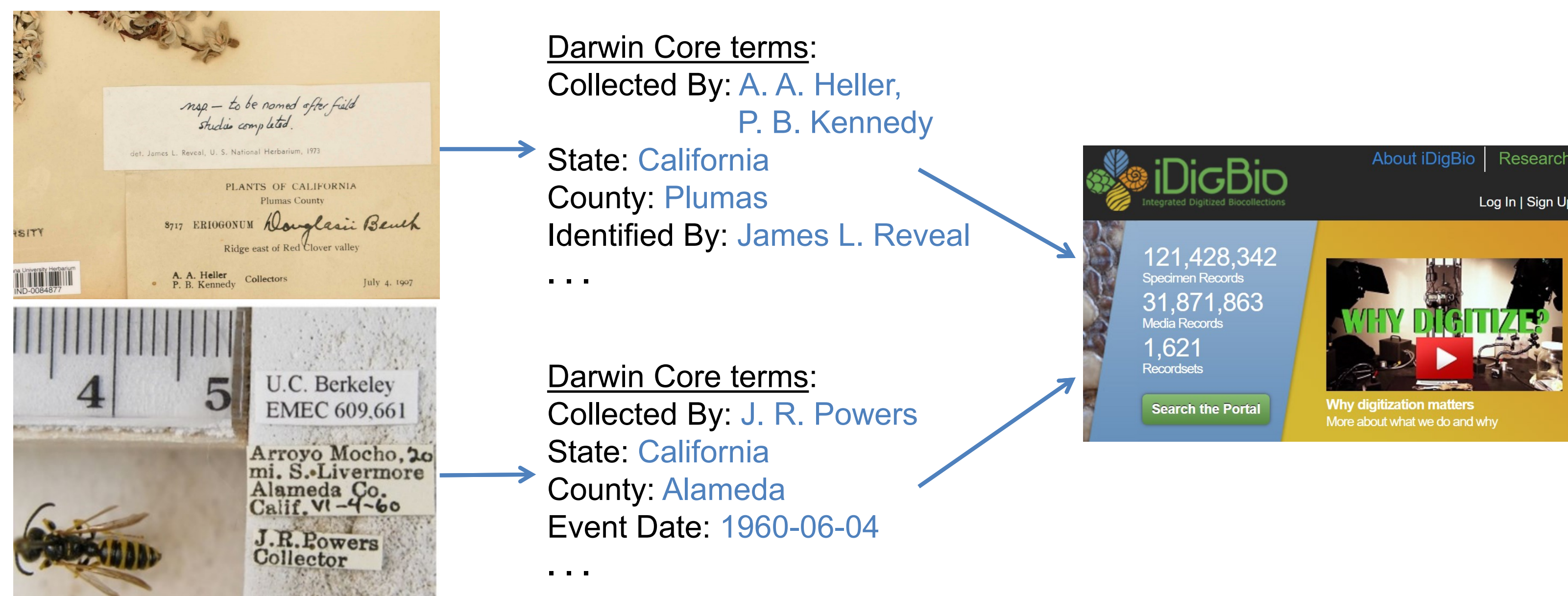


## Abstract

Biodiversity information extraction (IE) from imaged text in digitized museum specimen records is a challenging task due to both the large number of labels and the complexity of the characters and information to be extracted.

The HuMaIN project investigates software-enabled solutions that support the combination of machine and human intelligence to accelerate IE from specimen labels.

Among other contributions, the project proposed the use of self-aware workflows to orchestrate machines and human tasks (the SELFIE model), Optical Character Recognition (OCR) ensembles and Natural Language Processing (NLP) methods to increase confidence in extracted text, named-entity recognition (NER) techniques for Darwin Core (DC) terms extraction, and a simulator for the study of these workflows with real-world data. The software has been tested and applied on large datasets from museums in the USA and Australia.



**Darwin Core terms:**  
Collected By: A. A. Heller, P. B. Kennedy  
State: California  
County: Plumas  
Identified By: James L. Reveal  
...

**Darwin Core terms:**  
Collected By: J. R. Powers  
State: California  
County: Alameda  
Event Date: 1960-06-04  
...

## Task Design and Crowd Sentiment in IE from Biocollections

**Problem:** Volunteers are a scarce resource. How do interface's characteristics affect crowdsourcing result and crowd sentiment?

**Results:**

**Transcription vs. Selection:**

- Selection-based tasks generate results of 7.7% higher quality than transcription-based tasks.
- Users take 35% less time completing selection-based tasks than transcription-based tasks.
- Selection-based tasks are perceived as 15% more boring than the equivalent transcription-based tasks.

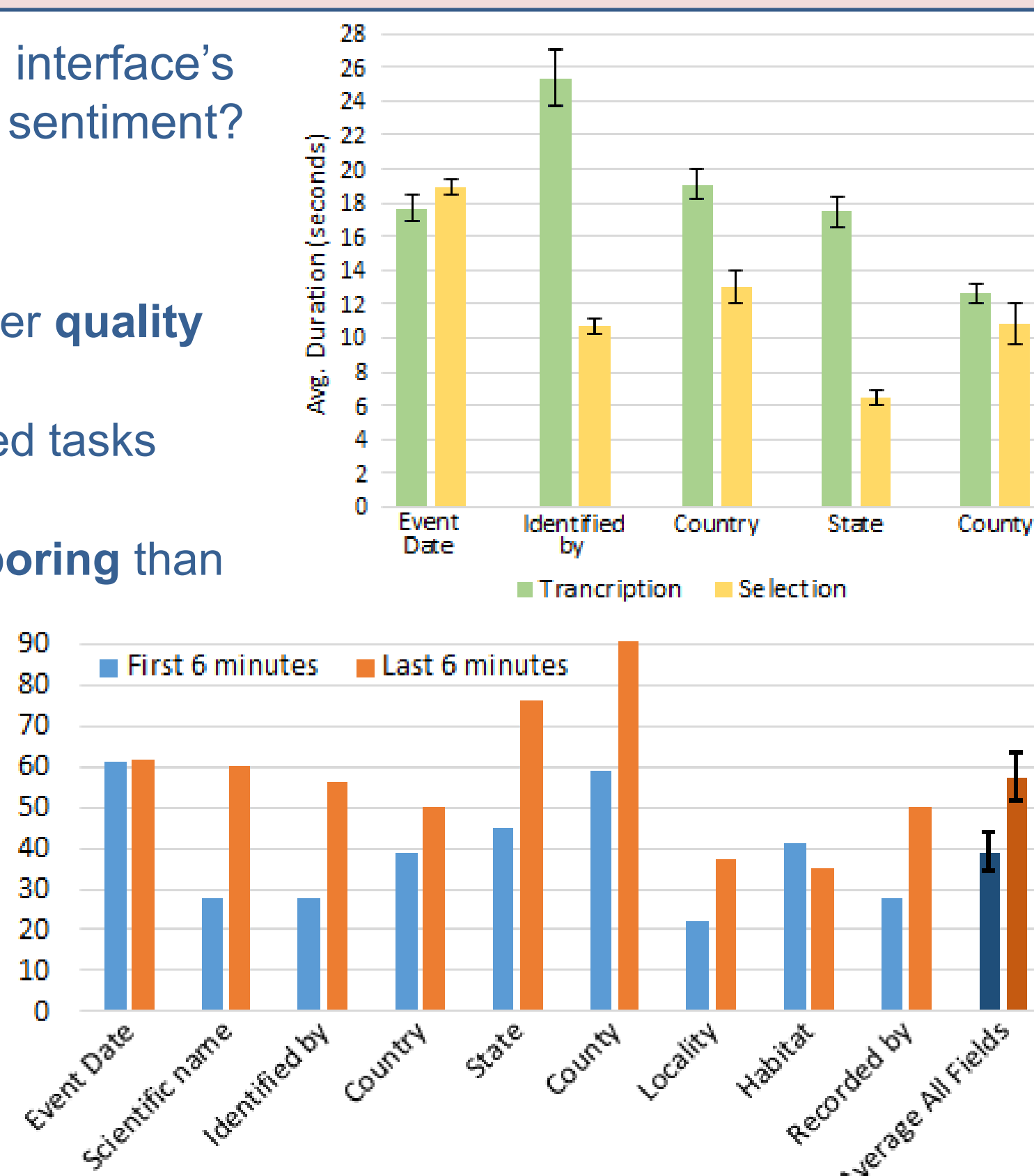
**Granularity: Twelve 1-field tasks vs. 12-fields task:**

- Twelve 1-field tasks improved the result's quality by 7.25% but required twice the time than a single 12-fields task.
- Users found 1-field tasks easier to complete than multi-fields tasks.

**Users' Learning Process: First vs. Last 6 minutes.**

- No difference found in the quality of the output.
- 1.5x more values are extracted in the last 6 minutes.

**Paper:** Task Design and Crowd Sentiment in Biocollections Information Extraction. Icaro Alzuru, Andréa Matsunaga, Maurício Tsugawa, and José A.B. Fortes. 3rd IEEE Collaboration and Internet Computing, 2017.

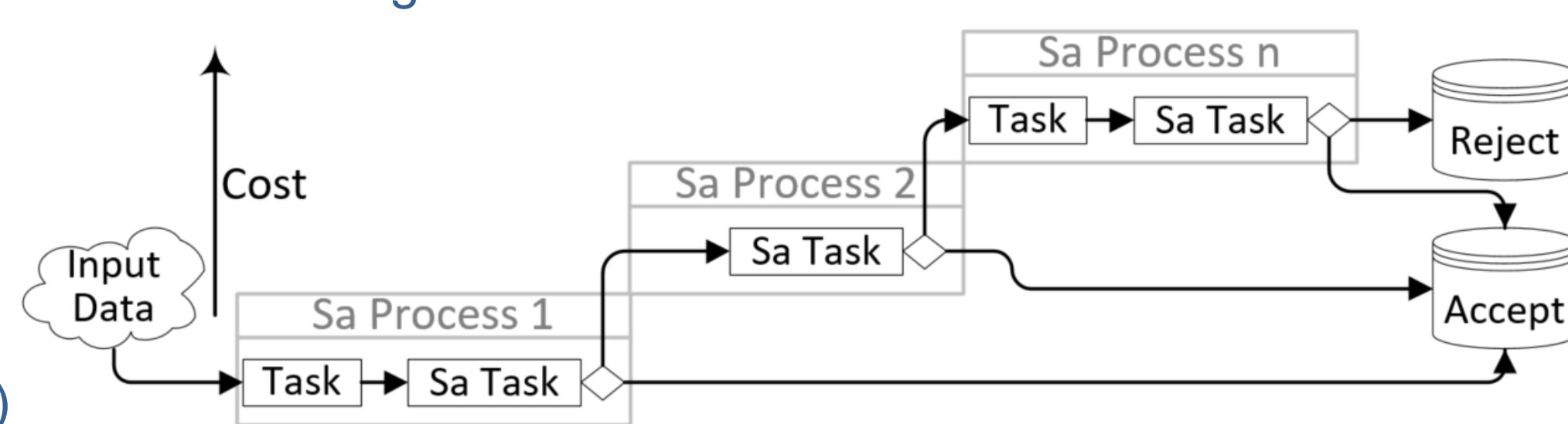


## Self-aware Information Extraction (SELFIE) from Biocollections

**Problem:** Data generated by automated methods for biocollections IE are not accepted because they contain many errors, only crowdsourcing is used.

**Proposed Solution:** Self-aware IE workflows. A SELFIE workflow is a sequence of self-aware human-machine processes (SaPs) organized in incremental-cost order. The cost is a function of performance variables defined by the workflow designer. Elements of a workflow:

- Input (e.g., specimens' images)
- Self-aware Processes (SaPs)
- Data processing Tasks and Self-aware Tasks (SaTs)
- Output: DC terms (Accept), Unknown images' terms (Reject)



SaTs extract and evaluate candidate values, taking the most appropriate action: accepting the own result or requesting processing by the next (more costly) SaP.

**Results (SELFIE Experiments):**

- Event date: Alphanumeric with defined patterns. NLP: Regular expressions.
- Scientific name: Textual known field. NLP: Suffixes (patterns), Sequential search.
- Recorded by: Textual unknown field. NLP: Dynamically created dictionary + search.

Table I. Results' improvement obtained by SELFIE, when compared to the Human-only IE approach.

	Event date	Scientific name	Recorded by	Avg.
Quality Improvement	-1.9%	1.6%	-0.6%	-0.3%
Duration Reduction	45.8%	15.3%	20.4%	27.2%
Crowdsourcing Reduction	48.0%	25.0%	23.5%	32.2%

**Paper:** SELFIE: Self-aware Information Extraction from Digitized Biocollections. Icaro Alzuru, Andréa Matsunaga, Maurício Tsugawa, and José Fortes. IEEE 13th eScience, 2017.

## OCR Ensembles for Text Extraction & Confidence Estimation

**Problem:** OCR engines generate errors (e.g. segmentation, recognition errors) that may propagate to extracted values.

**Proposed Solution:** An ensemble of three OCR engines (OCROpus, Tesseract, and the Google Cloud OCR) is used to estimate confidence in the extracted data.

**Characteristics of the Implementation:**

- Ensemble in a SELFIE workflow.
- Two hybrid human-machine approaches for crowdsourcing were tested.
- Dataset: representative group of 6 biocollections and ~ 3,000 images.

**Results:**

The ensemble labeled about 57.5% of the automatically extracted sentences as high-confidence transcriptions.

The two hybrid crowdsourcing techniques saved, on average, an additional 18.5% of crowdsourcing tasks, for total average saving of 76% in the crowdsourcing tasks.

**Paper:** Quality-aware Human-Machine Text Extraction for Biocollections using Ensembles of OCRs. Icaro Alzuru, Rhiannon Stephens, Andréa Matsunaga, Maurício Tsugawa, Paul Flemons, and José Fortes. IEEE 15th eScience, 2019.

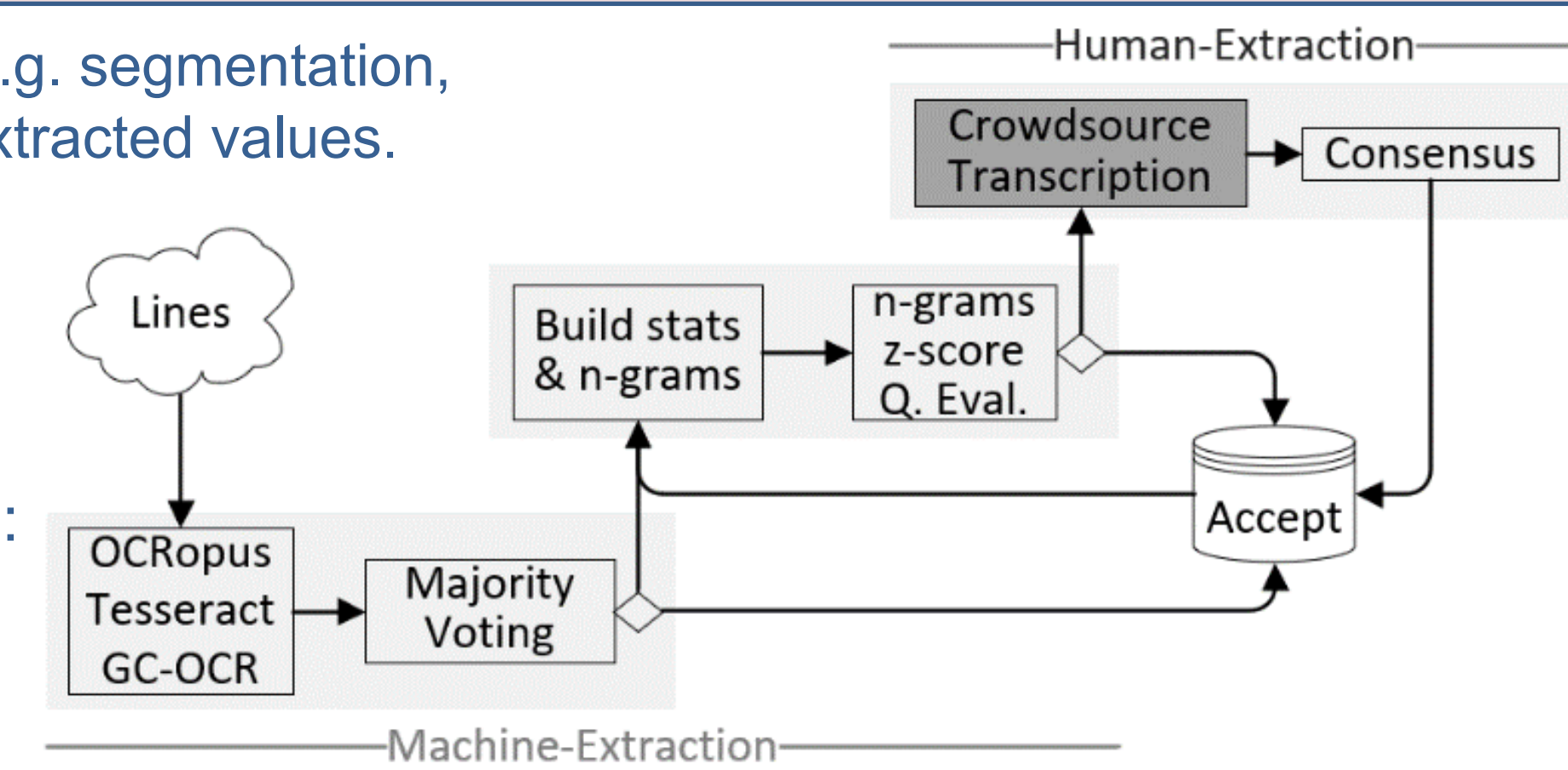


Table 4-5: Savings in the number of crowdsourcing tasks when the Ensemble of OCRs and two types of crowdsourcing are applied.

	Tasks required	Ensemble savings	Hybrid crowd. savings	Total savings
Dynamic Human-Machine Consensus	3 x nL	57.55%	15.801%	73.35%
Hybrid Transcriber /Reviewer	2 x nL	57.55%	21.225%	78.78%

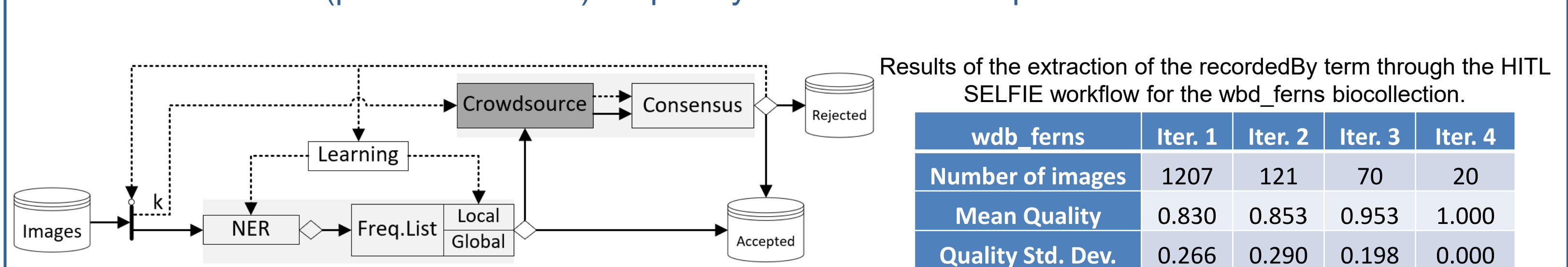
**Problem:** Proposed IE and confidence estimation methods rely on specific term's characteristics (e.g., suffixes, regular form). Can general IE and confidence estimation methods be created?

**Proposed Solution:** Reusing previously transcribed data to train general machine learning and statistical methods for the extraction and confidence estimation of DC terms.

**Implementation:**

- A named-entity recognition model was trained, by using 30% of the transcribed DC terms from 23,523 images of 29 biocollections, to extract ten DC terms. The remaining 70% of previously transcribed data were used for testing.
- By using local (dynamic - biocollection) and global (static - iDigBio) per-term frequency lists, high-confidence is assigned to automatically extracted values that exist in the local (crowdsourced) list or are found in the global (iDigBio) list 5 or more times.

**Results:** The NER model and the frequency lists were integrated in a Human-in-the-loop workflow for IE extraction. In every iteration, the transcription of the term in 50 images is crowdsourced, the NER model and the local (per-biocollection) frequency list are trained/improved with the new data.



## HuMaIN Simulator

**Problem:** Human-Machine IE requires images, crowdsourcing interfaces, volunteers, ground-truth values, scripts to process data, etc. Researchers invest a lot of time and resources validating an idea.

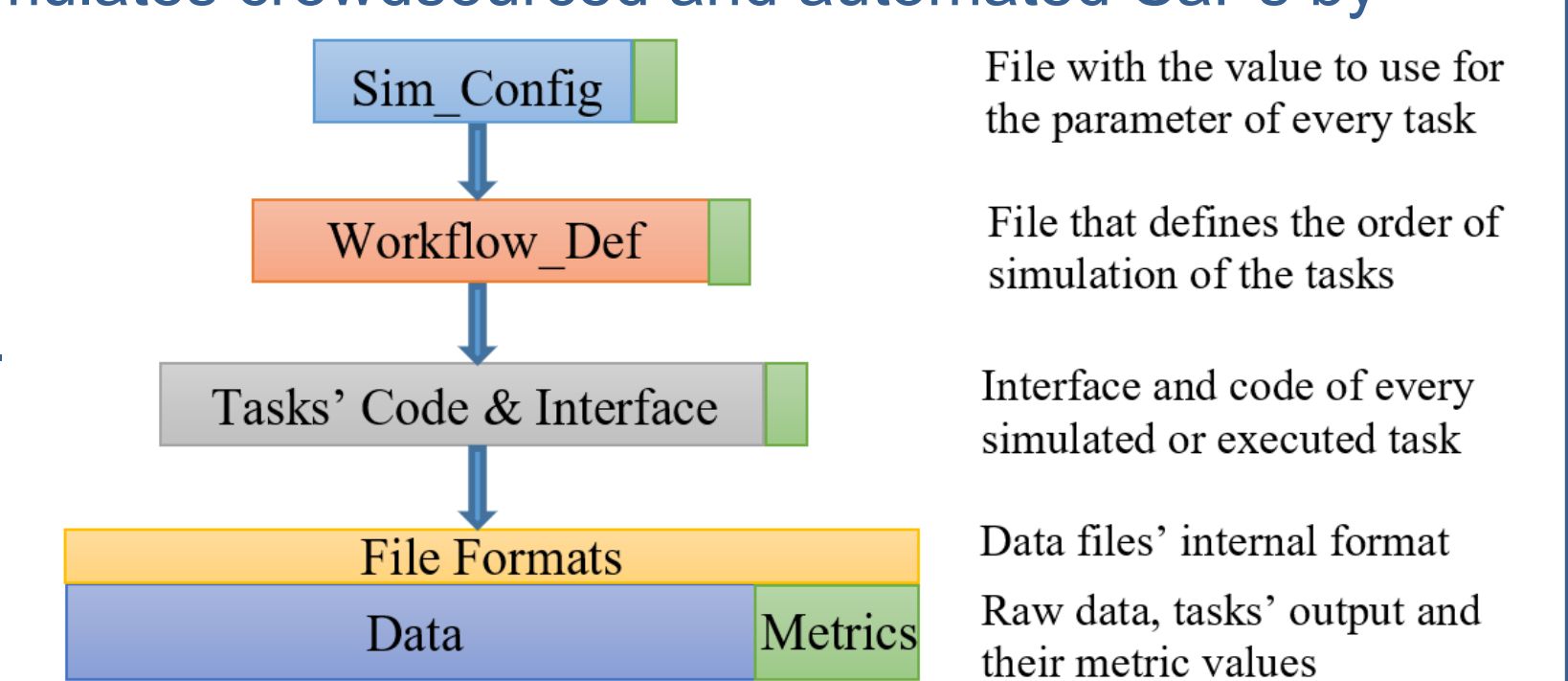
**Proposed Solution:** The HuMaIN Simulator. It emulates crowdsourced and automated SaPs by reusing previously extracted data.

**Characteristics:**

- Simulation engine and visualization capabilities.
- Hybrid workflows for the extraction of DC terms.
- Crowdsourcing and Automated IE Tasks' output
- Human-in-the-loop (iteration) support

**Results:** [github.com/acislab/HuMaIN\\_Simulator](https://github.com/acislab/HuMaIN_Simulator)

**Paper:** Human-Machine Information Extraction Simulator for Biological Collections. Icaro Alzuru, Aditi Malladi, Andréa Matsunaga, Maurício Tsugawa, and José Fortes. IEEE 3rd HMDData Workshop, 2019.



## Conclusions

- HuMaIN allowed the development of human-machine methods and software workflows for the extraction of DC terms from specimens' images. <https://github.com/acislab/HuMaIN>
- The hybrid method (SELFIE) and IE workflows were applied to real biocollections from iDigBio, the University of Australia, and WeDigBio FL Plants.
- SELFIE provides an approach to use automatically generated data, reducing the number of crowdsourcing sessions required while keeping a human-equivalent output quality.
- OCR Ensembles can be used to estimate confidence in the automatically extracted text.
- The data collected in biodiversity repositories can be used to generate automated extraction methods and creating confidence estimation methods for newly extracted data.
- The data and methods generated in this study can be tried and extended in the openly available HuMaIN Simulator.