



Enhancing Crowdsourcing using Text Analytics and Visualization

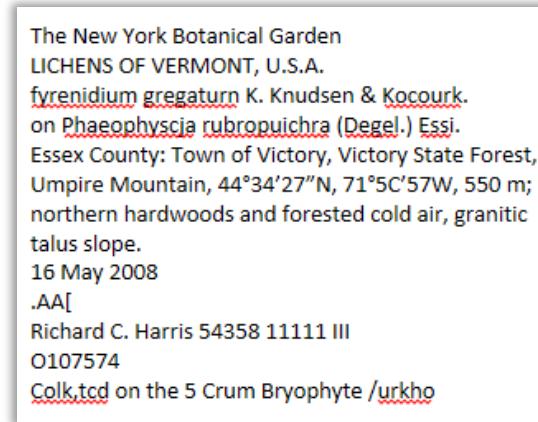
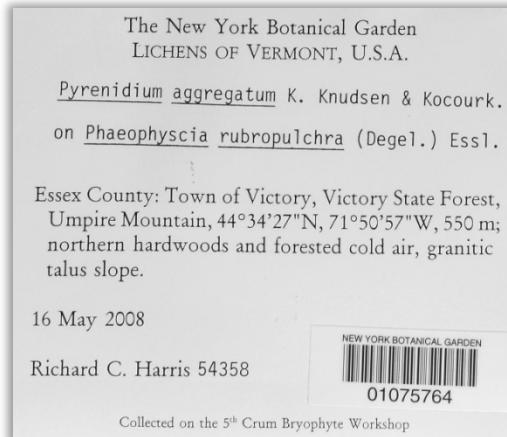
Deb Paul, Andrea Matsunaga, Miao Chen, Jason
Best, Reed Beaman, Sylvia Orli, William Ulate

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

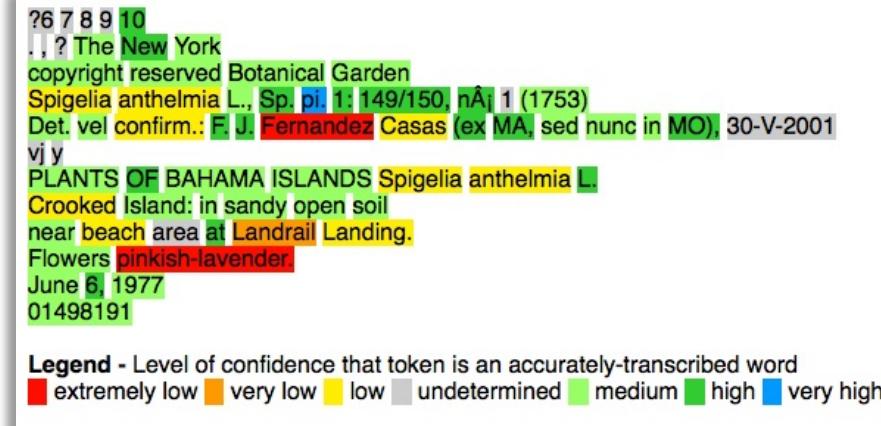
Why

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

Reasons for 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



User Stories

Who

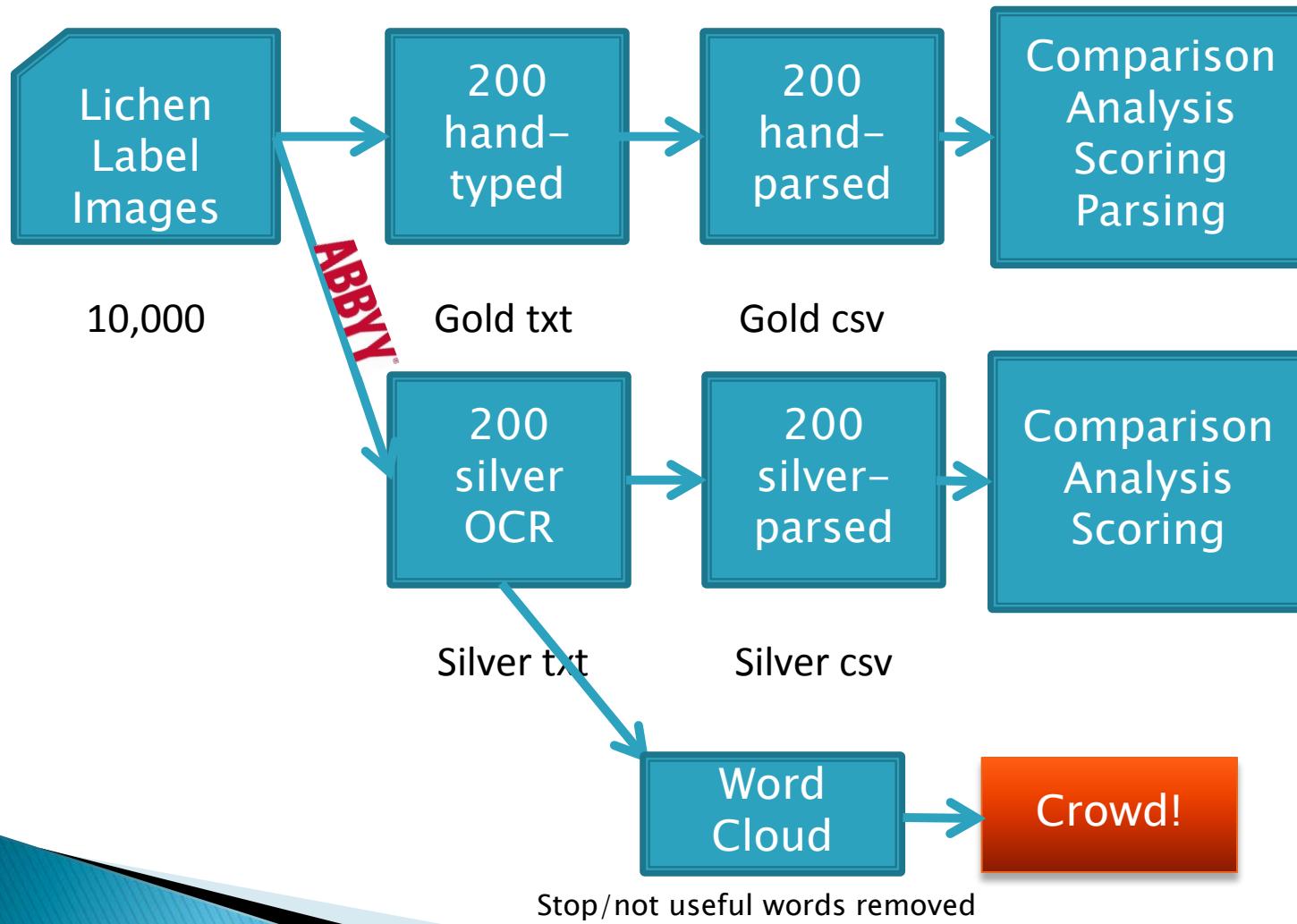
- ▶ Users like ordered datasets
- ▶ Transcription
 - **faster** with ordered/sorted sets
 - **less error prone** with sorted sets

Handwriting vs. Typed

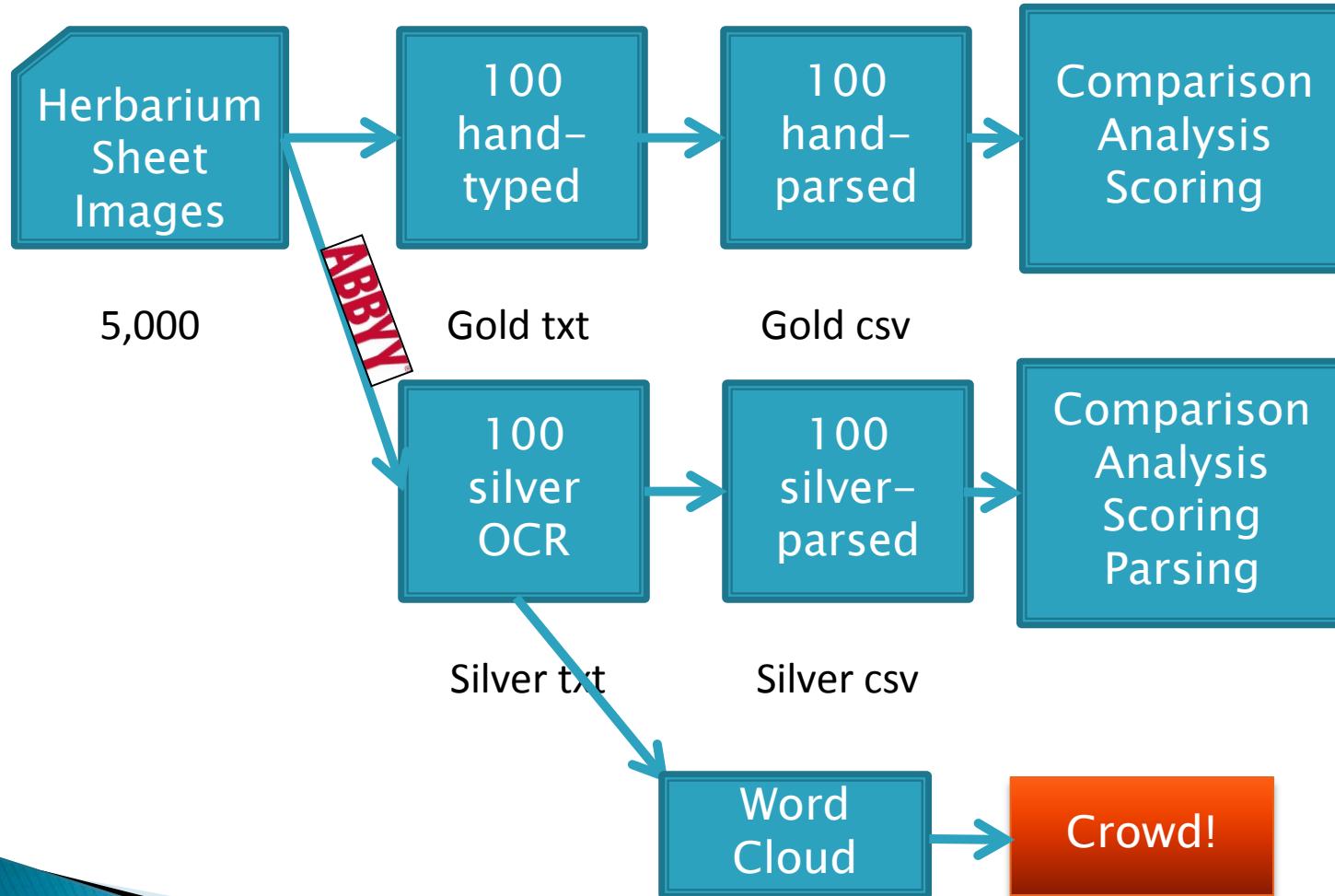
Threshold	Correct	False Positives	False Negatives
$T > 1$ and $N > 20\%$	82%	10 of 45	8 of 60
$T > 0$ and $N > 20\%$	84%	13 of 45	4 of 60
$T > 1$	79%	10 of 45	12 of 60
$N > 20\%$	75%	8 of 45	18 of 60
$N > 10\%$	81%	14 of 45	6 of 60

- ▶ 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
 - Humanity's community using now!
- ▶ Let transcriber choose label format
- ▶ *Typed?....go to word cloud workflow*

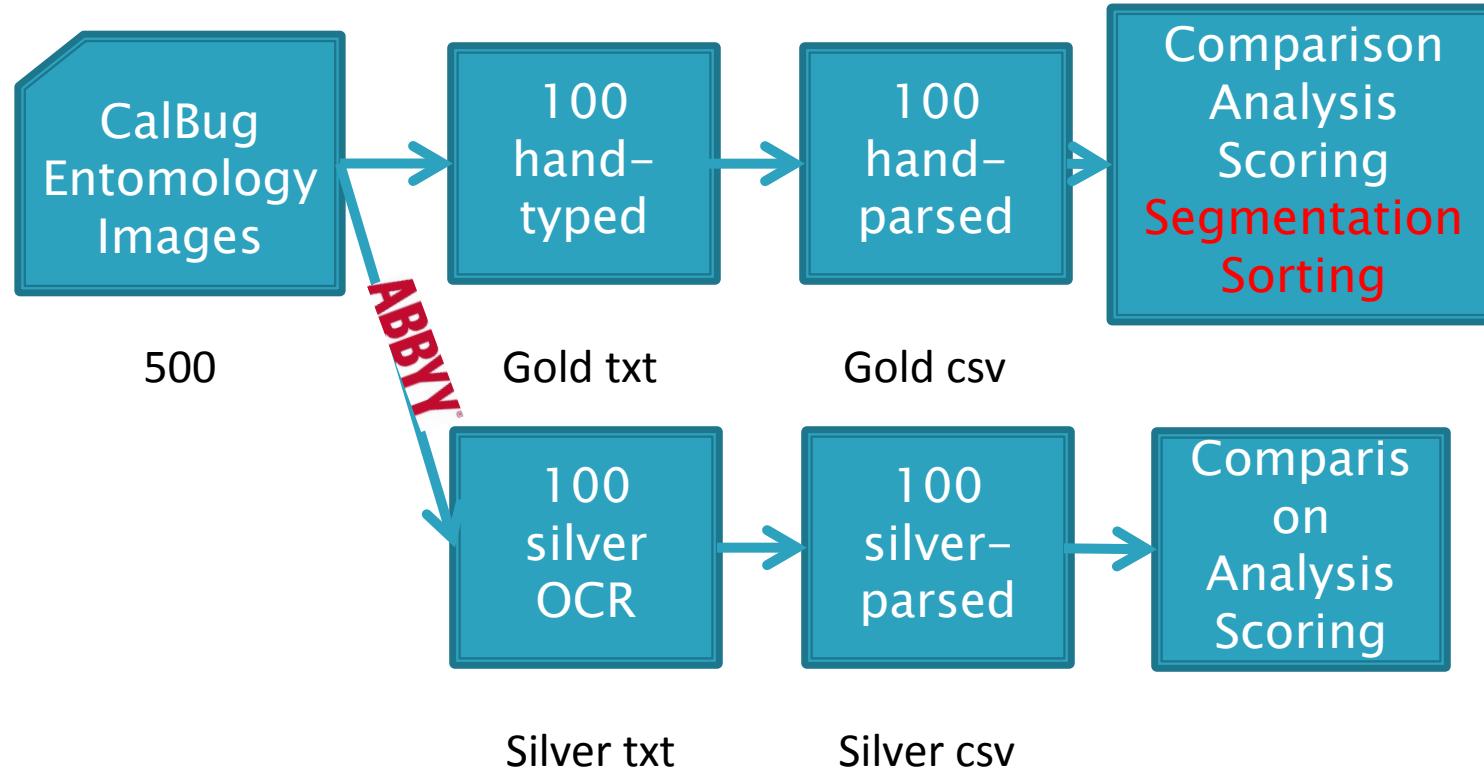
aOCR iDigBio – BRIT Hackathon Lichen Dataset



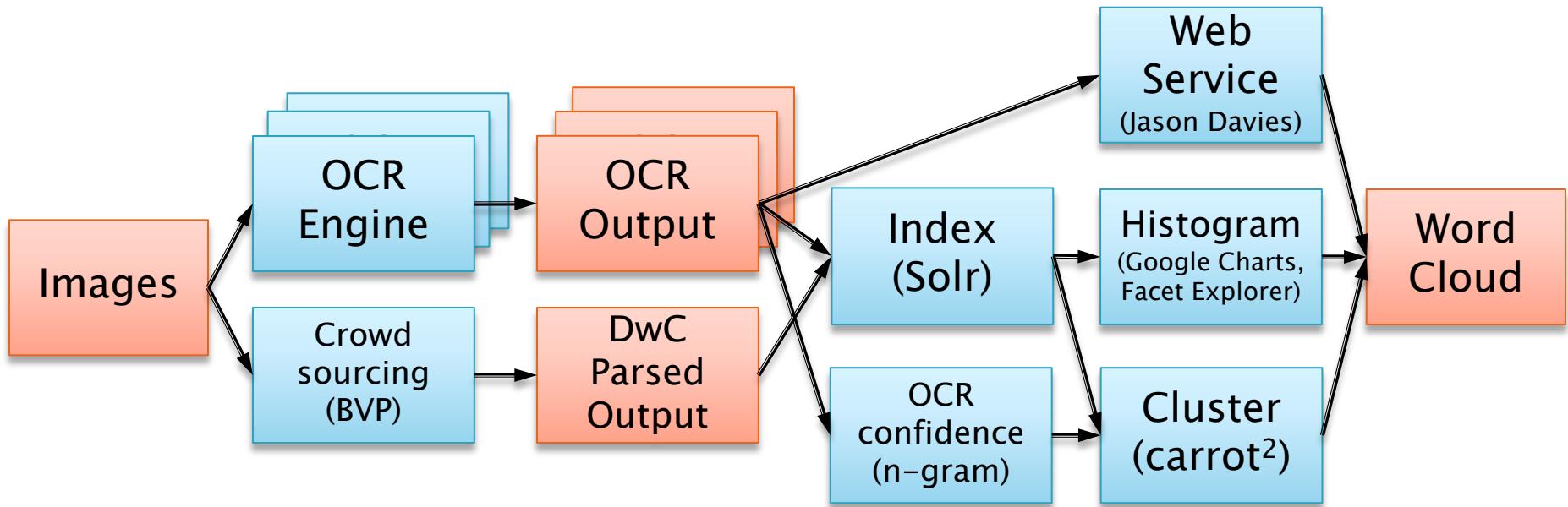
aOCR iDigBio – BRIT Hackathon NYBG Herbarium Dataset



aOCR iDigBio – BRIT Hackathon CalBug Entomology Dataset



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>

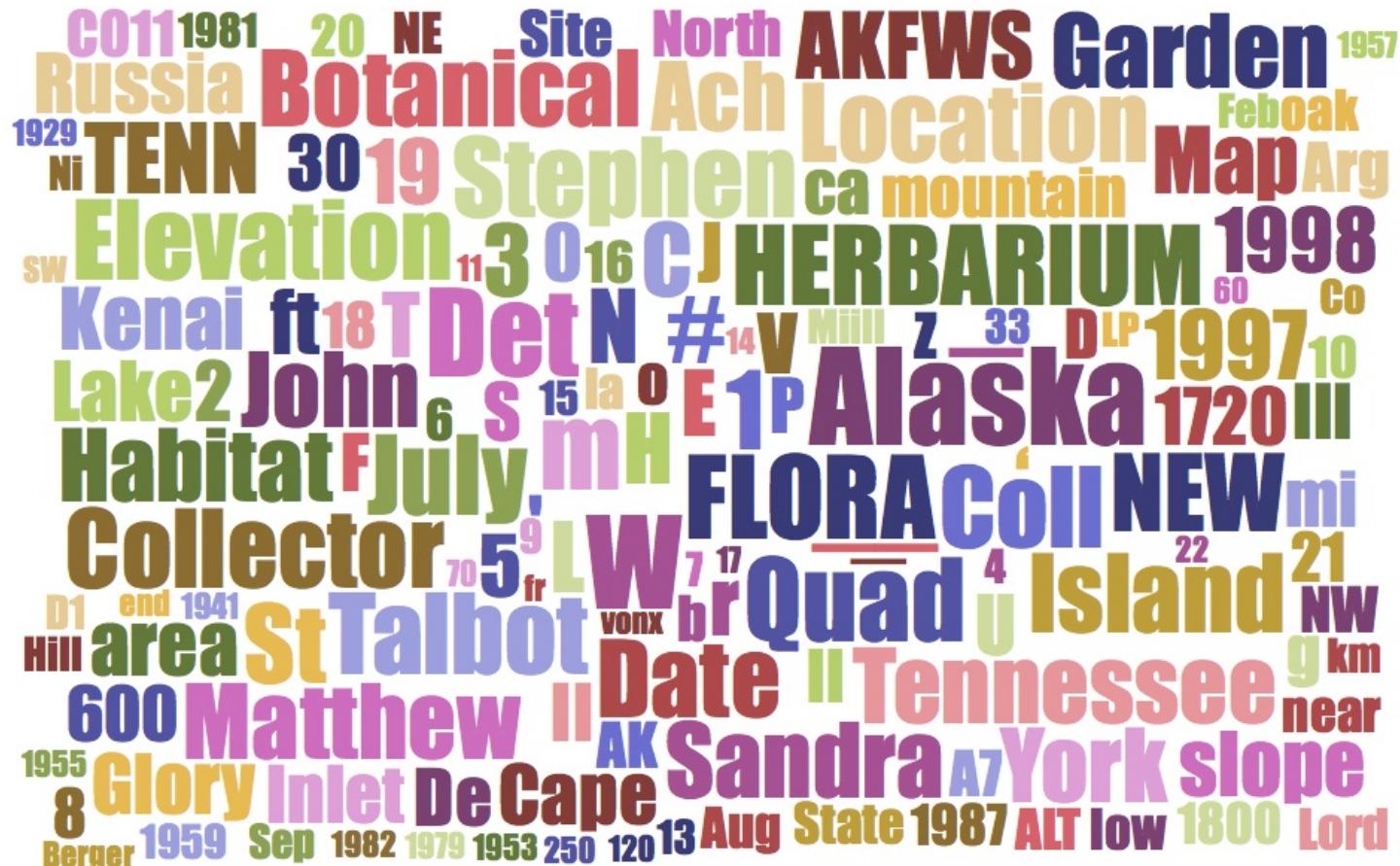
Jason Davies WC: <http://www.jasondavies.com/wordcloud/>

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

carrot²: <http://project.carrot2.org/>

Web Service–Based Word Cloud

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



Try:

<http://www.jasondavies.com/wordcloud/#http%3A%2F%2Focr1.acis.ufl.edu%2Fdatasets%2Flichens%2Fsilver%2Focr%2FWebrootDatasetsLichensSilverOcr.txt>

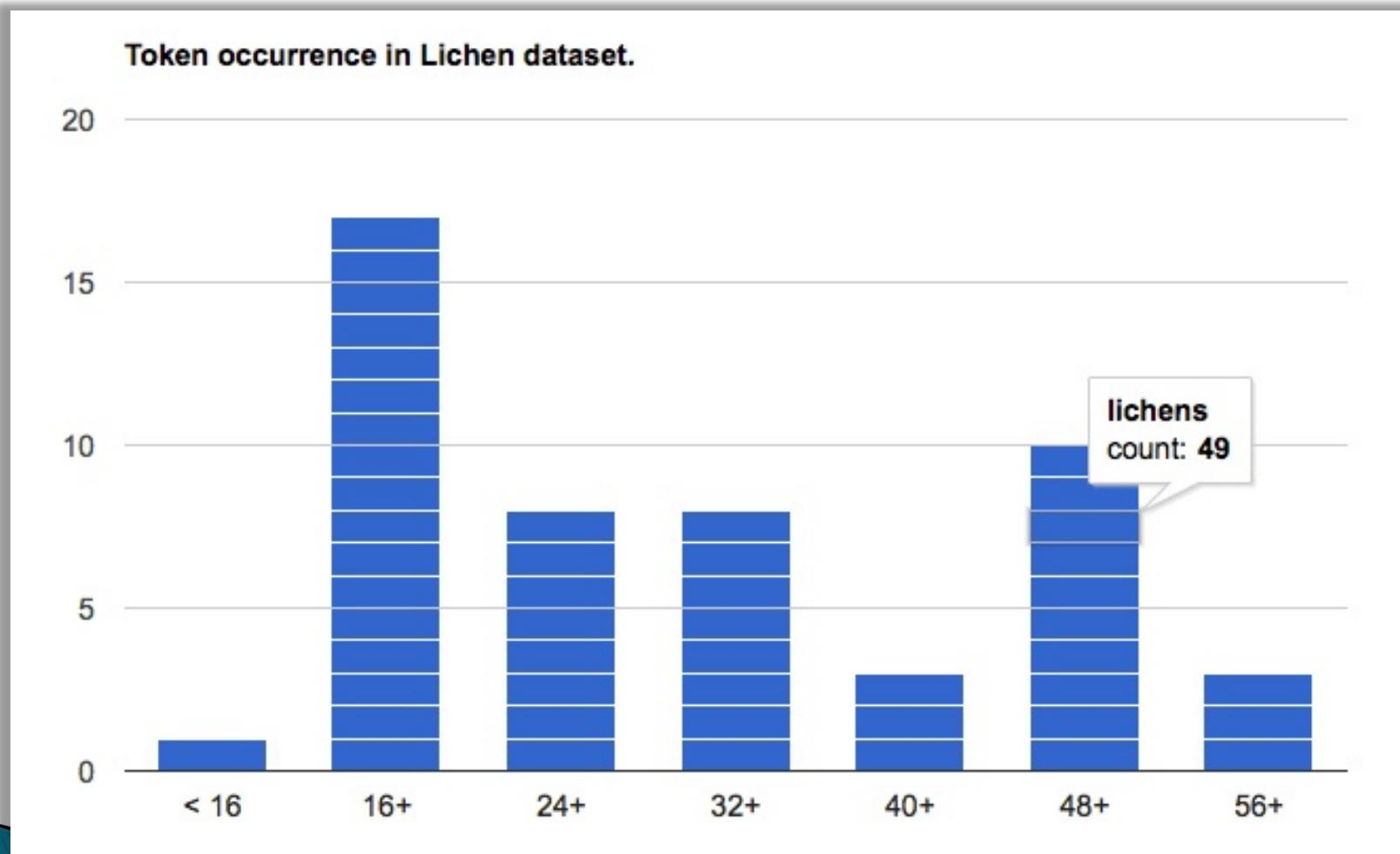
Apache Solr Indexing

► Datasets Indexed:

- 100 herb silver (OCR)
- 5,000 herb silver (OCR)
- 200 lichen silver (OCR)
- 10,498 lichen ABBY (OCR)
- 10,495 lichen ocropus (OCR)
- 10,498 lichen tesseract (OCR)
- 809 Smithsonian/BVP
 - catalogNumber, collector, country, eventDate,
 - fieldNotes, fieldNumber, scientificName,
 - stateProvince, transcriberNotes, validatorNotes,
 - verbatimElevation, verbatimLatitude,
 - verbatimLongitude, verbatimLocality

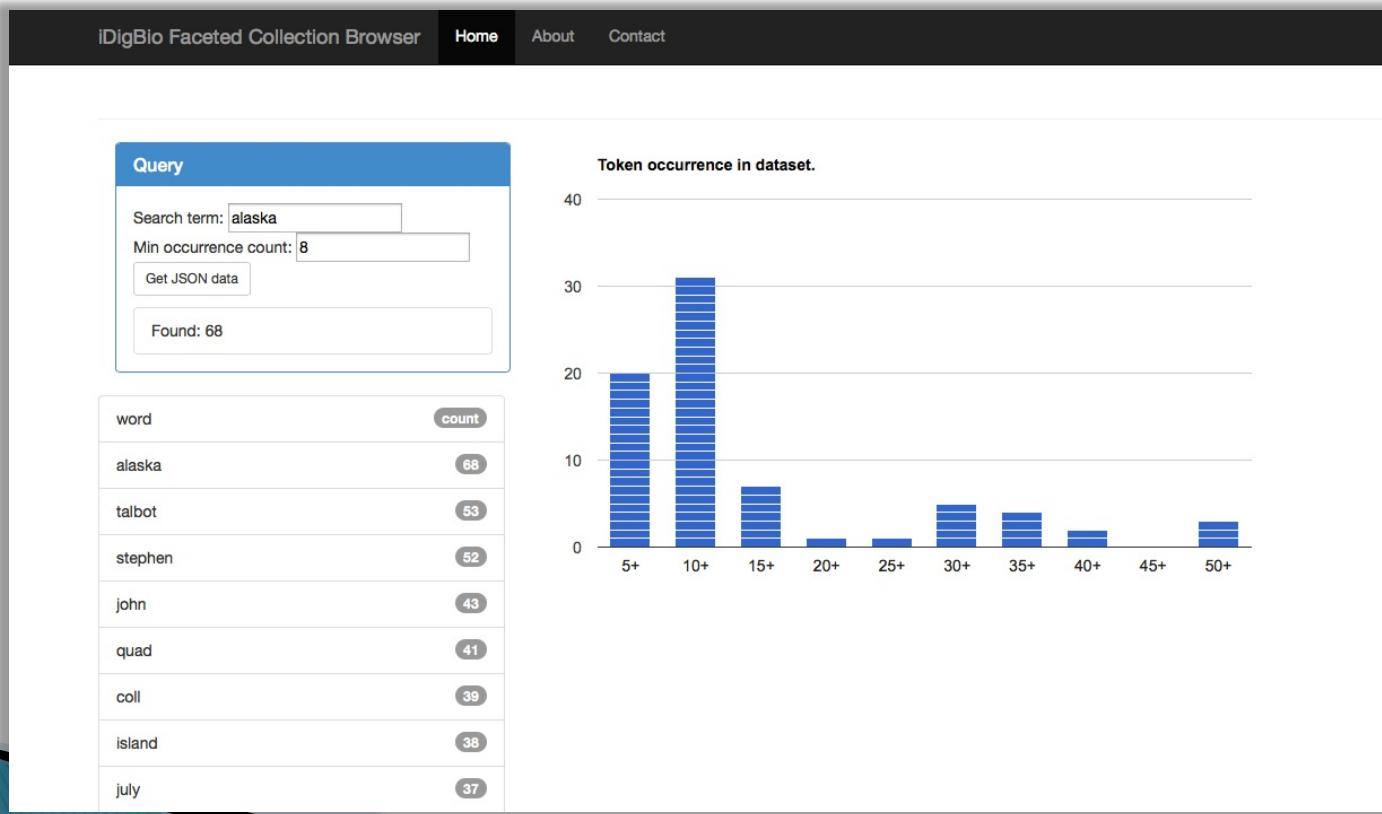
1000	/3201	Top-Terms:	?
68	alaska		
62	coll	county	
56	talbot		
55	ach	july	stephen
54	new		
53	garden	york	
51	tenn		
49	lichens	tennessee	
48	university		
47	john	island	
43	quad		
39	thomson		
35	ft		
34	sandra		
33	longitude	cinchonae	map
32	latitude	akfws	
29	mi	u.s.a	
28	19	forest	co
27	arg		
25	ll	lake	
23	bay		

Token Histogram – Google Charts



iDigBio Faceted Collection Browser

- ▶ discovery
- ▶ how many documents have this issue



OCR Confidence Estimation

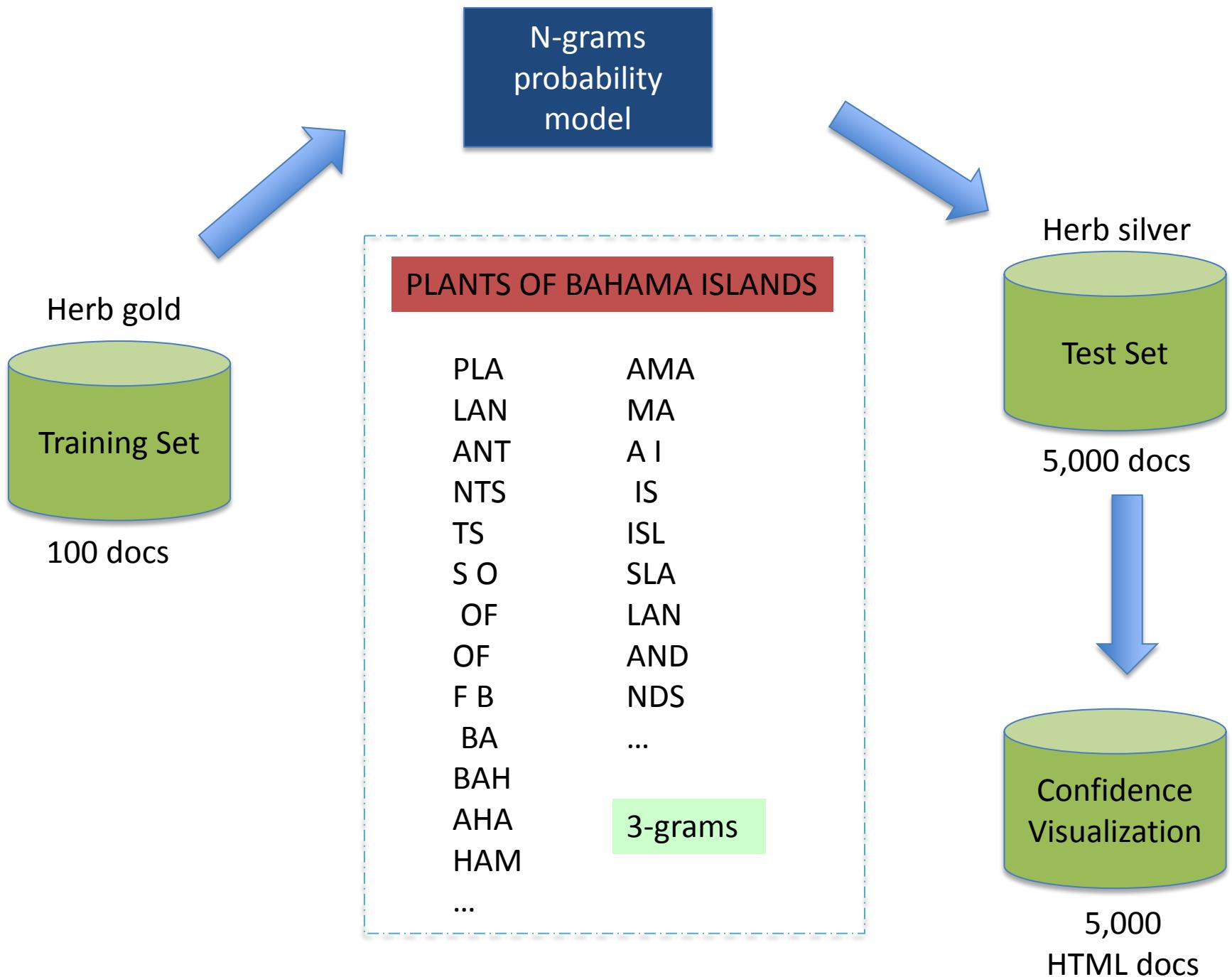
REPORT ANY REIDÍ TO THE INSTITU
SA*.f / b-fs
The Key York Botanical Garden INSTITUTE OF ECONOMIC BOTANY Plaits of Cononiealtli of
Dominica
uu.«./«. (Jici ARACEAE
Anthurium -iacant» d // T. 93 HooKer\ K'intU - M.4.r. T. Croq. <* ?3 WEST INDIES, Dominica. St.George. Above new steel bridge on road to Freshwater 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 |

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
garden -25.276526293980815

Based on the probability
of n-grams

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)

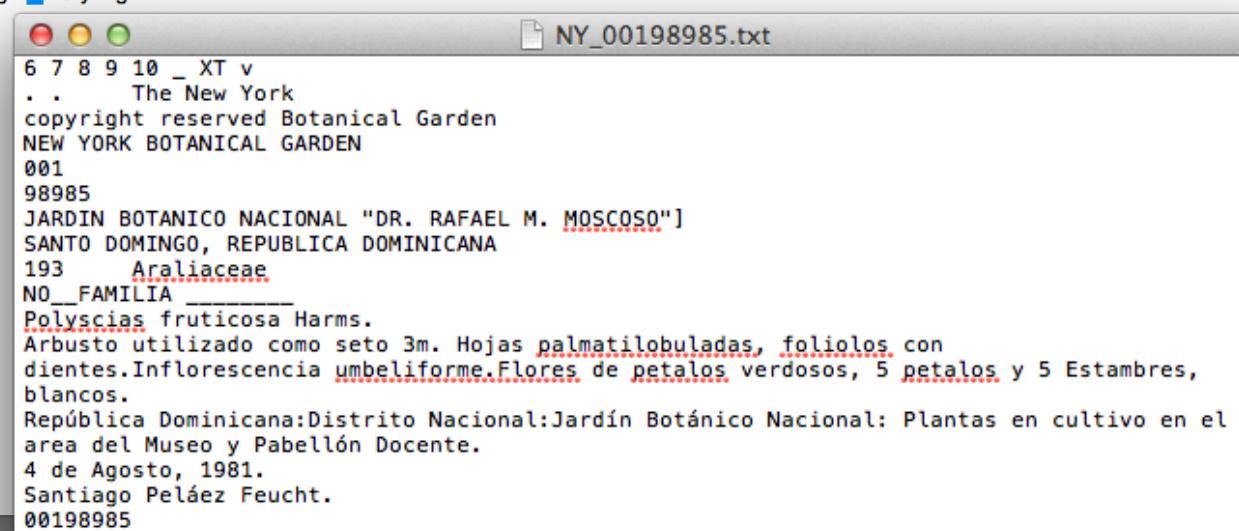


Visualizing OCR Confidence

<https://github.com/idigbio-citsci-hackathon/OCR-Text-Confidence-Visualization>

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.
Repœblica Dominicana:Distrito Nacional:Jard'n Botnico Nacional: Plantas en cultivo en el area 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



Word Cloud using Solr + Carrot²



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

*

Search More options

Carrot² organizes your search results into topics. With an instant overview of what's available, you will quickly find what you're looking for.

Example queries: [florida](#) | [tennessee](#)

.....

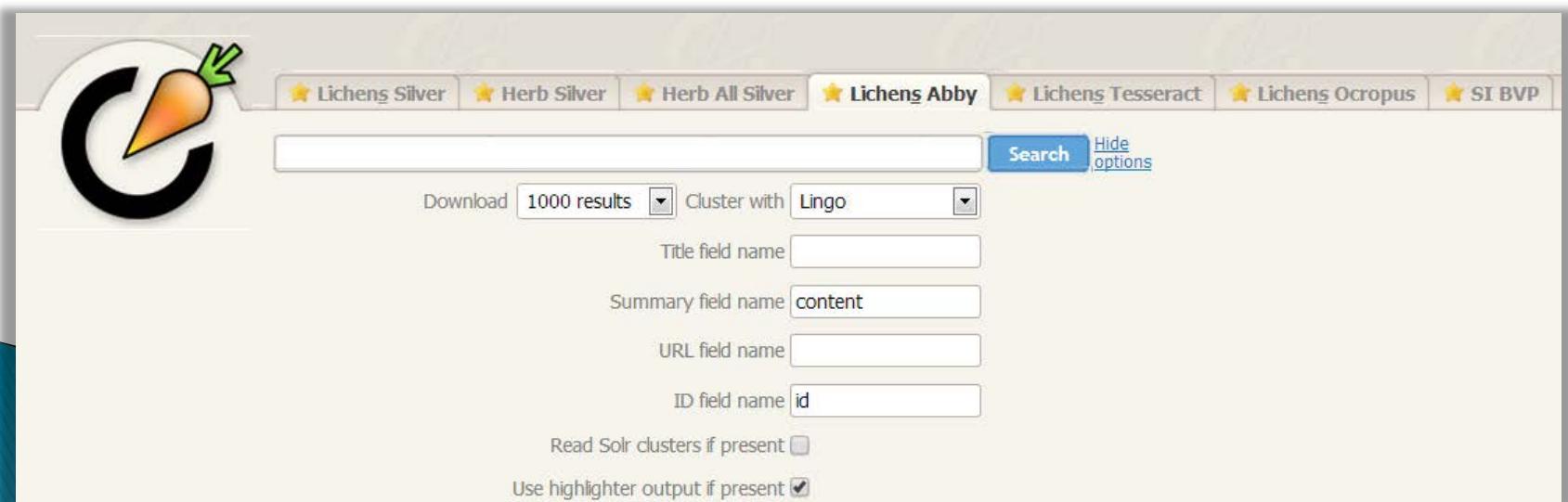
[About](#) | [New features!](#) | [Search plugins](#) | [Download](#) | [Contact](#)

| 2013-10-10 13:47:45 © 2002-2014 Stanislaw Osinski

<http://ammatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/>

Word Cloud using Solr + Carrot²

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



Folder View of Search



About | [New features!](#) | Search feeds | Search plugins | Download | Carrot Search | Contact
★ Lichens Silver ★ Herb Silver ★ Herb All Silver ★ Lichens Abby ★ Lichens Tesseract ★ Lichens Ocropolis

*

Search More options

Folders Circles FoamTree

All Topics (100)

- Cm 6 7 8 9 10 (22)
- TEX Capraria Biflora L. (18)
- F. J. (16)
- Britton Collectors (13)
- PLANTS OF BAHAMA ISLANDS (12)
- Puerto Rico (8)
- T. Zanoni (8)
- Spigelia Anthelmia L. (7)
- RAFAEL M. MOSCOSO SANTO DOMINGO (6)
- C. WRIGHT (5)

more | show all

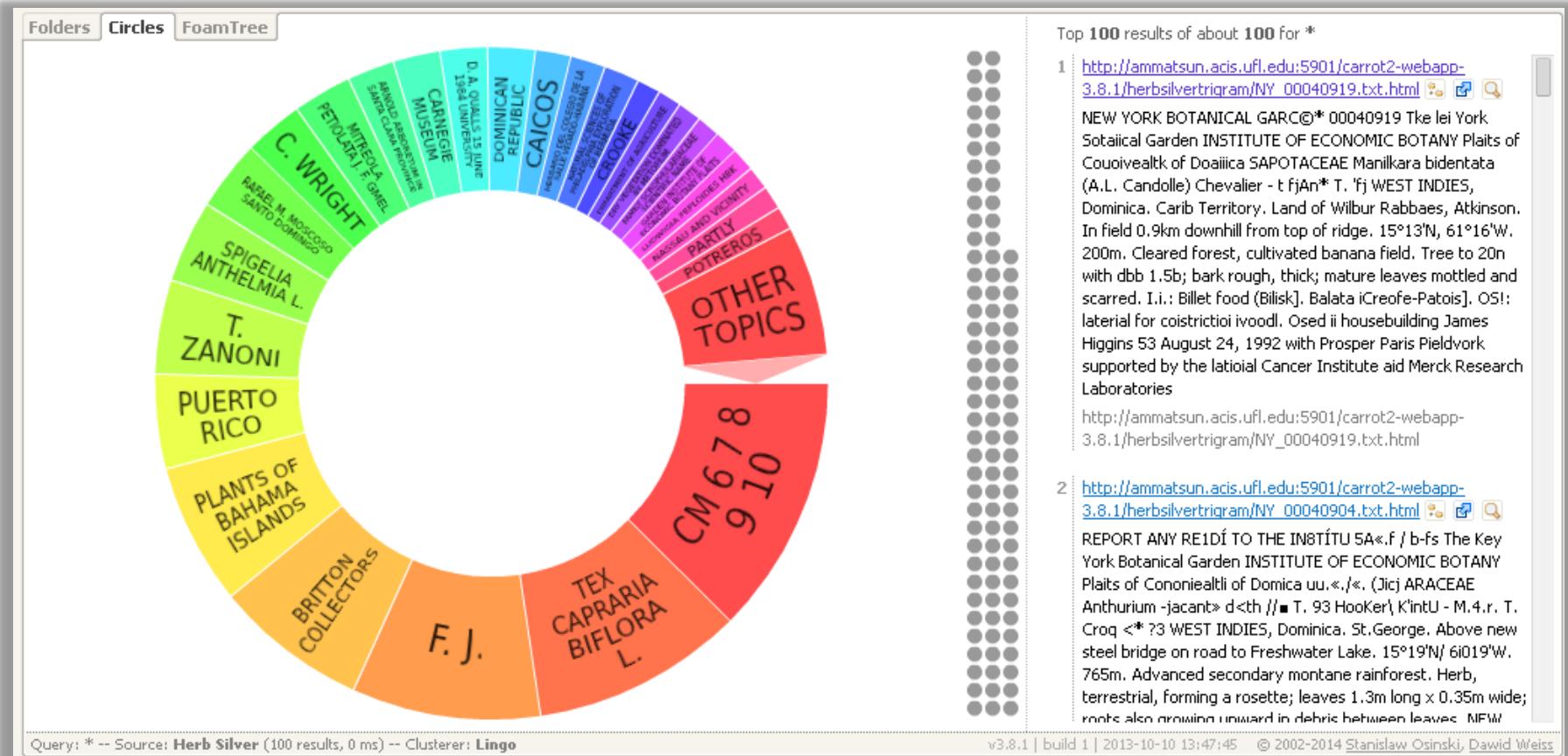
Top 100 results of about 100 for *

1 http://ammatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/herbsilvertrigram/NY_00040919.txt.
NEW YORK BOTANICAL GARDEN 00040919 The New York Botanical Garden INSTITUTE OF ECONOMIC
Commonwealth of Dominica SAPOTACEAE Manilkara bidentata (A.L. Candolle) Chevalier - t fjan* T.
Carib Territory, Land of Wilbur Rabbaes, Atkinson. In field 0.9km downhill from top of ridge. 15°19'N
Cleared forest, cultivated banana field. Tree to 20m with dbh 1.5m; bark rough, thick; mature le
I.i.: Billet food [Bilisk], Balata [Creole-Patois]. OS! lateral for costrictio iwoodl. Used in housebu
August 24, 1992 with Prosper Paris Fieldwork supported by the International Cancer Institute aid Merc
http://ammatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/herbsilvertrigram/NY_00040919.txt.

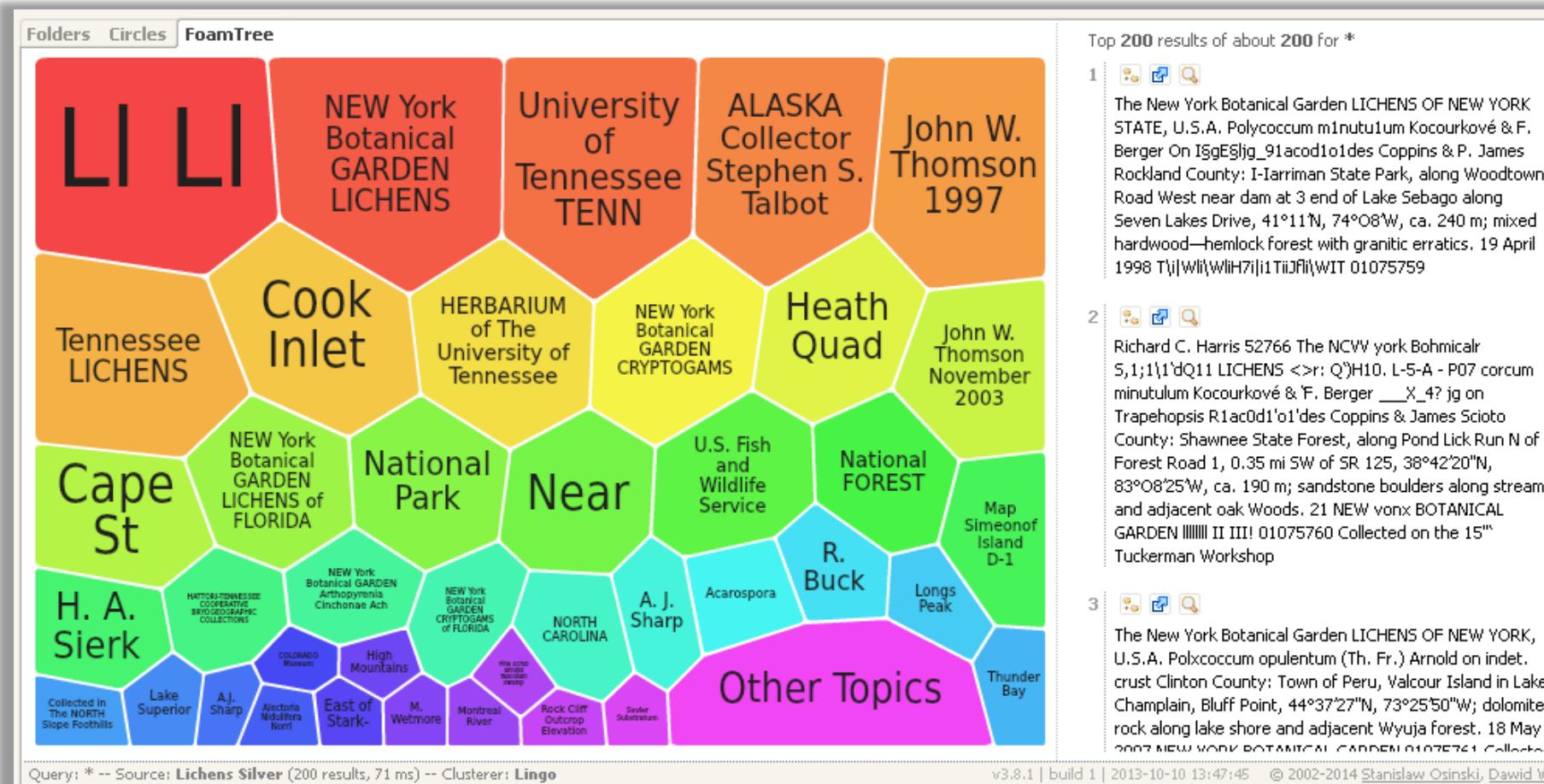
2 http://ammatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/herbsilvertrigram/NY_00040904.txt.
REPORT ANY REID TO THE INSTITUTO DE LA PLANTA / b-fs The Key York Botanical Garden INSTITUTE OF
Constitutional of Dominica u., «, /, (Jic) ARACEAE Anthurium -jacant» d //■ T. 93 Hooker, Kintu INDIES, Dominica, St. George. Above new steel bridge on road to Freshwater Lake. 15°19'N/ 61° secondary montane rainforest. Herby, terrestrial, forming a rosette; leaves 1.3m long x 0.35m w upward in debris between leaves. NEW YORK BOTANICAL GARDEN 00040904 Wm ya BOTAN Stijfhoorn 776 May 11, 1992 with G. A. Eidesen and H. T. Beck Fieldwork supported by the Inter r.....— 6 7 8 9 10 the New York copyright reserved botanical garden ■■ 00040904 http://ammatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/herbsilvertrigram/NY_00040904.txt. |

3 http://ammatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/herbsilvertrigram/NY_00191205.txt.
The New York Botanical Garden Institute of Economic Botany Plants of Dominica RHIZOPHORAC
(Sw.) Poir. det: M. Nee, 1998 NEWYORK BOTANICAL GARDEN 0019 205 BOTANICAL GARDE**

Circles Visualization of Search



Foam Tree Visualization of Search



Source of our LI LI team name!

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

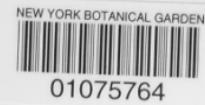
The New York Botanical Garden
LICHENS OF VERMONT, U.S.A.

Pyrenidium aggregatum K. Knudsen & Kocourk.
on Phaeophyscia rubropulchra (Dege1.) Essl.

Essex County: Town of Victory, Victory State Forest,
Umpire Mountain, 44°34'27"N, 71°50'57"W, 550 m;
northern hardwoods and forested cold air, granitic
talus slope.

16 May 2008

Richard C. Harris 54358



Collected on the 5th Crum Bryophyte Workshop

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

Carrot 2 Web Application

Search bar: *
Download: 1000 results Cluster with: Lingo
Title field name: url
Summary field name: content
URL field name: url
ID field name: id
Read Solr clusters if present:
Use highlighter output if present:

Folders Circles FoamTree

Top 1000 results of about 5000 for *

1 | <http://ammatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/herbalsilvertrogram/00040931.txt.html> 🔍

6 7 8 9 10 " XT " The New York copyright reserved botanical Garden NEW YORK BOTANICAL GARDEN 00040931 BOTANICAL GARDE** REPORT ANY REIDENTIFICATION OF THIS VOUCHER TO THE INSTITUTE OF ECONOMIC BOTANY, NY The Dev York Botanical Garden INSTITUTE OF ECONOMIC BOTANY Plants of Cononfrealth of Dominica APIACEAE Eryngium foetidum L. - IAMS, T.B.Coc.Cn WEST INDIES, Dominica, Carib Territory, Bataka, 300m up a trail heading W of feeder near field in Galaback. 15°13'N, 61°16'W. Flat ground, road. ---- 100m. Maintained in a garden, shady, moist soil. Herb to 0.3m with bitter odor; leaves 6-8 at base; flowers with sharp spiked petals. n.v.: Chardon Beni [Crcle-Patois]. DSE: Untrmissive llvsl. Infusion with Pluchea inti tussive for colds leavesflowers I lade with 3 leaves and 1 flower »Cl Sample codes: IMOCUHOOJ-O) Janes Higgins 12 with Prosper Paris synphytfolia Drink infusion Amount: 8 1002 Evidences generated by the

6 7 8 9 10 " XT "
The New York copyright reserved botanical Garden
NEW YORK BOTANICAL GARDEN
00040931
BOTANICAL
GARDE**
REPORT ANY REIDENTIFICATION OF THIS VOUCHER
TO THE INSTITUTE OF ECONOMIC BOTANY, NY
The Dev York Botanical Garden INSTITUTE OF ECONOMIC BOTANY Plants of Cononfrealth of Dominica
APIACEAE
Eryngium foetidum L.
IAMS, T.B.Coc.Cn
WEST INDIES, Dominica, Carib Territory
Bataka, 300m up a trail heading W of feeder
near field in Galaback. 15°13'N, 61°16'W
Flat ground,
road. ----
100m. Maintained in a garden, shady, moist soil.
Herb to 0.3m with bitter odor; leaves 6-8 at base; flowers with sharp spiked petals.
n.v.: Chardon Beni [Crcle-Patois].
DSE: Untrmissive llvsl. Infusion with Pluchea inti tussive for colds leavesflowers I lade with 3 leaves and 1 flower »Cl Sample codes:
IMOCUHOOJ-O)
Janes Higgins 12 with Prosper Paris synphytfolia Drink infusion Amount: 8 1002 Evidences generated by the
00040931

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



10,000+ records sample

About | New features

★ Lichens Silver ★ Herb Silver ★ Herb All Silver ★ Lichens Abby ★ Lichens Tesseract ★ Lichens Ocropus ★ SI BVP ★ Wiki

*

Download 10000 results Cluster with Lingo

Search Hide options

Title field name

Summary field name content

URL field name

ID field name id

Read Solr clusters if present

Use highlighter output if present

[Hide advanced options](#)

Folders Circles FoamTree

Top 10000 results of about 10498 for *

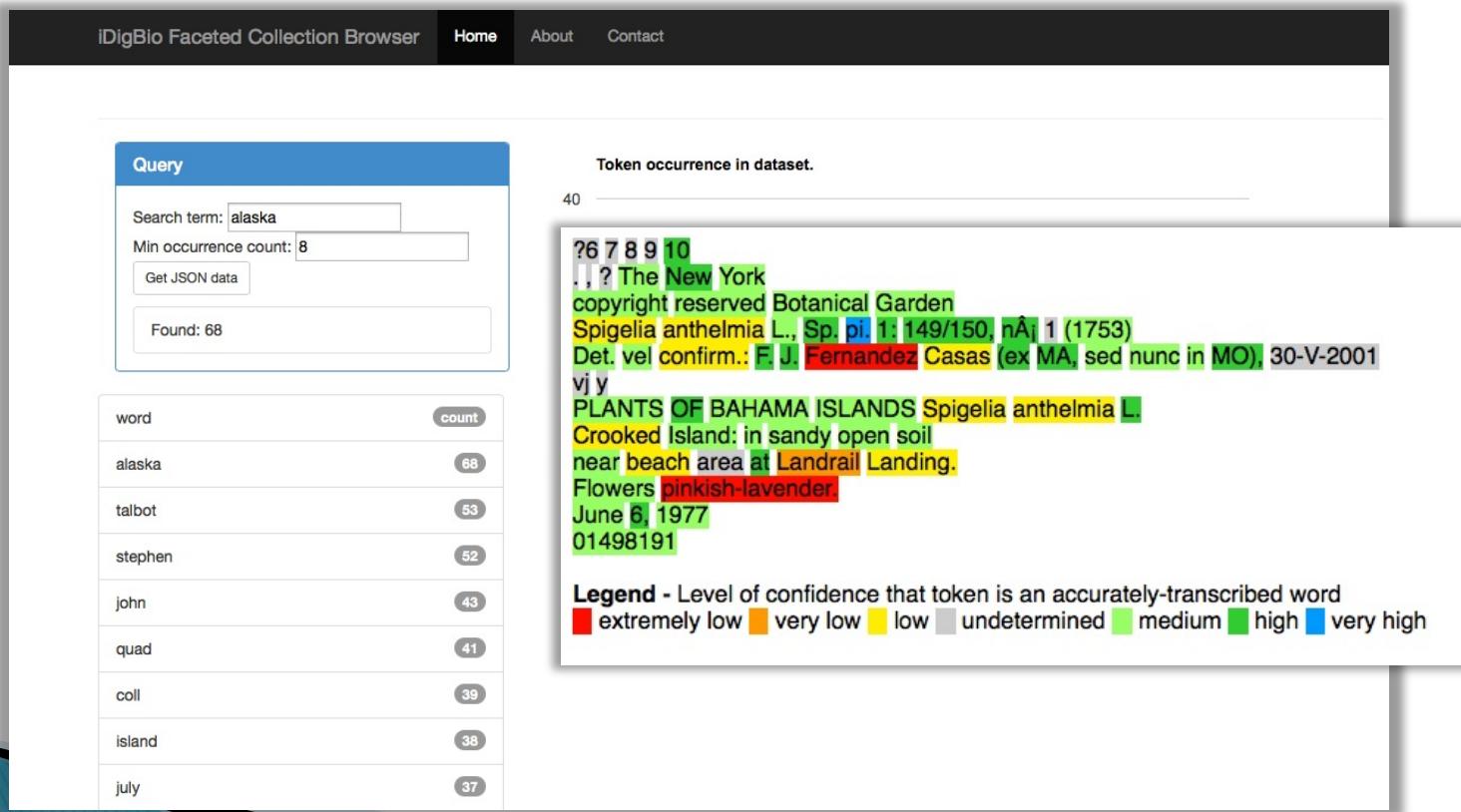
1 The New York Botanical Garden Cryptog County: Little Big Econlockhatchee State 3.3 mi NE of Co. Rd. 419 in Oviedo, 28° 10 January 1996 NEW YORK BOTANICA

2 The New York Botanical Garden Cryptog County: Little Big Econlockhatchee State 3.3 mi NE of Co. Rd. 419 in Oviedo, 28° 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



Imagine Integration with NfN/BVP

Let's make it happen!

GOT IT! LET ME TRANSCRIBE

A thick green arrow pointing to the right, indicating the direction of the next step.

GOT IT! LET ME CHOOSE MY TRANSCRIPTION GROUP

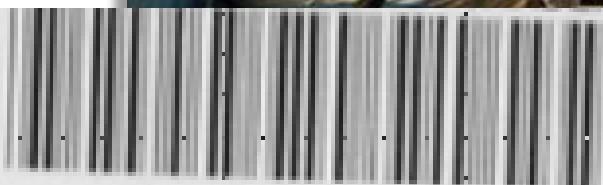
Choose your group:

Country

- ▶ Use for initial sort or validation



Using OCR output to enhance the transcription process



01075764

LI LI Team!

