Cooperative Human-Machine Data Extraction from Biological Collections

Icaro Alzuru, Andréa Matsunaga, Maurício Tsugawa, José A.B. Fortes

12th IEEE International Conference on e-Science
October 24th, 2016
Baltimore, Maryland, USA

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.

Outline

- Biological Collections and their Data Extraction challenges
- Data Extraction approaches
- HuMaIN
- Experimental setup
- Approaches' performance & Results
- Time, cost and quality
- Conclusions

Biological Collections

Plants, fungi, animals, bacteria, archaea, and viruses.

- Organizations and people from around the world have assorted biological materials and specimens for decades.
- The number of samples has been estimated in
  - 1+ Billion in the USA
  - 2+ Billions worldwide
- These collections have a potential enormous impact: new medicines, species conservation, epidemics, environmental changes, agriculture, etc.
- Digital Biological Collections
  - iDigBio (USA) – 72 million of specimen records.
  - ALA - Atlas of Living Australia
  - GBIF – Global Biodiversity Information Facility (Worldwide)

Photo by Jeremiah Trimble, Department of Ornithology, Museum of Comparative Zoology, Harvard University. doi:10.1371/journal.pbio.1001466.g002

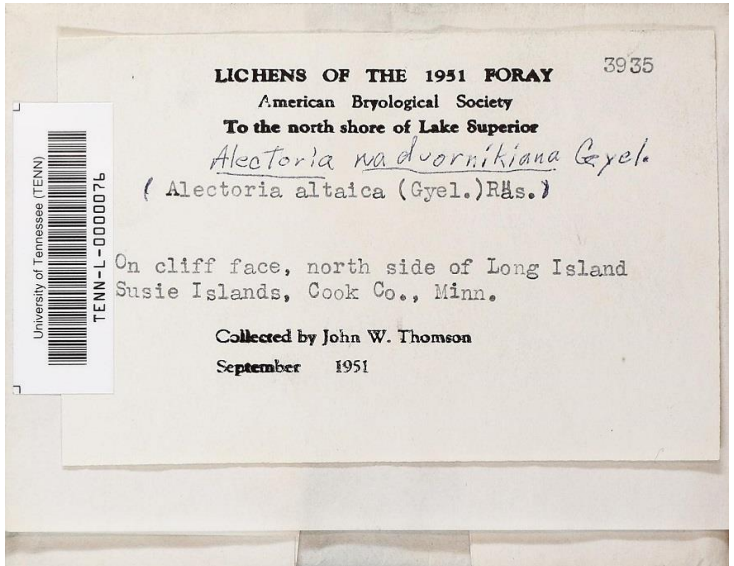
Data Extraction from Biocollections

- Goal: Getting the what, where, when, and who about the collected specimens.
- Data extraction challenges:
  - No standard format
  - Several languages
  - Multiple Font types and sizes
  - Tinted background
  - Multiple images qualities
  - Elements overlapping text

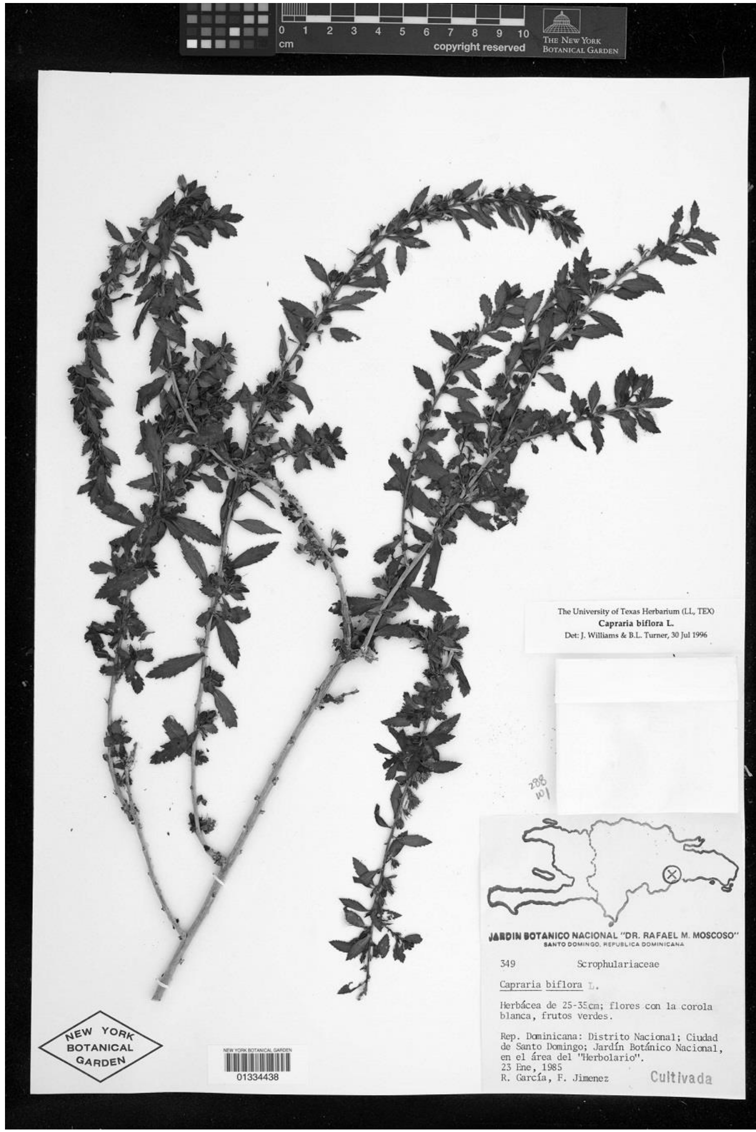
How to extract that information from this massive data source?

Entomology

Lichen



Bryophyte

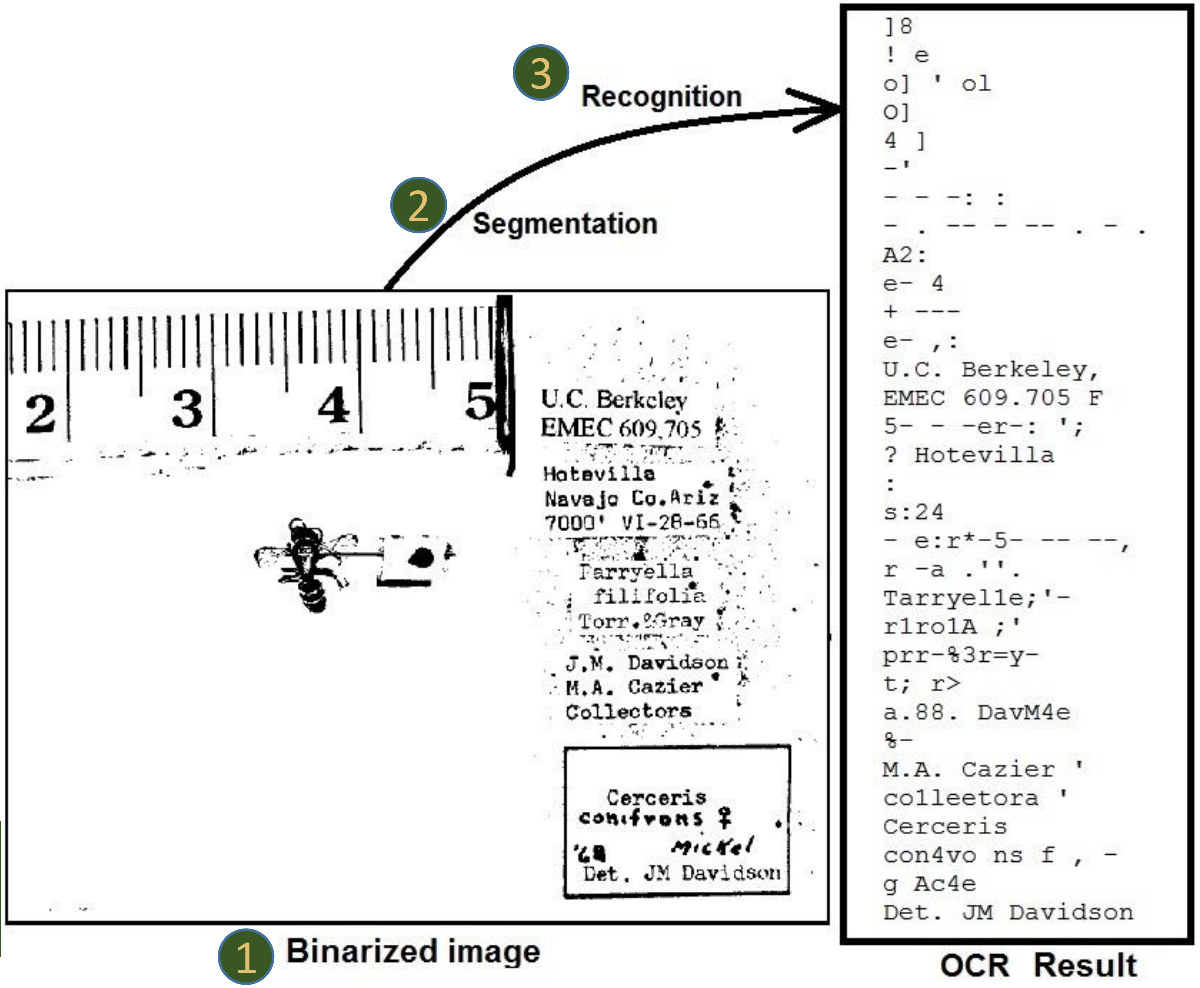


Machine-only approach

- Premises: Machines are fast, cheaper than humans, and perform repetitive tasks with less errors.
- Procedure:
  - **Optical Character Recognition (OCR)** software processes the images and extract the text.
  - A Natural Language Processing (NLP) algorithm could post-process the extracted data
- With so much variability, training-based algorithms are not worth.
- Bad results (No NLP tried, only OCR):
  - Accuracy between 0 % and 95 % for word recognition (In Lichens).
  - Average similarity: 0.42

1 **Best** – equal strings
0 **Worst** – totally different

OCR process



Human-only approach

Image by Justin Whiting

- Premises: Humans have good judgement, perception, induction, and detection capabilities.
- Procedure:
  - Volunteers or paid participants transcribe the labels or fields. Many humans: crowdsourcing.
  - Consensus need to be reached among the posted answers.
- Previous work[1] showed, in average, consensus was found in 86.7% of times with an accuracy of 91.1% => 79% of correct results.
- Assuming 1 Billion of specimens, and taking 1 minute/specimen digitization, we would take ~ 8,000 man-year

[1] "Reaching Consensus in Crowdsourced Transcription of Biocollections Information", A. Matsunaga, A. Mast, and J. A.B. Fortes.

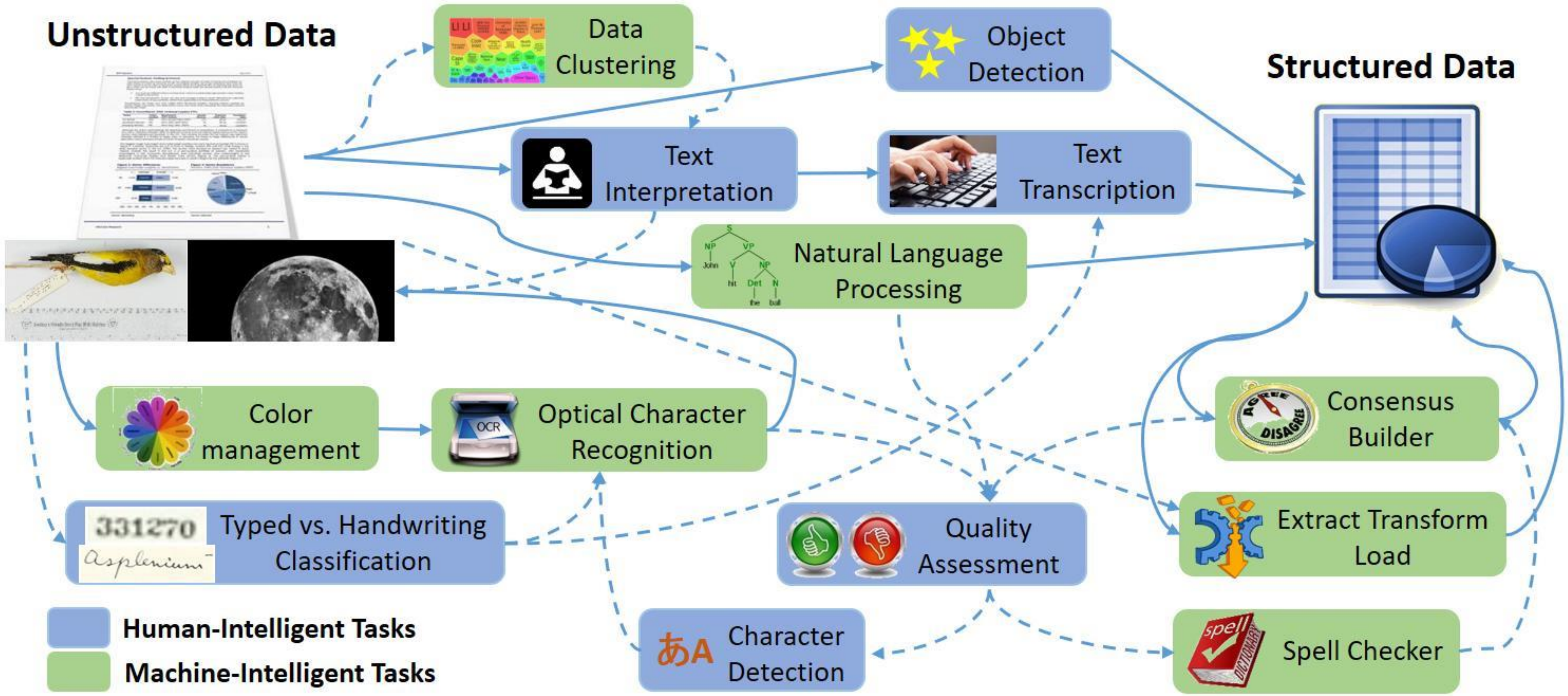
Hybrid approaches

- Using the strengths of humans and machines in a cooperative manner to improve data extraction results.
  - Improvements in terms of time, quality, or both.

- Our goal with this study is to demonstrate that hybrid approaches improve results when extracting data from biological collections.
- This study is part of the HuMaIN project.

HuMaIN

Human and Machine Intelligent Software Elements for Cost-Effective Scientific Data Digitization



https://humain.acis.ufl.edu

Experimental setup

- **Considered approaches**:
  0. Human-only (Previous study). Baseline.
  1. Machine-only – OCR whole image (no cropping). Baseline.
  2. Cooperative – Crop label (Humans), then OCR.
  3. Cooperative – Crop fields (Humans), then OCR.

- **Data Set**: 400 images prepared by the Augmenting OCR Working Group (A-OCR) of the iDigBio project.

https://github.com/idigbio-aocr/label-data

Specimen type	Number of images	Avg. Size (KB)	Dimension	Resolution (dpi)
Entomology	100	325	1600x1200	180
Bryophyte	100	1214	3744x5616	300
Lichen	200	153	1530x1128	96

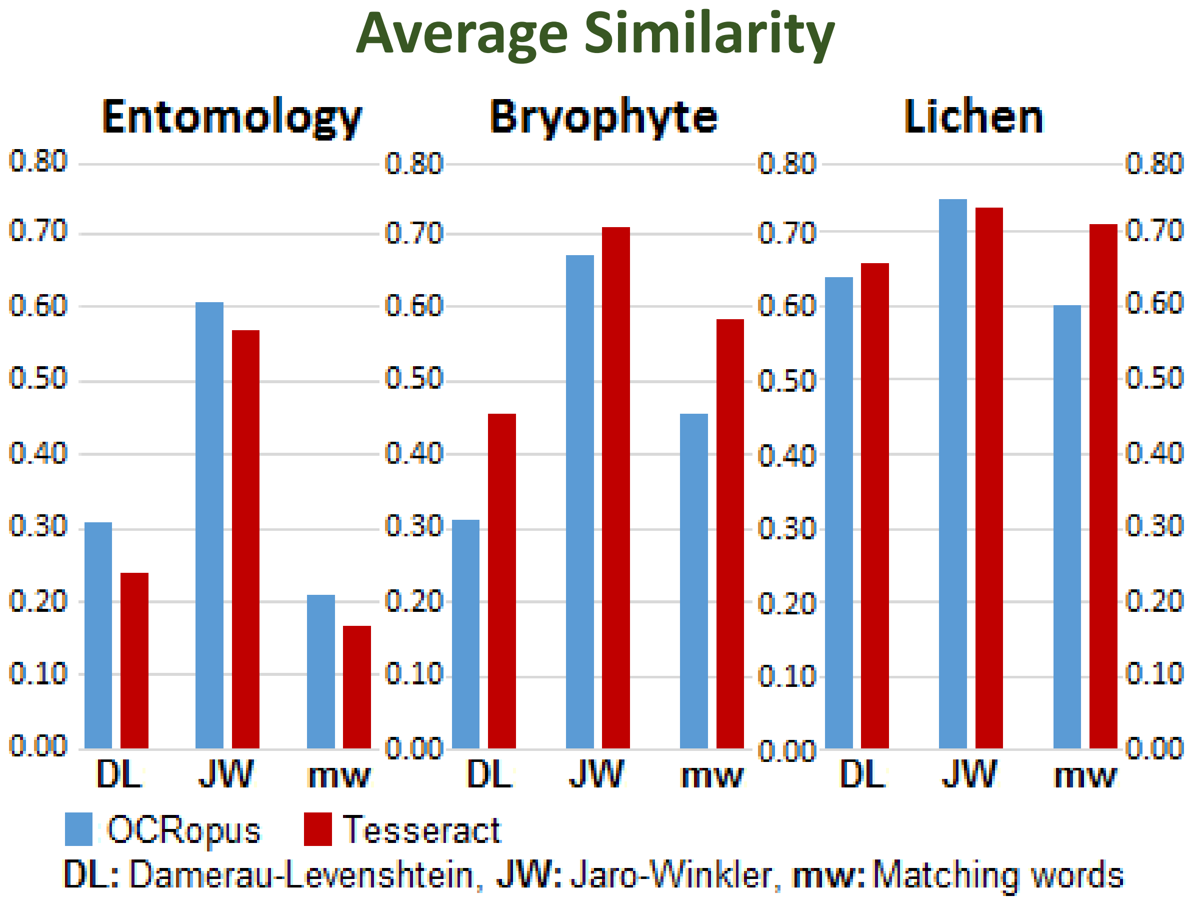
- **Optical Character Recognition technology**: OCRopus (OCRopy) and Tesseract

- **Metrics:**
  - Damerau-Levenshtein (DL) similarity
  - Jaro-Winkler (JW) similarity
  - Matched words (mw) rate

$$sim_{DL}(x,y) = 1 - \frac{DL\ distance(x,y)}{\max(|x|,|y|)}$$

$$mw(x,y) = \frac{|words\ in\ common\ between\ x\ and\ y|}{|x|}$$

A1. Machine-only Performance (OCR whole image)



- Avg.Sim. Lichen > Avg.Sim. Bryophyte > Avg.Sim. Entomology
- Similar recognition rate for OCRopus and Tesseract
- Jaro-Winkler is the most optimistic metric
- In Average, Tesseract was 18.5x faster than OCRopus

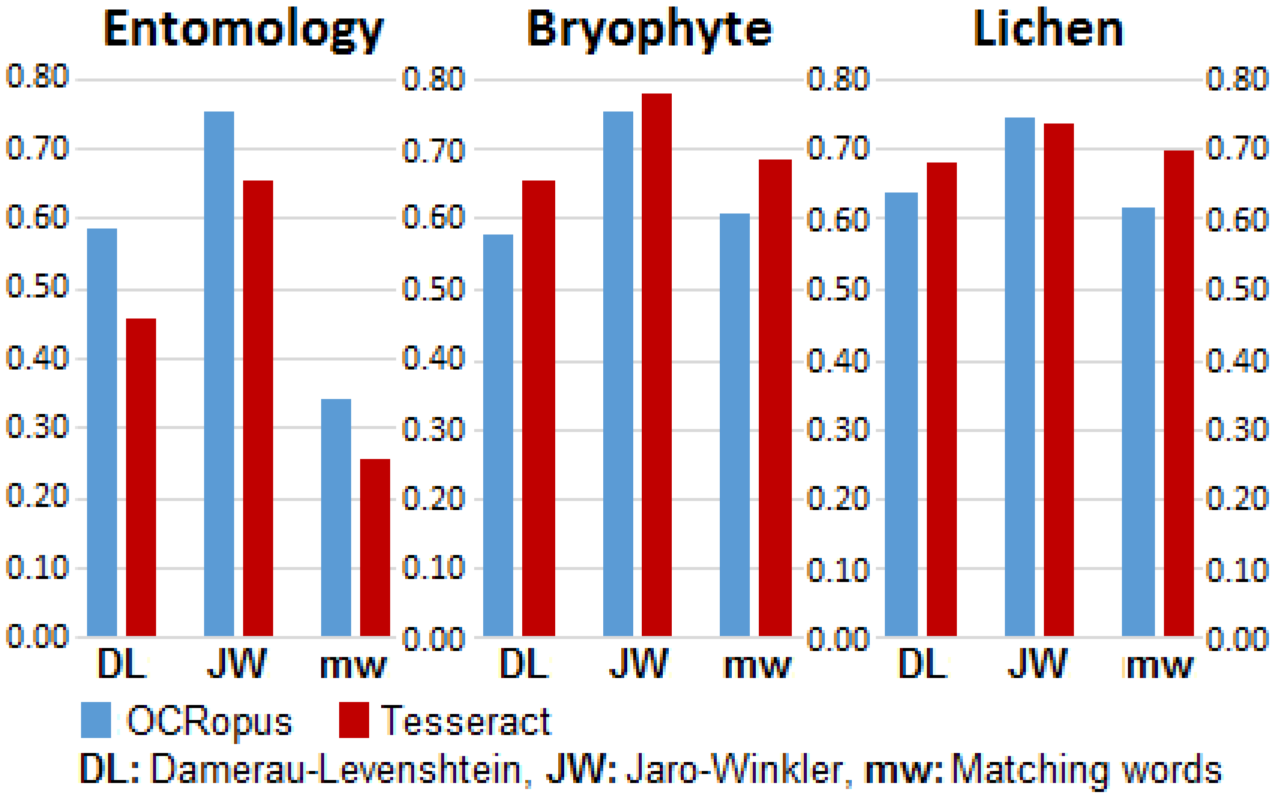
OCR's average execution time (s)

	Avg. Execution Time (s)	
Specimen type \ Tool	**OCRopus**	**Tesseract**
Entomology	28.36	3.60
Bryophyte	158.57	4.54
Lichen	30.46	1.95

A2. Hybrid performance (Crop Label + OCR)

Cropped labels

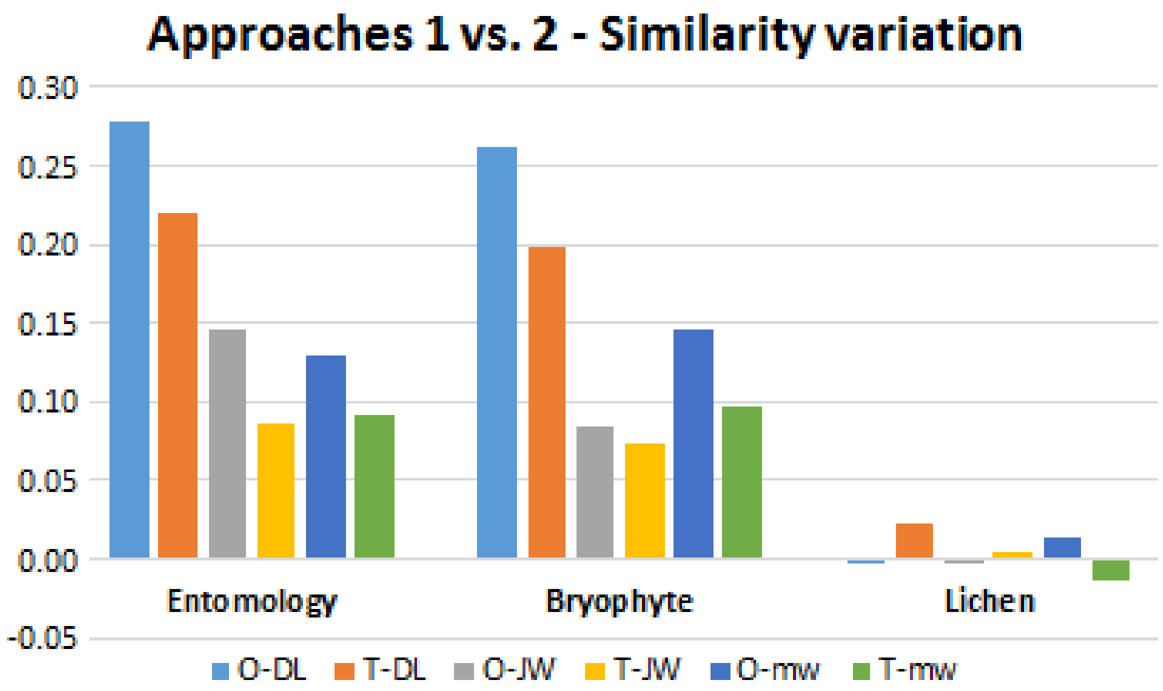
Average Similarity



- Avg.Sim. Lichen > Avg.Sim. Bryophyte > Avg.Sim. Entomology
- Similar recognition performance for OCRopus and Tesseract
- All the similarity values improved

Machine vs. Hybrid (Cropping Labels) approaches

- Entomology and Bryophyte:
  - Avg. similarity improvement of 0.15
  - Damerau-Levenshtein had a bigger improvement than the other two metrics
  - OCRopus had a higher improvement than Tesseract
- Lichen:
  - No improvement (Images = Labels)
- Execution Time with respect to A1:
  - Similar for OCRopus
  - 6.5x slower for Tesseract

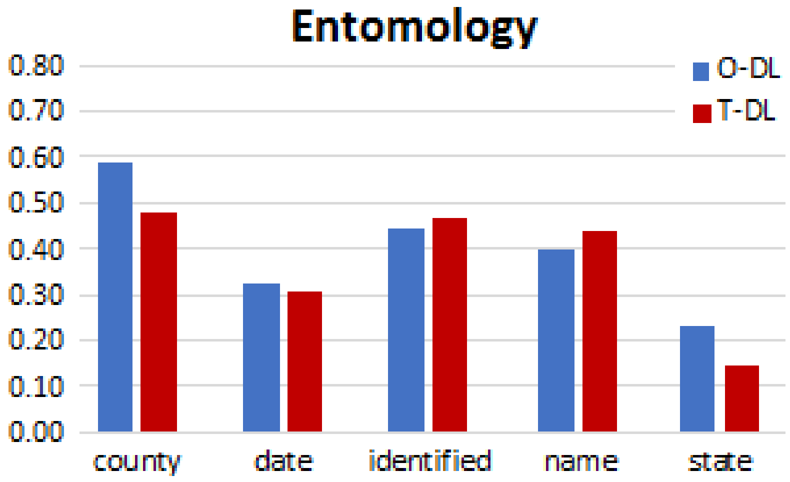


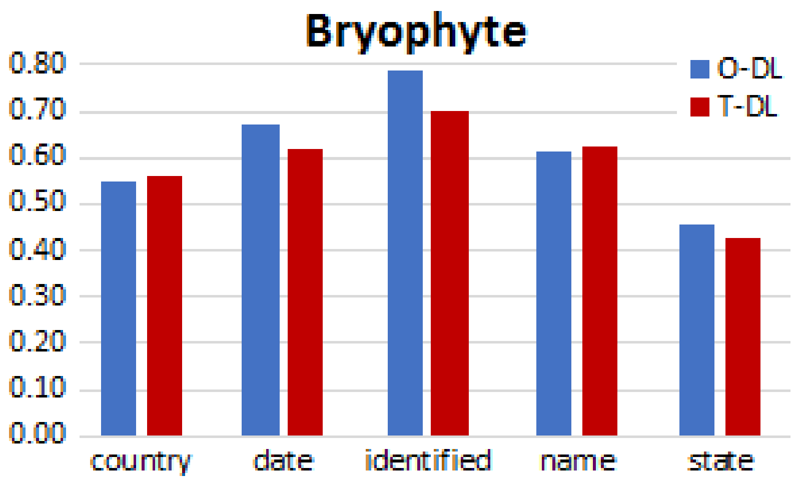
Approach 2 - Average execution time

	Execution time (s)						
Type \ Tool	Crop	Ocropus	Tesser.	Tot. Oc.	Tot. Te.	O2/O1	T2/T1
Entomology	15.36	15.65	2.47	31.01	17.83	1.09	4.95
Bryophyte	24.56	32.74	1.68	57.30	26.24	0.38	5.78
Lichen	15.13	25.52	1.82	40.65	16.95	1.33	8.69

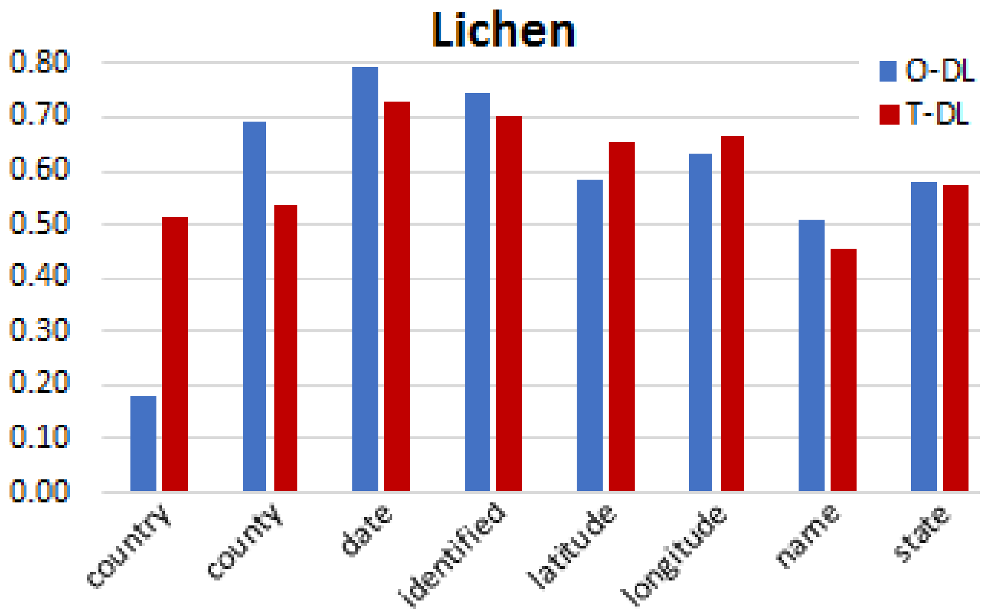
A3. Hybrid performance (Crop fields + OCR)

Cropped fields





Damerau-Levenshtein similarity



- Fields with few data or not verbatim were omitted for the calculations.
- Avg.Sim. Lichen > Avg.Sim. Bryophyte > Avg.Sim. Entomology
- Similar recognition performance for OCRopus and Tesseract, even inside the same collection.

Results

Average similarity and improvement with respect to A1

	Entomology	Bryophyte	Lichen
A1: whole image	0.27	0.38	0.64
A2: cropped label	**0.52 – 93%**	0.61 – 61%	**0.66 – 3%**
A3: cropped field	0.43 – 59%	**0.67 – 76%**	0.64 – 0%

- Hybrid approaches (A2 and A3) always improve similarity with respect to the machine-only approach (A1) up to a factor of 1.93.
- No improvement for Lichen images (because these images contain only text)
- Cropping fields eliminate the need of NLP, adding interpretation.

Estimated Time, Cost, & Quality for 1B specimens

- Machine-only shows the lowest price, is one of the fastest approaches, but has the worst quality.
- Human-only is the most expensive and slowest approach, but provides the best quality.
- Hybrid approaches are in the middle, providing similar execution time than Machine-only with a better data extraction quality.

Time, Cost, and Similarity

Approach	Human + Machine (Time in years)	Cost ($ in Millions)	Recognition rate or Similarity
0. Human-only	17123 + 0 (17123)	1500.00	**0.79**
1. Machine-only	0 + 1202 (1202)	**3.61**	0.43
2. Hybrid (Crop Label)	580 + 422 (**1002**)	52.10	0.60
3. Hybrid (Crop Fields)	6342 + 1218 (7560)	559.21	0.58

Assumptions:

- Sequential processing of 1 billion scientific images to process
- Total cost of ownership of a server = $3000 per year.
- Payment of $10 per hour to participants
- Averaging the behavior of OCRopus and Tesseract obtained in the experiments

Related Work

- Crowdsourcing platforms: allow the definition of crowdsourcing projects to be completed by the public.
  - **Notes from Nature** and other **Zooniverse** projects.
  - **DigiVol** and the **Atlas of Living Australia**.
  - **Les herbonautes** (Muséum National D'Histoire Naturelle), France.
  - **Amazon Mechanical Turk**.
- Hybrid Biocollections Apps: OCR, NLP, and humans correct the interpreted data.
  - **SALIX** (Semi-automatic Label Information Extraction system) and **Symbiota**.
  - **Apiary**: adds selecting areas and quality control. Includes **HERBIS**, a web app similar to SALIX.
  - **ScioTR**: Humans cropping, OCR, NLP, humans correcting.
- Hybrid platform: workflow of crowdsourcing and machine learning tasks
  - **CrowdFlower**.

Conclusions

- Cooperative approaches improved the OCR quality by a factor of 1.37 (37%), with respect to the machine-only approach, taking similar time, but at higher cost.
- The quality generated by cooperative approaches was 25% lower than the human-only approach, but is 4x faster and is cheaper.
- For complex images, the OCR's recognition rate was improved by at least 59% when cropping the text area.
- OCRopus and Tesseract showed a similar recognition rate, but Tesseract was, in average, 15x faster than OCRopus.
- Cooperative machine-human approaches are a balanced alternative to human-only or machine-only approaches.

Thank you!

Any question?

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.