Enhancing Crowdsourcing using Text Analytics and Visualization

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

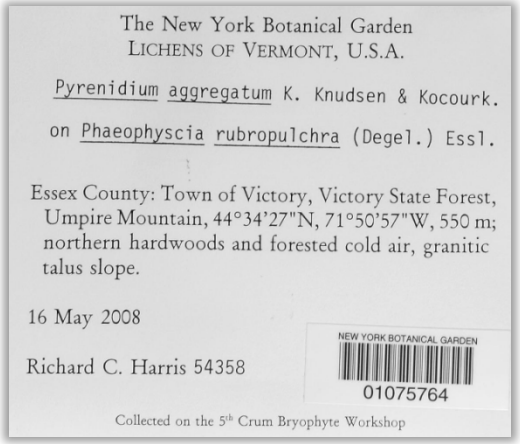
iDigBio – Notes From Nature Hackathon December 2013

Increasing Citizen Science Participation in Museum Specimen Digitization

Text Clusters

What

- Preprocess specimen label images with OCR



The New York Botanical Garden
LICHENS OF VERMONT, U.S.A.
fyrenidium gregaturn K. Knudsen & Kocourk.
on Phaeophyscja rubropuichra (Degel.) Essi.
Essex County: Town of Victory, Victory State Forest, Umpire Mountain, 44°34'27"N, 71°5C'57W, 550 m; northern hardwoods and forested cold air, granitic talus slope.
16 May 2008
.AA[
Richard C. Harris 54358 11111 III
O107574
Colk,tcd on the 5 Crum Bryophyte /urkho

- 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

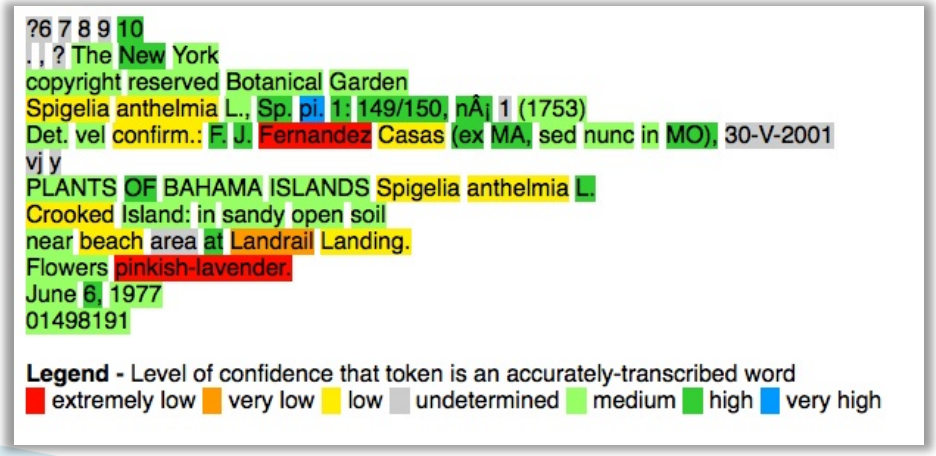
Why

- 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

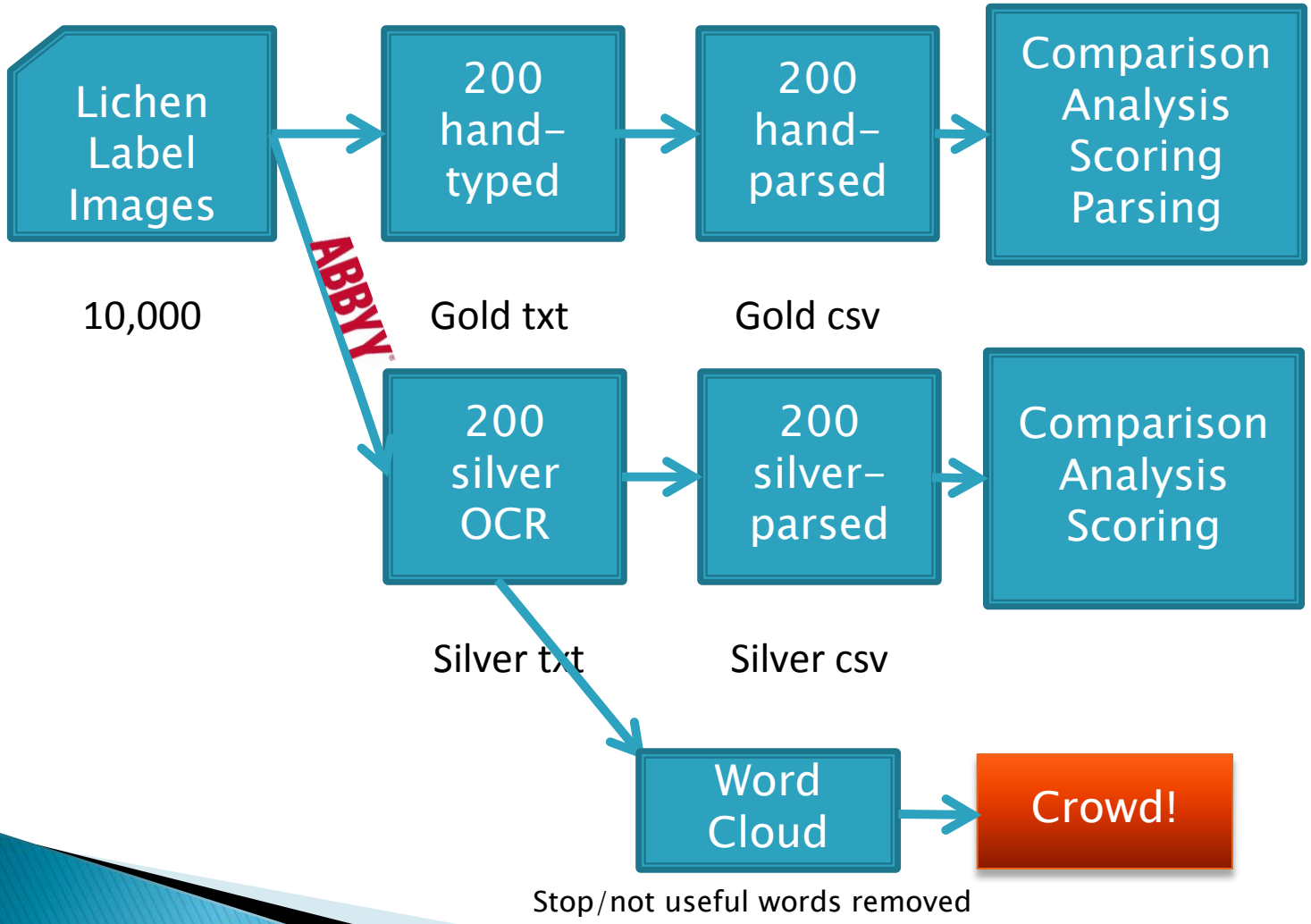
- Users like ordered datasets
- Transcription
  - **faster** with ordered/sorted sets
  - **less error prone** with sorted sets

Handwriting vs. Typed

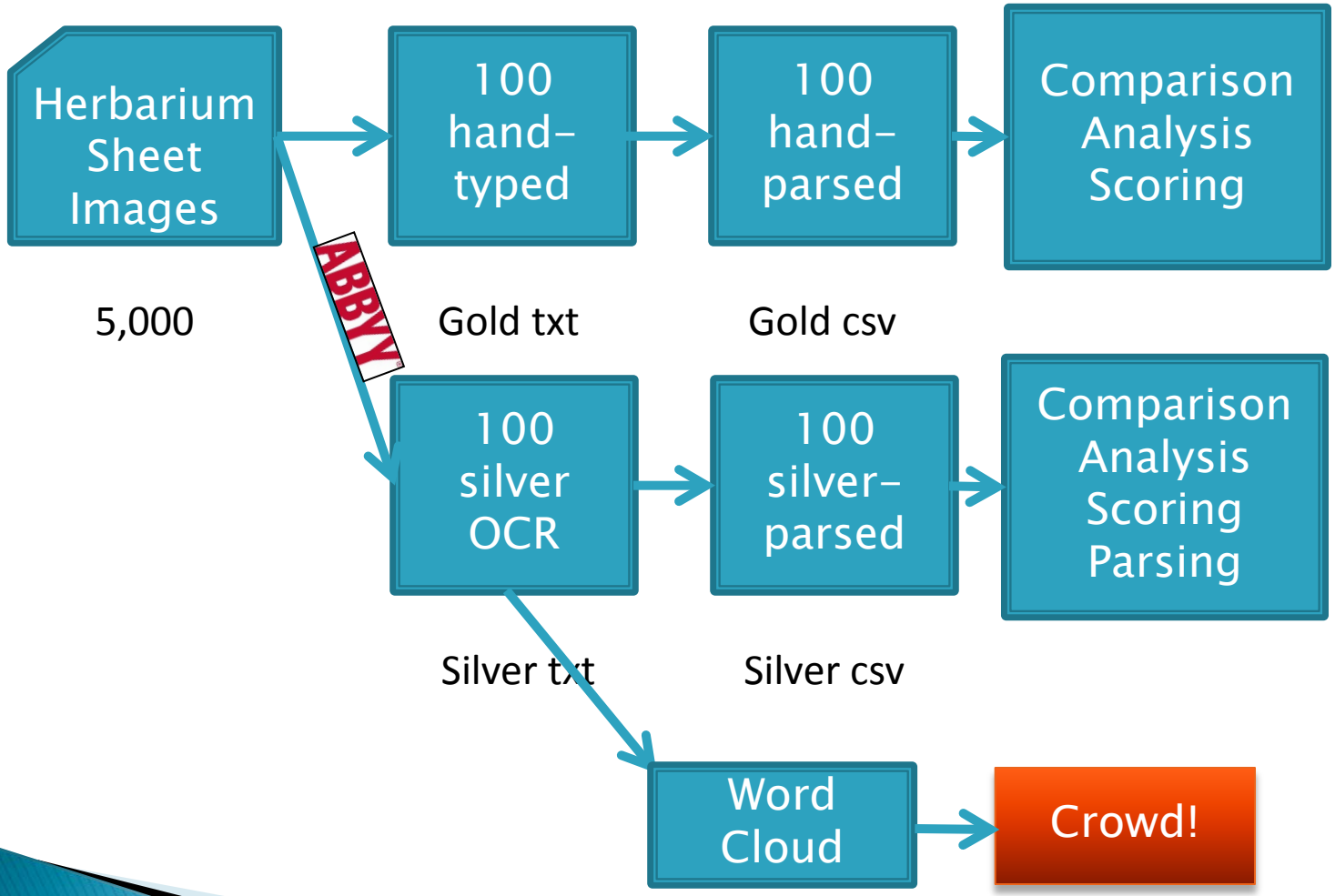
Threshold	Correct	False Positives	False Negatives
$T > 1$ and $N > 20\%$	82%	10 of 45	8 of 60
$T > 0$ and $N > 20\%$	84%	13 of 45	4 of 60
$T > 1$	79%	10 of 45	12 of 60
$N > 20\%$	75%	8 of 45	18 of 60
$N > 10\%$	81%	14 of 45	6 of 60

- **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*

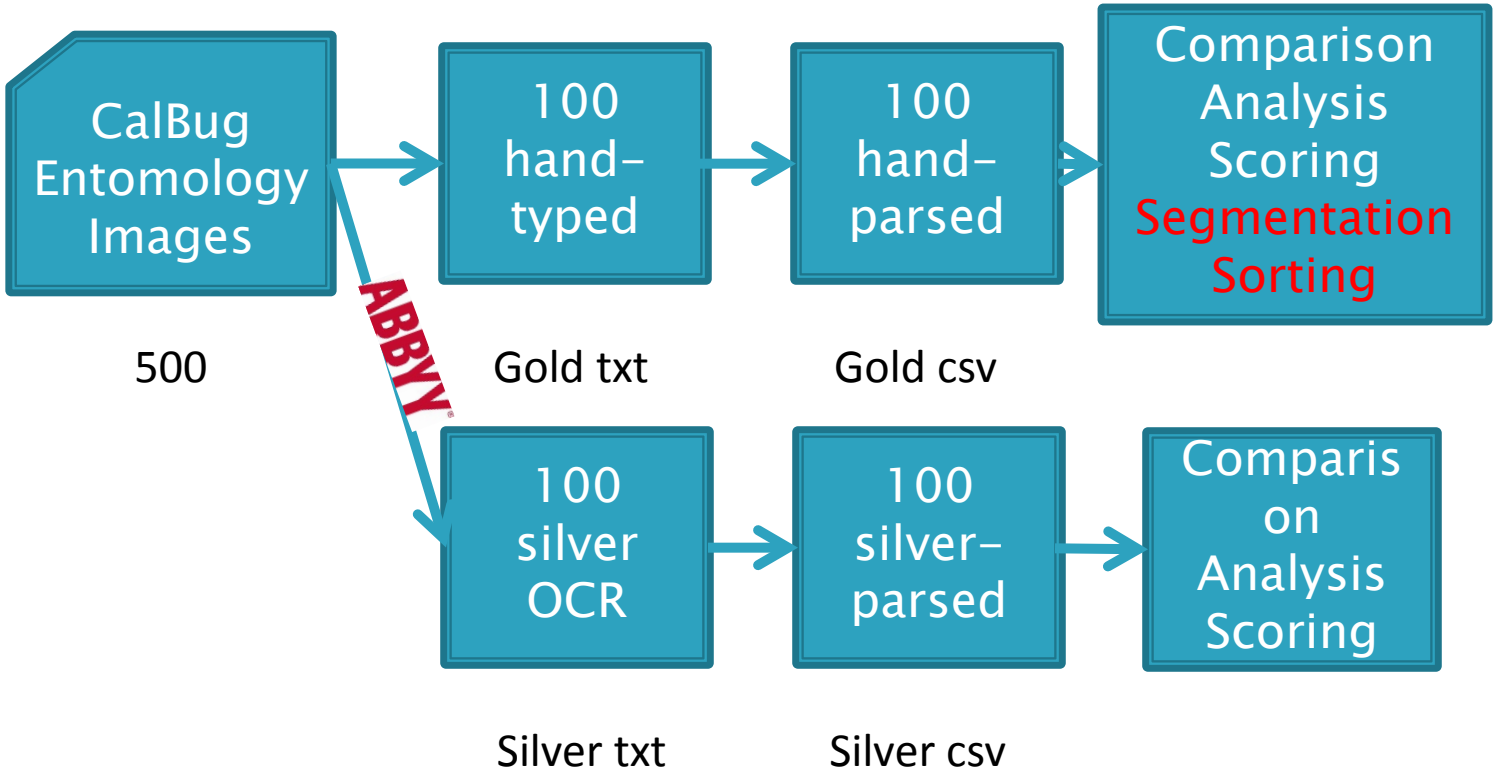
aOCR iDigBio – BRIT Hackathon Lichen Dataset



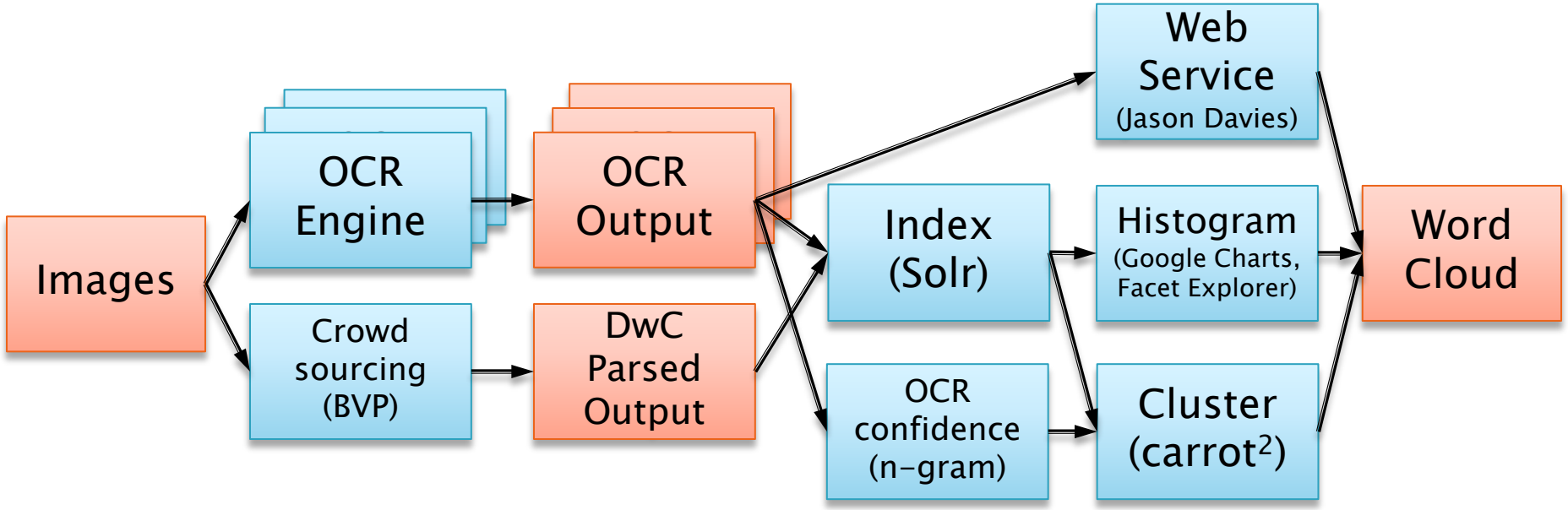
aOCR iDigBio – BRIT Hackathon NYBG Herbarium Dataset



aOCR iDigBio – BRIT Hackathon CalBug Entomology Dataset



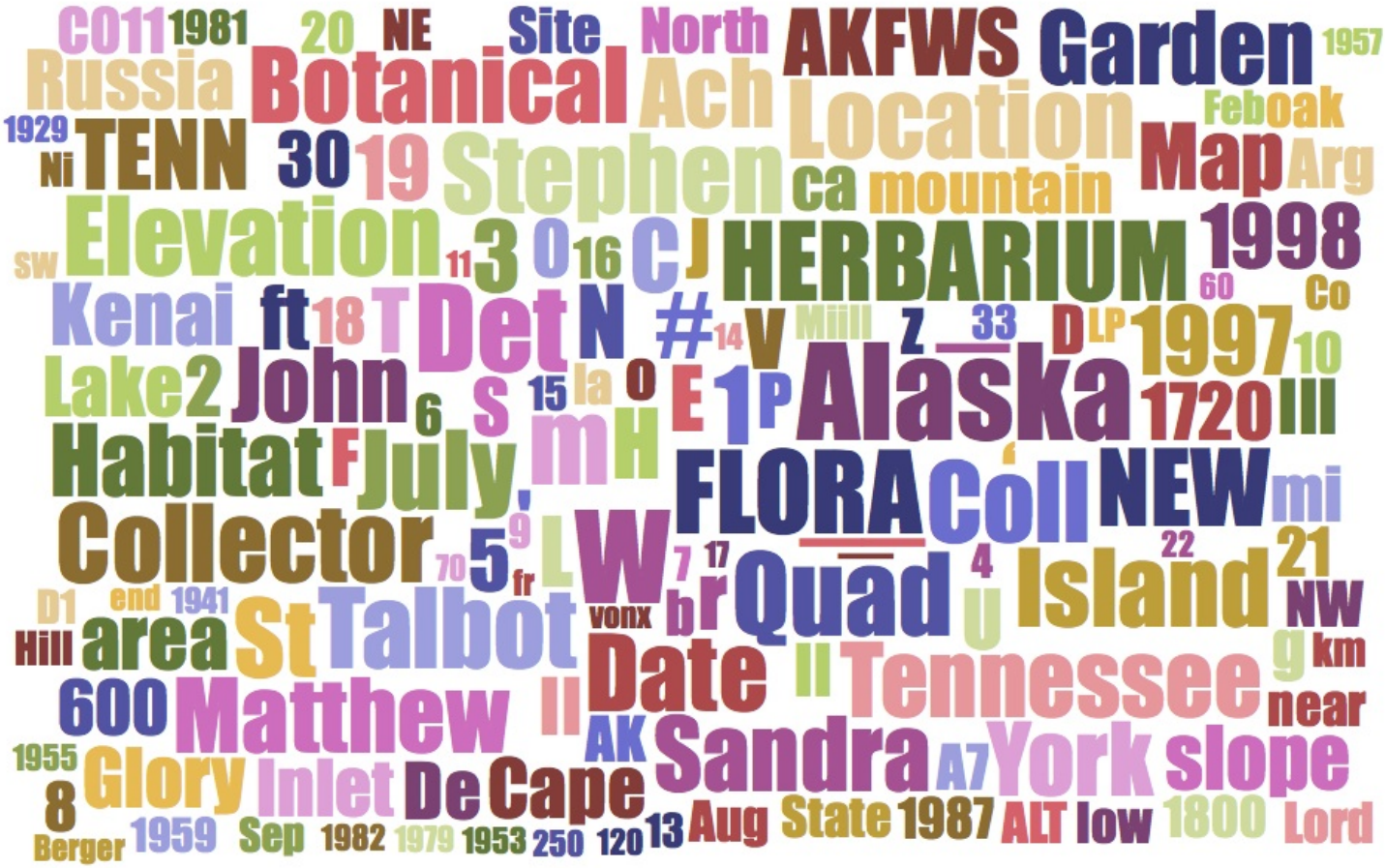
Overall Word Cloud Workflow



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

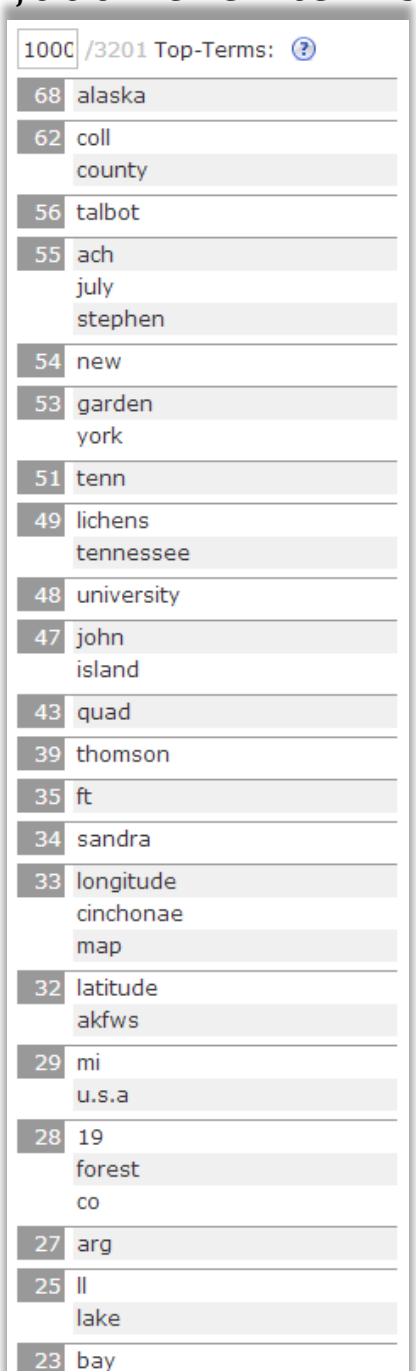


Try:
http://www.jasondavies.com/wordcloud/#http%3A%2F%2Faocr1.acis.ufl.edu%2Fdatasets%2Flichens%2Fsilver%2Focr%2FWebrootDatasetsLichensSilverOcr.txt

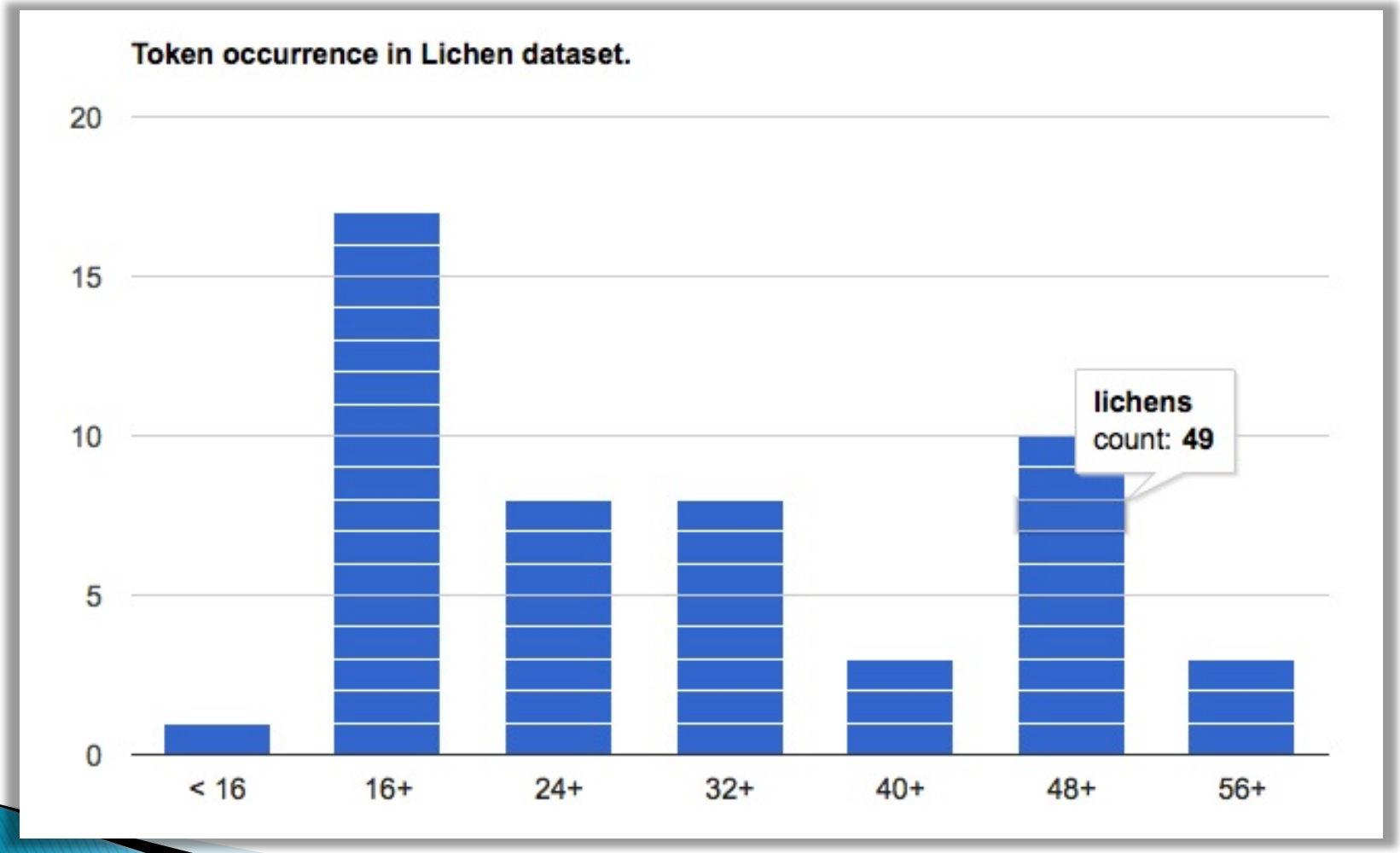
Apache Solr Indexing

- 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, verbatinLatitude,
    - verbatimLongitude, verbatimLocality

Top 1,000 lichen terms

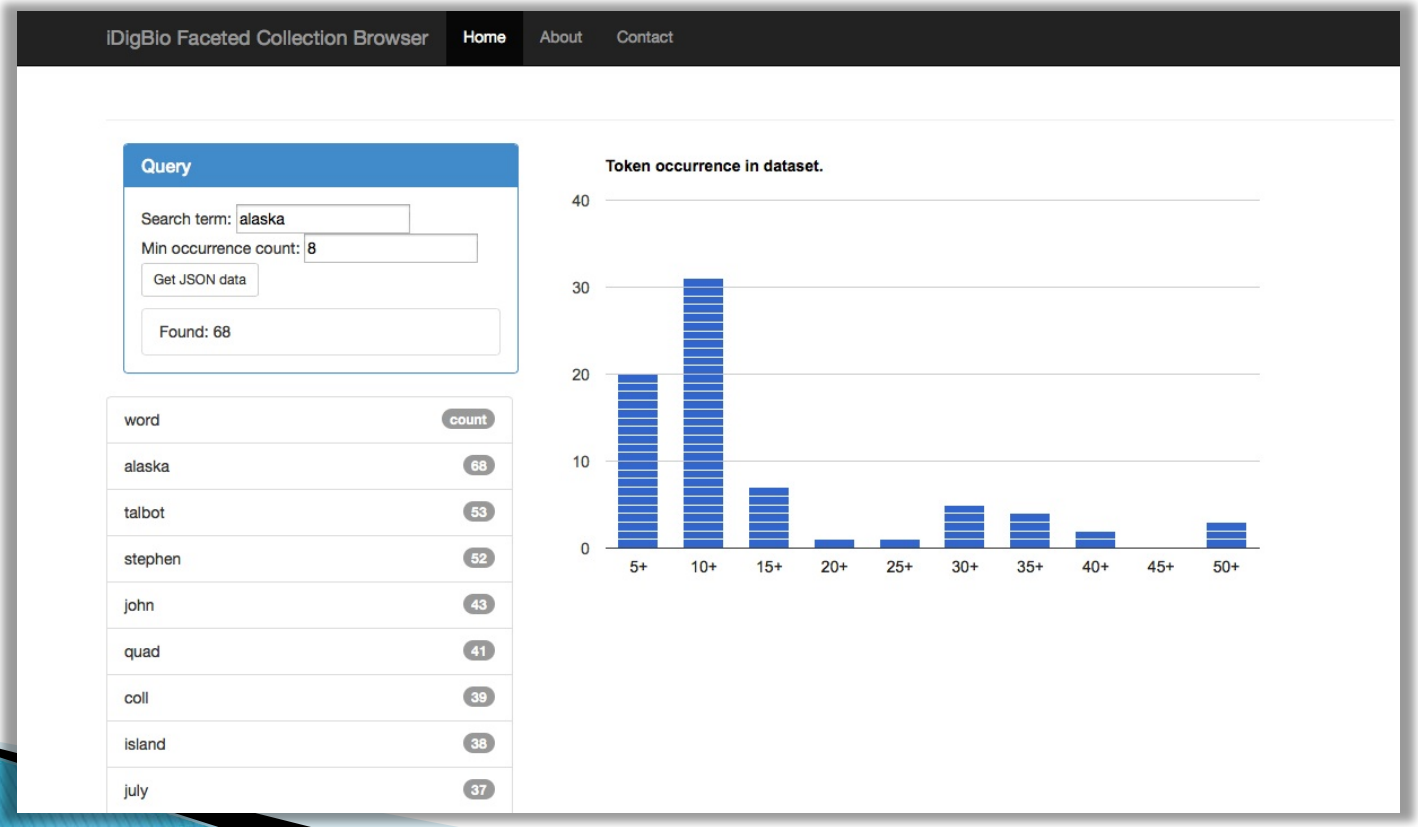


Token Histogram – Google Charts



iDigBio Faceted Collection Browser

- discovery
- how many documents have this issue



OCR Confidence Estimation

00040904.txt

```
REPORT ANY RE1DÍ TO THE IN8TÍTU
5A«.f / b-fs
The Key York Botanical Garden INSTITUTE OF ECONOMIC BOTANY Plaits of Cononiealtli of
Domica
uu.«./«. (Jici  ARACEAE
Anthurium -jacant» d<th //■ T.  93
HooKer\ K'intU - M.4.r. T. Crog <* ?3
WEST INDIES, Dominica. St.George. Above new steel bridge on road to Freshwater Lake.
15°19'N/ 6i019'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
WmÊÊ
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
```

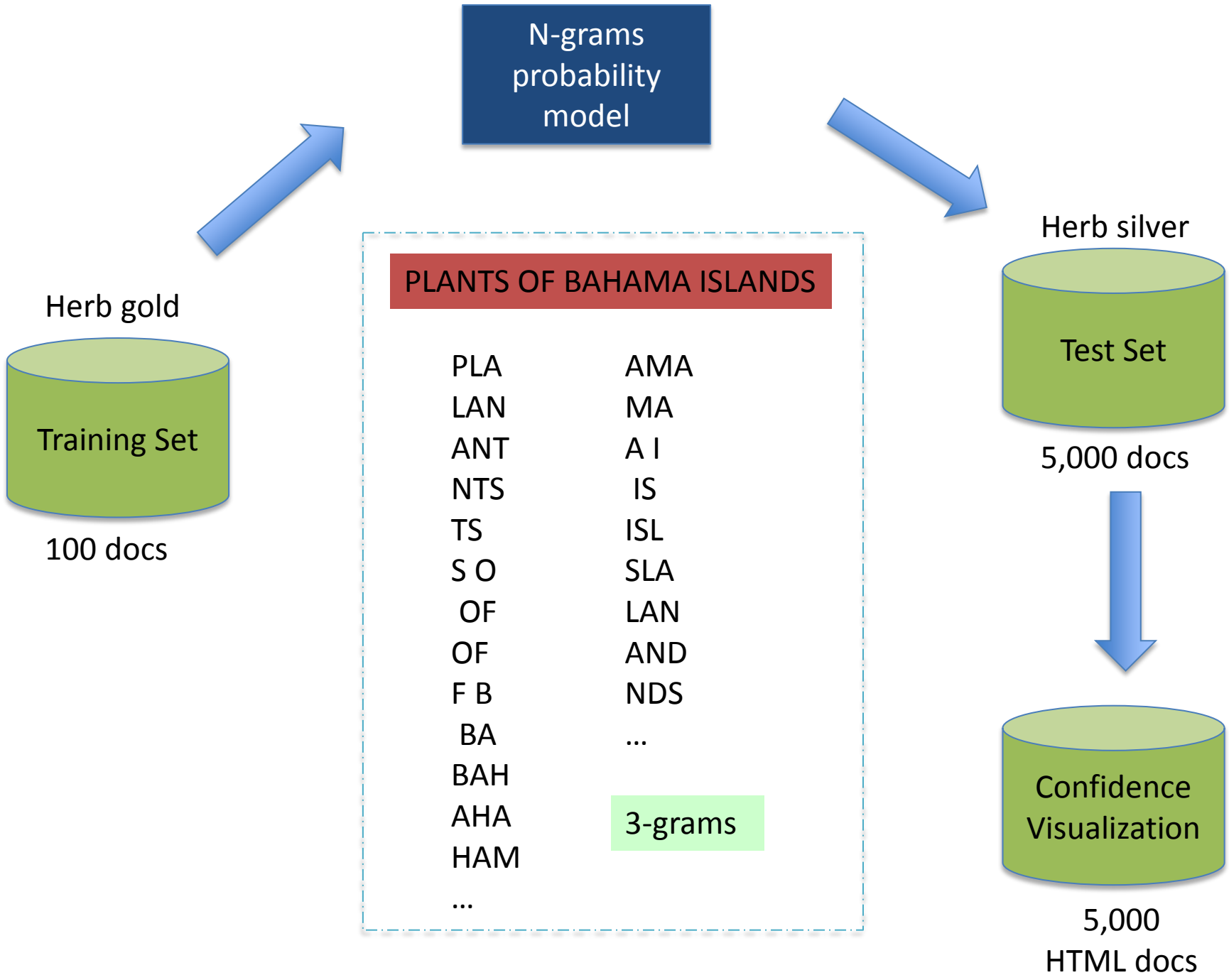
00040904.txt

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
Crog	-22.345097378723125
Garden	-22.453229125763336
Beck	-23.285538570829566
GARDEN	-23.61087758272402
Herb,	-24.92388863748758
garden	-25.276526293980815

Based on the probability of n-grams

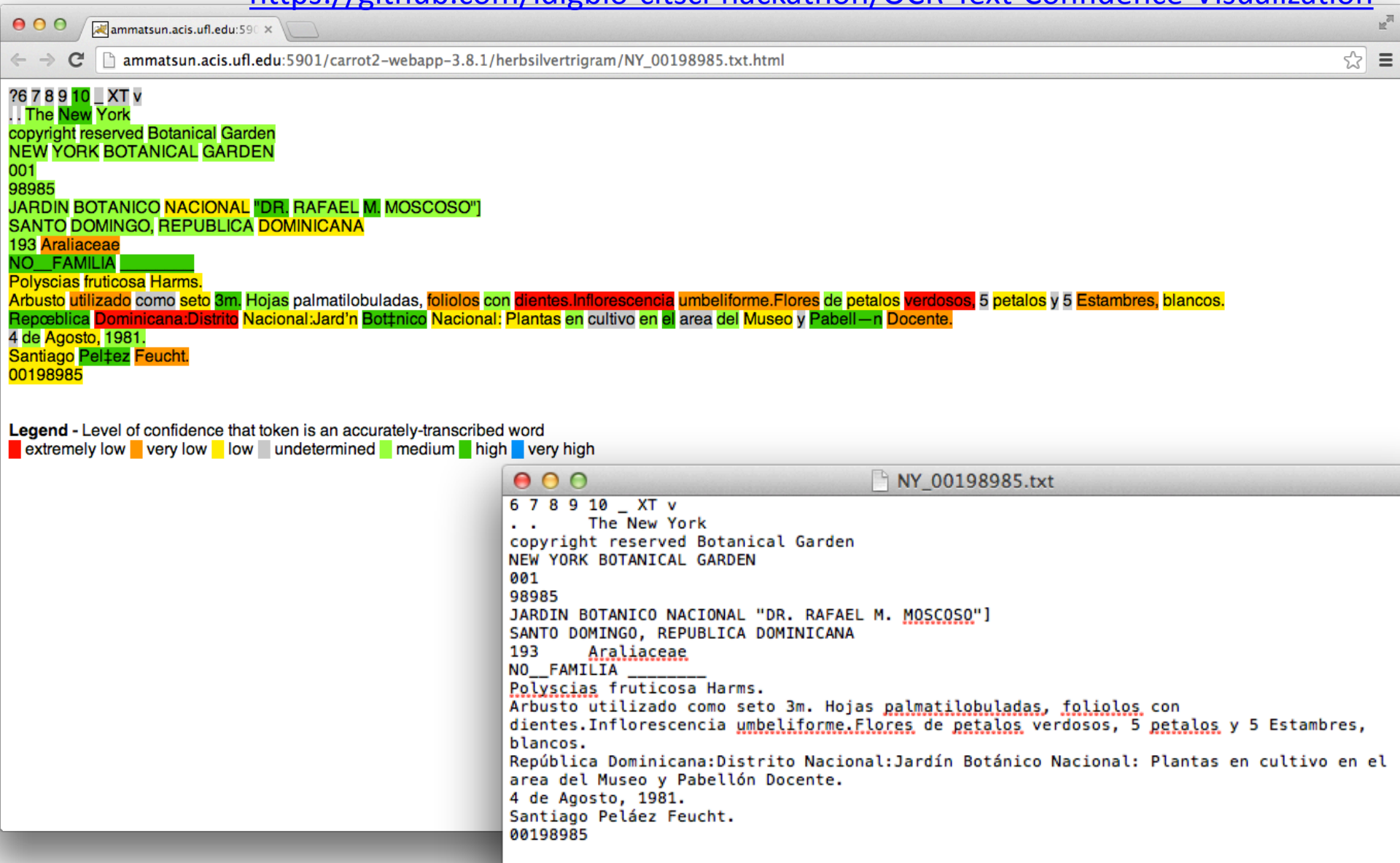
Estimating OCR confidence

- 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)

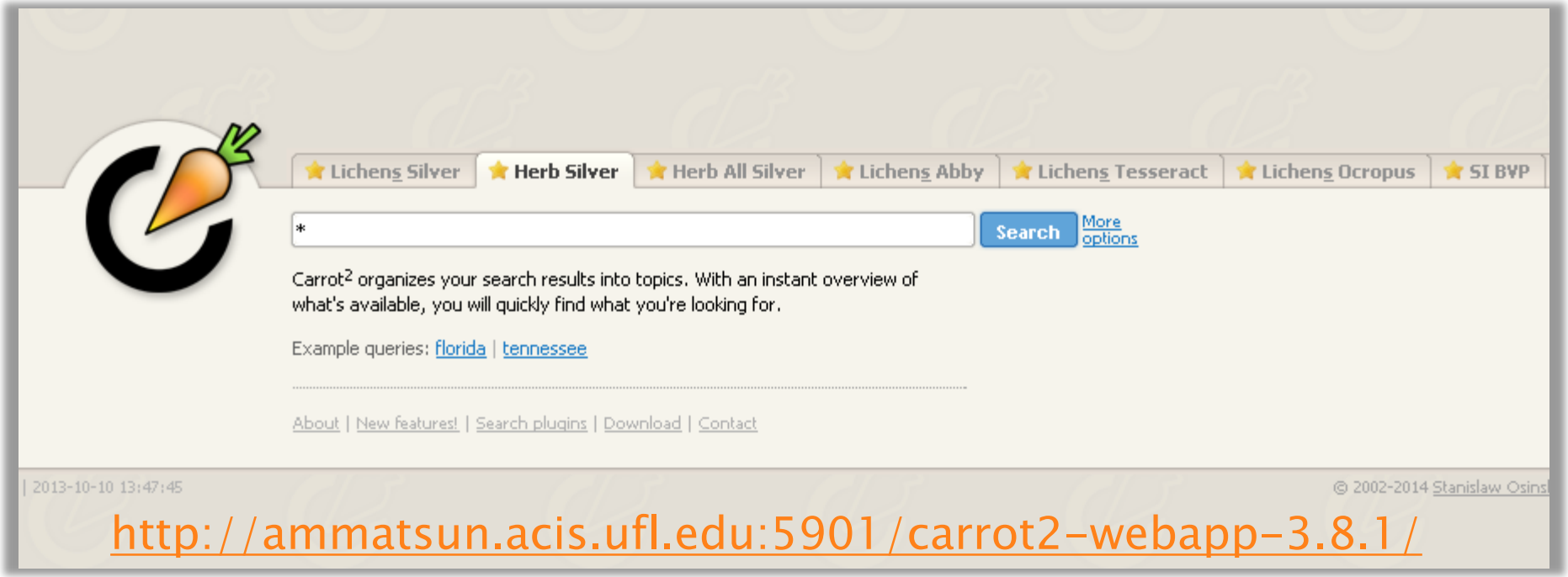

N-grams probability model
Herb gold
Training Set
100 docs
PLANTS OF BAHAMA ISLANDS
PLA
LAN
ANT
NTS
TS
S O
OF
OF
F B
BA
BAH
AHA
HAM
…
AMA
MA
A I
IS
ISL
SLA
LAN
AND
NDS
…
3-grams
Herb silver
Test Set
5,000 docs
Confidence Visualization
5,000
HTML docs

Visualizing OCR Confidence

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



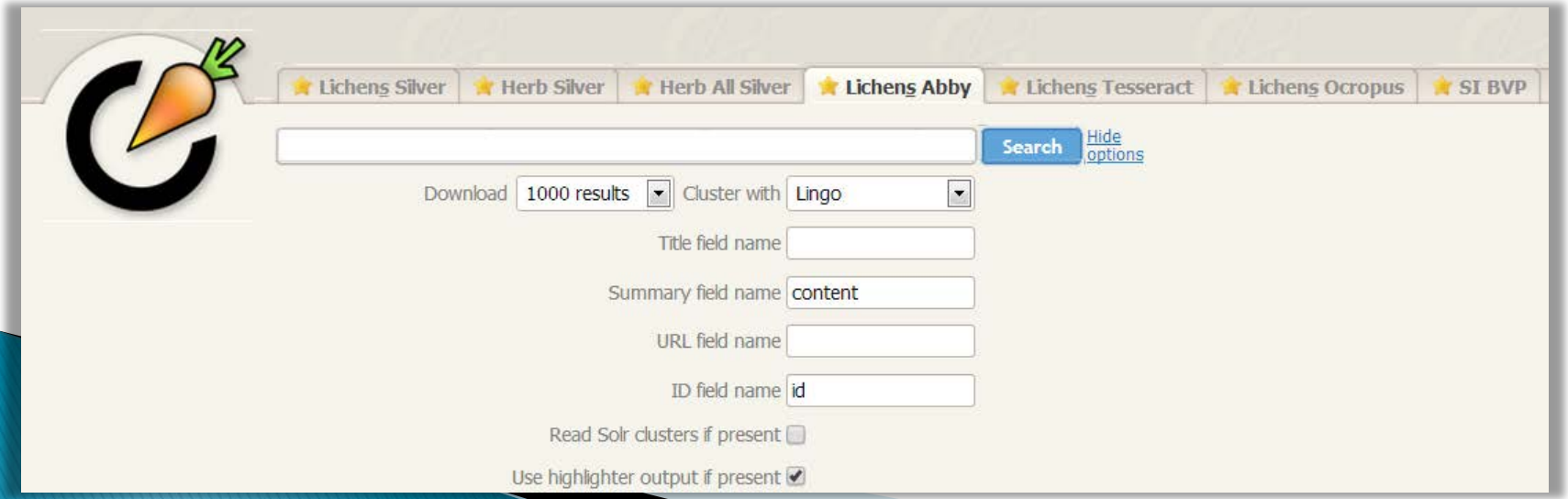
Word Cloud using Solr + Carrot²



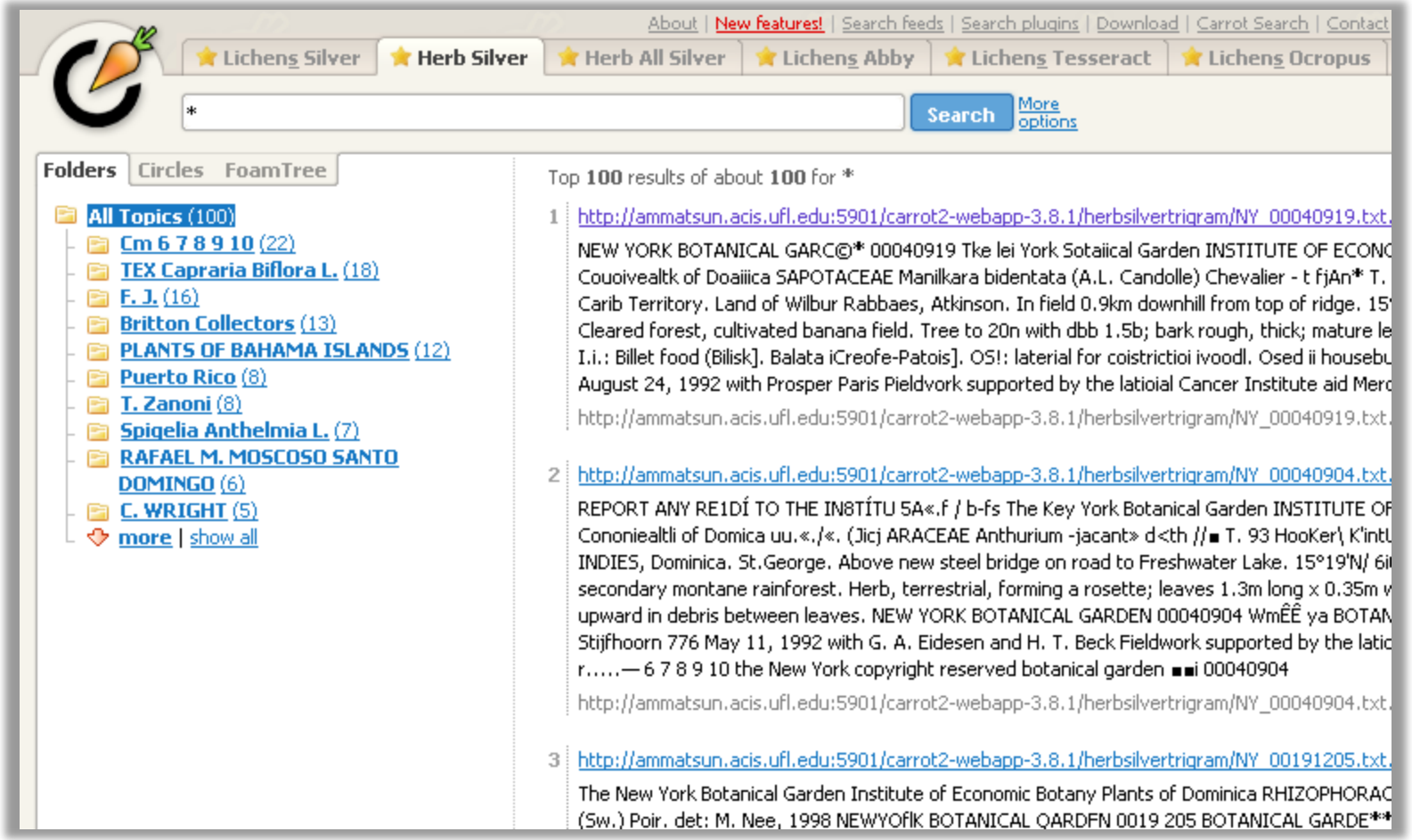
http://ammatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/

Word Cloud using Solr + Carrot²

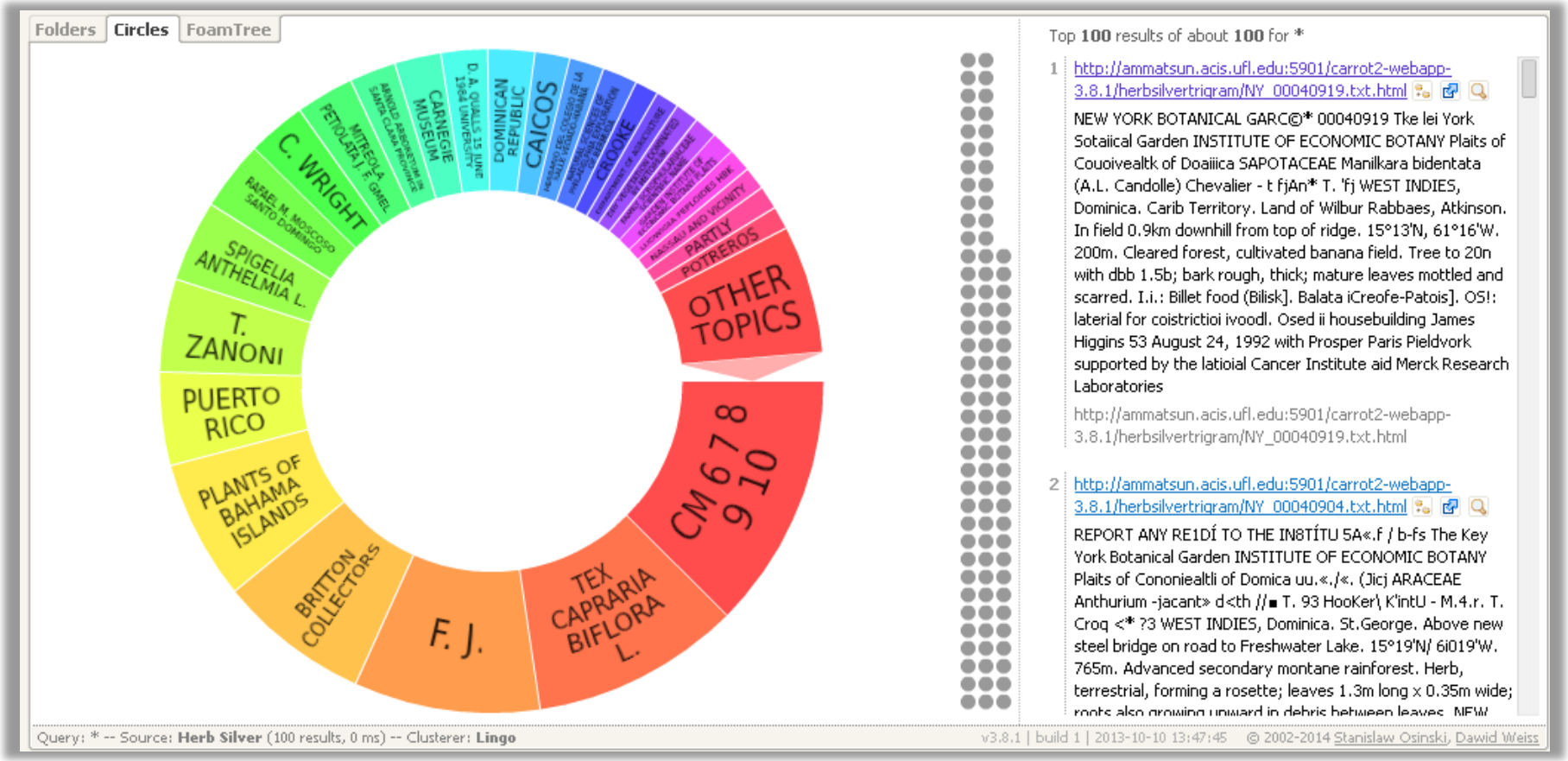
- Index OCR text output using SOLR
- http://ammatsun.acis.ufl.edu:5900/solr/#/lichenssilver/schema-browser?field=content
- Using Carrot² to visualize data http://ammatsun.acis.ufl.edu:5901/carrot2-webapp-3.8.1/



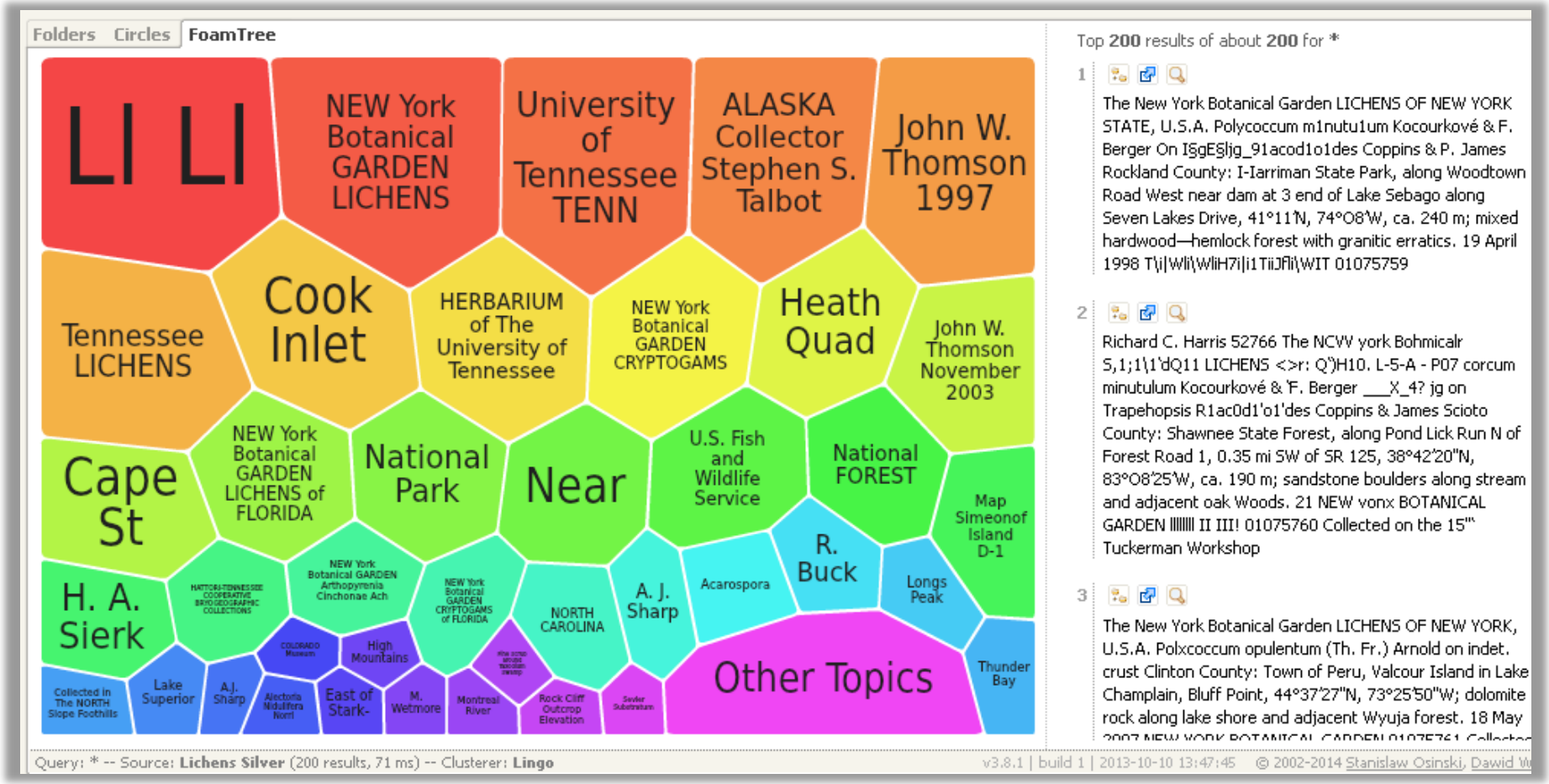
Folder View of Search



Circles Visualization of Search

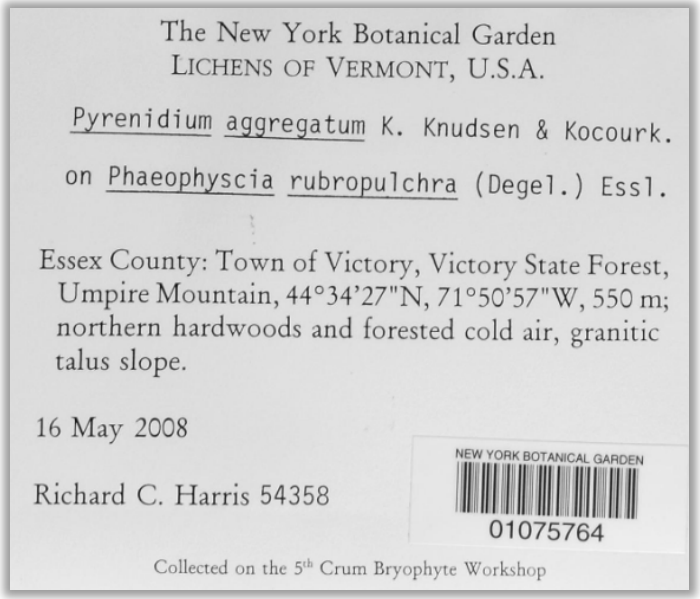


Foam Tree Visualization of Search



Source of our LI LI team name!

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


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

Pyrenidium aggregatum K. Knudsen & Kocourk.

on Phaeophyscia rubropulchra (Degel.) 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

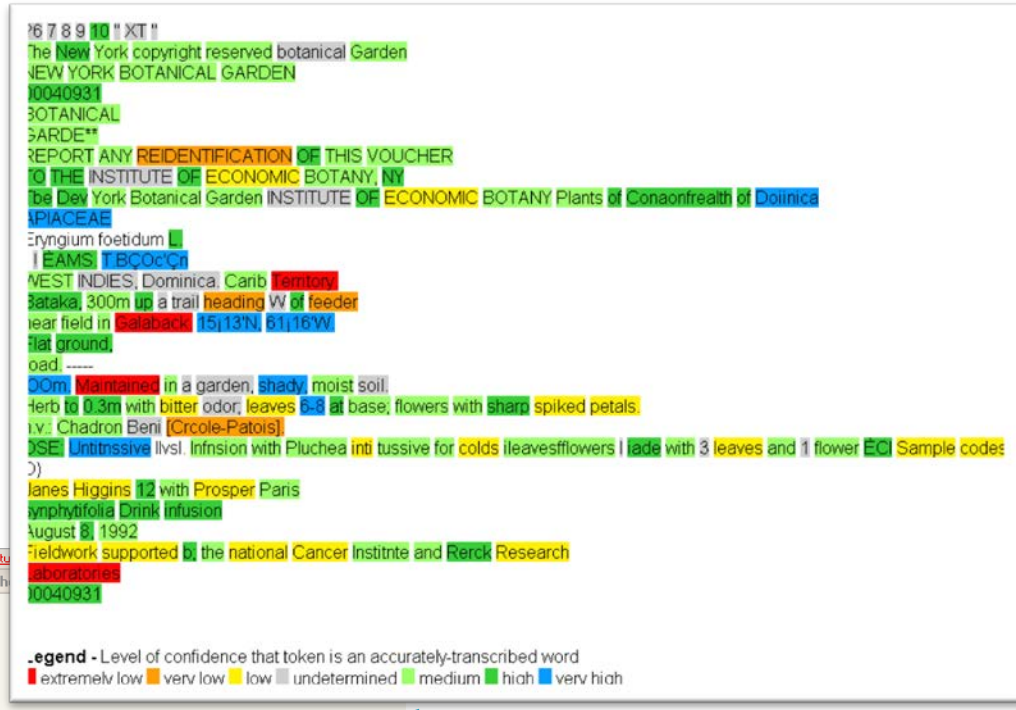
Richard C. Harris 54358

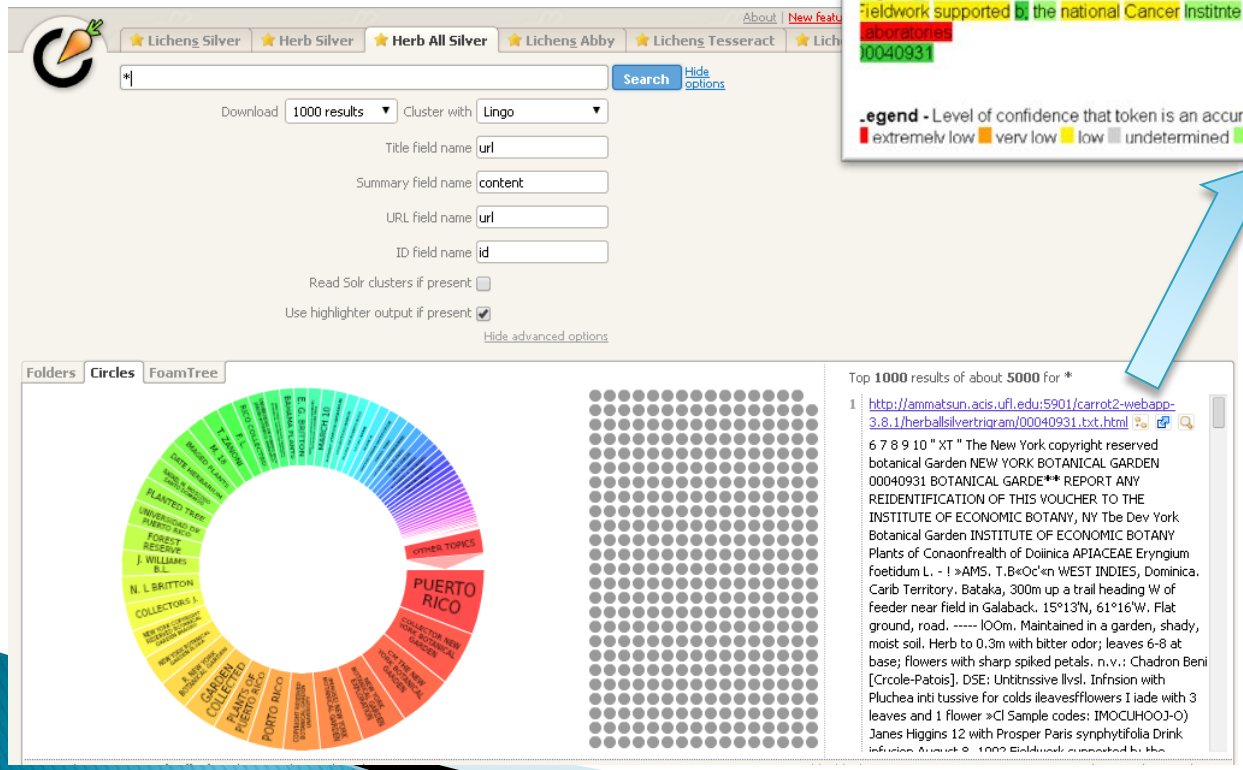
NEW YORK BOTANICAL GARDEN
01075764

Collected on the 5th Crum Bryophyte Workshop

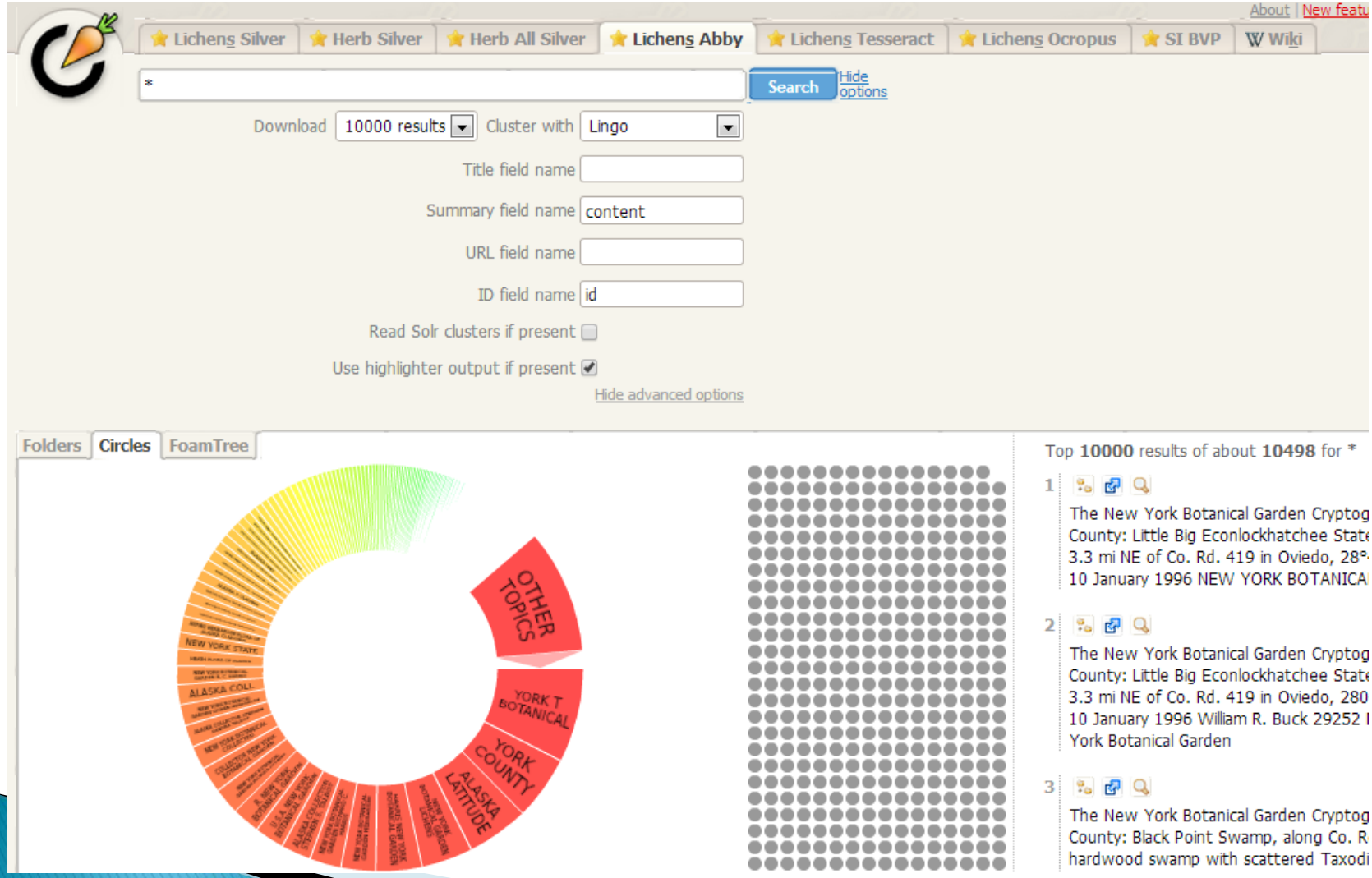
- Data Sets
  - silver (ABBYY) OCR output from 200 lichen packet images
  - all ABBYY OCR output from 10,000 lichen packet images
  - silver and gold (ABBYY, Tesseract) OCR output from 100 herbarium sheets
    - OCR output from 5000 herbarium sheets
  - SI BVP dataset

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



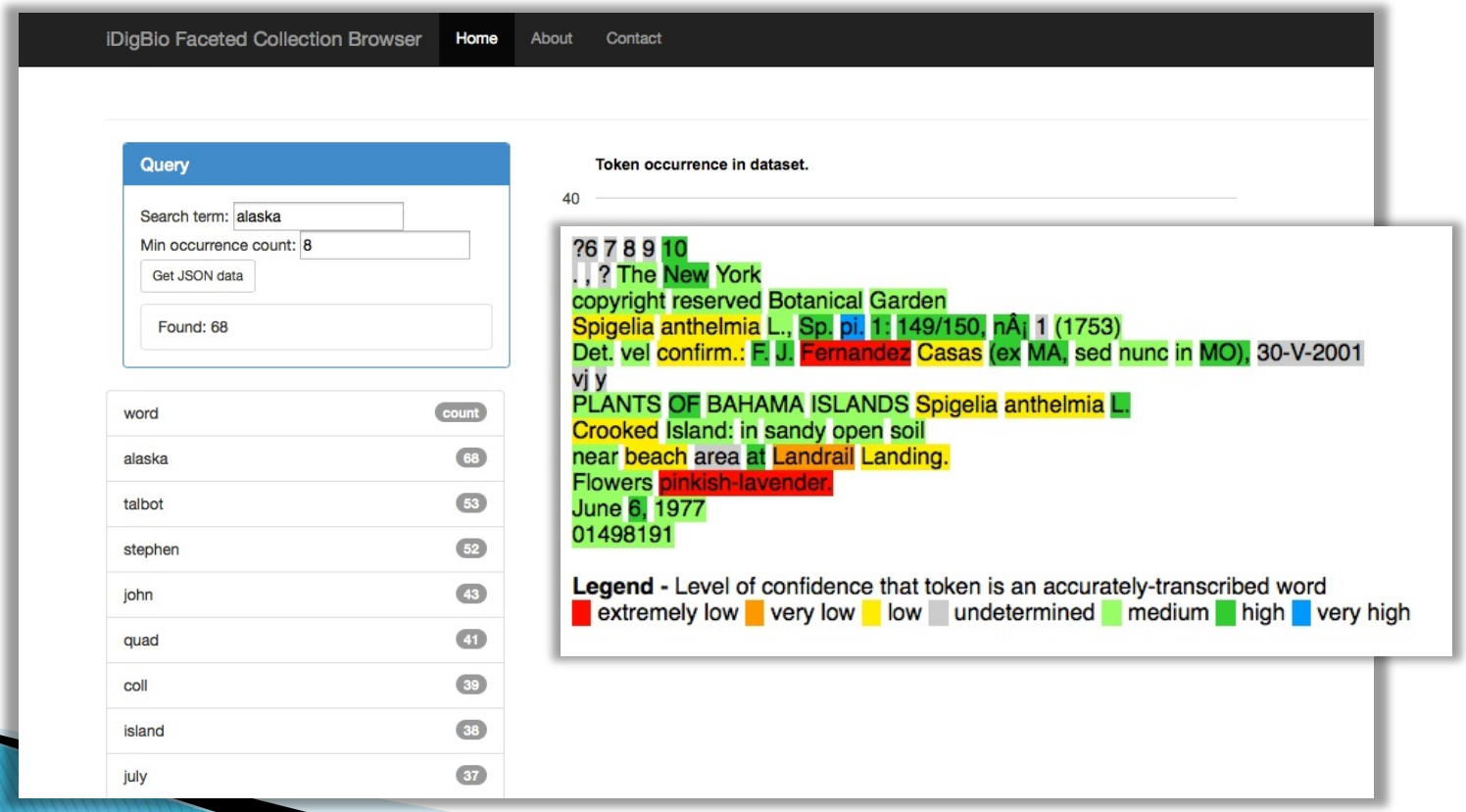


10,000+ records sample



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

GOT IT! LET ME CHOOSE MY TRANSCRIPTION GROUP

Choose your group:

Country

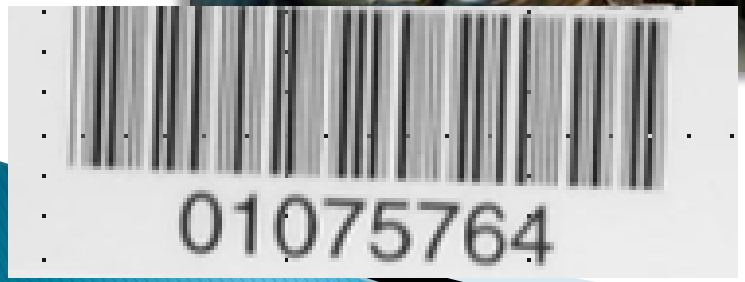
- Use for initial sort or validation

iDigBio
Integrated Digitized Biocollections

Using OCR output to enhance the transcription process

Transcribe museum records to
TAKE NOTES FROM NATURE

ATLAS OF LIVING AUSTRALIA
sharing biodiversity knowledge
Biodiversity Volunteer Portal | Expeditions | Tutorials
Biodiversity Volunteer Portal



LI LI Team!

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