SELFIE: Self-aware Information Extraction from Digitized Biocollections

Icaro Alzuru, Andréa Matsunaga, Maurício Tsugawa, and José A.B. Fortes
Advanced Computing and Information Systems (ACIS) Laboratory
University of Florida, Gainesville, USA

***Abstract*—Biological collections store information with broad societal and environmental impact. In the last 15 years, after worldwide investments and crowdsourcing efforts, 25% of the collected specimens have been digitized; a process that includes the imaging of text attached to specimens and subsequent extraction of information from the resulting image. This information extraction (IE) process is complex, thus slow and typically involving human tasks. We propose a hybrid (Human-Machine) information extraction model that efficiently uses resources of different cost (machines, volunteers and/or experts) and speeds up the biocollections' digitization process, while striving to maintain the same quality as human-only IE processes. In the proposed model, called SELFIE, self-aware IE processes determine whether their output quality is satisfactory. If the quality is unsatisfactory, additional or alternative processes that yield higher quality output at higher cost are triggered. The effectiveness of this model is demonstrated by three SELFIE workflows for the extraction of Darwin-core terms from specimens' images. Compared to the traditional human-driven IE approach, SELFIE workflows showed, on average, a reduction of 27% in the information-capture time and a decrease of 32% in the required number of humans and their associated cost, while the quality of the results was negligibly reduced by 0.27%.**

Keywords—information extraction; self-awareness; digitization; human-machine; biocollections

I. INTRODUCTION

The biodiversity research community is vigorously pursuing efforts towards the digitization of biocollections [1]. Thousands of volunteers, workers, and initiatives are extracting metadata from collected specimens and making them available to the scientific community and the general public.

The metadata visible in Figure 1 can be used to better understand pests, biodiversity, climate change, species invasions, historical natural disasters, diseases, and other environmental issues [3]. The potential consumers of these data go far beyond scientists, to include decision makers in agriculture, food security, public health, genomics, bioprospecting, and many other areas [2].

From 2012 through 2017, iDigBio has aggregated over 105 million digitized records of over 200 million specimens [4]. Around the globe, national projects such as Atlas of Living Australia [5], or Les Herbonautes in France [6], are amongst the many institutions contributing worldwide biodiversity data to the Global Biodiversity Information Facility (GBIF). GBIF, in existence since 2001, currently reports 740 million occurrences in its database [7].

These and other ongoing successful digitization efforts are still short of meeting the daunting challenge of digitizing all private and public biological collections, whose specimens have been estimated at 1 billion in the US [3] and between 2.5 and 3 billion in the whole world [8]. Digitization tasks done by humans, of which information extraction (IE) is an example, are particularly slow by comparison with automated tasks. Using only human-driven IE, the digitization of all the biological collections could take decades. Similar challenges are faced by other collections (e.g., geological and paleontological) [9][10].

Figure 1. EMEC609675 Cerceris conifrons – Entomology Collection

Nowadays, workflows for biocollections mass-digitization usually require the participation of volunteers who transcribe Darwin core terms [11] (such as *scientificName*, *recordedBy*, or *eventDate*) and experts who review the digitized information to guarantee its quality. Some computational tools can be used to assist volunteers and accelerate IE workflows [13][14]. Our previous work [12] discusses the impact of using hybrid (human-machine) systems on the duration and the quality of biocollections' information extraction; it shows that, in this domain, human work usually generates output of better quality, but at a lower rate than machine processing.

This paper considers the following question: How can biocollections' information extraction be accelerated and made more efficient while keeping the quality of the results similar to what capable humans can provide? Towards answering this question, a hybrid (human-machine) Self-aware Information Extraction model, called SELFIE, is proposed. SELFIE includes tasks that assess themselves to determine whether their results have an acceptable quality. Self-aware tasks trigger other tasks

978-1-5386-2686-3/17 $31.00 © 2017 IEEE
DOI 10.1109/eScience.2017.19

when they are unable to produce results that meet acceptance criteria. Confidence estimations are computed for each extracted value, determining its acceptance or, instead, the processing of the original input by an extraction method that has better quality at higher cost. The objective of SELFIE is to minimize the use of costly resources while maintaining output quality similar to the human-driven approach. The model also helps the planning and organization of IE work done by data scientists.

Three experiments are reported in this paper to explain and demonstrate the capabilities of the model. The experimental results show, on average, a reduction of about 32% in the number of required humans and associated crowdsourcing cost, a reduction of 27% in the duration of the IE process, and a negligible decrease of 0.27% in the IE quality.

The data and code utilized in this paper can be found at https://github.com/acislab/HuMaIN_Self-aware_Information_Extraction.

II. Related Work

The advent of the information era has condemned to extinction the traditional sequential and per-specimen cataloging process of museum collections described in [15].

In 2012, the Workshop for Developing Robust Object-to-Image-to-Data (DROID) Workflows, using the experience of 28 digitization programs, initiated the construction of a collections digitization workflow. Nelson et al. [16] identify three dominant digitization workflows: in the first two, metadata are transcribed directly from the physical labels; in the third (Image to Data to Distribution), the image is collected and, in a posterior process, the information is extracted from it. This is the type of workflow studied in this paper, because it can lead to partial automation of the information extraction process.

In 2015, during the Herbarium Workflows Workshop at Valdosta State University (Georgia, USA), a collections data extraction workflow that leverages three years of experience was presented. This customizable specimen processing pipeline has 14 modules [19]:

(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

The image capture result, an example of which is shown in Figure 1, is done by modules 7 through 9 of this workflow. The information extraction model proposed in this paper (SELFIE) is relevant to the Data Capture (11) and Transcription Blitz modules (12).

As it can be inferred by the utilization of Transcription Blitzes, which are "*short periods of intense effort involving more than the average number of people involved in digitization*", the information extraction from large numbers of specimens relies on crowdsourcing, possibly online, thanks to initiatives like Notes from Nature [17] and the Zooniverse platform [18].

Nowadays, many data capture processes are intrinsically organized around a software-based system that facilitates transcription, in a semi-automatic IE process. Examples of these applications are Symbiota [25], used for the digitization of the New York Botanical Garden; and SALIX [26], used in several digitization projects at the Arizona State University Herbarium. The latter is reported to accelerate the entry process in nearly 30%, in comparison to typing. These software products are very specific to the type of collection for which they were created, some are tied to proprietary software and, more importantly, they help accelerate, not replace, human work. Volunteers and experts still have to verify each value extracted by the Optical Character Recognition software. SELFIE proposes to rely on computed confidence to avoid the human verification and editing of the extracted values. All these semi-automatic tools could be included in the human IE processes of a SELFIE workflow.

Extracting information from biocollections is complex because of its hybrid (human-machine) nature, the heterogeneity of the technologies to use, the amount of data to process, its interdisciplinary nature, and its multiple optimization goals (maximum quality, minimum duration, and effective resource utilization). Complexity is the main reason of the rise, in the last two decades, of self-awareness and advanced autonomous behavior in computing systems [27].

There is no definition or model of self-aware computing that applies to all domains. In general, "*computing systems are self-aware if they possess the capability to learn and exploit the models of themselves and the environment in which they are situated so as to act in accordance with high-level goals*" [31].

Self-aware principles have been applied to robotics [32], agent theory, and other areas, but we believe it is the first time they are applied to IE from biocollections. In hardware design, [28] defines a SElf-awarE Computing (SEEC) framework to meet conflicting goals (e.g., high performance with low energy consumption), using dynamically scheduling actions. In IE from biocollections, there are conflicting goals (e.g., low cost with high quality) and we search for the optimal orchestration of tasks with similar functionality but different cost and quality, in order to carry-out an IE job. Heterogeneity (humans & machines) of the processes adds additional complexity. SELFIE is a workflow model, but it is not a workflow management system (WMS) as Pegasus, Triana, Taverna, or Kepler [39]. Its implementation could use one of these WMS.

SELFIE is inherently hybrid: machine and human dynamics are combined to improve the IE result. In [12], it was shown that hybrid workflows using Optical Character Recognition (OCR) can improve the accuracy of information extraction processes by more than 42%. It was also observed that the OCR tool Tesseract

is, on average, 25 times faster than OCRopus (another open source OCR tool) [33], which explains why Tesseract is chosen for OCR tasks in our experiments.

III. Self-aware Information Extraction Model

Information extraction (IE) is the process of finding and linking relevant information from unstructured and semi-structured machine-readable sources [22][23]. We will use a workflow of data processing tasks to represent IE models (see Figure 3).

The quality of the IE process is commonly controlled and verified by data scientists. In order to make the model self-aware, elements capable of assessing themselves are required.

"*A self-aware system has knowledge of itself and its experiences, permitting reasoning and intelligent decision making to support effective autonomous adaptive behavior*" [24]. In this paper, we will follow this definition, which includes both capabilities: assessing and acting, when referring to an entity as self-aware. Other authors understand self-awareness as just knowing the current state and use self-expressiveness to refer to the ability to adapt [36].

A. Self-aware Task (SaT)

In the context of this paper, Self-aware Tasks (SaTs) are data processing tasks capable of assessing their confidence on the quality of their outputs, and deciding whether they should be accepted as final results or not. If the output is accepted, the processing of the subject's field ends. If it is rejected, the SaT "adapts" the SELFIE workflow by deciding where to send the candidate value or original unstructured data for further processing.

In general, a SaT can be represented by a tuple consisting of (at least) input type, output type, an adaptable script or program, and an adaptable acceptance method with the corresponding actions to take. Possible actions include accepting an output and selecting an alternate task for further processing. The acceptance method could, for example, be implemented using artificial intelligence (AI) algorithms, logic rules, or ad-hoc tests. In general, both the script/program and the acceptance method can change or learn over time on the basis of observed output values or other feedback. However, in the cases considered in this paper, the adaptation of the script/program is limited to deciding where to send its output based on feedback from the acceptance method. The reported experiments do not consider the case when the acceptance method is adaptable. Future work will consider SELFIE workflows where more complex adaptations of both the script/program and the acceptance method are possible.

Figure 2 exemplifies the components of a SaT. The "*Adaptable Acceptance Method*" section shows that candidate values with quality between b and 1 will be accepted, while values with quality lower than b will be sent to Task y.

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*

Figure 2. Components of a Self-aware Task (SaT).

B. Self-aware Processes (SaP) and the SELFIE Model.

Generally, data scientists can use and combine different methods to extract the information from the data source: crowdsourcing, optical character recognition (OCR), natural language processing, AI-based methods, experts' transcription, etc. Each of these methods has different implications on utilization of resources, running time, monetary cost, and quality. For example, using domain experts can generate high-quality results, but it can be very costly in comparison to other options (and very slow when few domain experts are available).

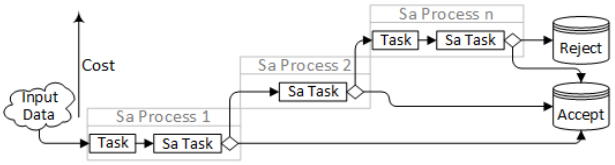

Figure 3. Generic SELFIE workflow. The tasks are sequenced according to their (increasing) cost.

The idea of the Self-aware Information Extraction (SELFIE) model (Figure 3), is to opt for using information extraction methods in incremental cost order (i.e., least expensive method first, followed by more expensive only if previous method has poor-quality output). The term "Self-aware Process" is used to refer to a sequence of tasks that perform an information extraction method.

The basic principle in the workflow design is the following: if a process P has a higher extraction cost than process Q and generates results of lower or equal quality, process P should be discarded and not considered to be part of the workflow. This principle can be repeatedly applied among the IE alternatives to build a workflow similar to Figure 3.

The cost of a SELFIE workflow (and its processes and tasks) is a function of one or more metrics that the implementer of the model decides as appropriate, e.g., execution time, required investment, number of volunteers, or any combination thereof. One example is the cost of using resources for the time needed to execute the workflow, which can be computed by the product of the cost of the resource per unit of time, the number of resources, and the execution time. Such a model could be used, for example, to capture the cost of using cloud resources.

A SaP is a logical group of tasks that collectively implement an information extraction method, while a SaT is a machine or human activity that is part of a SaP. The last task of a SaP is always a SaT, while previous tasks can be conventional tasks. A SaP can be as simple as a single SaT. The required last SaT of a SaP includes the decision to either accept output candidate or invoke another task to generate better outputs.

The model in Figure 3 represents *n* different SaPs. Potentially, more than one SaP could be active at a certain moment in time, typically working on different data. When a candidate output of sufficient quality is produced by a given process, it is accepted as a final result and the processing finishes for the correspondent input. If the minimum quality requirement is not satisfied after a given SaP, the candidate output and/or input is passed to a SaP of higher cost, which is expected to generate a result with better quality.

The SELFIE approach is conceived and evaluated for extraction of information from biocollections data, but can be generalized to other domains. The SELFIE model adaptation can occur at different levels: SaT (adapting the program and acceptance method), SaP (adapting its tasks), and SELFIE model (adapting the SaPs).

C. Nomenclature

Figure 4 shows the graphical elements of SELFIE workflows. Besides Self-aware Tasks and Self-aware Processes, already explained, the following objects are used:

Unstructured Data: the group of elements to be analyzed, which must be in a digital format to allow their use in crowdsourcing and/or machine tasks. Internally, the data required to extract a particular information item is not explicitly demarked. Audio files, images, and unstructured documents are common examples of unstructured data in biocollections. Unstructured data is usually specified with a logical address from where the elements can be retrieved.

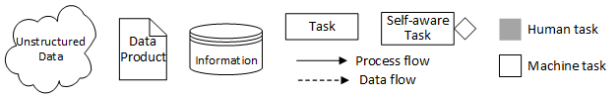

Figure 4. SELFIE's nomenclature

Information: processed data with some degree of organization, common examples being database tables or columnar files. Their processing is typically straightforward. The objective of the information extraction process is to provide the metadata in some structured format. In biocollections, outputs are formatted using Darwin Core terms. The SELFIE model adds a confidence estimation for each output value. Two storage buckets collect the output: one stores the accepted values and the other one collects the list of specimens for which a value with sufficient certainty could not be identified.

Data Product (DP): data objects that support the execution of tasks during the IE process – e.g., dictionaries, OCR's confusion matrix, or frequency lists. DPs can be created by an external entity or by a workflow's task. A DP is specified through its name, location, and internal structure.

Tasks: data processing steps, which can be executed by humans (gray-shaded in the figures) or machines (unshaded) and can be interconnected to form a workflow. A task is defined by a name, a script or interface to call, its parameters, and the output format. Crowdsourcing tasks require the specification of the number of users per subject that are expected to complete the task.

D. Hybrid SELFIE for Biocollections

In biocollections, the images from which the metadata need to be extracted are very diverse, as it can be observed in Figure 5. This variability makes it challenging for artificial intelligence methods and automated information extraction to produce high quality results. Human intervention is always needed, therefore crowdsourcing methods and experts input have been historically preferred.

We propose a hybrid approach where human and machine dynamics are integrated in a single model that efficiently uses the strength of each type of task.

As machine-only workflows are typically faster than crowdsourcing methods, the key is to identify when machines produce results of higher or equivalent quality to that provided by humans. It is hereon assumed that SELFIE models are hybrid; if that is not the case, it will be explicitly indicated.

IV. Experimental Setup

A. Dataset

Between 2011 and 2014, the iDigBio's Augmenting OCR Working Group (A-OCR) [29] carried out several initiatives to generate content and tools for the scientific digitization community. One of the results was a dataset with 400 images distributed in three collections: 100 insects (entomology), 100 herbs, and 200 lichens images [30]. The dataset includes experts' transcription of about 25 Darwin Core terms (fields) for these specimens' images. This dataset was used to evaluate and verify the proposed SELFIE model.

Figure 5. Image excerpts from the iDigBio-AOCR dataset. Left: image of the lichens collection. Right: image from the entomology collection (contains different labels, the specimen, a ruler, and graphical characteristics that make its automatic IE processing more difficult).

Figure 5 shows excerpts of two different images in this dataset. Specimens, rulers and graphs are included in the pictures, which can have different quality, background, language, fonts, font sizes, symbols and types of writing (handwriting and/or typed).

The inputs used in the first two experiments were a subset of 100 images: 34 insects, 33 herbs, and 33 lichens, randomly selected from their respective collections. In the third experiment, all the 400 images of the dataset were used as inputs.

B. Crowdsourcing

At the ACIS Lab of the University of Florida, a crowdsourcing experiment was conducted for the extraction of 12 fields (Darwin Core terms), using the subset of 100 images mentioned above. The processed fields are: *Event date*, *Scientific name*, *Identified by*, *Country*, *State*, *County*, *Latitude*, *Longitude*, *Elevation*, *Locality*, *Habitat*, and *Recorded by*.

Thirty-eight (38) participants, most of them undergraduate students, were asked to transcribe the fields. At least three different participants transcribed each field of each image. Participants were paid at a rate of $10 per hour, being allowed a maximum of two hours of work. Each hour of crowdsourcing

was divided in three parts: 5-15 minutes of training, 40 minutes of work, and 5 minutes to answer a survey.

The "Crowdsource Transcription" tasks found below in the experiments' workflows and their generated data correspond to this crowdsourcing activity.

C. Damerau-Levenshtein (DL) similarity.

Quality estimations for the transcribed (crowdsourced) and machine extracted values were computed using the normalized DL algorithm, which calculates the distance of two strings as the minimum amount of insertions, deletions, substitutions, and transpositions of two adjacent characters, required to convert one string into the other [21].

Symbols were excluded, leading and trailing spaces were removed, internal double spaces were converted to a single space, and strings were lowercased before computing the DL similarity; which is defined as the complement of the normalized DL distance:

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

D. Hardware/Software Platform

The experiments were executed in an ASUS N46J laptop (CPU: Quad core i7 and 12 GB of RAM), running Ubuntu 16.04 Desktop.

The Optical Character Recognition software product utilized was Tesseract [20] version 3.04.01. The information extraction scripts were developed in Python and executed using Anaconda 4.3.14, Python 3.6.0, and GCC 4.4.7. The Geoffrey Fairchild's implementation of the Damerau-Levenshtein normalized distance algorithm [38] was used.

V. Experiments & Results

Three information extraction experiments using images from biocollections were completed. In addition to validating and showing the usability of the SELFIE model, they present IE alternatives for different types of text fields.

A. Experiment 1: Event-Date Extraction

The main purpose of this experiment is to find out whether the use of the SELFIE model to process biocollections' images can reduce the information extraction time while maintaining a similar quality result.

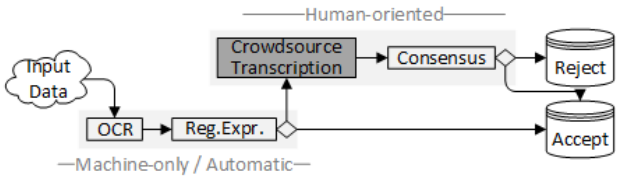

Figure 6. *Event date*'s SELFIE model

The SELFIE model depicted in Figure 6 was used to extract the *Event date* (collection date) of the specimens. In Figure 5, for example, the *Event dates* are "March 12, 1931" (left) and "VI-4-60" (right).

Images from biocollections can contain several dates, but the *Event date* is usually the oldest one, because it is the date when the specimen was collected. There can be other dates in the image labels, indicating other cataloging events, but these are posterior to the collection date.

The *Event date*'s SELFIE model has 2 processes: a process that tries to automatically extract the *Event date*, which uses an Optical Character Recognition (OCR) algorithm on the image and then extracts the dates via regular expressions; and a process that uses humans to transcribe the *Event date* (crowdsourcing experiment). In this last case, the crowd is asked to transcribe the *Event date* of a specimen only if the automatic process failed to find it or there is not enough confidence on the extracted value.

Each image was transcribed by three participants, and a consensus algorithm was used to reconcile possible mistakes and/or differences in opinions. The consensus algorithm selects one of the dates based on the average similarity to the other candidates. Each workflow task is described below.

OCR: Whole images (Input Data) are processed by the Optical Character Recognition software (Tesseract), which generates a plain text file with presumably all the textual information of the image (Output of the task).

Reg. Expr.: The text generated by the OCR (Input) is scanned in search of patterns with a date format, e.g., Month DD, YYYY. Several date patterns are tested. Because the OCR process can generate some garbage characters and omit others, only long dates, with month in textual format, are considered. The confidence is implicitly obtained when the sequence of words matches a date pattern with reasonable length. Because an image can contain more than one date, the script returns the oldest one. In each image, the oldest identified date is considered the Accepted *Event date* for the specimen, which are sent to the repository of accepted values. If no date is identified in the image, it is sent to the next process for further processing. The regular-expression analysis, as all the other scripts utilized in this research, can be reviewed and downloaded from the GitHub repository of the paper [35].

Crowdsource Transcription: Figure 7 shows the Web interface used by the participants to complete this crowdsourcing experiment, which can be tried online at [34]. After a short training, participants completed the *Event date* transcription task. Users typed in a box the date found in the image and clicked on "Save and Next" to proceed with the next image. At least three volunteers processed each image. If there is no date in the image, the text box must be left blank; and if there are more than one date, workers are instructed to write the oldest one. The transcription must be verbatim ("write exactly what you see, no interpretations or completions").

The output of this task are three strings per specimen (image), which can be blank. These candidate *Event dates* are sent to the consensus task.

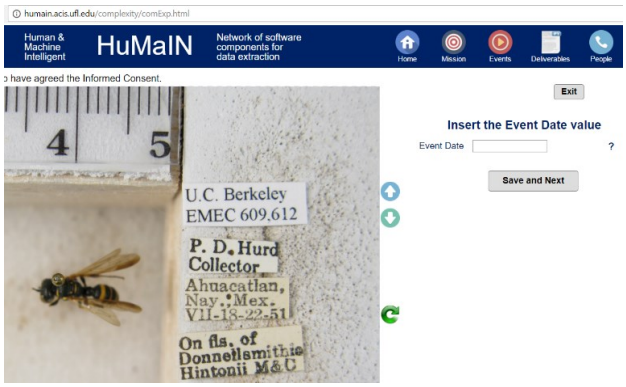

Figure 7. *Event date* Transcription interface

Consensus: This SaT receives the three candidate dates and computes the Damerau-Levenshtein similarity among them. If two or more candidate outputs have the same value, this value is chosen as the winner and the returned similarity is 1.0. If the three values are different, the candidate with the highest average similarity is the winner. In order to be accepted, the winner must have an average similarity of 0.75 (in a range from 0 to 1) or higher; otherwise, the three candidate dates are rejected.

After running the first SaP (machine-only), a date was found for 48 of the 100 images. The remaining 52 images were sent to be extracted by the crowd. From these 52 specimens, one image was rejected because consensus was not reached.

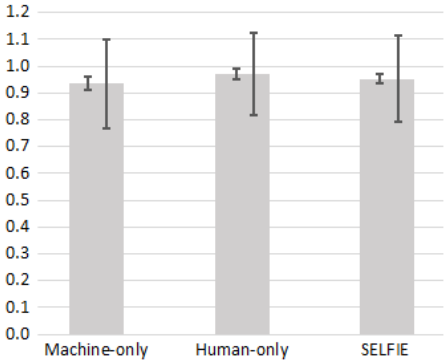

Figure 8. Average similarity (0.0 - 1.0) of the *Event date* IE SaPs & SELFIE. The standard error to the mean (left) and the standard deviation (right) are shown as error bars for each process and for the whole workflow.

The generated dates were compared to the experts' transcription using the Damerau-Levenshtein similarity. Figure 8 shows the average similarity, standard error of the mean (left, small range), and standard deviation (right range) for the two IE processes and for the entire workflow (SELFIE). The values represented in Figure 8, are detailed in Table I.

TABLE I. SIMILARITY TO EXPERTS' TRANSCRIPTION – EXPERIMENT 1

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

The crowdsourcing process obtained a 3.7% higher quality (similarity to experts' output) than the automatic (machine-only) SaP, with a relatively small error but considerable variability in both cases.

The SELFIE workflow delivered a quality of 0.953, which is 1.8% lower than the human-only approach (assuming the human-only approach would keep the same quality when processing all the 100 images).

Figure 9 shows the average time required to generate an accepted output, assuming a sequential execution and disregarding the time required for programming, systems setup, advertising campaign, event scheduling, etc. needed for setting and executing machine and human IE processes.

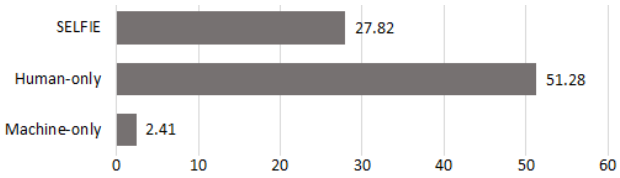

Figure 9. Average required time (seconds) per accepted *Event date*.

The machine-only process is about 21 times faster than the human-only IE approach. The overall (combined) IE workflow took 54.25% of the time it would have taken a human-only IE process. The hybrid SELFIE model execution was 1.84 times faster than the traditional approach for data capture.

Therefore, using the SELFIE model for processing biocollections' images, it was possible to reduce the IE time, while maintaining a result with a quality similar to the common information extraction approach. For 99 of 100 images, the *Event date* was collected with an average similarity of 0.953, with respect to experts' transcription. The number of humans required for the crowdsourcing experiment and the cost related to the crowdsourcing activity were reduced by 48% because 48 *Event dates* were automatically extracted. Other observations about the experiment are as follows:

- Since Tesseract was not modified to improve its performance on the experiment's dataset and no improvements were made to the images or the OCR process, the amount of erroneous characters was high. A better or improved OCR process could decrease the errors of the extraction script, increase the amount of automatically extracted dates and reduce the overall execution time.
- We used a relatively simple script for regular-expression analysis. More robust data mining and machine learning mechanisms could be used to potentially improve the quality, recall, and execution time of the task.
- In SELFIE, it is key to create SaTs with an accurate quality assessment function. In this experiment, all the errors of the automatic process happened because the OCR was not able to recognize parts of the text in the image. This impacted negatively the accuracy of the acceptance criteria and the whole experiment success.

B. *Experiment 2: Scientific-Name Extraction*

The first experiment demonstrated how to extract *Event dates*, a mostly numeric field with a relatively fixed pattern. However, in biocollections, there are fields with values that do not follow an obvious pattern. In the last two experiments, we show how these other field types can be extracted.

In this second experiment, we want to check whether it is possible to extract a complex field (*Scientific name)* using a SELFIE model to increase time and resource efficiency when compared to traditional human-oriented extraction process, while producing results of similar quality.

Scientific names consist of two parts: the genus (first name) and the species (second name). One genus can have many species. These names, of Latin and Greek roots, have an internal structure with some specific meanings; for example, in some names, the suffixes "a", "us", and "um" are used to indicate the gender of the collected specimen, as feminine, masculine, and no gender, respectively [37]. Similarly, suffixes -oidea, -idae, and -inae are used to indicate the size or type of the family-group. Taking advantage of these suffixes, we implemented the SELFIE model as depicted in Figure 10 for the extraction of the *Scientific name* field.

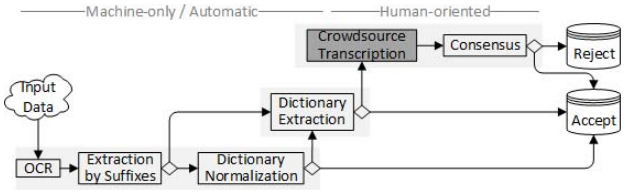

Figure 10. *Scientific Name*'s SELFIE model

Existing taxonomies can be used as a dictionary to validate the extracted values, while suffixes can help identify candidate values and reduce the amount of processing. The workflow was divided into three SaPs:

1. The OCR is run on each specimen's image and the output text is processed by a "Extraction by Suffixes" script, which tries to identify *Scientific name* candidates. If the script does not generate any candidate for an image, its OCR's output is sent to the second process. Using the Damerau-Levenhstein similarity, candidate values are compared to the *Scientific names* of a Dictionary. The *Scientific name* of the highest similarity to the candidate value is accepted if the DL similarity is greater than 0.9. Otherwise, the OCR's output is sent to the next process.
2. The second process scans the text provided by the OCR, searching for genus and species with high similarity to one of the *Scientific names* in the dictionary. If the similarity is higher than 0.9, the *Scientific name* of the dictionary is accepted. Otherwise, it is sent to the next process. The similarity comparison performed in this and the previous SaPs enables the correction of small errors in the candidates values.
3. In the last SaP, the images are processed by humans. After receiving a short training, at least three different users transcribed the *Scientific name* value in each image. A consensus algorithm, based on similarity among the candidate values picks one winner or sends the image to the rejected images container. The consensus criteria consists on selecting the candidate with the highest average similarity to the other candidates; the average needs to be higher than 0.85 for the output to be accepted.

Table II shows the number of accepted *Scientific name* values and the similarity (quality) with respect to the experts' transcription. Standard error of the mean and standard deviation are provided. SELFIE was unable to generate an acceptable output for only nine of 100 images.

TABLE II. SIMILARITY TO EXPERTS' TRANSCRIPTION – EXPERIMENT 2

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

The quality of the automatic processes was 5.6% better than the human-only SaP, which generated a small improvement in the quality (i.e., the similarity with respect to the experts' transcription) of the whole SELFIE model, when compared to the traditional human-only approach.

Figure 11 compares the average *Scientific name* extraction time per accepted output for the different IE processes and the SELFIE model. The average duration is the ratio of the total time required to process all images by the number of accepted values. For example, the Machine 2 process (Dictionary extraction) was used to extract the *Scientific name* from 85 images, taking 1098.82 seconds in total, but only 10 values were accepted; therefore, the graph shows 109.88 seconds, and not 1098.82/85 = 12.93 seconds.

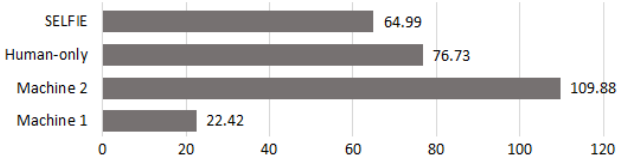

Figure 11. Average duration (seconds) per accepted *Scientific name* for the SELFIE model and the 3 processes of Experiment 2.

The SELFIE model reduced in 15.3% the time required to extract the *Scientific name* of the 100 images compared to the traditional human-only approach, while slightly increasing (1.5%) the quality of the result. Because 25% of the values were extracted using only machines, the human resources required to complete this information extraction process and the associated cost were reduced in this same proportion.

In this experiment, the implementation of two automatic SaPs relied on an external data structure (a dictionary) to compute confidence. Unfortunately, these processes could not extract a large portion of the *Scientific names* in the images, mainly because the OCR was not able to recognize part of the text in the images (some of the images have the *Scientific name* in handwritten text). This generated a low efficiency in the

Machine 2 (Dictionary Extraction) process (see Figure 11), which generated only 10 accepted values from 85 analyzed text files.

C. Experiment 3: Recorded-by Extraction

For the *Scientific name* field, it was possible to use a pre-existing dictionary, but not all the fields have a known list of possible values. This last experiment deals with textual fields for which possible or valid values are not known. It entails executing a human information-extraction process that allows the collection of some of the valid values, and then using these values for automatic extraction.

The success of this method depends on how repetitive the values are in the specific field. A machine learning algorithm could also be used instead of a dictionary in order to identify candidate values, converting this first human extraction process into a training process.

Experiment 3 evaluates if a SELFIE model can reduce the required time and resources to extract a field that has unknown values, while maintaining a quality equivalent to the human-oriented approach. The field to retrieve is the *Recorded-by* Darwin Core term, which is the name of the person or people who collected the specimen at some location.

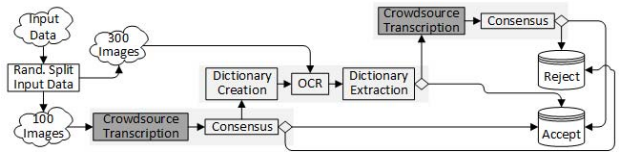

Figure 12. SELFIE model of the *Recorded by* field.

Figure 12 shows the proposed SELFIE model for the extraction of *Recorded by*. The first task, which can be considered part of the first SaP, divides the dataset in two subsets: a subset of 100 images to be transcribed by the crowd, and a larger subset (300 images) to be processed by machine algorithms.

The workflow has three SaPs:

1. The *Recorded-by* values of 100 images are transcribed by the crowd. Each image is transcribed by at least three users. The consensus algorithm computes the similarity among the candidate values. The candidate with the maximum average similarity, which has to be greater than 0.85, is accepted.
2. Using the *Recorded-by* values that were accepted in the consensus process, a dictionary or list of valid values is created. The remaining 300 images are processed by the OCR, generating 300 labels or text files. These files are scanned by an extraction script that uses the created dictionary. The sequence of strings that correspond to an entry in the dictionary, are accepted, while specimens for which no *Recorded by* value are found are sent to the third process.
3. This SaP is the same as the first SaP but using the remaining images for which no *Recorded by* value was identified in the second SaP. This process was not implemented, we assume that its quality is the same as the first crowdsourcing process.

Table III shows the similarity to the experts' transcription obtained by the human and machine processes of the experiment. The quality of the machine-only SaP was 3.8% lower than the quality of the human-only process.

TABLE III. SIMILARITY TO EXPERTS' TRANSCRIPTION – EXPERIMENT 3

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		

A total of 94 values were automatically extracted, which reduced by 23% the humans required in crowdsourcing and the cost of it. Assuming that the human IE process for the remaining 206 images will have a similar quality to the first "training" human IE process, the SELFIE workflow can deliver results that are only 0.5% worse than the human-only approach.

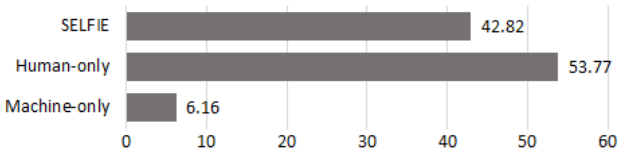

Figure 13. Average duration (in seconds) for extracting accepted "*Recorded by*" values for the SELFIE model and the processes of Experiment 3.

The overall duration of the extraction process was reduced by 20.36% (see Figure 13) compared to the human-only approach. The machine process is much faster than the human process, but only a small subset of values was automatically extracted. If this subset increases, the speedup will also increase.

Nevertheless, 23% of the *Recorded by* values were automatically extracted thanks to the creation of a dictionary from the real data. It would have been much harder, even impossible, to extract the field values using regular expressions or other basic techniques.

VI. ERROR AND COST ANALYSIS

The SELFIE model is highly affected by the accuracy of the self-aware tasks in recognizing incorrect candidate values. If this self-assessment fails, the quality of the whole process is compromised.

If the designer of the workflow sets a high-quality threshold to accept the extracted values, high execution time and utilization of costly resources will be expected. On the other hand, if the thresholds are set to a low value, quality cannot be guaranteed.

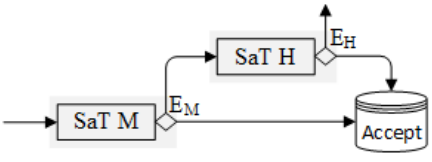

Figure 14. SELFIE's basic components.

Consider a SELFIE workflow (see Figure 14) with the following characteristics:

- Two SaTs: Machine (M) and Human (H). The analysis applies regardless of the nature of each task.
- Each task has an estimator (E) of its level of confidence in its output being correct.
- When the estimator predicts a correct output, the task output is routed to the Accept repository. The remaining input subjects are forwarded to the next information extraction process.

Error Analysis:

A self-aware task has two components: an information extraction function and an estimator. The extraction function has the following associated probabilities:

P(C) = Probability of extracting the correct value.
P(I) = 1-P(C) = Probability of extracting an incorrect value.

Similarly, the estimator has the following associated probabilities:

P(A) = Estimator's probability of accepting an extracted value.
P(R) = 1 – P(A) = Probability of the estimator of rejecting an extracted value.

Four conditions can occur and each has an associated probability. They are:

P(A|C) = Probability of accepting an extracted value given it is the correct one (True positive)
P(A|I) = Probability of accepting an extracted value given it is incorrect (False positive). This is the probability of an error that affects SELFIE's output quality.
P(R|C) = Probability of rejecting an extracted value given it is the correct one (False negative). This is the probability of an error that affects SELFIE's cost.
P(R|I) = Probability of rejecting an extracted value given it is incorrect (True negative).

These probabilities depend on the accuracies associated with the implementations of the IE program and the estimator. In order to increase the quality of the process, we want to minimize *P(A, I)*, the probability of error all over the inputs. That is to say, we want:

$$\min P(A,I) = \min\{P(A|I).P(I)\}$$

In other words, we need to minimize the false positive conditions by using estimator and IE program that are as accurate as possible.

Cost analysis:

Let C denote total cost, C_M the cost of the task M and C_H the cost of task H. Let N denote the number of inputs to be processed. The total cost is the cost of processing all inputs by task M and the cost of processing the subset of inputs forwarded to H. The probability of an input being sent to H is the probability *P(R)*, i.e., *P(R|C) + P(R|I)*. The formula of the total cost would be:

$$C = C_M * N + C_H * N * [\,P(R|C) + P(R|I)]$$

In order to minimize the cost, the expression $P(R|C) + P(R|I)$ should be minimized, for which a good estimator and information extraction tasks are again crucial.

Cost analysis is shown in general terms because of the high variability in hardware infrastructure and crowdsourcing implementations that can be used. For example, we decided to pay $10 to the participants of the crowdsourcing activity, but options like Zooniverse, which is free, could also be used: modifying importantly the structure of costs. A more detailed analysis will be undertaken in future follow-up work.

Integrated cost and accuracy analysis:

The SELFIE approach relies on task types such that the larger their cost the more accurate they are. Since we consider the combination of human and machine tasks, it is reasonable to assume the following: (1) human tasks are the most accurate and also much more expensive than machine tasks, and (2) compared to human tasks, machine tasks have much lower costs that may grow with accuracy requirements. In future follow-up studies, task cost models will be analyzed (e.g., constant, linear, and exponential), with the objective of making predictions. For example, how high must the accuracy of certain machine task be in order to accelerate the extraction time by a factor of ten with a bounded increase in cost? Similar analysis and questions will be addressed for the quality estimators.

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.

ACKNOWLEDGMENT

This work is supported in part by the National Science Foundation (NSF) grants No. ACI-1535086, EF-1115210, EF-1547229, and the AT&T Foundation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or the AT&T Foundation.

REFERENCES

[1] G. Nelson, "iDigBio: The U.S. National Science Foundation's National Resource for Digitization of Biological and Palobiological Collections." Geological Society of America, Annual Meeting, Vancouver, Canada, October, 2014.

[2] B. W. Bishop and C. Hank, "Data curation profiling of biocollections," Proceedings of the Association for Information Science and Technology, vol. 53, pp. 1-9, 2016.

[3] J. Hanken, "Biodiversity online: toward a network integrated biocollections alliance," Bioscience, vol. 63, pp. 789-790, 2013.

[4] Integrated Digitized Biocollections (iDigBio). [Online]. Available: https://www.idigbio.org/. [Accessed: 07-Jul-2017]

[5] Atlas of Living Australia. [Online]. Available: http://www.ala.org.au/. [Accessed: 07-Jul-2017]

[6] Les herbonautes. [Online]. Available: http://lesherbonautes.mnhn.fr/. [Accessed: 07-Jul-2017]

[7] Global Biodiversity Information Facility. [Online]. Available: http://www.gbif.org/. [Accessed: 07-Jul-2017]

[8] A. H. Ariño, "Approaches to estimating the universe of natural history collections data," Biodiversity Informatics, vol. 7, 2010.

[9] J. H. Beach, "Conceptualizing and managing paleontological collection data with Specify Software," in GSA Annual Meeting in Vancouver, British Columbia, Advancing the Digitization of Paleontology and Geoscience Collections: Projects, Programs, and Practices II, session, vol. 342, 2014.

[10] Museum of Geology, Digitized Collections. South Dakota School of Mines & Technology. [Online]. Available: http://www.sdsmt.edu/Academics/Museum-of-Geology/MPRL/Digitized-Collections/. [Accessed: 07-Jul-2017]

[11] J. Wieczorek, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais. "Darwin Core Terms: A quick reference guide." Date Modified: 2015-06-02. [Online]. Available: http://rs.tdwg.org/dwc/terms/. [Accessed: 07-Jul-2017]

[12] I. Alzuru, A. Matsunaga, M. Tsugawa and J. A. Fortes, "Cooperative human-machine data extraction from biological collections," in e-Science (e-Science), 2016 IEEE 12th International Conference on, pp. 41-50, 2016.

[13] D. Lafferty and L. Landrum, "SALIX, the Semi-automatic Label Information Extraction System," 1st ed. Tempe: School of Life Sciences, Arizona State University, 2012.

[14] C. Gries, E. Gilbert and N. Franz, "Symbiota–a Virtual Platform for Creating Voucher-Based Biodiversity Information Communities," Biodiversity Data Journal, 2, e1114, 2014.

[15] P. S. Humphrey and A. C. Clausen, "Automated Cataloging for Museum Collections: a model for decision and a guide to implementation," Association of Systematics Collections, pp. 79, 1976.

[16] G. Nelson, D. Paul, G. Riccardi and A. Mast, "Five task clusters that enable efficient and effective digitization of biological collections," Zookeys, vol. 209, pp. 19, 2012.

[17] Notes from Nature, Transcribe Museum Records. [Online]. Available: https://www.notesfromnature.org/. [Accessed: 07-Jul-2017]

[18] Zooniverse. [Online]. Available: https://www.zooniverse.org/. [Accessed: 07-Jul-2017]

[19] G. Nelson, P. Sweeney, L. E. Wallace, R. K. Rabeler, D. Allard, H. Brown, J. R. Carter, M. W. Denslow, E. R. Ellwood, C. C. Germain-Aubrey, E. Gilbert, E. Gillespie, L. R. Goertzen, B. Legler, D. B. Marchant, T. D. Marsico, A. B. Morris, Z. Murrell, M. Nazaire, C. Neefus, S. Oberreiter, D. Paul, B. R. Ruhfel, T. Sasek, Joey Shaw, P. S. Soltis, K. Watson, A. Weeks, and A. R. Mast, "Digitization workflows for flat sheets and packets of plants, algae, and fungi," Applications in Plant Sciences, vol. 3, issue 9, pp. 1-9, 2015.

[20] Tesseract Open Source OCR Engine. [Online]. Available: https://github.com/tesseract-ocr/tesseract. [Accessed: 07-Jul-2017]

[21] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," International Journal of Computer Applications, vol. 68, 2013.

[22] J. Cowie and W. Lehnert, "Information extraction," Commun ACM, vol. 39, pp. 80-91, 1996.

[23] A. McCallum, "Information extraction: Distilling structured data from unstructured text," Queue, vol. 3, pp. 48-57, 2005.

[24] P. R. Lewis, A. Chandra, F. Faniyi, K. Glette, T. Chen, R. Bahsoon, J. Torresen and X. Yao, "Architectural aspects of self-aware and self-expressive computing systems: From psychology to engineering," Computer, vol. 48, pp. 62-70, 2015.

[25] B. M. Thiers, M. C. Tulig and K. A. Watson, "Digitization of the New York Botanical Garden Herbarium." Brittonia, vol. 68, pp. 324-333, 2016.

[26] A. Barber, D. Lafferty and L. R. Landrum, "The SALIX Method: A semi-automated workflow for herbarium specimen digitization," Taxon, vol. 62, pp. 581-590, 2013.

[27] P. R. Lewis, "Self-aware computing systems: From psychology to engineering," in 2017 Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 1044-1049, 2017.

[28] H. Hoffmann, M. Maggio, M. D. Santambrogio, A. Leva, and A. Agarwal, "SEEC: a general and extensible framework for self-aware computing," Technical Report MIT-CSAIL-TR-2011-046, MIT, 2011.

[29] Augmenting OCR. [Online]. Available: https://www.idigbio.org/wiki/index.php/Augmenting_OCR. [Accessed: 04-Jul-2017]

[30] Label-data. [Online]. Available: https://github.com/idigbio-aocr/label-data. [Accessed: 04-Jul-2017]

[31] S. Kounev, P. Lewis, K. L. Bellman, N. Bencomo, J. Camara, A. Diaconescu, L. Esterle, K. Geihs, H. Giese and S. Götz, "The notion of self-aware computing," in Self-Aware Computing Systems, Springer, pp. 3-16, 2017.

[32] A. Gorbenko, V. Popov and A. Sheka, "Robot self-awareness: Exploration of internal states," Applied Mathematical Sciences, vol. 6, pp. 675-688, 2012.

[33] OCRopy. [Online]. Available: https://github.com/tmbdev/ocropy. [Accessed: 05-Jul-2017]

[34] Complexity. [Online]. Available: http://humain.acis.ufl.edu/complexity/. [Accessed: 06-Jul-2017]

[35] Self-aware Information Extraction Model. [Online]. Available: http://github.com/acislab/HuMaIN_Self-aware_Information_Extraction. [Accessed: 07-Jul-2017]

[36] T. Becker, A. Agne, P. R. Lewis, R. Bahsoon, F. Faniyi, L. Esterle, A. Keller, A. Chandra, A.R. Jensenius and S.C. Stilkerich, "EPiCS: Engineering proprioception in computing systems," in Computational Science and Engineering (CSE), IEEE 15th International Conference on, pp. 353-360, 2012.

[37] R. M. Coleman. Welcome to Introduction to Scientific Names. [Online]. Available: http://www.csus.edu/faculty/c/rcoleman/natural%20history%20museums/sacramento_state_online_natural_history_museum/introduction%20to%20scientific%20names.html. [Accessed: 07-Jul-2017]

[38] G. Fairchild, "pyxDameraуLevenshtein," GitHub, 2015. Available: https://github.com/gfairchild/pyxDameraуLevenshtein. [Accessed: 07-Jul- 2017].

[39] J. Yu and R. Buyya, "A taxonomy of scientific workflow systems for grid computing," ACM Sigmod Record, vol. 34, pp. 44-49, 2005.