Insection

An entomology specimen databasing pipeline

Suhas Gupta

Neha Kumar

Shweta Sen

Apik Zorian

Insects play many key ecological roles

Decomposers

Insects play many key ecological roles

Pollinators

Insects play many key ecological roles

Pest Controllers

EMEC1027379 Lipeurus sp.jpg

Millions of Labeled Specimen spanning 100+ years

How do we harness this unstructured information for ecological studies?

EMEC1025003 Austromenopon transversum.jpg

Our Client

Client: Essig Museum of Entomology (Berkeley, CA)

Dataset Size: ~350k undatabased images
(to be applied on 5M images)

Current Method: Manual Annotations

At the current rate, Essig's collection would not be databased for another

12 years

Our solution can do the job in about

3 days

Our Mission

Digitize and database insect specimen labels in a scalable manner

Input

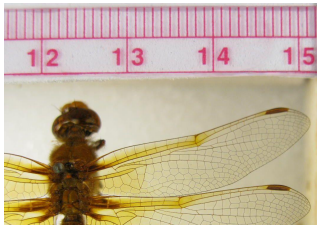
Battus philenor hirsuta

Output

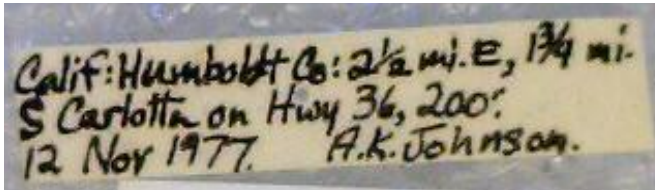
ID	EMEC464559
Genus	Battus
Species	Philenor hirsuta
Collector	Bertram C. Walker Collection
Location	Lake Merced San Francisco Cal.
Date	5-1-45

Insect photographs vary tremendously

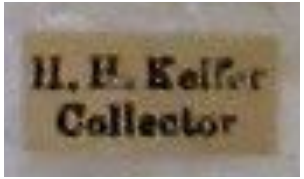
Date format with Roman Numerals



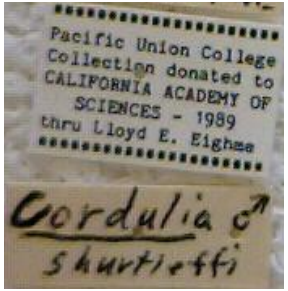
Ruler Interference



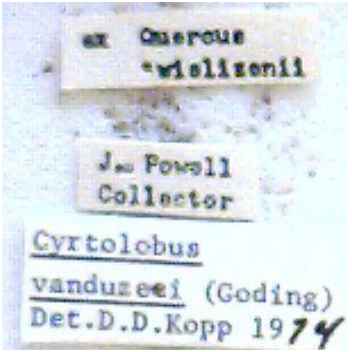
Abbreviations, Fractions



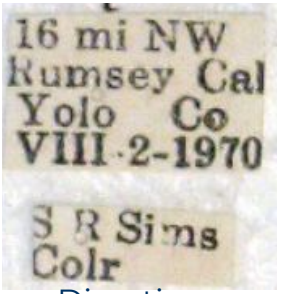
Poor Image Quality



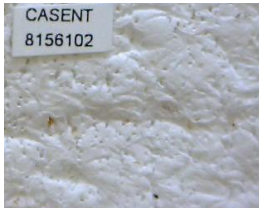
Underlines, Symbols, Borders



Text Blur, Non-Dictionary Words

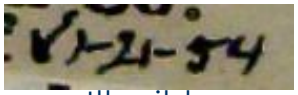


Directions / Locations



Textured Backgrounds

QR Codes



Illegible Handwriting

Stitch images

Optical Character Recognition

Information Extraction

Extract ID, genus, species

OCR with Google Cloud Vision

Extract collector

Deconstruct output

Extract date

Spell check to flag poor output

Extract location

Build ⟵ ⟶ *Buy*

Tesseract

```
INED CRED| ]
U - , `
i o . Yenin
S , )
et e - . :
gt Al N
AT i -
. STy P CASENT BESsrves
3 peectRiativEE. . . { 8156102 %
o
e L 1 ;
o EmEmei A
e S
Ber R 5 b > <
e einum s L P dg:;rt"\""%- T
L SR S —
W o BN T e
Cordulia shurtlefffi -
```

Cordulia shurtlefffi

Google Cloud Vision API

```
A130 640x480 2012/02/07 12:08:43
G. Keszler
Plaskett Calif.
1965
July
CASENT
8156102
CERduiiNAE
Pacific Union College
Collection donated to
CALIFORNIA ACADEMY OF
SCIENCES - 1989
thru Lloyd E. Eighse
```

Google's Cloud Vision API

5 million images = 5 million API hits = $7500

How can we better scale costs?

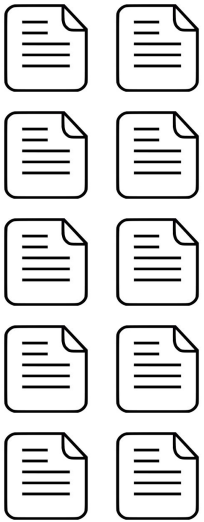
https://cloud.google.com/vision/pricing

Developing a pipeline to stitch images together reduced the number of API hits, **driving down cost 10X**

Stitch Images Together → *Send Images to API for OCR* → *Deconstruct Output*

Also contains:

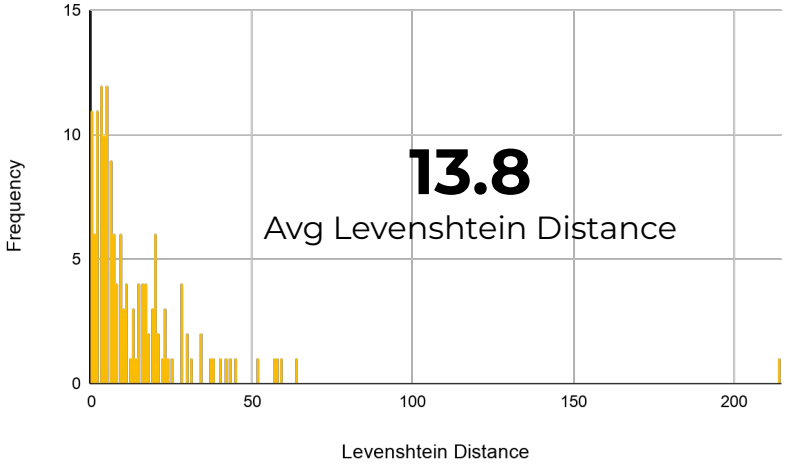
- Bounding boxes
- File sizes
- File Names



We implemented a scalable method to flag Inaccurate Outputs

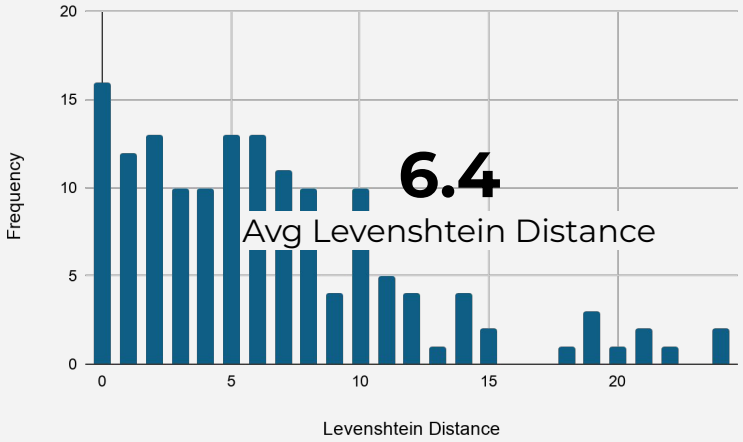
More Accurate, Less Scalable → *Less Accurate, More Scalable*

Manual Check



* n = 150 with 144 character average per sample

Spell Check with Pychant



Levenshtein Distance: Text similarity measure that compares two strings and returns a numeric value representing the distance between them

It takes a village...

We employed a myriad of methods of varying complexities to achieve our task

Genus / Species / ID

Pulled from filename

100% Population

Battus philenor hirsuta

Collector Name

Lookup + Regex

78% Population

Collector Extraction

56% completed via a lookup

22% completed via regex

78% of records populated, with over 80% accuracy

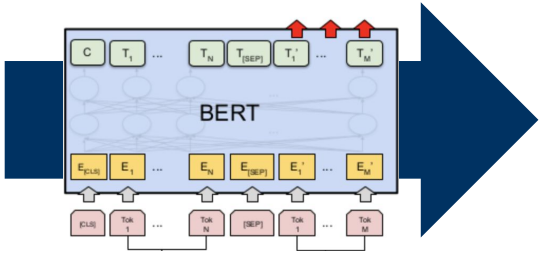
Location
Question Answering
100% Population

Location Extraction

Approached as a Named Entity Recognition problem

Staphylus ceos

UC Berkeley EMEC 451828 011 12 13 Teques quetengo Morelos, Mexico D.H.Janzen Collector Staphylus ceos



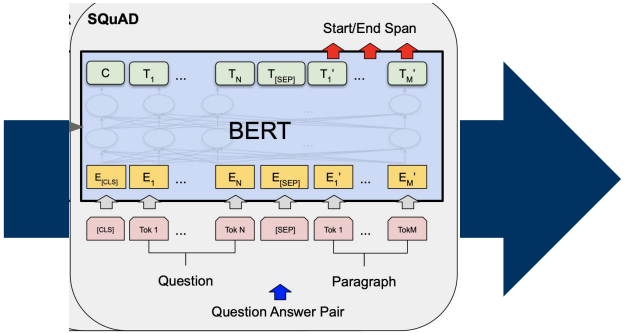
13 Mexico Staphylus ceos

Location Extraction

Approached as a **Question Answering** problem

Staphylus ceos

UC Berkeley EMEC 451828 011 12 13 Teques quetengo Morelos, Mexico D.H.Janzen Collector Staphylus ceos



Where was the specimen collected?

Teques quetengo Morelos, Mexico

60% of extracted locations were sufficiently granular

Date Collected

Regex + Question Answering

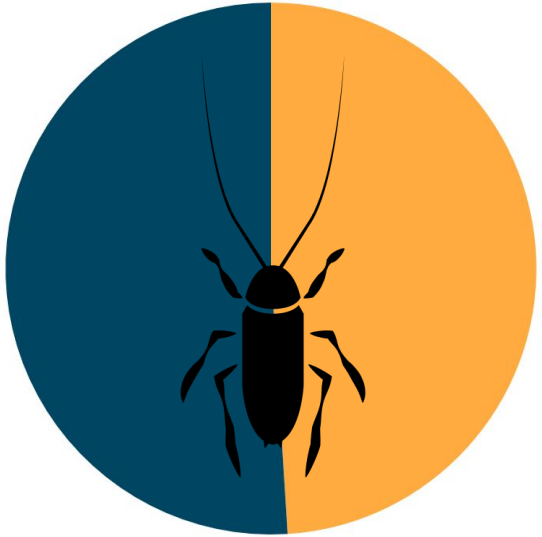
100 % Population

Date Extraction

Step 1: Regex Extraction → *Step 2: Question Answering*

51%

Of records extracted, with 90% accuracy



49%

Of records extracted, with 65% accuracy

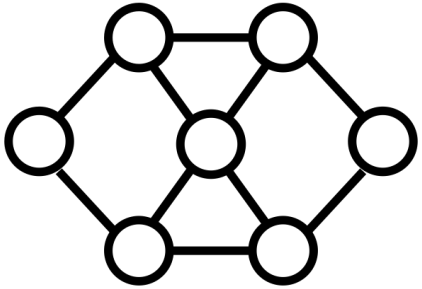
A score provides model confidence for BERT-extracted dates

EMEC1027379 Lipeurus sp.jpg

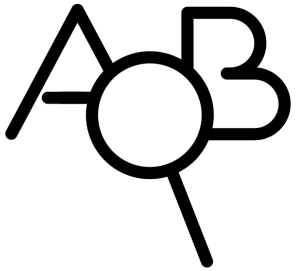
Score	0.999997	0.513901
OCR Output	1740 Lipeurus squaridus N. Dafila acuta, Bezzi Italy. V. L. KELLOGG, STANFORD UNIVERSITY. EMEC1027379 Lisperusu sp. jpg	26 27 28 29 Sacramento Sacto. Co. Calif. VH-12-1969 R.P.Meyer Collector UC BME P 0204864 Cicindela oregona oregonal

Collector *Location* *Date*

Recommendations for Further Research



Fine tune BERT on databased records for Question Answering



Incorporate a more expansive entomology vocabulary in the spell check dictionary

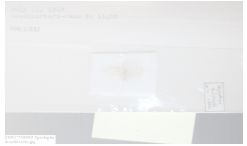
Establish further QC checks

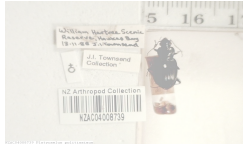
Impact

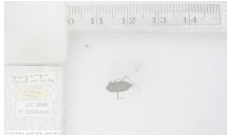
Empower entomologists to make key ecological insights

Thank you!

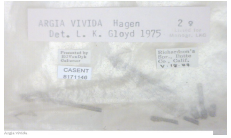
EMEC1027379 Lipeurus sp.jpg

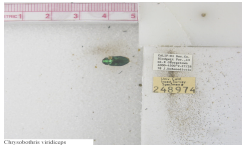






EMEC1025003 Austromenopon transversum.jpg





Lestes disjunctus

References

Papers

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Data Sources

https://essigdb.berkeley.edu/

http://bnhmipt.berkeley.edu/resource?r=essig

Tools

Google Cloud Vision API

OCRopus

Tesseract

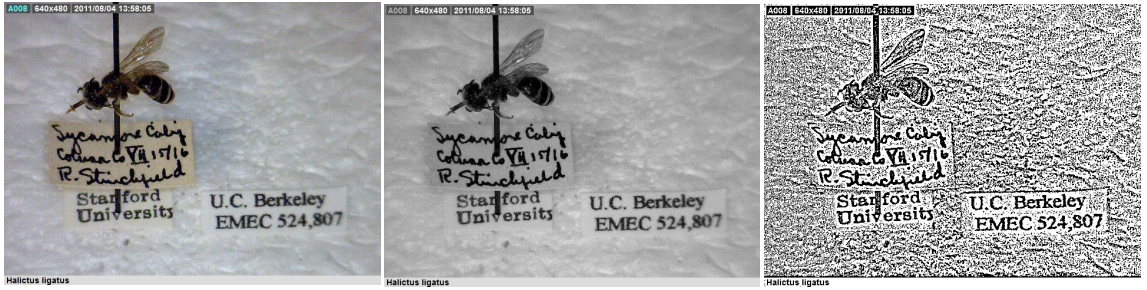
Levenshtein Distance Online Calculator for initial exploration (later used the nltk “edit_distance” function)

Icons

The Noun Project

Appendix

Levenshtein distance (or edit distance)



	Color	Grayscale	Binary
1 Image	0	30	306
10 Stitched Images	119	104	398
50 Stitched Images	116	101	394

** I Image, Color used as baseline*
***1338 characters in reference sequence*

Levenshtein Distance Calculator: https://planetcalc.com/1721/

Changing Resolution

- Right now we take the smallest image (by width) in a batch and resize all the images so they have the same width (height gets scaled accordingly). Max width is set at 1600 pixels
- Most images have a 4:3 aspect ratio
- Some images are 4000 by 3000 pixels. We looked at how resizing these can affect the OCR Performance and output

	Levenshtein Dist	Avg Time per Image
4000 X 3000	0*	7.24s
1600 X 1200	24	2.16s
640 X 480	55	1.64s

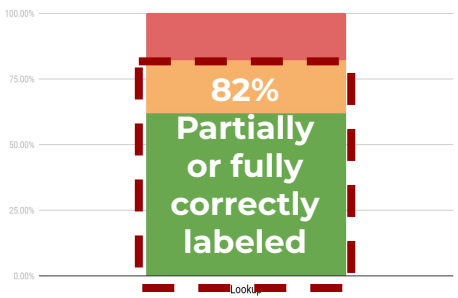
**4000 by 3000 pixels is the baseline reference*
*** This test took 3 images since we would hit the 75 megapixel limit from cloudvision with a full size batch*
**** Reference sequence is 651 characters*

78% of Collectors Extracted with >80% Accuracy

Lookup Method

8218 Distinct Collector Values from Databased Records

56% nondatabased records matched

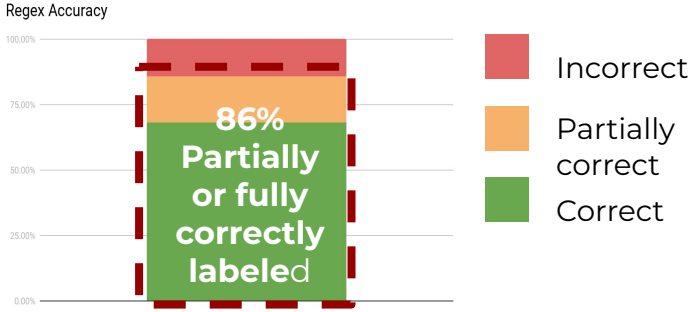


Spotcheck n size = 50

Regex Method

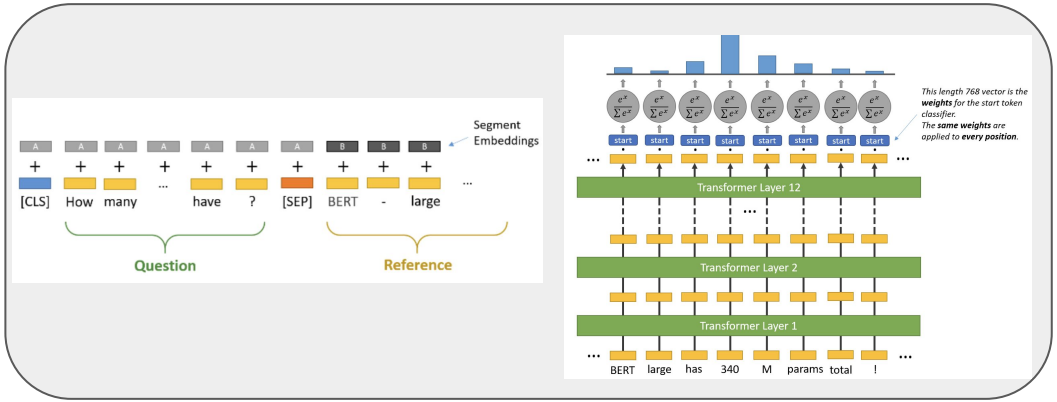
Use Regex to find the word preceding “coll”

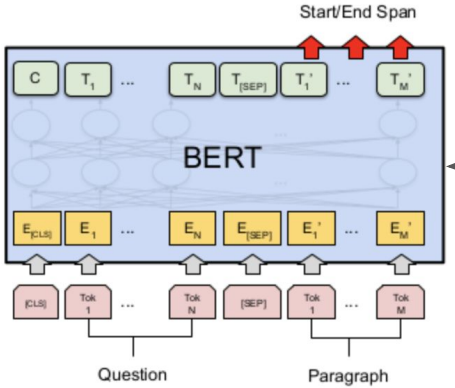
22% nondatabased records matched (half of remaining records)



Spotcheck n size = 50

Location Extraction: Anatomy of a QA Task with BERT





SQuAD2.0
The Stanford Question Answering Dataset

Location Extraction

Cicindela ocellata rectilatera

- BERT NER fined tuning / statistical models don't capture the contextual phrases that define a *location* in the label text
- This is a reading comprehension problem asking "*Where was the sample collected ?*" from a list of *paragraphs*

Finetuned BERT

Rio I-ORG
Guayalejo O
+ B-ORG
Hiway B-MISC
85 B-ORG
VIII-13 O
-1959 B-PER
MEXICO, B-LOC
Tamps. O
A.S.Menke O
L.A.Stange B-ORG
Collectors I-PER
5 B-LOC
26 B-LOC
27 B-LOC
28 B-LOC
2 B-LOC
Cicindela I-ORG
ocellata I-ORG
rectilatera B-LOC
Chd I-ORG
det. O
N.L.Rumpp O
UC I-ORG
BME I-ORG
P I-PER
0209159 O
Cicindela I-ORG
ocellata I-ORG
rectilatera B-LOC

Spacy 3.0

85 CARDINAL
-1959 MEXICO GPE
Tamps PERSON
26 CARDINAL
2 CARDINAL
Cicindela PERSON
rectilatera Chd det PERSON
BME P ORG
0209159 DATE
Cicindela PERSON

No Good

GPE: Geo-political entity
CARDINAL: Numbers that don't fall in other categories
Spacy corpus : OntoNotes 5.0

Reframing the Question

Earlier: Extracting named entities

```
Alice lives in California
 PER    O    O     LOC
```

Now: Question Answering

?

Where does Alice live?

California

!

ocr_output
11 12 13 14 9 mi SE Medora Billings Co.,N.D. VII-7-1975 Mike Brand,
0. 0'11 12 13 14 Dickinson Stark Co..N.D. VII-21-1975 Mike Brand,Co
11 12' 13 Tex. Prionus rissicornis 14090. Hald UC Berkeley EMEC 91
11 12 13 14 Ft.Collins Colo.7-16-03 VanDuzee Collector UC Berkeley
0. 11 12' 13 W.T.Mewaldt 8-18-60 Danville Va.188 UC Berkeley EMEC
0. 11 12 13 Pine Ridge Shannon Co. S.Dak. W.S,Cook Collector UC E
0'11 12 13 1. okla, Co. Sept. 1955 UC Berkeley EMEC 916999 Prionu
11 12 13 14 Volga SDaa UC Berkeley EMEC 917000 Prionus fissicorn
Pine Ridge Shannon Co. S.Dak. 11' 12 13 1 W.S.Cook Collector fissig
0'11 12 13 1 Dickinson Stark Co.,N.D. VII-15 1975 Mike Brand,Collr U

Using Question Answering, 67% of extracted locations were sufficiently granular

%	Classification	OCR Output	Extracted Location
10%	Captures most granular location available	0'11 12 UC Berkeley EMEC 663896 CALIFORNIA: Alameda Co. Albany, U.C. Gill Tract 26 September 89 K.S. Hagen 789 ex alfalfa Aphidius ervi	CALIFORNIA: Alameda Co. Albany, U.C. Gill Tract
29%	Captures regional info	U.C. Berkeley EMEC 626,182 11 12 13 nr.Little Toad Lake Becker Co., MN VII-2-1994 J.R.Powers,collr. Ammophila sp.	Little Toad Lake
18%	Captures county info	6. 7 8. 1 U.C. Berkeley EMEC 653,060 CALIF:Mendocino Co. Mendoc ino Natl.For. Little Doe Capgd. VIII-11/13-1975 J.A.&J.M.Chensak Sphex sp	Mendocino Co
9%	Captures state / country info	U.C. Berkeley EMEC 631,133 0. o 11 12 MEX:Baja Calif.Norte 4 mi S.El Socorro 111-25-1973 at light J. Powell Ammophila sp.	Baja Calif.Norte
19%	Incorrectly captures location	2 23 24 2. Calif:S L.O.Co., Santa Margarita, 5mi. NE.VI-13-63 Clarkia speciosa speciosa J.W.MacSwain Collector 9:30- 9:45 UC Berkeley EMEC 976754 Megachile gravita 23 24	J.W.MacSwain

5% had no answer