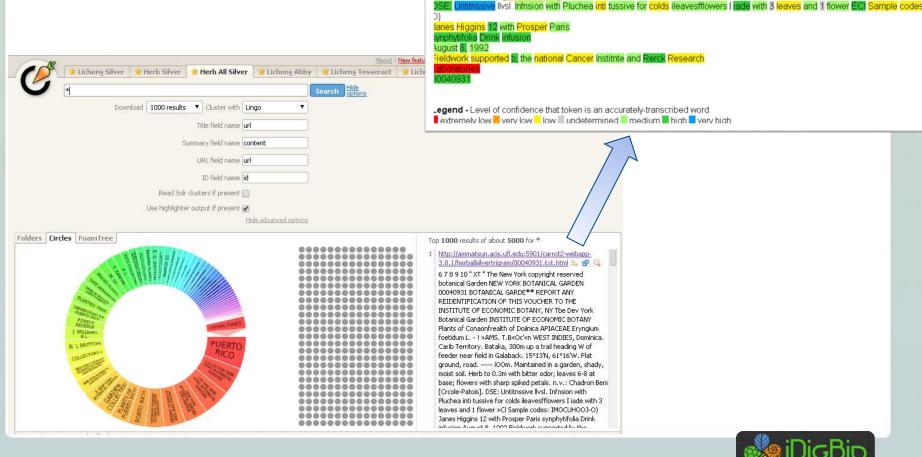
Jason Davies WC: http://www.jasondavies.com/wordcloud/

Apache Solr: http://lucene.apache.org/solr/

carrot²: http://project.carrot2.org/







The New York copyright reserved botanical Garden

REPORT ANY REIDENTIFICATION OF THIS VOUCHER
TO THE INSTITUTE OF ECONOMIC BOTANY, NY

ined in a garden, shady, moist soil.

erb to 0.3m with bitter odor; leaves 6-8 at base; flowers with sharp spiked petals.

ev York Botanical Garden INSTITUTE OF ECONOMIC BOTANY Plants of Conaonfrealth of D

NEW YORK BOTANICAL GARDEN

VEST INDIES, Dominica. Carib <mark>Ferril</mark> Lataka, 300m up a trail heading W of

Eryngium foetidum

Imagine Integration with current software

GOT IT! LET ME TRANSCRIBE

GOT IT! LET ME CHOOSE MY TRANSCRIPTION GROUP

Choose your group:

 Use for initial sort or validation







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 vg!)	
	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
	Install OCR software	Ensure that hardware meets requirements of the OCR software Ensure sufficient memory capacity	

- Setting up OCR
- RunningOCR

- Machine Learning
- Natural Language Processing



OCR use, a bit more...

Got Handwriting?

- 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



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















