SELFIE: Self-aware Information Extraction from Digitized Biocollections

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

Advanced Computing and Information Systems (ACIS) Laboratory

University of Florida, Gainesville, USA

13th IEEE International Conference on e-Science

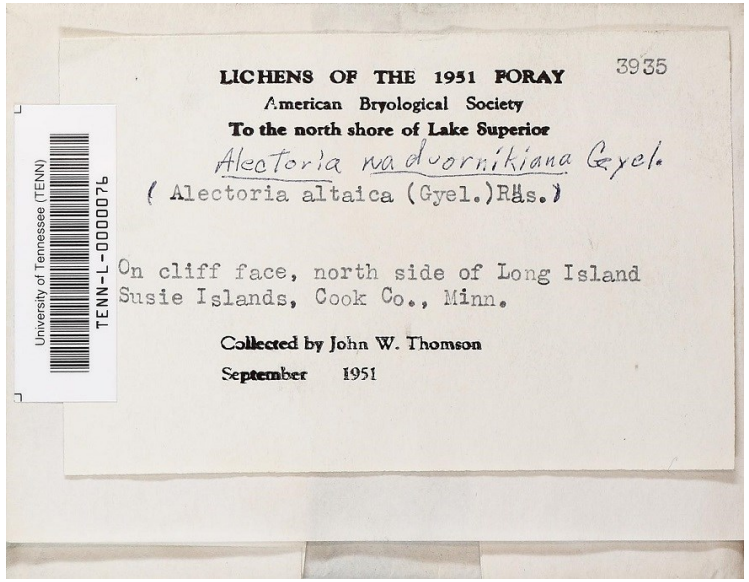
October 24th, 2017

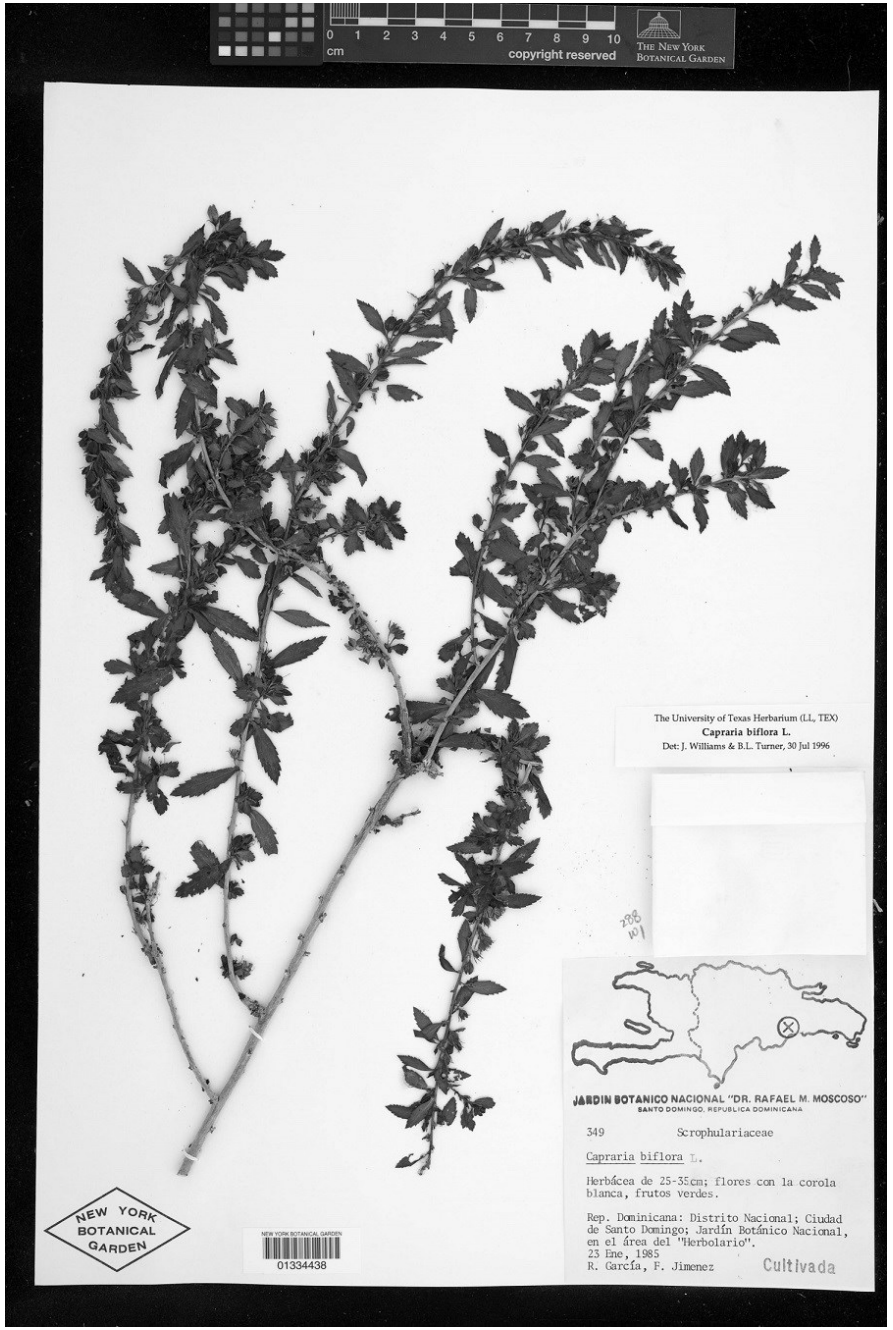
Auckland, New Zealand

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.

Biological Collections

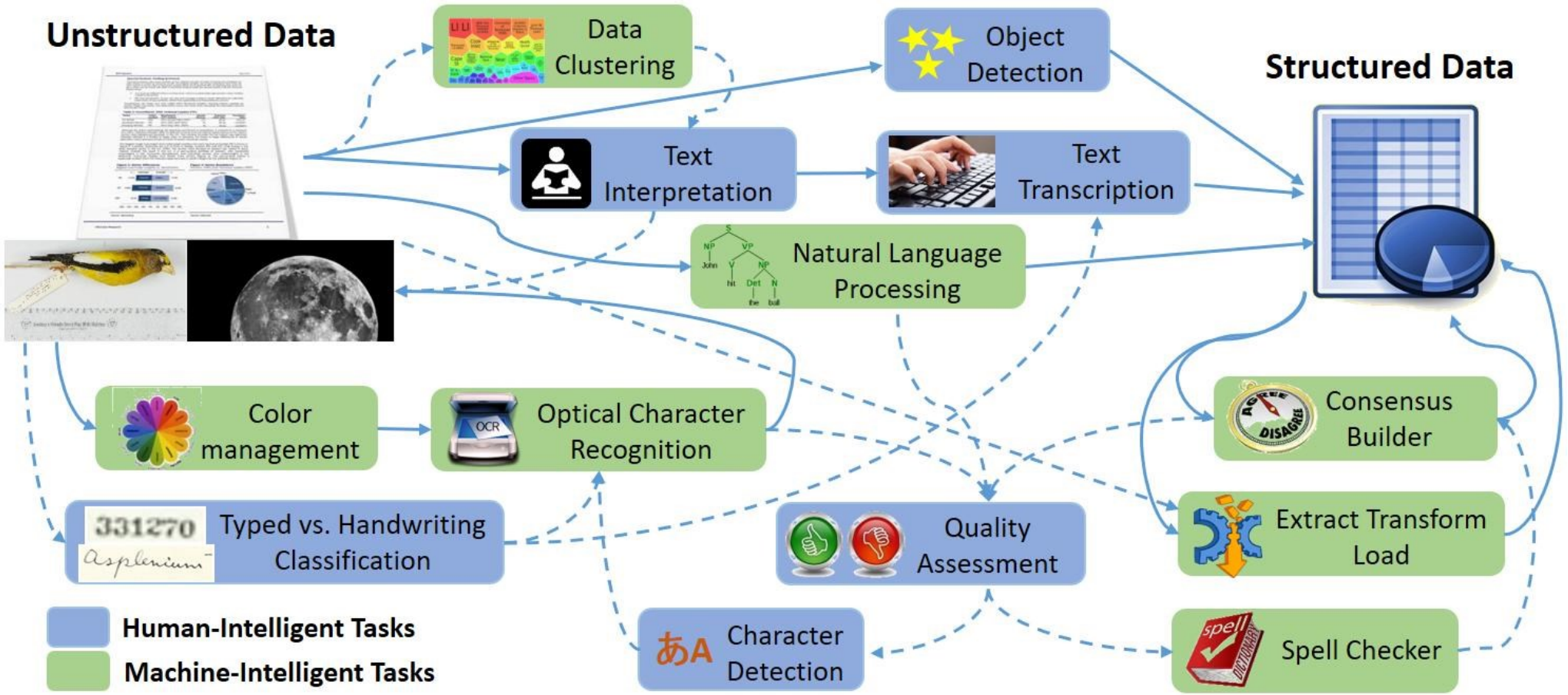
- Extracted metadata can be used to better understand pests, biodiversity, climate change, species invasions, historical natural disasters, diseases, and other environmental issues.
- In USA, iDigBio, since 2012, has aggregated 105 M. digitized records.
- Worldwide, GBIF accumulates 740 M. records in its database.





HuMaIN

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



The problem

- Specimens in biocollections have been estimated at 1 Billion in the USA, and almost 3 Billions worldwide.
- Automatic Information Extraction (IE) generates errors
- Crowdsourcing is relatively slow.
- **How can biocollections' IE be accelerated while keeping the quality of the results similar to what capable humans can provide?**

Related work – Specimen processing pipeline

1. Pre-digitization Curation
2. Selecting Components for an Imaging Station
3. Imaging Station Setup, Camera/Copy Stand
4. Imaging Station Setup, Light Box
5. Imaging Station Setup, Scanner
6. Imaging
7. Image Processing
8. Organizing and Implementing a Public Participation Imaging Blitz
9. Imaging Archiving
10. Selecting a Database
11. **Data Capture**
12. **Organizing and Implementing a Public Participation Transcription Blitz**
13. Georeferencing
14. Proactive Digitization

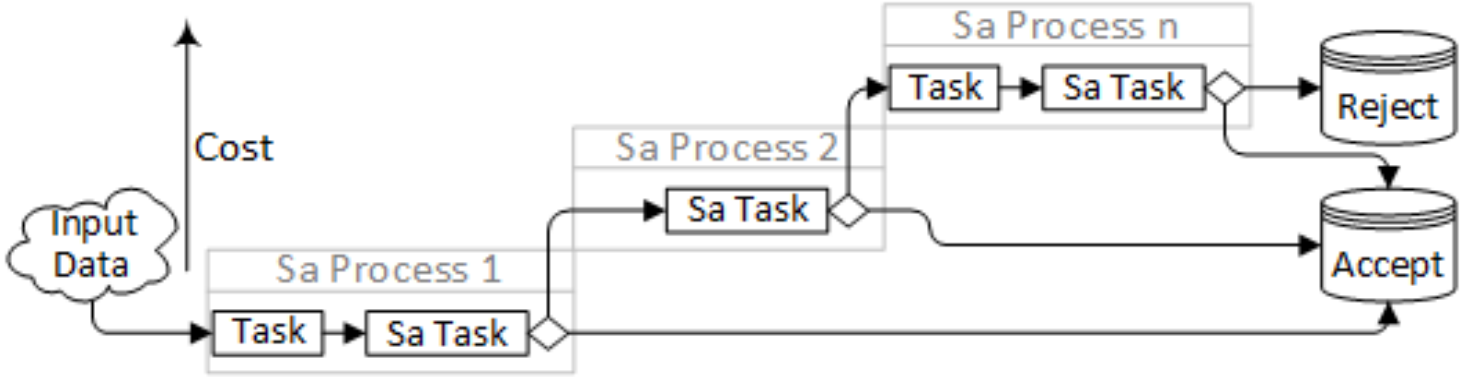
Related Work

- Crowdsourcing platforms: Notes from Nature, Zooniverse.
- IE Applications: Symbiota, SALIX

  Help accelerate but not replace human.
- Workflow Management Systems (WMS): Pegasus, Triana, Taverna, Kepler. They could be used to implement SELFIE.
- In a previous study, were demonstrated the benefits of hybrid (human-machine) IE solutions.
- Self-awareness studies. Applied to other fields, mainly robotics and agents.

SELFIE: Self-aware Information Extraction

Workflow
Task
Self-aware Task (SaT)
Self-aware Process (SaP)

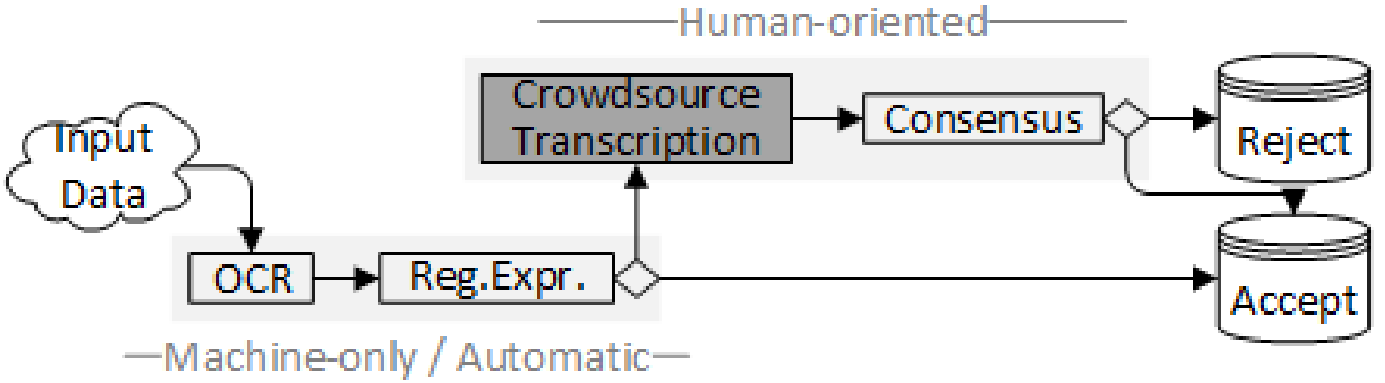


Self-aware Process (SaP)

Part	Input	Adaptable Script/program	Adaptable Acceptance Method	Outputs
Example	*Image x*	*/path/script1.py*	*[0,b) -> Task y* *[b,1] -> Accept*	*Image x* *Value, Confidence*

Experiment 1 – Event date

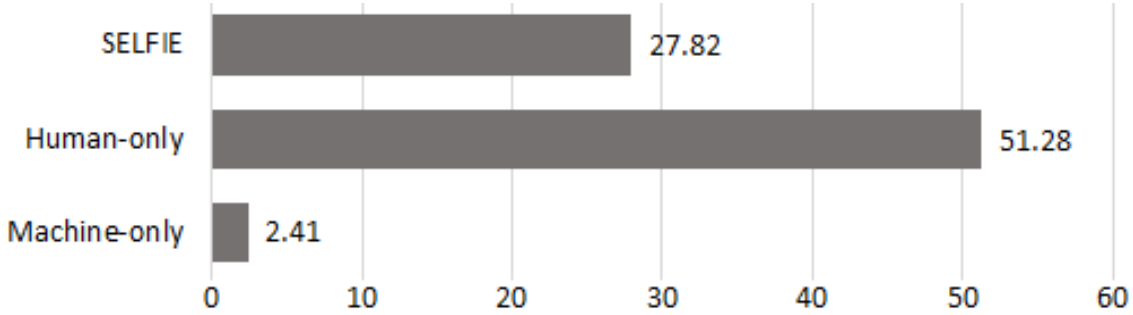
- Regular expressions utilized to extract dates.
- Longer the pattern, higher confident.
- Values with identifiable pattern



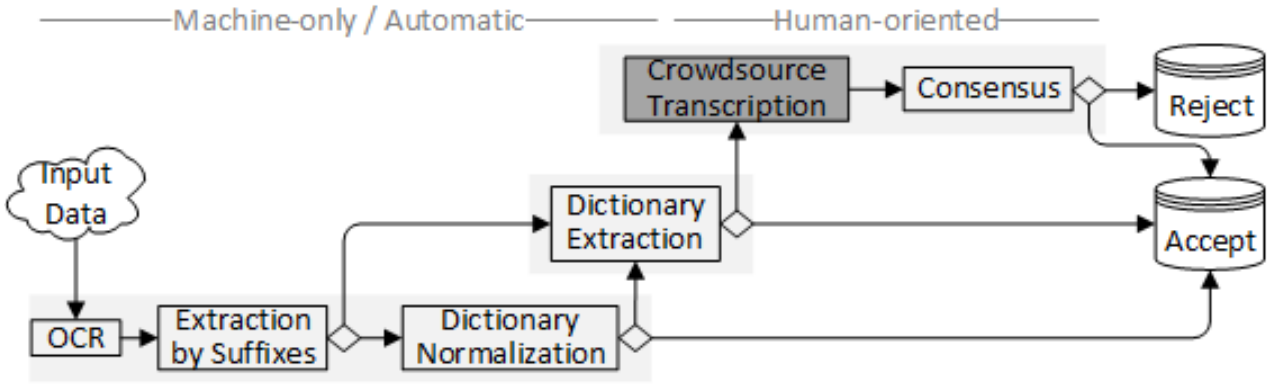
Similarity/Quality

SaP/SELFIE	# Accepted	Similarity	SEM	Std. Dev.
Machine-only	48	0.934	0.024	0.167
Human-only	51	0.971	0.022	0.155
SELFIE	99	0.953	0.016	0.162

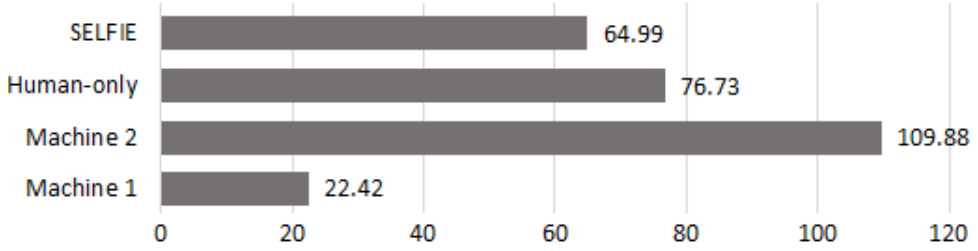
Average required time (seconds) per accepted date:



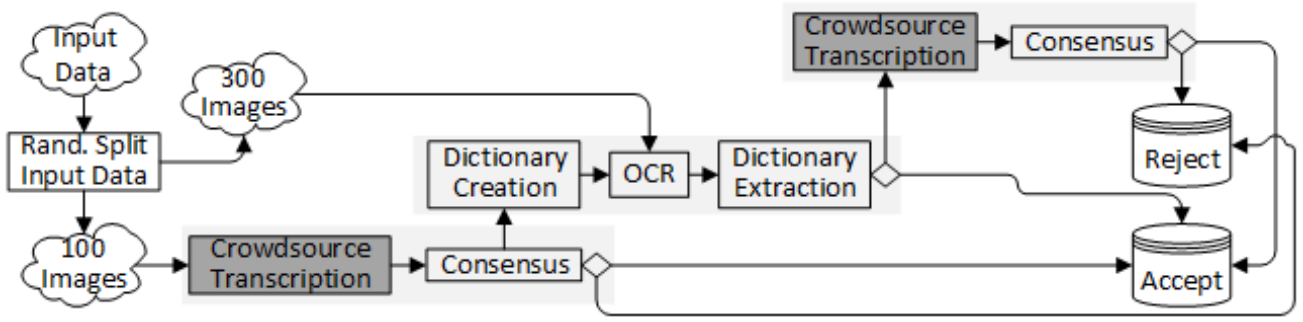
Experiment 2: Scientific name



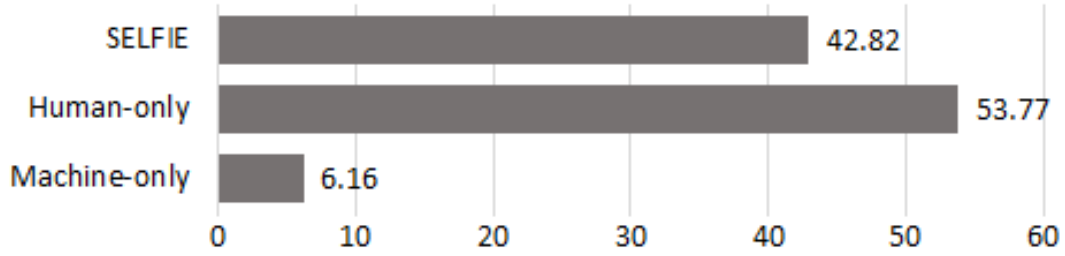
SaP/SELFIE	# Accepted	Similarity	SEM	Std. Dev.
1. Suffixes	15	1.0	0.00	0.00
2. Dict. Ex.	10	1.0	0.00	0.00
3. Crowd	66	0.944	0.026	0.214
SELFIE	91	0.959	0.019	0.183



Experiment 3: Recorded by



SaP/SELFIE	# Accepted	Similarity	SEM	Std. Dev.
1. Human 100i	92/100	0.900	0.030	0.288
2. Machine-only	94/300	0.862	0.027	0.262
3. Human 300i	191/206	0.900		
SELFIE	375/400	0.895		



Conclusions

- The paper proposes SELFIE, a hybrid (human-machine) IE model for biocollections. SELFIE is based on the execution of a cost-ordered sequence of IE processes and the use of self-aware tasks which can evaluate the quality of their results and decide whether to accept the values or to send the input to be analyzed to a higher quality process.
- Three experiments following the proposed SELFIE model showed that it is possible to extract information from biocollections datasets using less time, human resources, and monetary cost than the human-only IE alternative without significantly degrading quality.
- On average, when using the SELFIE model, the time required to extract an accepted value was reduced by 27.14%. This estimated reduction considers only the tasks execution time and the processing time of the data. It does not consider the time needed to organize crowdsourcing activities and developing or setting the required software infrastructure. Likewise, it was not considered the time spent on programming the IE scripts.
- On average, the number of required human-hours and other crowdsourcing costs were reduced by 32% when using the SELFIE model, while the quality negligibly decreased by 0.27%.
- Three different types of fields, commonly found in biocollections were used in the experiments to demonstrate that self-aware tasks can be created for a wide variety of cases. One case considers field values that are easily identifiable. Another case illustrates a method to create dictionaries from real data in order to enable automatic IE.

Thank you!

Any 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.