



Enhancing NfN using Text Analytics and Visualization

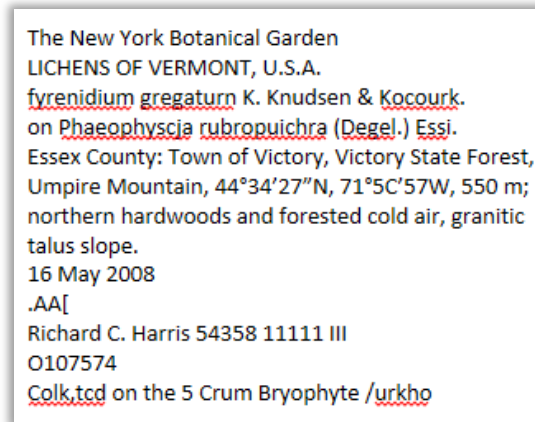
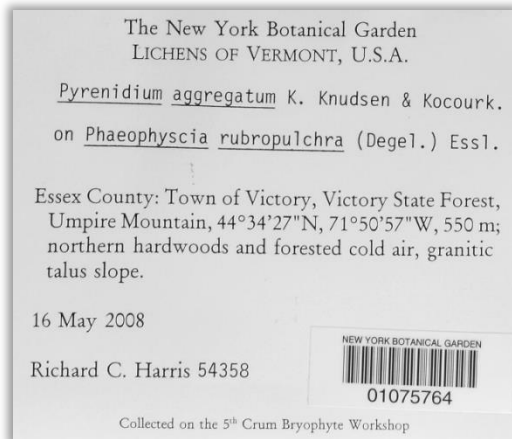
Deb Paul, Andrea Matsunaga, Miao Chen, Jason
Best, Reed Beaman, Sylvia Orli, William Ulate

iDigBio – Notes From Nature Hackathon December 2013
Increasing Citizen Science Participation in Museum Specimen Digitization

Text Clusters

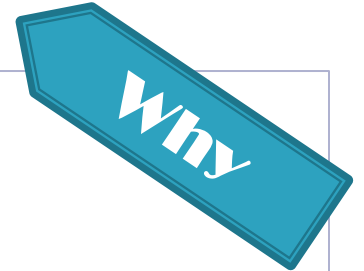
What

- ▶ Preprocess specimen label images with OCR



- ▶ Remove (and use!) noise from text
- ▶ Utilize OCR text
 - create word cloud linked to record ids
 - differentiate hand-written from typed labels
- ▶ Allow transcribers to choose terms from word cloud to create individual sets
- ▶ Allow validators to choose sets to clean

Reasons for Cluster Methodology



- ▶ Enhance user experience
- ▶  **User Happiness!**
- ▶ Leverage user expertise
- ▶ Improve speed
- ▶ Reduce Errors
- ▶ Enables ditto function

User Stories

Who

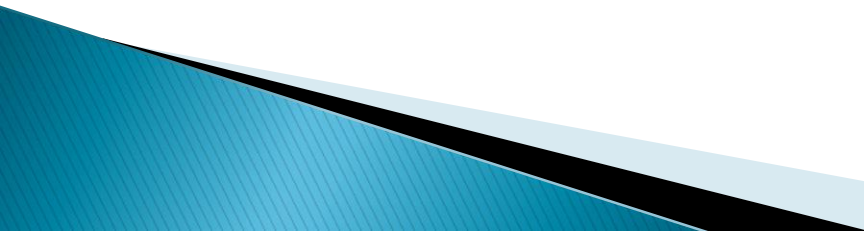
- ▶ Users like ordered datasets
- ▶ Transcription
 - **faster** with ordered/sorted sets
 - **less error prone** with sorted sets

Handwriting vs. Typed

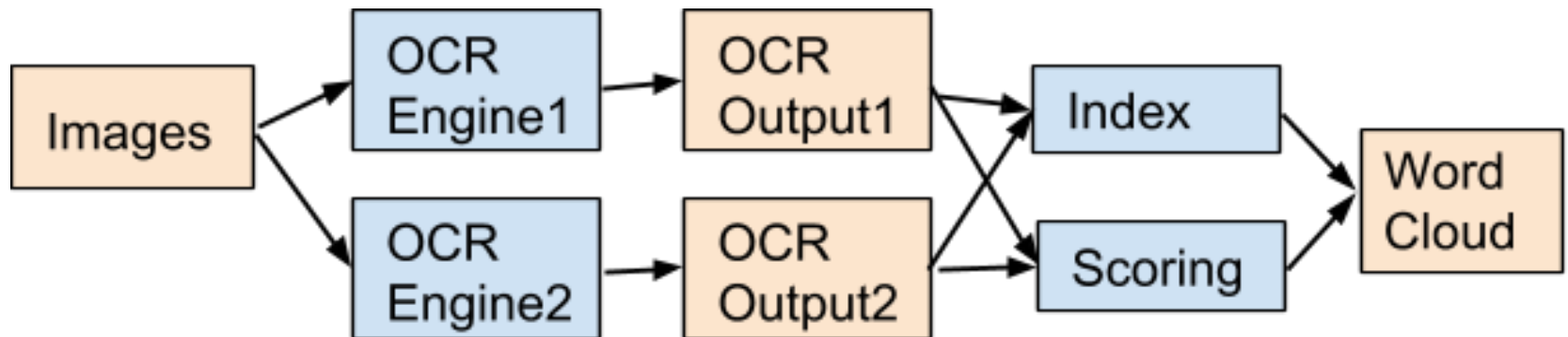
Threshold	Correct	False Positives	False Negatives
$T > 1$ and $N > 20\%$	82%	10 of 45	8 of 60
$T > 0$ and $N > 20\%$	84%	13 of 45	4 of 60
$T > 1$	79%	10 of 45	12 of 60
$N > 20\%$	75%	8 of 45	18 of 60
$N > 10\%$	81%	14 of 45	6 of 60

- ▶ **Segregate hand-written from typed labels**
- ▶ Ben Brumfield code uses **regex** to sort out garbage (higher garbage = higher likelihood hand-written)
 - Read all about it at [Ben's blog!](#)
 - Code is at GitHub
 - Humanities community using now!
- ▶ Let transcriber choose label format
- ▶ *Typed?....go to word cloud workflow*

Estimating OCR confidence

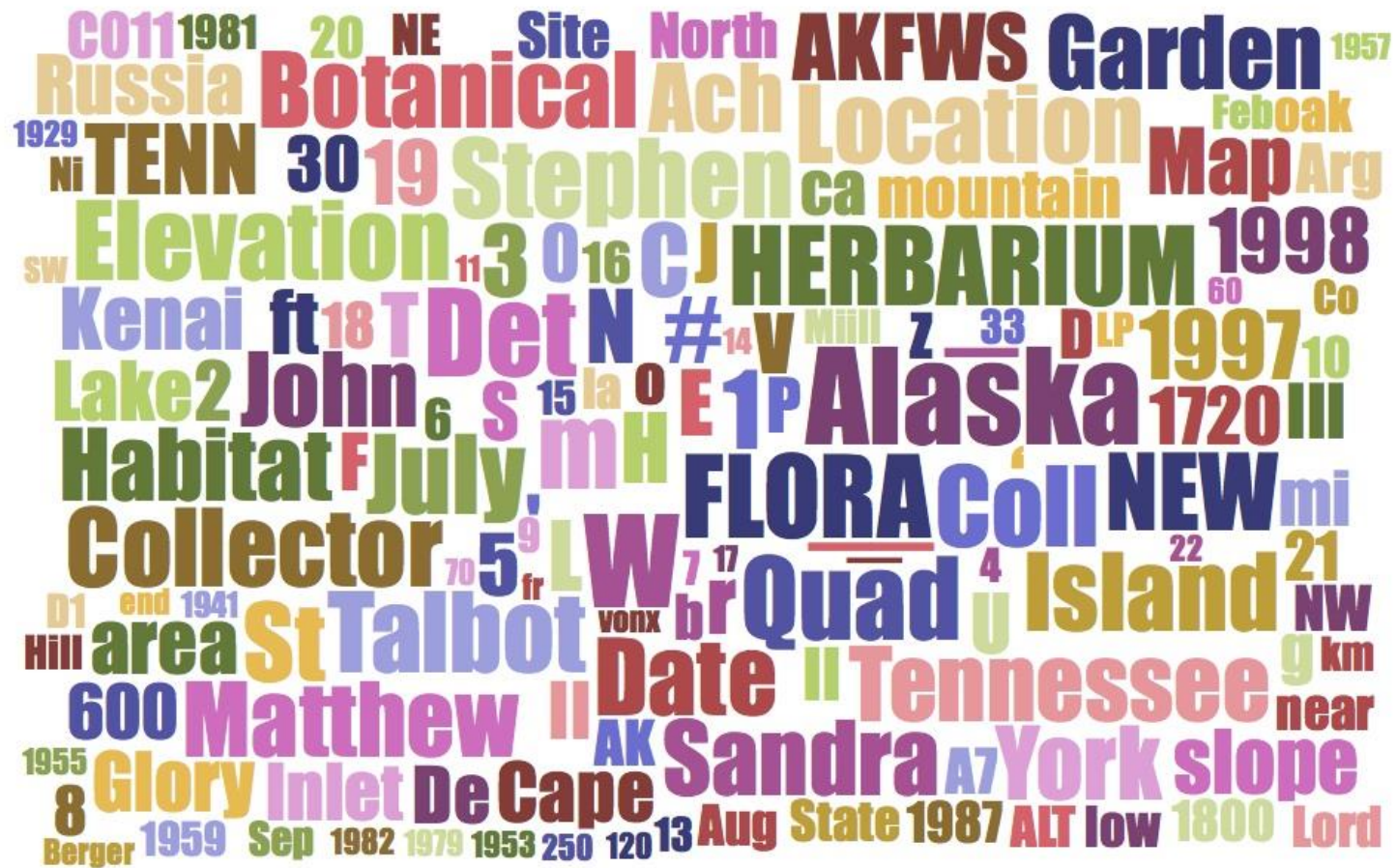
- ▶ extract character-level n-gram from a corpus (the OCR corpus + an external good corpus, ideally)
 - ▶ obtain a list of character-level n-gram
 - e.g. bi-gram looks like th, sh, ph ...
 - ▶ given a word, compute its probability based on the n-gram probability
 - ▶ this is used for computing the final OCR confidence score
 - ▶ can use standard dictionary in computing the score (if time allows)
- 

Word Cloud Workflow

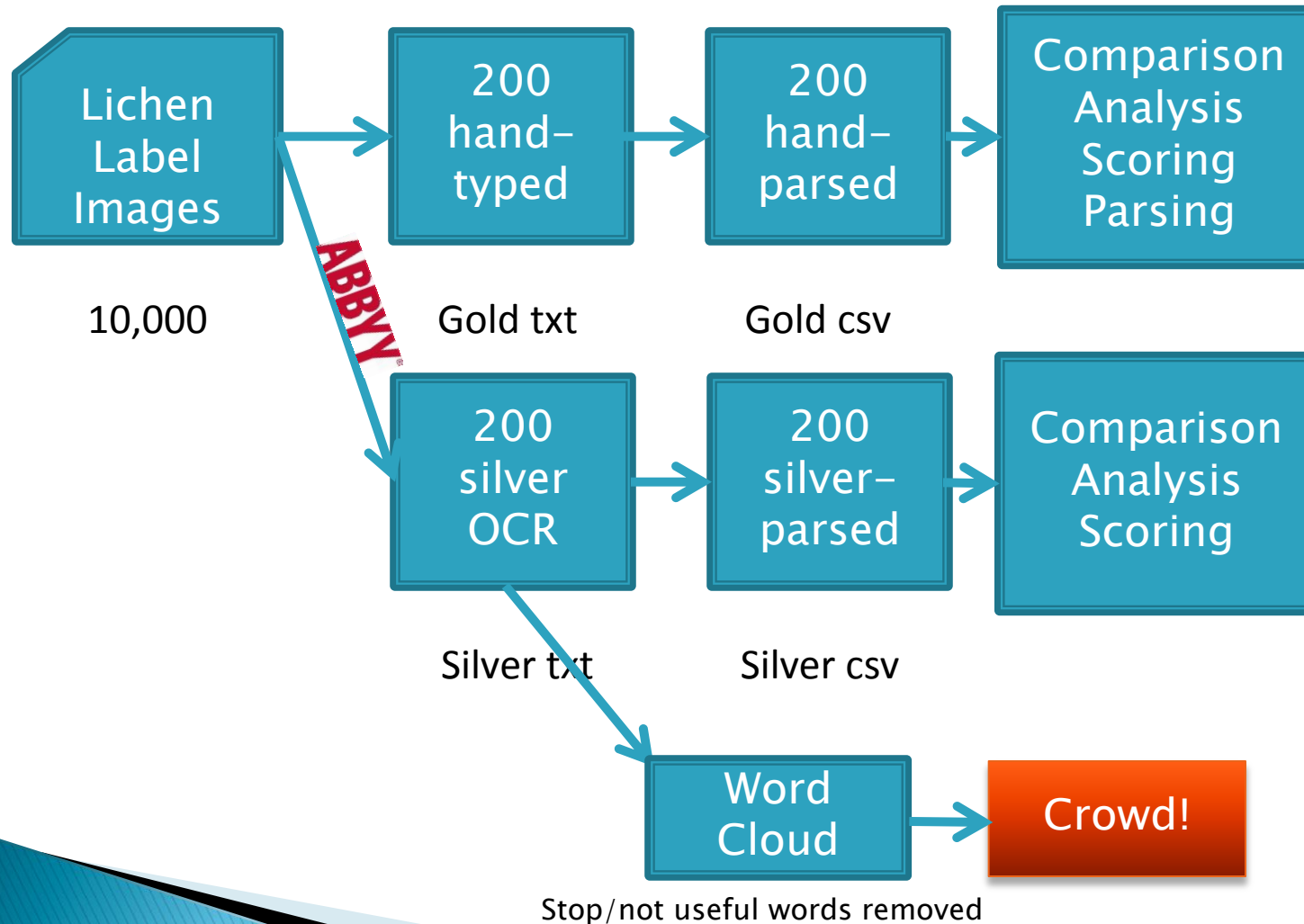


Word Cloud

<http://aocr1.acis.ufl.edu/datasets/lichens/silver/ocr/WebrootDatasetsLichensSilverOcr.txt>



aOCR iDigBio – BRIT Hackathon Lichen Dataset



1000 /3201 Top-Terms: ?

68 alaska

62 coll
county

56 talbot

55 ach
july
stephen

54 new

53 garden
york

51 tenn

49 lichens
tennessee

48 university

47 john
island

43 quad

39 thomson

35 ft

34 sandra

33 longitude
cinchona
map

32 latitude
akfws

29 mi
u.s.a

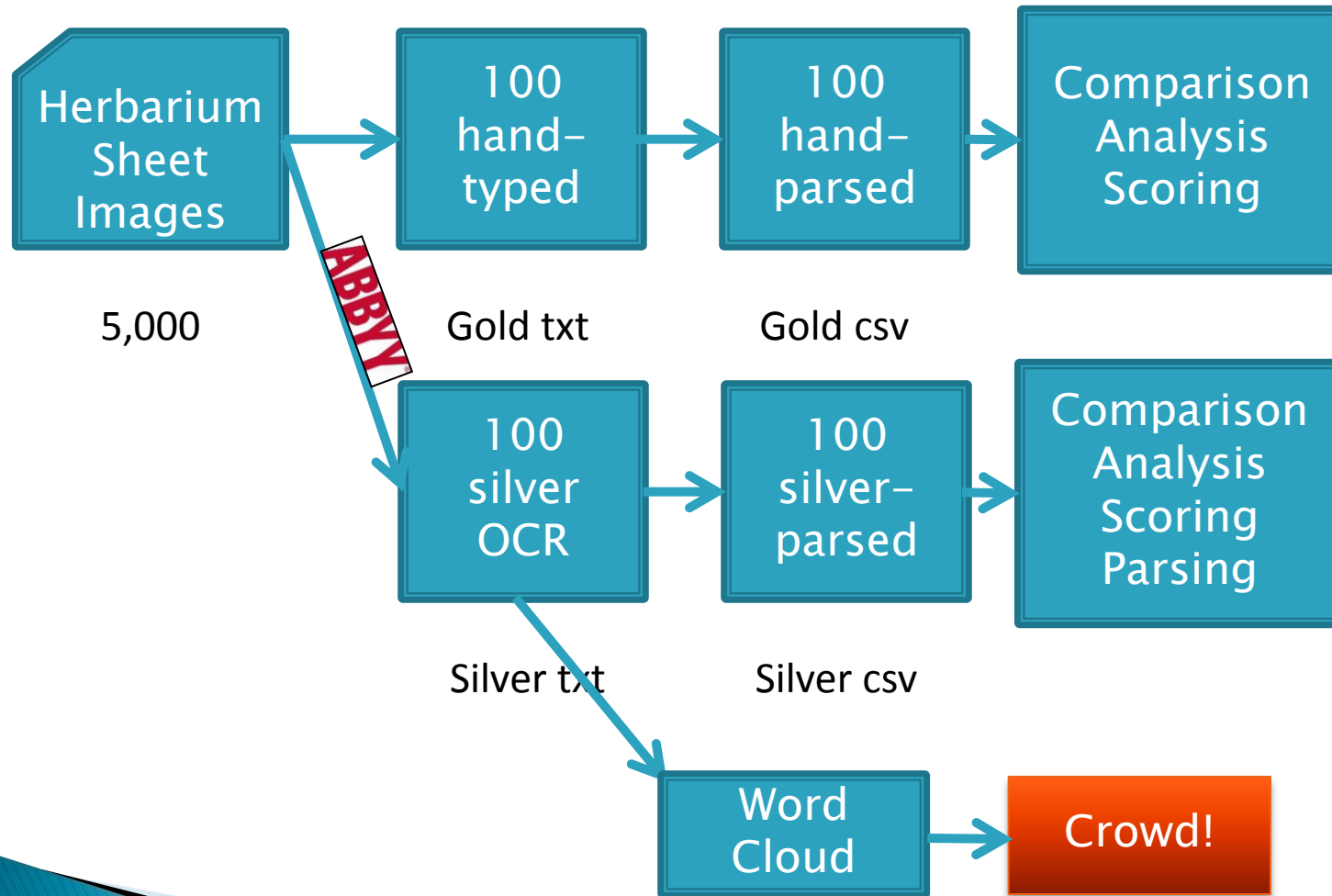
28 19
forest
co

27 arg

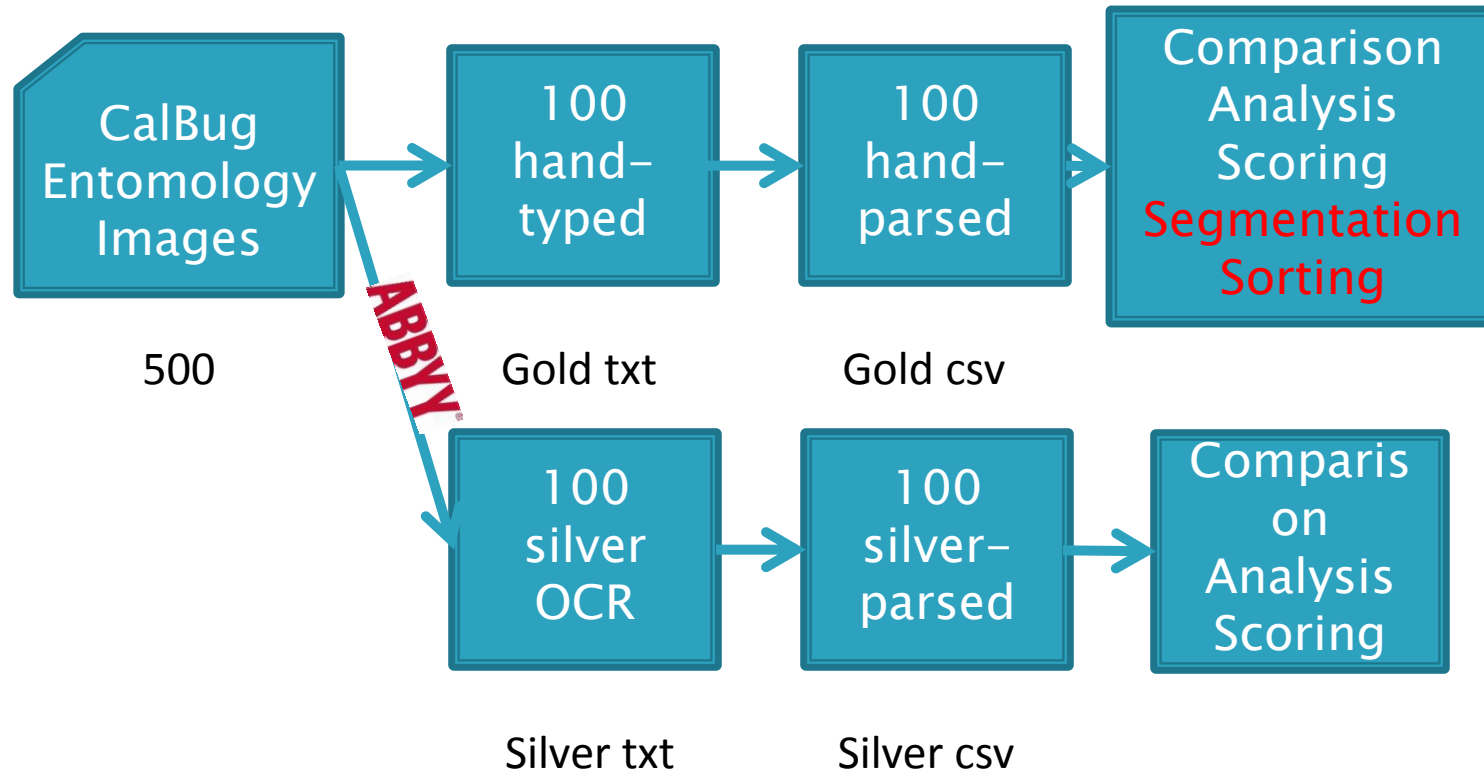
25 ll
lake

23 bay

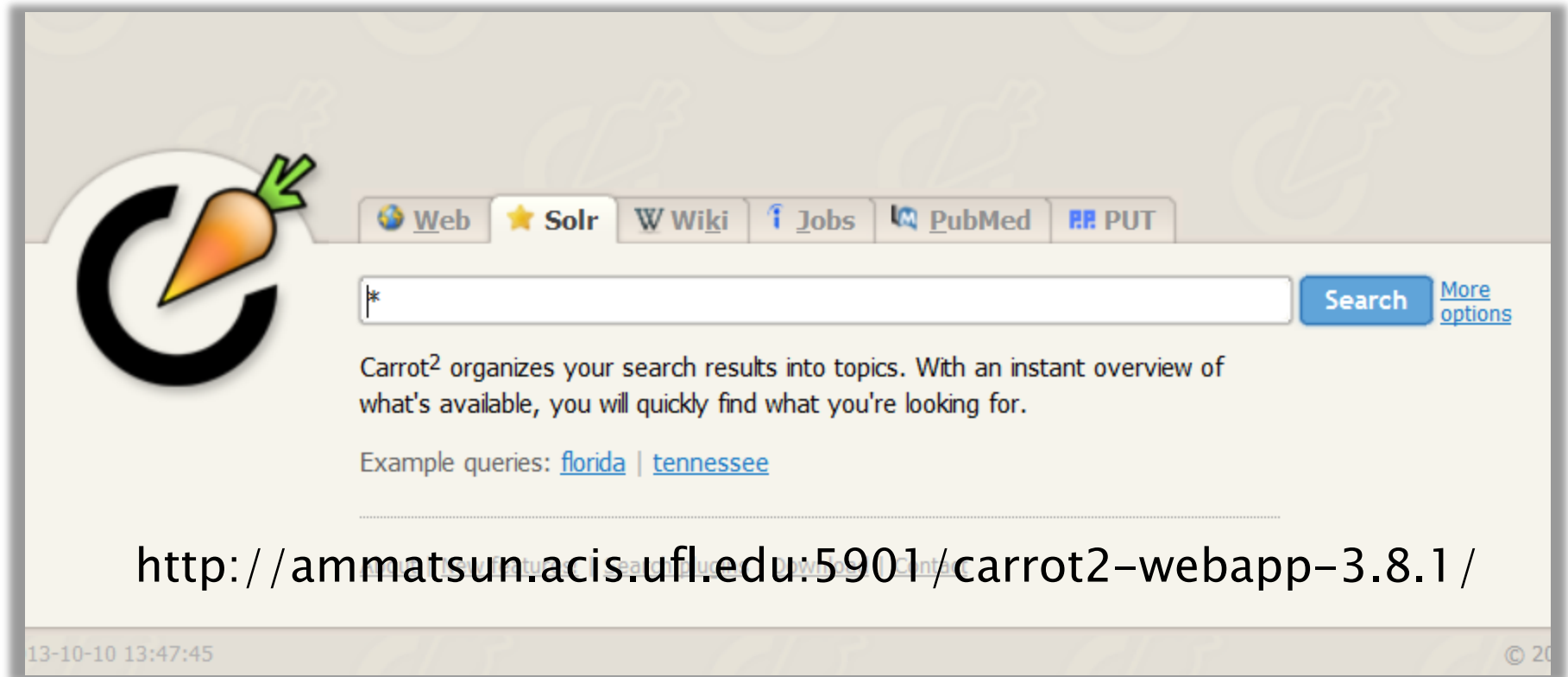
aOCR iDigBio – BRIT Hackathon NYBG Herbarium Dataset



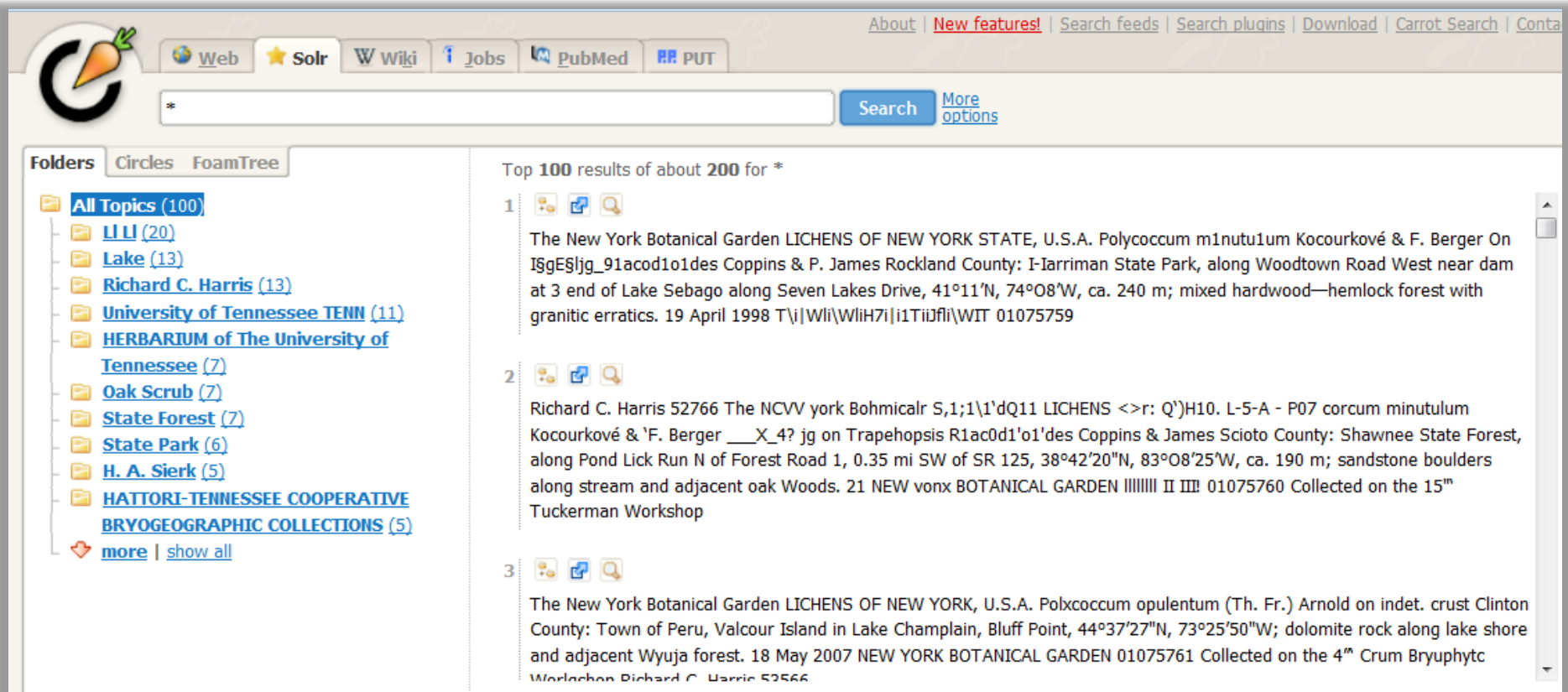
aOCR iDigBio – BRIT Hackathon CalBug Entomology Dataset



Word Cloud using Solr + Carrot²



Folder View of Search



The image shows a screenshot of the Carrot Search web application. The interface includes a navigation bar with links for Web, Solr, Wiki, Jobs, PubMed, and PUT. A search bar is present with a search button and a link to more options. The main content area is divided into two sections: Folders and Search Results. The Folders section on the left lists various categories like All Topics, Lake, Richard C. Harris, University of Tennessee, and others. The Search Results section on the right displays the top 100 results for a search query, showing details for three specific results related to lichens in New York State.

Carrot Search interface showing a folder view of search results.

Navigation links: [About](#) | [New features!](#) | [Search feeds](#) | [Search plugins](#) | [Download](#) | [Carrot Search](#) | [Contact](#)

Search bar: [Search](#) [More options](#)

Folders: [Circles](#) [FoamTree](#)

All Topics (100)

- [All Topics \(100\)](#)
- [Lake \(13\)](#)
- [Richard C. Harris \(13\)](#)
- [University of Tennessee TENN \(11\)](#)
- [HERBARIUM of The University of Tennessee \(7\)](#)
- [Oak Scrub \(7\)](#)
- [State Forest \(7\)](#)
- [State Park \(6\)](#)
- [H. A. Sierk \(5\)](#)
- [HATTORI-TENNESSEE COOPERATIVE BRYOGEOGRAPHIC COLLECTIONS \(5\)](#)
- [more](#) | [show all](#)

Top 100 results of about 200 for *

-   
The New York Botanical Garden LICHENS OF NEW YORK STATE, U.S.A. Polycoccum minutulum Kocourková & F. Berger On IsgEsljg_91acod1o1des Coppins & P. James Rockland County: I-Iarriman State Park, along Woodtown Road West near dam at 3 end of Lake Sebago along Seven Lakes Drive, 41°11'N, 74°08'W, ca. 240 m; mixed hardwood—hemlock forest with granitic erratics. 19 April 1998 T\i|Wli\WliH7i|i1Tiüfli\WIT 01075759
-   
Richard C. Harris 52766 The NCVV york Bohmicalr S,1;1'dQ11 LICHENS <>r: Q')H10. L-5-A - P07 corcum minutulum Kocourková & 'F. Berger ____X_4? jg on Trapehopsis R1ac0d1'o1'des Coppins & James Scioto County: Shawnee State Forest, along Pond Lick Run N of Forest Road 1, 0.35 mi SW of SR 125, 38°42'20"N, 83°08'25'W, ca. 190 m; sandstone boulders along stream and adjacent oak Woods. 21 NEW vonx BOTANICAL GARDEN IIIIIII II III 01075760 Collected on the 15th Tuckerman Workshop
-   
The New York Botanical Garden LICHENS OF NEW YORK, U.S.A. Polxcoccum opulentum (Th. Fr.) Arnold on indet. crust Clinton County: Town of Peru, Valcour Island in Lake Champlain, Bluff Point, 44°37'27"N, 73°25'50"W; dolomite rock along lake shore and adjacent Wyuja forest. 18 May 2007 NEW YORK BOTANICAL GARDEN 01075761 Collected on the 4th Crum Bryophyte Workshop Richard C. Harris 52566

Circles Visualization of Search

Circles FoamTree



Top 100 results of about 200 for *

1

The New York Botanical Garden LICHENS OF NEW YORK STATE, U.S.A. Polycoccum minutulum Kocourkové & F. Berger On IşgEşljg_91acod1o1des Coppins & P. James Rockland County: I-Iarriman State Park, along Woodtown Road West near dam at 3 end of Lake Sebago along Seven Lakes Drive, 41°11'N, 74°08'W, ca. 240 m; mixed hardwood—hemlock forest with granitic erratics. 19 April 1998 T\i\Wli\WliH7i|i1Tiüfli\WIT 01075759

2

Richard C. Harris 52766 The NCVV york Bohmicalr S,1;1\1'dQ11 LICHENS <>r: Q')H10. L-5-A - P07 corcum minutulum Kocourkové & 'F. Berger ____X_4? jg on Trapehopsis R1ac0d1'o1'des Coppins & James Scioto County: Shawnee State Forest, along Pond Lick Run N of Forest Road 1, 0.35 mi SW of SR 125, 38°42'20"N, 83°08'25"W, ca. 190 m; sandstone boulders along stream and adjacent oak Woods. 21 NEW vonx BOTANICAL GARDEN IIIIIIII II IIII 01075760 Collected on the 15th Tuckerman Workshop

3

The New York Botanical Garden LICHENS OF NEW YORK, U.S.A. Polxcoccum opulentum (Th. Fr.) Arnold on indet. crust Clinton County: Town of Peru, Valcour Island in Lake Champlain, Bluff Point, 44°37'27"N, 73°25'50"W; dolomite rock along lake shore and adjacent Wyuja forest. 18 May 2007 NEW YORK BOTANICAL GARDEN 01075761 Collected on the 4th Crum Bryophytic Workshop Richard C. Harris 53566

Foam Tree Visualization of Search



OCR Confidence Estimation

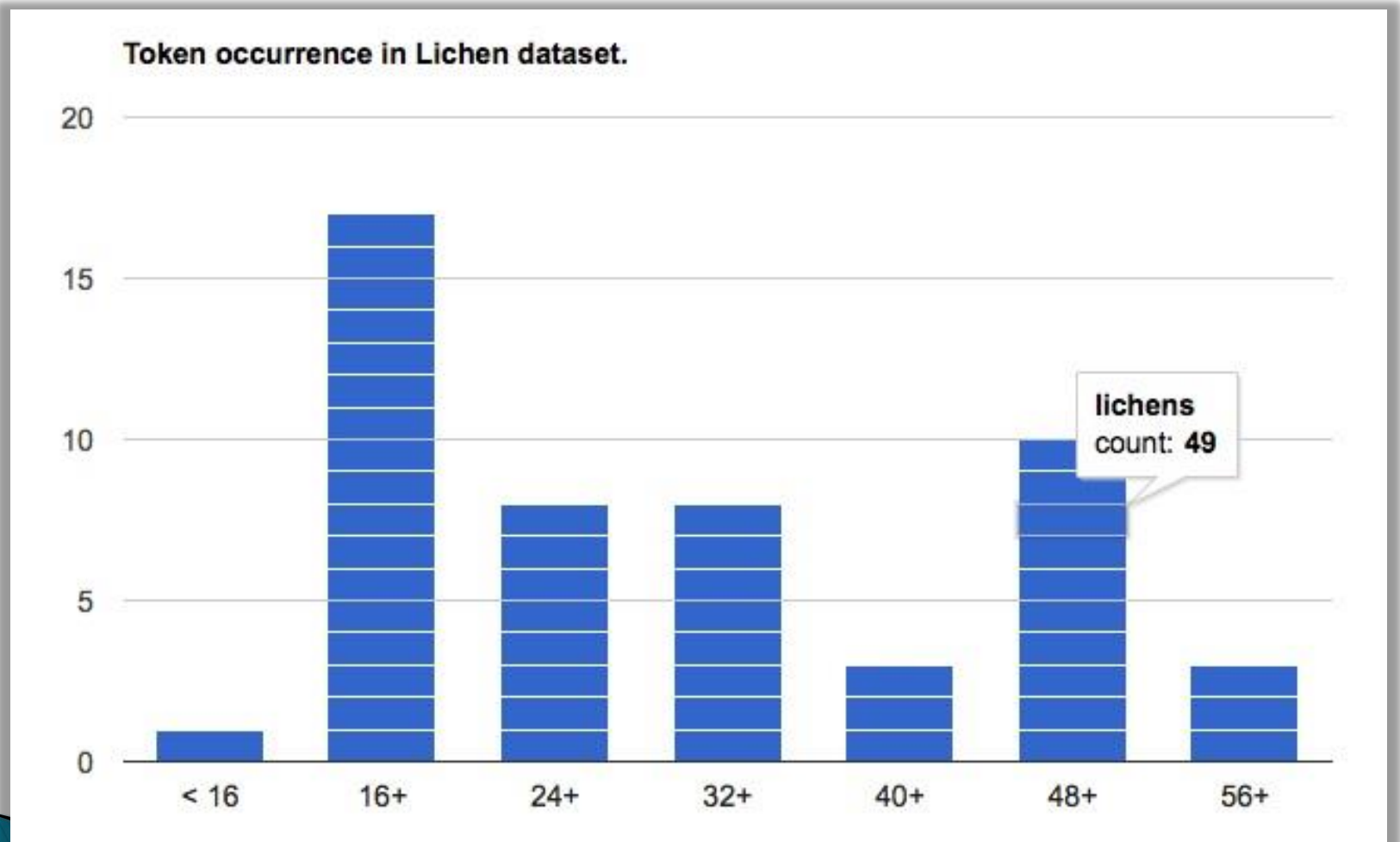
00040904.txt

REPORT ANY REID TO THE INSTITUTE
SA«.f / b-fs
The Key York Botanical Garden INSTITUTE OF ECONOMIC BOTANY Plaits of Cononiealtli of
Domica
uu.«./«. (Jicj ARACEAE
Anthurium -iacant» d<th //■ T. 93
Hooker\ K'intU - M.4.r. T. Croq <* ?3
WEST INDIES, Dominica. St.George. Above new steel bridge on road to Freshwater Lake.
15°19'N/ 61°19'W. 765m. Advanced secondary montane rainforest.
Herb, terrestrial, forming a rosette; leaves 1.3m long x 0.35m wide: roots also growing
upward in debris between leaves.
NEW YORK BOTANICAL GARDEN
00040904
WmEE
ya
BOTANICAL
GARDE**,
Eirik Stijfhoorn 776 May 11, 1992
with G. A. Eidesen and H. T. Beck
Fieldwork supported by the lational Cancer Institute
fc~T7 r.....-
6 7 8 9 10 the New York copyright reserved botanical garden
■i
00040904

00040904.txt

T.	0.0
in	0.0
New	-5.79936762254632
The	-6.168029170890025
NEW	-6.224911973560464
THE	-6.665716448113937
and	-6.879143737870735
the	-7.215488756121444
May	-8.971389551796559
ANY	-9.908514370573556
11,	-10.080364627500215
new	-10.32722470543174
776	-11.103753494930737
??i	-11.600190381244628
York	-11.630983777700038
YORK	-12.402768064352813
//?	-12.852953349739996
Key	-13.140635422191778
long	-16.23802134947227
with	-16.30193397956399
road	-16.647676382489575
WEST	-17.480018803132992
1992	-18.01196937397072
also	-19.392698384818193
Croq	-22.345097378723125
Garden	-22.453229125763336
Beck	-23.285538570829566
GARDEN	-23.61087758272402
Herb,	-24.92388863748758
qarden	-25.276526293980815

Token Histogram – Google Charts



Using OCR output to enhance the transcription process



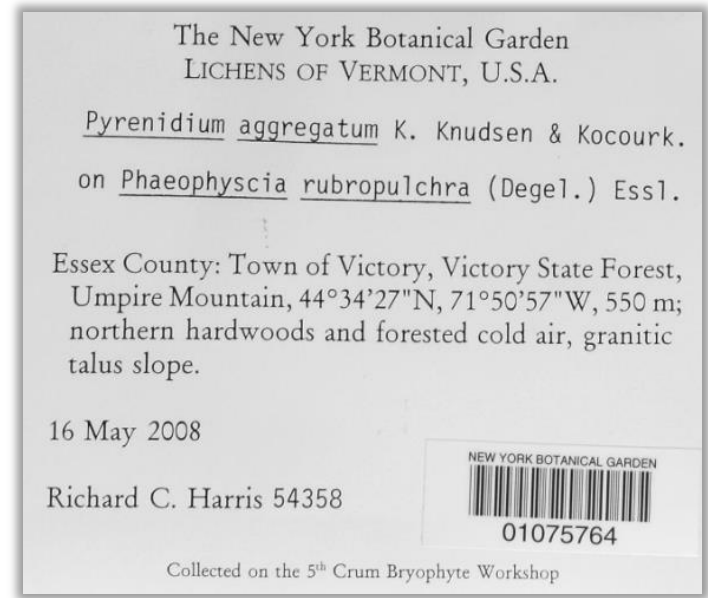
01075764

LI LI Team!

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

► Data Sets

- silver (ABBYY) OCR output from 200 lichen packet images
- all ABBYY OCR output from 10,000 lichen packet images
- silver and gold (ABBYY, Tesseract) OCR output from 100 herbarium sheets
 - OCR output from 5000 herbarium sheets
- SI BVP dataset



Word Cloud using Solr + Carrot²

- ▶ Index OCR text output using SOLR
- ▶ <http://ammatsun.acis.ufl.edu:5900/solr/#/lichenssilver/schema-browser?field=content>
- ▶ Using Carrot² to visualize data
<http://ammatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/>



★ Lichens Silver ★ Herb Silver ★ Herb All Silver ★ Lichens Abby ★ Lichens Tesseract ★ Lichens Ocropsus ★ SI BVP

[Hide options](#)

Download Cluster with

Title field name

Summary field name

URL field name

ID field name

Read Solr clusters if present ☐

Use highlighter output if present ☒

Imagine Integration with NfN/BVP

GOT IT! LET ME TRANSCRIBE

GOT IT! LET ME CHOOSE MY TRANSCRIPTION GROUP

Choose your group:

Country

- Use for initial sort or validation



10,000+ records sample



[★ Lichens Silver](#) [★ Herb Silver](#) [★ Herb All Silver](#) [★ Lichens Abby](#) [★ Lichens Tesseract](#) [★ Lichens Ocropus](#) [★ SI BVP](#) [Wiki](#)

Search [Hide options](#)

Download

10000 results

 Cluster with

Lingo

Title field name

Summary field name

URL field name

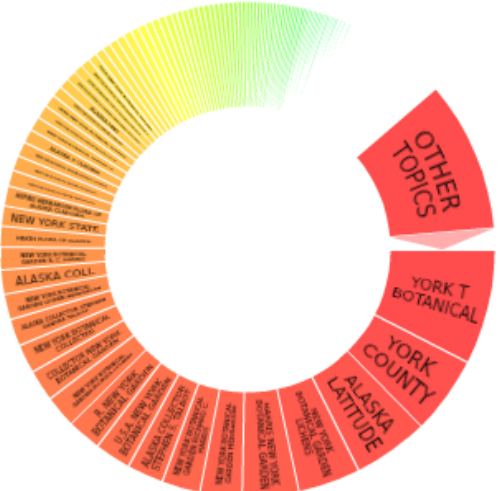
ID field name

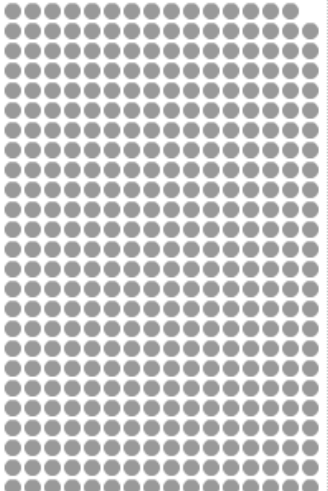
Read Solr clusters if present ☐

Use highlighter output if present ☒

[Hide advanced options](#)

FoldersCirclesFoamTree





Top **10000** results of about **10498** for *

1



The New York Botanical Garden Cryptog County: Little Big Econlockhatchee State 3.3 mi NE of Co. Rd. 419 in Oviedo, 28° 10 January 1996 NEW YORK BOTANICA

2



The New York Botanical Garden Cryptog County: Little Big Econlockhatchee State 3.3 mi NE of Co. Rd. 419 in Oviedo, 280 10 January 1996 William R. Buck 29252 I York Botanical Garden

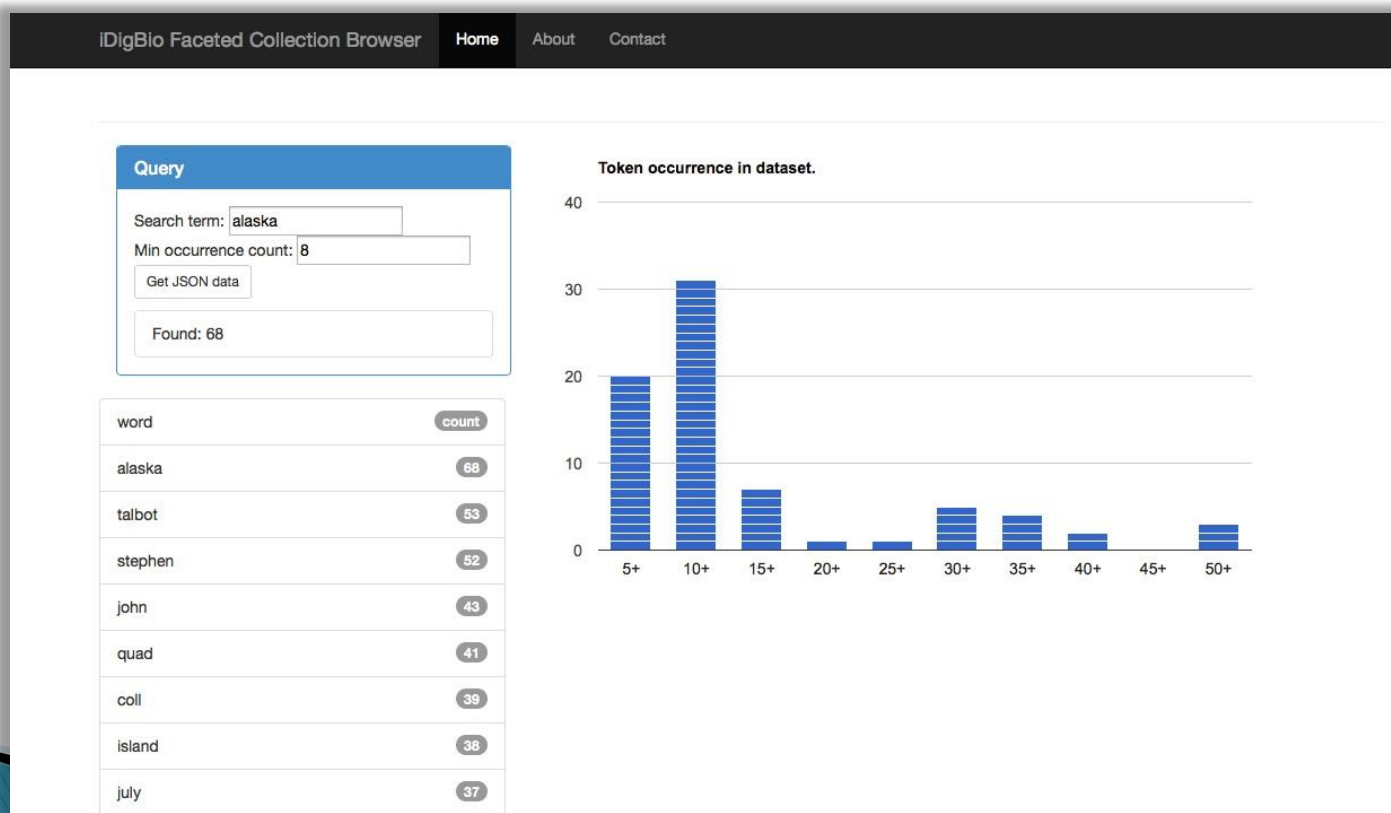
3



The New York Botanical Garden Cryptog County: Black Point Swamp, along Co. R hardwood swamp with scattered Taxodi

iDigBio Faceted Collection Browser

- ▶ discovery
- ▶ how many documents have this issue



iDigBio Faceted Collection Browser

- ▶ discovery
- ▶ how many documents have this issue

The screenshot displays the iDigBio Faceted Collection Browser interface. At the top, there is a navigation bar with links for Home, About, and Contact. The main content area is divided into two sections. On the left, a 'Query' box contains a search term 'alaska', a minimum occurrence count of 8, and a 'Get JSON data' button. Below this, a table lists words and their counts. On the right, a 'Token occurrence in dataset' section shows a list of tokens with their occurrence counts and a legend for confidence levels.

Query

Search term:

Min occurrence count:

Found: 68

word	count
alaska	68
talbot	63
stephen	52
john	43
quad	41
coll	39
island	38
july	37

Token occurrence in dataset.

40

? 6 7 8 9 10

?, ? The New York
copyright reserved Botanical Garden
Spigelia anthelmia L., Sp. pl. 1: 149/150, n. 1 (1753)
Det. vel confirm.: F. J. Fernandez Casas (ex MA, sed nunc in MO), 30-V-2001
vj y
PLANTS OF BAHAMA ISLANDS Spigelia anthelmia L.
Crooked Island: in sandy open soil
near beach area at Landrail Landing.
Flowers pinkish-lavender
June 6, 1977
01498191

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

■ extremely low ■ very low ■ low ■ undetermined ■ medium ■ high ■ very high

Visualizing OCR Confidence

- ▶ early triage
 - score each document
 - score transcriptions
 - low scores to human
 - high scores to automated parsing
- ▶ humans check
- ▶ human correct
 - transcription errors
 - ocr errors

Why

?6 7 8 9 10
The New York
copyright reserved Botanical Garden
Spigelia anthelmia L., Sp. pl. 1: 149/150, n. 1 (1753)
Det. vel confirm.: F. J. Fernandez Casas (ex MA, sed nunc in MO), 30-V-2001
vj y
PLANTS OF BAHAMA ISLANDS Spigelia anthelmia L.
Crooked Island: in sandy open soil
near beach area at Landrail Landing.
Flowers pinkish-lavender
June 6, 1977
01498191

Legend - Level of confidence that token is an accurately-transcribed word
extremely low very low low undetermined medium high very high

N-grams
probability
model

Herb gold



Training Set

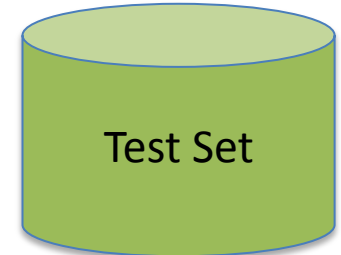
100 docs

PLANTS OF BAHAMA ISLANDS

PLA	AMA
LAN	MA
ANT	A I
NTS	IS
TS	ISL
S O	SLA
OF	LAN
OF	AND
F B	NDS
BA	...
BAH	
AHA	
HAM	
...	

3-grams

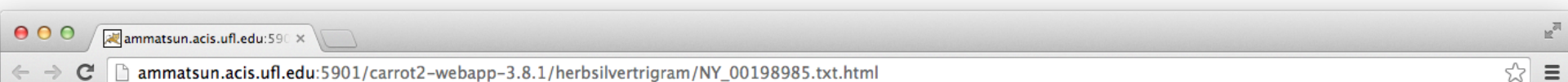
Herb silver



Test Set

100 docs

Visualizing OCR Confidence



6 7 8 9 10 _ XT v
.. The New York
copyright reserved Botanical Garden
NEW YORK BOTANICAL GARDEN
001
98985
JARDIN BOTANICO NACIONAL "DR. RAFAEL M. MOSCOSO"
SANTO DOMINGO, REPUBLICA DOMINICANA
193 Araliaceae
NO_FAMILIA
Polyscias fruticosa Harms.
Arbusto utilizado como seto 3m. Hojas palmatilobuladas, foliolos con dientes. Inflorescencia umbeliforme. Flores de petalos verdosos, 5 petalos y 5 Estambres, blancos.
República Dominicana: Distrito Nacional: Jardín Botánico Nacional: Plantas en cultivo en el área del Museo y Pabellón Docente.
4 de Agosto, 1981.
Santiago Peláez Feucht.
00198985

Legend - Level of confidence that token is an accurately-transcribed word
■ extremely low ■ very low ■ low ■ undetermined ■ medium ■ high ■ very high

NY_00198985.txt

6 7 8 9 10 _ XT v
.. The New York
copyright reserved Botanical Garden
NEW YORK BOTANICAL GARDEN
001
98985
JARDIN BOTANICO NACIONAL "DR. RAFAEL M. MOSCOSO"
SANTO DOMINGO, REPUBLICA DOMINICANA
193 Araliaceae
NO_FAMILIA
Polyscias fruticosa Harms.
Arbusto utilizado como seto 3m. Hojas palmatilobuladas, foliolos con
dientes. Inflorescencia umbeliforme. Flores de petalos verdosos, 5 petalos y 5 Estambres,
blancos.
República Dominicana: Distrito Nacional: Jardín Botánico Nacional: Plantas en cultivo en el
area del Museo y Pabellón Docente.
4 de Agosto, 1981.
Santiago Peláez Feucht.
00198985

5000 OCR html files

See it for yourself!

- ▶ herb_allsilver_trigram_html.zip

?6 7 8 9 10
. , ? The New York
copyright reserved Botanical Garden
Spigelia anthelmia L., Sp. pi. 1: 149/150, nÂl 1 (1753)
Det. vel confirm.: F. J. Fernandez Casas (ex MA, sed nunc in MO), 30-V-2001
vj y
PLANTS OF BAHAMA ISLANDS Spigelia anthelmia L.
Crooked Island: in sandy open soil
near beach area at Landrail Landing.
Flowers pinkish-lavender
June 6, 1977
01498191

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

 extremely low  very low  low  undetermined  medium  high  very high

Let's make it happen!

GOT IT! LET ME TRANSCRIBE

GOT IT! LET ME CHOOSE MY TRANSCRIPTION GROUP

Choose your group:

Country

- Use for initial sort or validation



Happy HoLLidays!



Transcribe museum records to
TAKE NOTES FROM NATURE

ATLASofLIVING AUSTRALIA
Sharing biodiversity knowledge
Biodiversity Volunteer Portal Expeditions Tutorials
Biodiversity Volunteer Portal



NEW YORK BOTANICAL GARDEN

LI LI Team!

01075764

