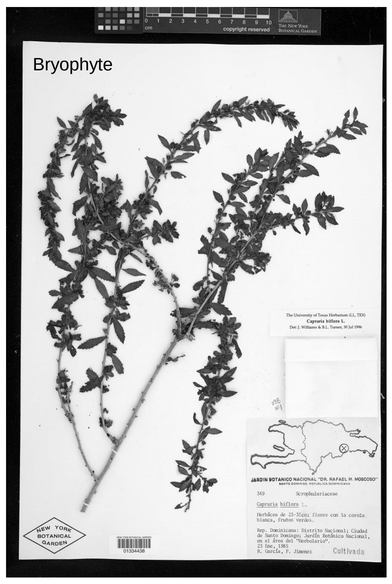
Cooperative Human-Machine Data Extraction from Biological Collections

Ícaro Alzuru, Andréa Matsunaga, Maurício Tsugawa , and José A.B. Fortes

ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

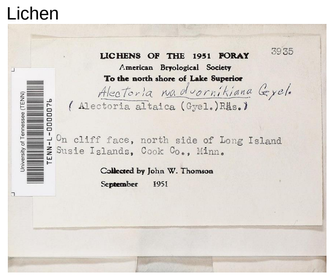
Biological Collections

- Biological materials and specimens have been assorted for decades.
- The number of samples to digitize has been estimated in
  - 1+ Billion in the USA
  - 2+ Billions worldwide
- Enormous potential impact: new medicines, environment, species conservation, epidemics, agriculture, etc.



Entomology

Lichen



Data Extraction

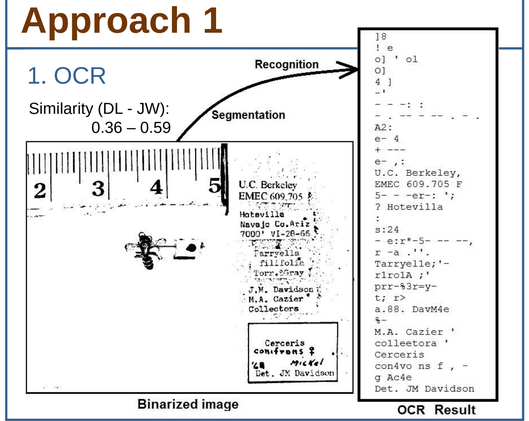
- Getting the what, where, when and who about the specimens.
- Challenges: No standards, mix of languages, fonts, quality, and tinted background.

Experimental Setup

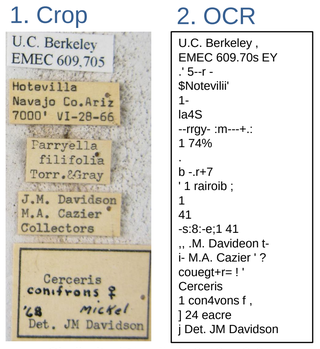
- Considered approaches:
  0. Human-only ("Reaching Consensus in Crowdsourced Transcription of Biocollections Information", Matsunaga et. all)
  1. Machine-only – OCR whole image (no cropping). Baseline.
  2. Cooperative – Crop label (Humans), then OCR.
  3. Cooperative – Crop Fields (Humans), then OCR.
- Optical Character Recognition (OCR) software: OCRopus (OCRopy) and Tesseract
- Metrics:
  - Damerau-Levenshtein (DL) similarity
  - Jaro-Winkler (JW) similarity
  - Matched words (mw) rate

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

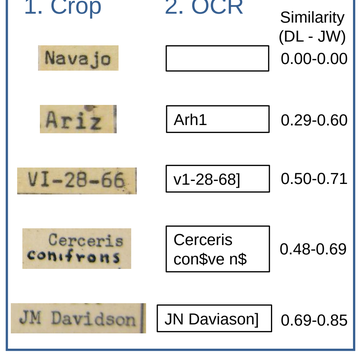
Approach 1



Approach 2

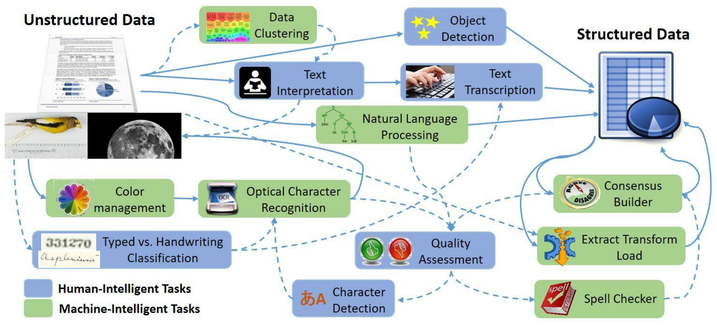


Approach 3



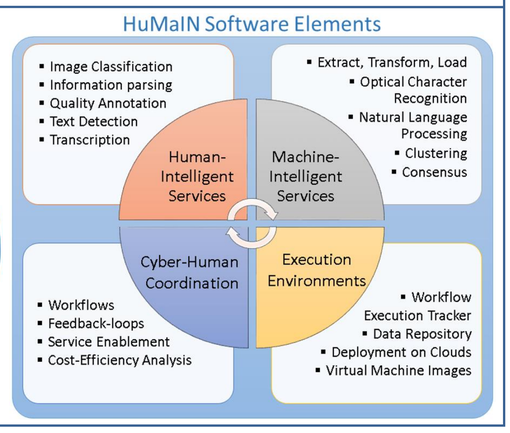
HuMaIN

Human and Machine Intelligent Network



http://humain.acis.ufl.edu

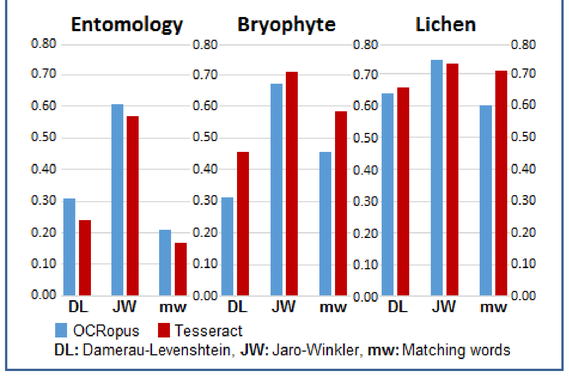
HuMaIN Software Elements



Related Work

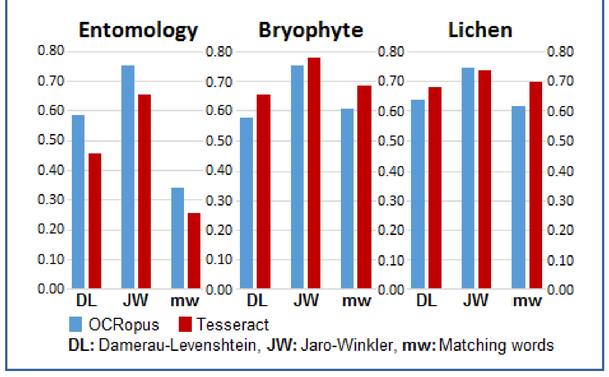
- **Notes from Nature** and **Zooniverse** projects: Define project, then crowd work. **DigiVol** and the **Atlas of Living Australia**; **Les herbonautes** (Muséum National D'Histoire Naturelle), France.
- **SALIX** (Semi-automatic Label Information Extraction): OCR, NLP, humans correct extracted data.
- **Apiary**: Selecting areas, OCR, Transcription, Quality control, ingestion. Includes HERBIS (~SALIX).
- **ScioTR**: Human cropping, OCR, NLP, Human correcting.
- **CrowdFlower**: Information extraction company with a crowdsourcing platform, which also integrates machine learning tasks.

Machine-only



- Sim. Lichen > Bryophyte > Entomology
- JW is the most optimistic metric
- Similar recognition performance for OCRopus and Tesseract

Hybrid (Cropping Labels)

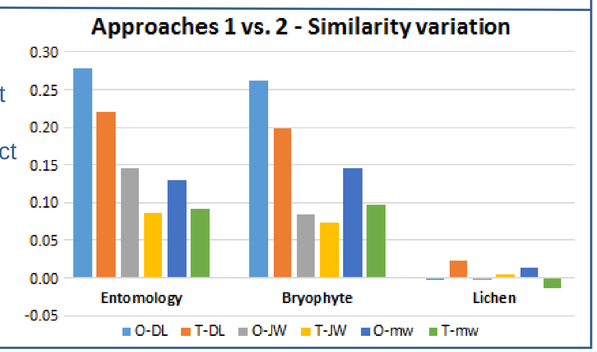


Hybrid (Cropping Labels) - Average Execution Time

Type \ Tool	Execution time (s)						
	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

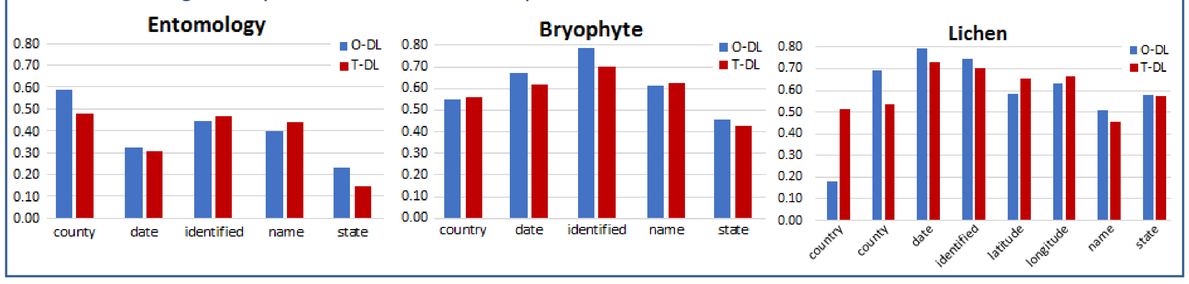
Machine-only vs. Cropping Labels

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



Hybrid (Cropping Fields)

- 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

- 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 does not require NLP, adding interpretation.

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%

Estimations

- 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 provide a balance: similar execution time than Machine-only with better quality.

Time, Cost, & 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

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.

12th IEEE International Conference on e-Science: October 24th, 2016

This material is based in part upon work supported by the National Science Foundation under Grant No. 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.

UF UNIVERSITY of FLORIDA