Quality-aware Human-Machine Text Extraction for Biocollections using Ensembles of OCRs

Ícaro Alzuru, Rhiannon Stephens, Andréa Matsunaga, Maurício Tsugawa, Paul Flemons, and José A.B. Fortes

15th eScience International Conference
September 26th, 2019

HuMaIN is funded by a grant from the National Science Foundation's ACI Division of Advanced Cyberinfrastructure (Award Number: 1535086). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

AGENDA

- Digitization of Biological Collections (Biocollections)
  - Problem
- Proposed Solution
  - Human-Machine Self-aware Information Extraction workflow
  - Ensemble of OCR Engines as the Self-aware Process
  - Hybrid Human-Machine Crowdsourcing
- Experiments & Results
- Related Work
- Conclusions

ACIS
ADVANCED COMPUTING
& INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

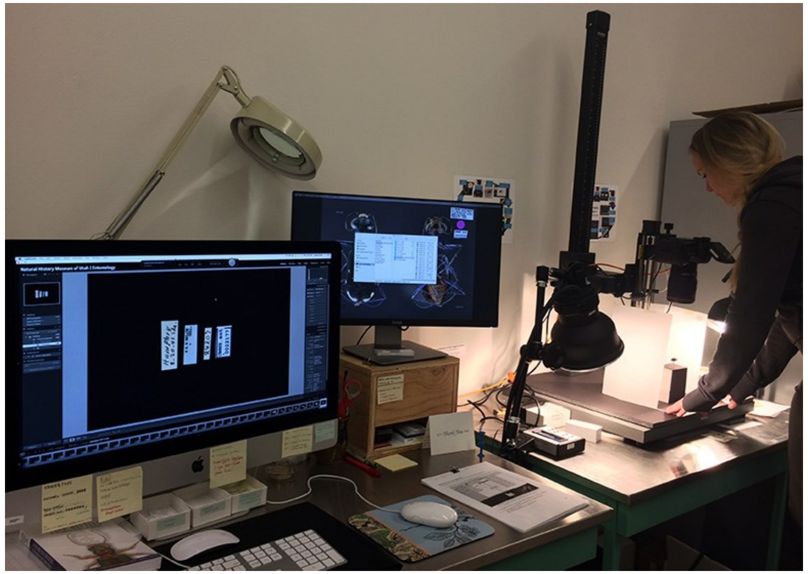
Digitization of Biocollections

- Information in biocollections can be used to understand pests, biodiversity, climate change, natural disasters, diseases, and other environmental issues.
- There are about 1 Billion specimens in Biocollections in the United States and about 3 Billion in the whole World (Estimated).
- NSF's Advancing Digitization of Biodiversity Collections (ADBC) program.

Photo by Chip Clark. U.S. National Herbarium at the Smithsonian Institution's National Museum of Natural History. Featured researchers: Dr. James Norris (right, front), research assistant Bob Sims (left, front), and associate researcher, Katie Norris (left, back).

Photo by Chip Clark. Bird Collection, Dept. of Vertebrate Zoology, Smithsonian Institution's National Museum of Natural History. In the foreground: Roxie Laybourne, feather identification expert.

Digitization Process



Digitization:

1. Photograph of the specimen and its correspondent labels.
2. Transcription of the metadata in a database (commonly performed by volunteers)

- Global Problem: How can we **accelerate** (make more efficient) the digitization process?
- General Answer: *Partial or total **automation** of the transcription process.*

The Challenge of Automated Information Extraction

Automated IE: Optical Character Recognition + Natural Language Processing

- Biocollections’ images are problematic for OCR engines
- OCR result is not perfect. Handwritten text is especially problematic.

Specific Problem: Can we generate trust in the text extracted by the OCR engine?

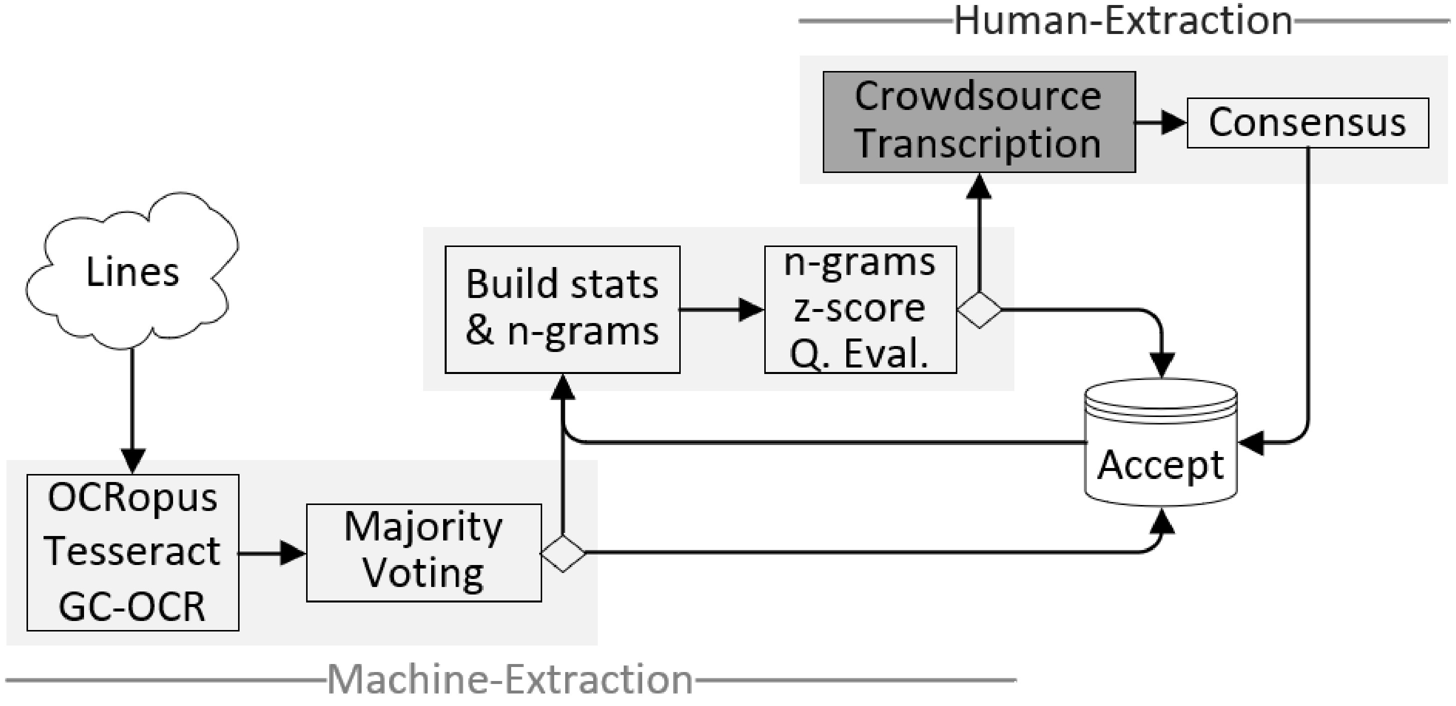
OCRopus 1.3.3	Tesseract 4.0	Google Cloud OCR
g 92--- ---		17.16'S 145.25'E. (GPS)
. : .-- "-	Carhrunag,, o	Mit Baldy Loop Rd, nr
C T, --	Mare;	Atherton
-+., S ; . A	R, ¢ Leth)	jerberton Range, 1097m .
7.l651452SE [6P9]		9 MAY 2011
Wf Baldy loop d, nr	Det. DCF Rept 2011	CF Rentz, B. Richardson,
Athenon		Stop14
ierhenton ange. 1097m		Carbrunnnia
'9MMY 2011		Marci
tCH RentL 3.		s Roth).
Rihordson, S!op 14		Det. DCF Rentz 2011
A		Australian Museum
s-a t5rend6yi 44		K 482255
tria rc		10 mm
e d 44 Y		
WOMJ		
DDet. CF Rert: ed		
Australian Museum		
K 4ocdbo		
g53rw P s		
10 rmm		
r-yr--;		
M- Lai ----- --- LuA.--		

ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

Proposed Solution

- We propose a SELFIE (Self-aware IE) workflow model for the transcription of biocollections' labels (https://doi.org/10.1109/eScience.2017.19)
- The challenge in SELFIE workflows is the confidence estimation method.
- Inspired by crowdsourcing, we use redundancy: an Ensemble of OCR engines.

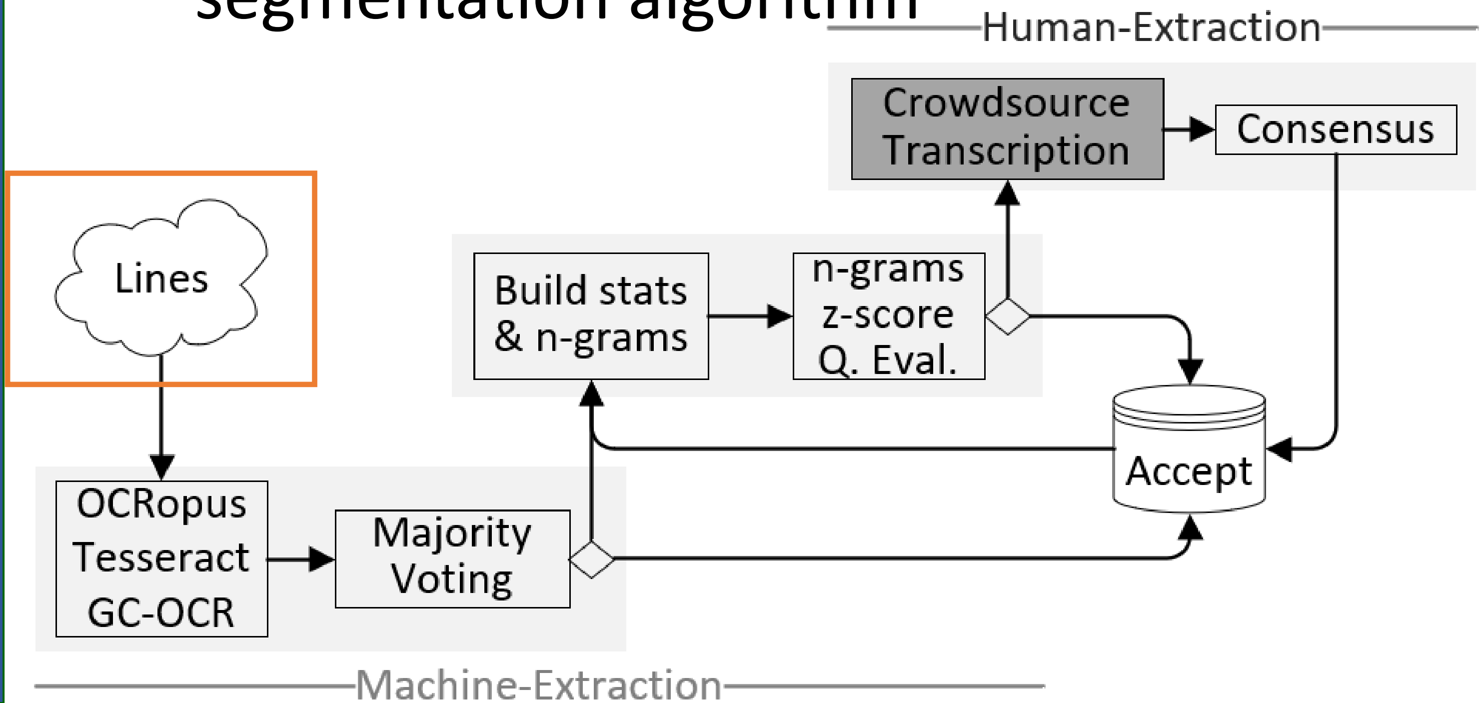


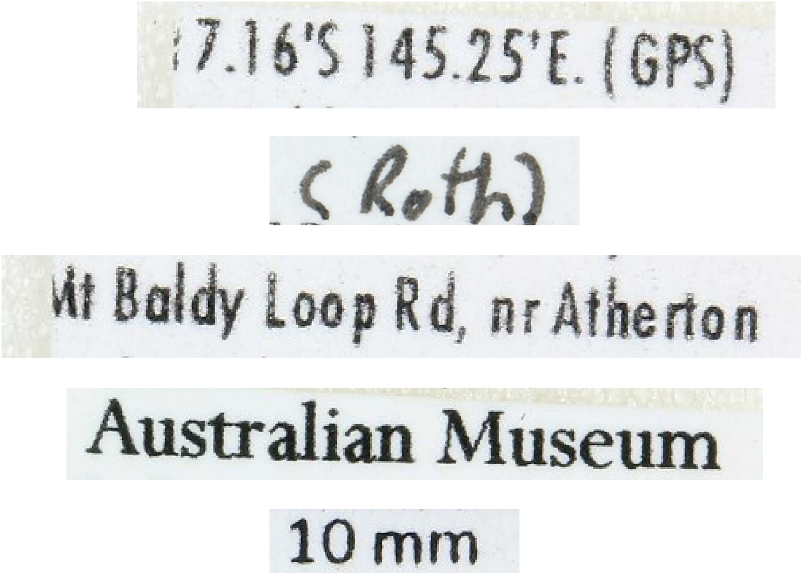
ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

Ensemble of OCR Engines – Lines Extraction

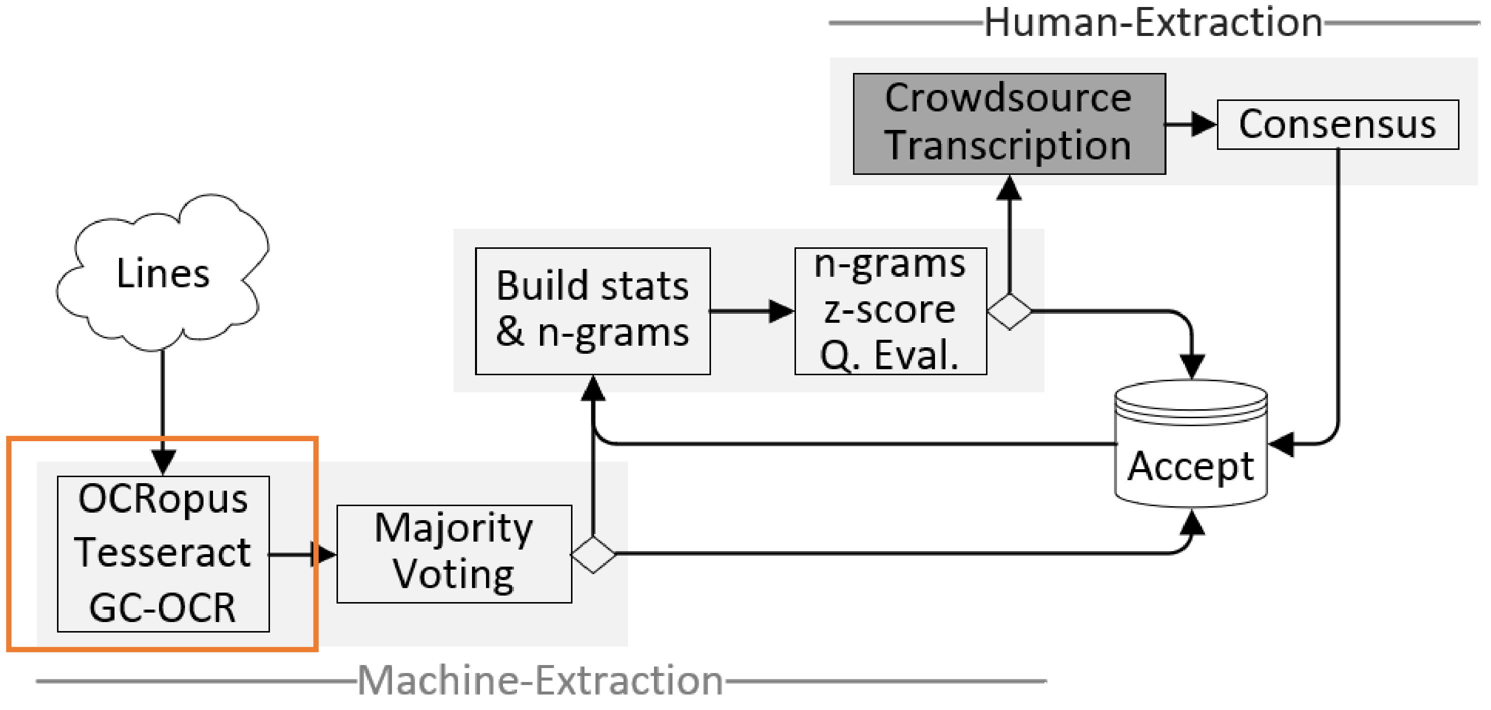
- OCR steps: binarization, segmentation, and recognition
- To compare the results provided by OCRopus, Tesseract, and the Google Cloud OCR (GC-OCR), we need a common text unit: **Lines**
- OCRopus and Tesseract segmentation introduce many errors.
- The GC-OCR character information was used to create a new segmentation algorithm



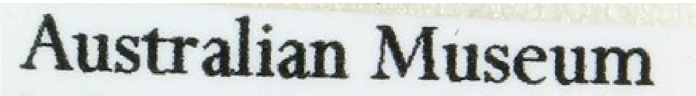


Ensemble of OCR Engines - OCR

- OCRopus, Tesseract, and the GC-OCR were run on each line.
- The per-character probability (confidence) was collected.



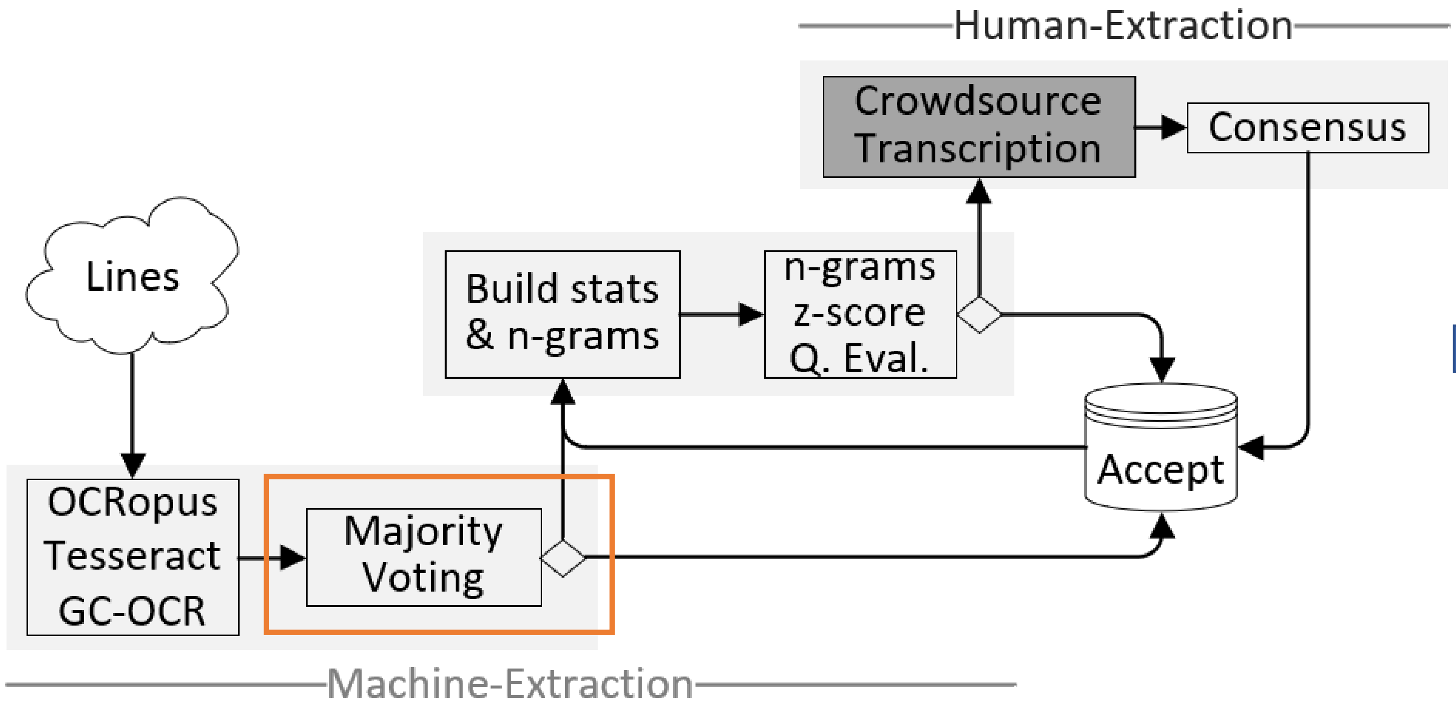

```
OCRopus:   c Rofhl
Tesseract: C RoHn)
GC-OCR:    C Roth)

     c 0.78
       0.94
     R 0.89
     . . .
```



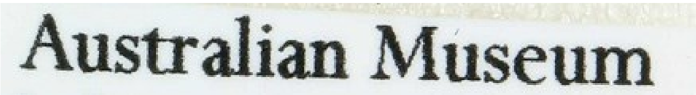

```
OCRopus:    Aushra1ian Museum
Tesseract:  Australian Museum
GC-OCR:     Australian Museum
```

Ensemble of OCR Engines – Majority Voting

- If three OCR engines agree, the text is accepted as correct
- If two OCR engines agree and their average per-character probability is greater than 0.8, the text is accepted as correct.



OCRopus: c Rofhl
Tesseract: C RoHn)
GC-OCR: C Roth)



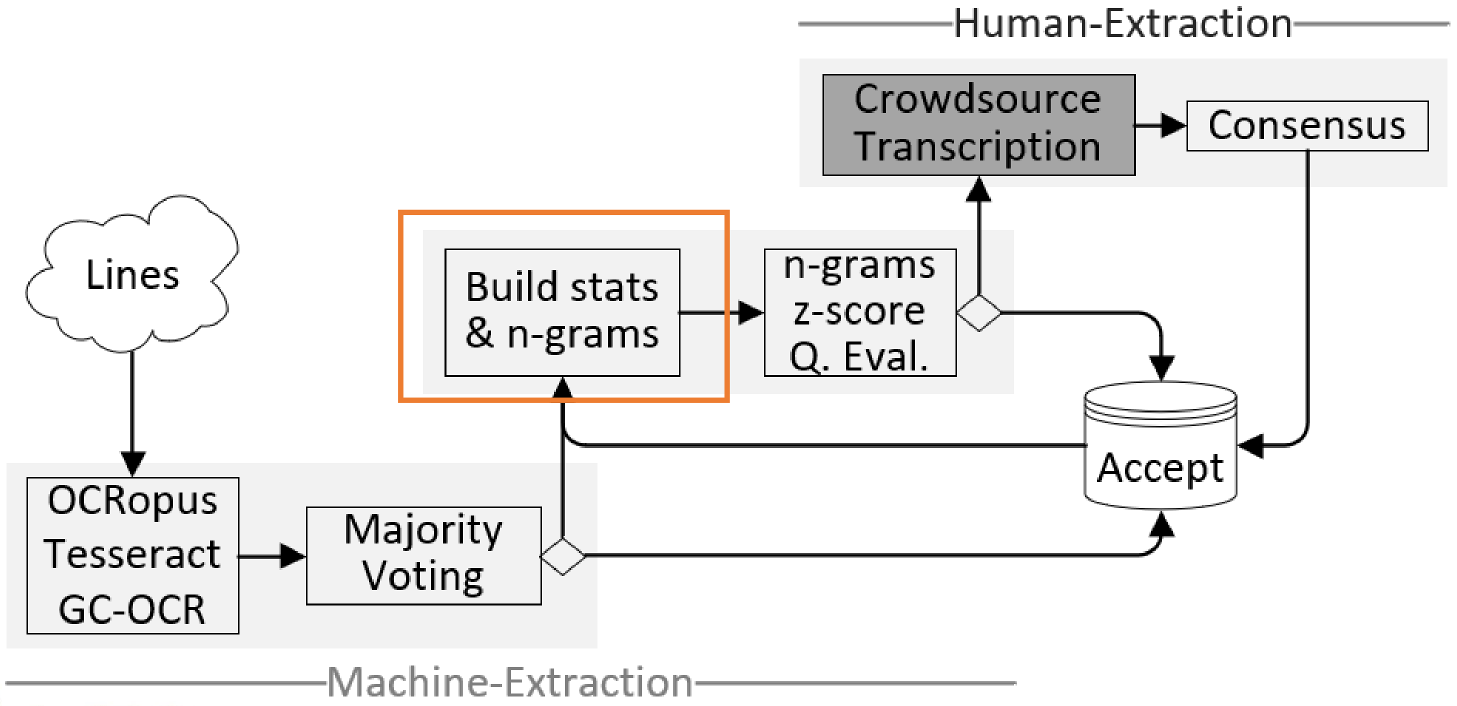
OCRopus: Aushra1ian Museum
Tesseract: Australian Museum
GC-OCR: Australian Museum

ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

Ensemble of OCR Engines – Support Structures

- Using the text in the accepted lines, two support structures are built:
  - Unigram (1-gram) model or word count. The words that appear less than 3 times are discarded.
  - The per-character probability average and standard deviation, per OCR engine (OCRopus, Tesseract, and GC-OCR)



1-gram or word count:

GPS, 10

Baldy, 3

Museum, 34

…

Per-character statistics:

character, mean, standard deviation

a, 0.78, 0.0456

b, 0.84, 0.0899

1, 0.92, 0.0919

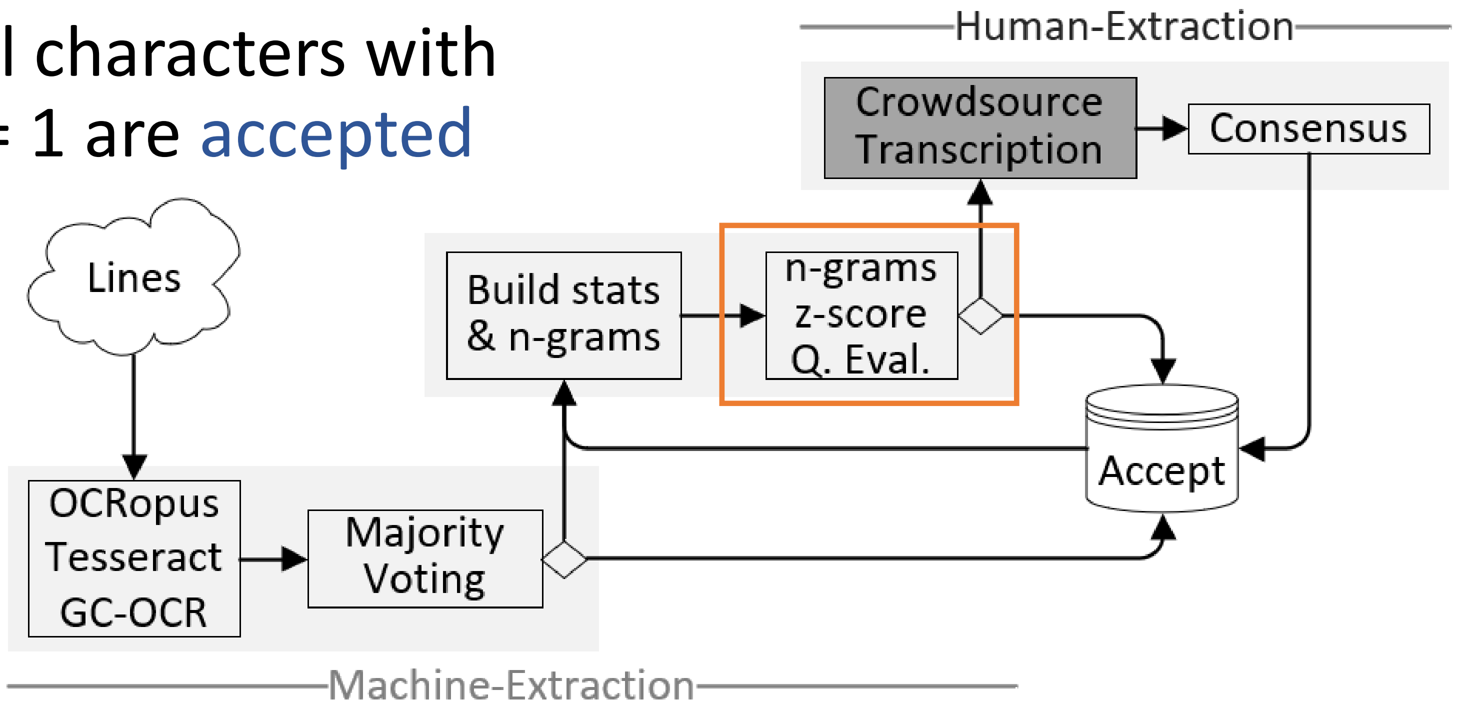
…

ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

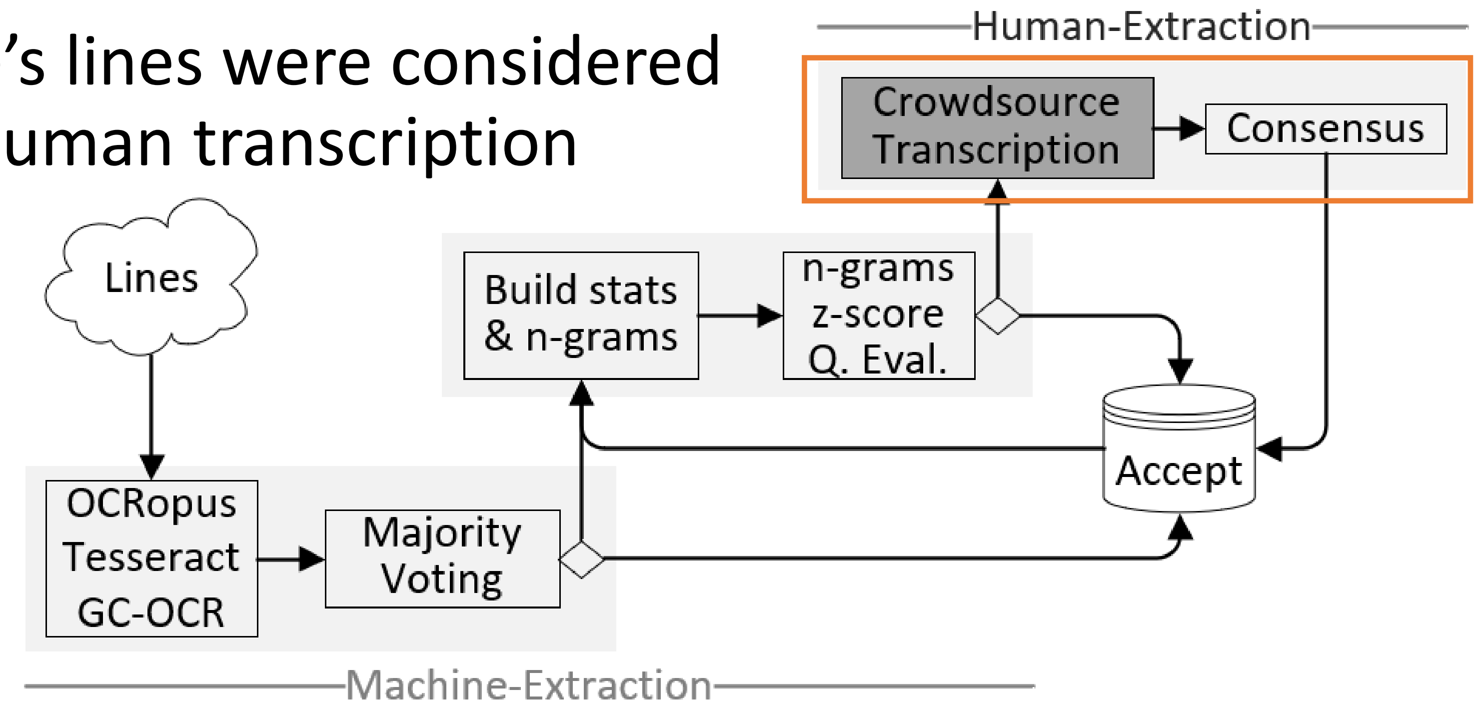
Ensemble of OCR Engines – Per-character Eval.

- Lines are scanned and those characters which belong to the words in the 1-grams are considered correct (confidence = 1).
- For the characters that do not belong to any n-gram:
  - Per line, the characters of the text extracted by the three OCR engines are aligned.
  - If at least OCR 2 engines extract the same character, it is considered correct.
  - If consensus is not reached, the character extracted by the GC-OCR is selected.
- Lines with all characters with confidence = 1 are accepted



Ensemble of OCR Engines - Crowdsourcing

- There are two common crowdsourcing approaches:
  - WeDigBio: Three transcribers + Consensus
  - DigiVol: One transcriber + One reviewer
- Volunteers of the Australian Museum were asked to transcribe lines from the remaining (rejected) lines.
  - Independent transcriptions were made to cover both crowdsourcing approaches.
- Ensemble's lines were considered the first human transcription



ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

Datasets and Segmented Lines

- Six collections were utilized in the experiments:
  - **A-OCR**: Augmenting-OCR Working Group (iDigBio), https://github.com/idigbio-aocr/label-data
  - **DV**: DigiVol – Australian Museum, https://digivol.ala.org.au/

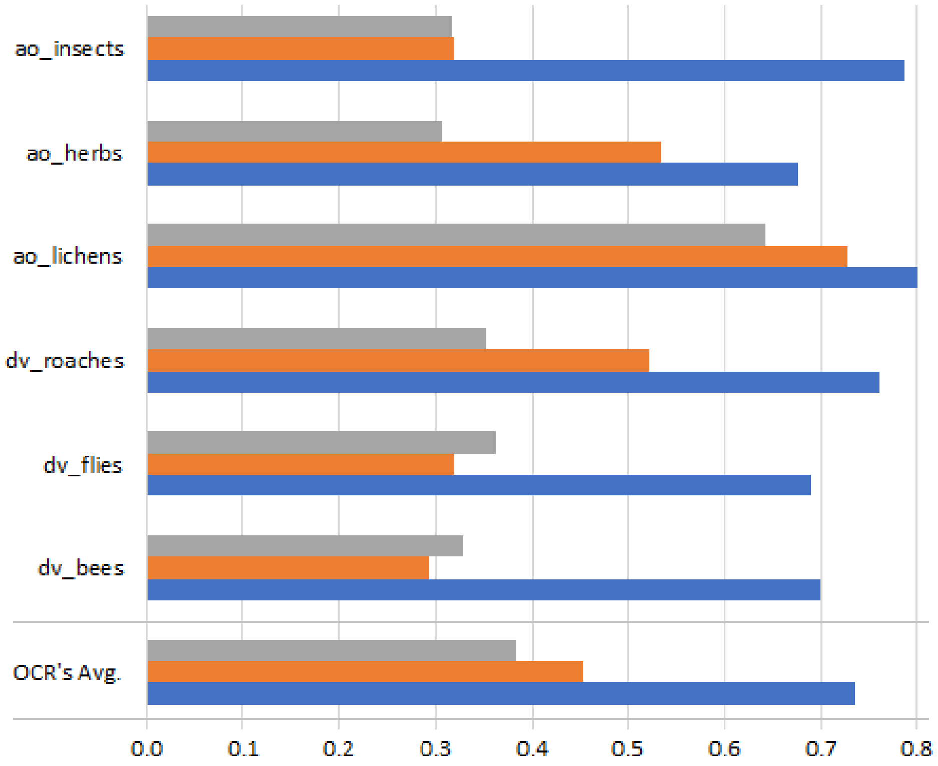
Dataset	# Images	# Lines
A-OCR Insects	100	1,132
A-OCR Herbs	100	3,192
A-OCR Lichens	200	2,618
DV-Roaches	1,117	10,002
DV-Flies	1,054	7,821
DV-Bees	395	3,053
Total	**2,966 Images**	**27,818 Lines**

ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

Results – Out-of-the-box Accuracy

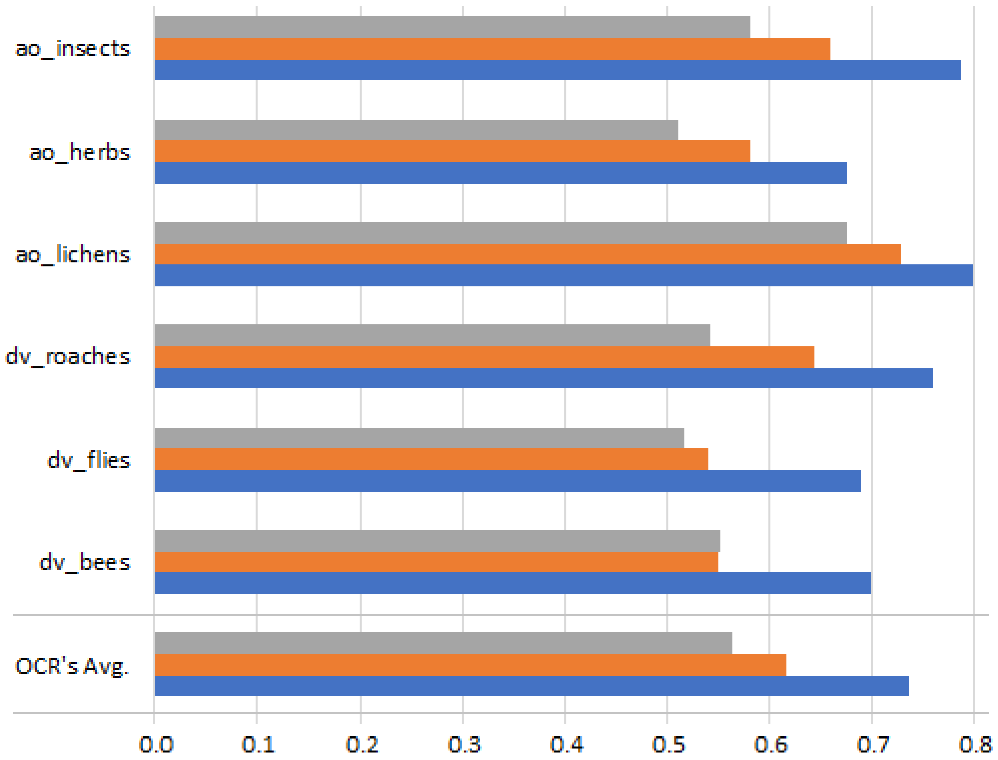
OCR Engine’s Own Segmentation



	ao_insects	ao_herbs	ao_lichens	dv_roaches	dv_flies	dv_bees	OCR's Avg.
OCRopus	0.316	0.307	0.643	0.353	0.362	0.328	0.385
Tesseract	0.318	0.534	0.728	0.523	0.318	0.294	0.452
GC-OCR	0.787	0.676	0.799	0.760	0.691	0.699	0.735

OCRopus Tesseract GC-OCR

GC-OCR’s Segmentation



	ao_insects	ao_herbs	ao_lichens	dv_roaches	dv_flies	dv_bees	OCR's Avg.
OCRopus	0.582	0.510	0.675	0.542	0.516	0.552	0.563
Tesseract	0.660	0.582	0.727	0.645	0.541	0.549	0.617
GC-OCR	0.786	0.676	0.799	0.760	0.689	0.699	0.735

OCRopus Tesseract GC-OCR

- Compared to the ground truth transcriptions of the entire text in the images.
- The segmentation algorithm improved the OCRopus’ and Tesseract’s output quality.

ACIS
ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

Results – Ensemble of OCRs

	Images	Lines	Accepted	To Crowd	% Accepted
ao_insects	100	1,132	711	421	62.81%
ao_herbs	100	3,192	1,657	1,535	51.91%
ao_lichens	200	2,618	1,639	979	62.61%
dv_roaches	1,117	10,002	5,831	4,171	58.30%
dv_flies	1,054	7,821	4,372	3,449	55.90%
dv_bees	395	3,053	1,800	1,253	58.96%

- 57.55% (16,010) of the 27,818 lines were accepted using the ensemble-of-OCRs algorithm.
- <u>Quality of the accepted data</u>:
  - Volunteers were asked to edit 600 lines.
  - Of the 10,081 characters in the 600 lines, volunteers made changes, insertions, or deletions in only 10 characters. This means that the accepted lines have a CER of 0.001 and an accuracy of 99.9%.

ACIS
ADVANCED COMPUTING
& INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

Results - Total Savings

	Tasks required	Ensemble savings	Hybrid crowd. savings	Total savings
Dynamic Human-Machine Consensus	3 x nL	**57.55%**	15.80%	**73.35%**
Hybrid Transcriber /Reviewer	2 x nL	**57.55%**	21.23%	**78.78%**

ACIS
ADVANCED COMPUTING
& INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

Related Work

- Crowdsourcing platforms:
  - **Symbiota** (flora/fauna)
  - **Zooniverse**
    - **Notes from Nature**, for biodiversity metadata transcription.
- IE Applications: Augment but not replace humans
  - SALIX
  - APIARY (workflow & tools)
- Parsers
  - LBCC, SALIX (Frequency tables!) – Included in Symbiota.
- NY Botanical Garden, Drinkwater et. al.

Conclusions

- This research proposed the use of a SELFIE workflow for the transcription of the biocollections' images, using an Ensemble of OCR engines to generate confidence and hybrid crowdsourcing to save tasks.
- About 58% of the text could be validated using the Ensemble of OCRs. The text extracted presented an accuracy of 99.9%.
- Two common crowdsourcing approaches for the generation of the final value were tested. The use of the Ensemble's transcription in these approaches save, in average, 44% of the crowdsourcing tasks.
- In total, the text extraction approach reduced, in average, 76% the number of crowdsourcing tasks.
- The code developed and utilized during the research is available at https://github.com/acislab/HuMaIN_Text_Extraction

ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA

Thank you

Questions?

HuMaIN is funded by a grant from the National Science Foundation's ACI Division of Advanced Cyberinfrastructure (Award Number: 1535086). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

ACIS
ADVANCED COMPUTING
& INFORMATION SYSTEMS

UF UNIVERSITY of FLORIDA