# Cooperative Human-Machine Data Extraction from Biological Collections

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

***Abstract*—Historical data sources, like medical records or biological collections, consist of unstructured heterogeneous content: handwritten text, different sizes and types of fonts, and text overlapped with lines, images, stamps, and sketches. The information these documents can provide is important, from a historical perspective and mainly because we can learn from it. The automatic digitization of these historical documents is a complex machine learning process that usually produces poor results, requiring costly interventions by experts, who have to transcribe and interpret the content. This paper describes hybrid (Human- and Machine-Intelligent) workflows for scientific data extraction, combining machine-learning and crowdsourcing software elements. Our results demonstrate that the mix of human and machine processes has advantages in data extraction time and quality, when compared to a machine-only workflow. More specifically, we show how OCRopus and Tesseract, two widely used open source Optical Character Recognition (OCR) tools, can improve their accuracy by more than 42%, when text areas are cropped by humans prior to OCR, while the total time can increase or decrease depending on the OCR selection. The digitization of 400 images, with Entomology, Bryophyte, and Lichen specimens, is evaluated following four different approaches: processing the whole specimen image (machine-only), processing crowd cropped labels (hybrid), processing crowd cropped fields (hybrid), and cleaning the machine-only output. As a secondary result, our experiments reveal differences in speed and quality between Tesseract and OCRopus.**

***Keywords—Digitization; human-machine; data extraction; biological collections; optical character recognition; crowdsourcing***

## I. INTRODUCTION

The extraction of information from historical data sources, like medical records and scientific collections, is a challenging task. Standards were not used or have changed since these paper documents were created, and standards will continue to evolve. These data sources mix typed, printed, and handwritten text on paper that in some cases already turned yellow or stained.

Nevertheless, the information stored in these documents is a valuable heritage, which helps us understand the past, and more importantly could allow us to forecast and improve our future. Biological collections, for example, could help us model past and future environmental changes, develop new medicines, control agricultural pests, and understand or avoid epidemics, among many other benefits [1].

Governments and institutions have recognized the value of these biological collections and the importance of providing access to the cataloged specimens not only to researchers but to the general public. Projects like the Integrated Digitized Biocollections (iDigBio: https://www.idigbio.org), the Global Biodiversity Information Facility (GBIF: http://www.gbif.org), and the Atlas of Living Australia (ALA: http://www.ala.org.au), are stable providers of universal access to millions of specimens.

The challenge is that the actual number of specimens to digitize, as stored in collections worldwide, has been calculated in more than a billion [4]. Considering the current digitization rate, which is on the order of minutes per specimen, these data could take several decades to be processed [3]. This time does not account for the time to train the personnel who perform the digitization or the domain expert's time to manage the process and validate results. For mass-digitization of specimens, the recommended approach has been to divide the process into image capture and metadata transcription stages [5]. This alleviates the need for institutions to prioritize only the "most important" collections or specimens, and focuses the effort on curation and scanning. The transcription of the text can be performed at a later time, using the captured image, which opens the possibility of crowdsourcing [6][7]. In this work, we follow this mass-digitization approach, and assume that there are millions of specimen images that require data extraction.

Some researchers believe "there is no point in collecting complete metadata if these are not going to be used for any purpose" [5], but in this Big Data era, which has been defined as "collect now, sort out later" [9], we should not decide what metadata is important and what is not. The digitization process needs to be accelerated in order to digitize all historical data sources, ensuring the best result quality possible, and using experts only when strictly needed (mainly for verification). Experts should ensure the digitization's quality, while crowds and machines make high volume data processing possible. It is this mix of human- and machine-intelligent processes where we believe is the ideal solution. The goal of this study is to demonstrate that a good balance of these processes, in a single workflow, can lead to an improved overall process.

Today, a pure machine-intelligent process to transcribe text from biological specimen images using Optical Character Recognition (OCR) tools does not produce a good result. This kind of historical data source includes a wide range of challenges for the OCR tools, whose recognition algorithm is based on training. The variety of font types (printed and typed) and sizes, languages, background colors, lines, stains, and other elements in the image affect the recognition rate of the OCR. The other alternative, pure human-intelligent approaches [7][8], like crowdsourcing, can get better accuracy, but deal with a different set of challenges, like training the crowd and handling consensus among the diverse set of results.

978-1-5090-4273-9/16/$31.00 © 2016 IEEE

In this paper, we propose to improve the overall output quality by combining the OCR execution (a machine-intelligent process) with the help of humans (crowdsourcing), who crop the text areas that the OCR has to process. The cropping reduces the complexity, noise and size of data that the OCR tool has to binarize, segment, and recognize; leading to higher data quality and time-saving. The crowdsourcing step has been simplified to only require cropping of text, which does not demand a complex training and requires less effort than full text transcription. Crowdsourcing tasks were performed using a cropping web application developed for the HuMaIN (Human- and Machine-Intelligent Network) project, see http://humain.acis.ufl.edu.

To demonstrate that the cooperative human-machine data extraction improves the quality of the OCR process, getting closer to the ideal (expert-equivalent) result, we use 400 specimen images of the iDigBio project, for which the experts' transcription of the label and fields are known. These images cover three specimen types: Entomology (i.e. Insect), Bryophyte, and Lichen. Four approaches are evaluated:

1. Machine-only: the OCR tool is run on the whole original image, and the result is compared to the experts' label transcription. This establishes a comparison baseline.
2. Hybrid: humans crop the entire labels (using the HuMaIN interface [32]) from the specimen images and the OCR is run on them. The results are compared to approach 1.
3. Hybrid: considering a higher granularity level, humans crop individual fields (using the HuMaIN interface) from the specimen images and the OCR is run on them. The result is compared to the two previous approaches.
4. Improved machine-only: due to the amount of extra characters and digital noise generated by the OCR on the whole image, we add a simple cleaning algorithm on the machine-only results to compare between explicit noise removal, through cropping, and implicit noise removal, through elimination of non-interpretable set of characters.

Our results show that OCR's recognition rate improves by at least 42% for any of the two approaches where the text areas are cropped. In order to guarantee that the obtained results are independent of the used software or metric, two OCR tools (OCRopus and Tesseract) and three string similarity metrics (Damerau-Levenshtein, Jaro-Winkler, and the rate of matching words) were used. The total execution time of the cooperative data extraction process (machine + human) was reduced when using OCRopus, but increased when using Tesseract. This is due to the fact that Tesseract is faster than OCRopus while both provide similar recognition quality. None of the OCR tools were trained or tuned, i.e., the default version of the English language dictionary that these tools provide was used.

## II. Related Work

The Human-Computer Cooperation field is broad. In the HuMaIN project, and specifically in this study, we show the benefits of this interaction and how it can be applied to improve data extraction. There are data sources for which an automated data extraction process is sufficient to get the information we need. But for other data sources, like scientific collections, the data extraction must still rely on humans.

In 2011, iDigBio created the Augmenting OCR Working Group (A-OCR) with the goal of generating tools that improve the OCR process, either in output quality, speed, cost, or efficiency [10]. The group has organized several events which have generated a number of software solutions. One of these tools is the Semi-automatic Label Information Extraction system (SALIX) [11][12]: using a friendly graphical interface, the user can run the OCR and SALIX automatically assign text into individual fields. Corrections can be applied to the result. We agree with SALIX in the semi-automatic (machine and human) nature of the solution, considering how difficult it is to obtain a perfect OCR output for this type of data source. Nevertheless, in this work, our goal is to improve the output of the OCR, which would be the input of SALIX, i.e., we do not create a natural language post-processing tool. Moreover, our focus is on open source tools, while SALIX is mainly tested with ABBYY©, a proprietary OCR tool.

The Apiary project (http://www.apiaryproject.org), of the Botanical Research Institute of Texas (BRIT), has developed a "High-Throughput Workflow for Computer-Assisted Human Parsing of Biological Specimen Label Data" [14]. HuMaIN follows the same spirit as Apiary, with the difference that even though we are using biological collections as a use case, our final goal is a general platform for the definition of hybrid data extraction workflows in any area, generalizing the hybrid (Human-Computer) concept. Apiary includes a web application, called HERBIS, which inspired SALIX and works similarly.

A good amount of scientific projects (Apiary, the Atlas of Living Australia, Les Herbonautes, and Symbiota [34], among many others) have developed products and conducted research to digitize specimens, creating workflows that integrate crowdsourcing and machine-intelligent tools. This study uses a human task (cropping images) to improve the result of a machine-intelligent process (OCR), reinforcing the idea that we need both, humans and machines, to get an optimal result.

In the business world, several crowdsourcing vendors serve as a link between companies and crowds, like computer programmers, designers, or transcribers. One interesting case is CrowdFlower, which employs crowdsourcing in data extraction workflows to ensure or improve data quality, an area where crowdsourcing has been especially successful. Nevertheless, our initiative points to an open, customer managed, platform.

Several open source OCR products are available: Tesseract, JOCR (GOCR - http://jocr.sourceforge.net/), and OCRopus (OCRopy), among others. In [15] a comparative study between Tesseract and GOCR conclude that Tesseract has better accuracy and precision than GOCR. Creators of OCRopus, show in [16] mixed results for Tesseract and OCRopus error rates. OCRoRACT [13] uses an iterative method, as it trains Tesseract (segmentation based OCR) with the data from OCRopus (segmentation-free OCR) and then trains OCRopus with the data generated by Tesseract. After few iterations, the method is able to reduce the misinterpretation rate from 23% to 7%. OCRoRACT requires an expert who provides the unique set of characters for the documents and their Unicode representation. Nevertheless, in the case of biological collections, with the diverse mix of typed, printed, and handwritten text, this training oriented approach would not be the most appropriate.

## III. Domain Data Sets and Software Tools

This section describes the main characteristics of the images used in the study, the OCR tools, the Web application used to crop labels and fields, and the metrics employed to measure string similarity or quality of the results.

### A. Data Set

Between 2011 and 2014, the Augmenting OCR Working Group (A-OCR) of the iDigBio project, organized several events to generate content and tools for the digitization community. Among their results, there is a dataset with 400 images of cataloged specimens [27] and their corresponding whole label transcription as well as information parsed into individual fields by domain experts' transcription. Individual fields correspond to Darwin Core standard terms [25]. These images of biocollection specimens belong to 3 specimen types: 100 insects, 100 bryophytes, and 200 lichens. Distinctive characteristics of the images are:

- Entomology images (i.e., insect images): include a picture of the insect and a ruler, which helps measure the size of the insect. Instead of a single metadata label, there are several pieces of paper varying in size, color, and border. Some images have a scientific name written at the bottom left corner, which is an annotation made by the current institution, not the original one. See upper right image of Figure 1.
- Bryophyte images: are the largest images of the experiment (generating longer OCR times). Besides the labels, they include the specimen and other elements like stamps, maps, and bar codes. There can be segments of handwritten text and labels can have different orientations, but the background and image are mostly clear. See left side of Figure 1.


Figure 1. Bryophyte, Entomology, and Lichen images.

- Lichen images: do not include the specimen, and are basically big labels, but the resolution and contrast of the image are not the best. There is a stamp and a bar code in most of them. See lower right image of Figure 1.

Challenges in automatically extracting data from these images are: unformatted text, mixed with pictures, maps, stamps, and bar codes; different fonts and sizes; several languages; different background colors and resolutions; handwritten and underlined text. Some of their technical characteristics are specified in the following table:

Table 1. Number, size, and resolution of the specimen types

| 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 |

### B. Optical Character Recognition (OCR) Technology

The OCR process is the extraction, in machine encoded-format, of the typed, printed, and handwritten text of an image [21]. This process consists of a sequence of machine learning steps. In order to add generality and robustness to our results, two OCR tools are used:

OCRopus (OCRopy): is a group of open-source document analysis programs, which can be integrated as a Character Recognition System [17]. Its modularity makes it ideal for code reuse and teaching purposes. It requires running three steps or programs in sequence: the binarization (creates a black & white version of the image), the segmentation (divides the image in multiple text segments), and the recognition (applies to each segment the Recurrent Neural Network for text identification). We created a script to perform the execution of the three OCR steps [22]. In our study, OCRopy v1.0 and its ad-hoc English model were used, no further training or configuration was done. OCRopus recommends a 300dpi resolution as a minimum, which only bryophyte images comply.

Tesseract: is an open-source OCR engine initially developed by HP (between 1984 and 1995). It is sponsored, since 2006, by Google [18] [19]. Tesseract recommends a minimum resolution of 300dpi. Its artificial intelligence model must be trained, we used the default English trained model available in Tesseract 3.04. Tesseract is executed in a single step or program.

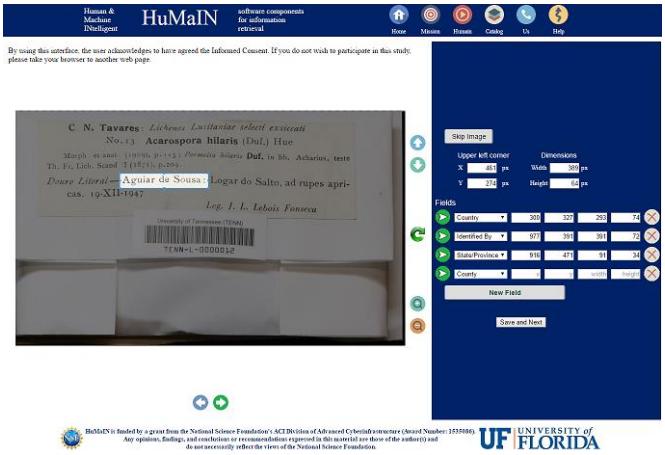


Figure 2. HuMaIN interface for fields cropping.

### C. HuMaIN

The Human- and Machine-Intelligent Network (HuMaIN – http://humain.acis.ufl.edu) is a project in development by the Advanced Computing and Information Systems (ACIS) laboratory (https://www.acis.ufl.edu) of the University Florida. It is funded by the National Science Foundation, with the goal

of researching and creating hybrid (crowdsourcing and machine learning) workflows of software components for data extraction.

Figure 2, shows the interface developed for cropping the fields of the images. The user picks the field name from one of the list boxes, selects the area where the value of the field is, and clicks on the green arrow button of that field. After selecting all the fields of a specimen, the user clicks on the "Save and Next" button to store the coordinates in the database. Later, these coordinates are used to generate cropped images of each field.

The label cropping interface works in a similar way, but the user only needs to select one text area. These two web applications were used to get the data for approaches 2 and 3. They are available at http://humain.acis.ufl.edu/app.html.

### D. *Metrics*

Comparing strings is a common Data Science problem. Many similarity metrics are token-based, which make them unsuited to compare sentences. Token-based metrics, require strings to have the same length and would penalize too much the characters inserted or omitted by the OCR, being that these events modify the absolute position of the characters that follow. Due to these reasons, it was decided to use edit-based metrics.

- Damerau-Levenshtein (DL) similarity: The DL distance of two strings is the minimum amount of insertions, deletions, substitutions, and transpositions of two adjacent characters, required to convert one string into the other [31]. The DL similarity is computed as the complement to 1 of the normalized DL distance:

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

  For this study, the Geoffrey Fairchild's DL normalized distance implementation of the algorithm [33] is used.

- Jaro–Winkler (JW) similarity: The JW algorithm considers the number of matching characters and adjacent transpositions, giving better rating to the letters that match at the beginning of the string [24]. The JW distance is not a metric in the mathematical sense [20], while its results are normalized (range 0 - 1), they do not represent a real distance. In our study, it was used the JW implementation available at the jellyfish 0.5.3 library [23].

- Matched words (mw) rate: This is an empirical metric, *mw(x, y)* is equal to the number of words of x that are in y, divided by the number of words in x:

$$mw(x,y) = \frac{|words\ in\ common\ between\ x\ and\ y|}{|x|} \quad (2)$$

  The order and frequency of words are not considered [22].

For these three metrics, a 0 (minimum value) means totally different strings, while 1 (maximum value) implies the strings are exactly the same (or "it is included in" for mw).

## IV. Experimental Setup, Results and Analysis

In this section, the experimental setup is detailed and the four approaches explained in Section I are evaluated with regard to consistency of outputs and the effectiveness of the human-machine cooperation.

### A. *Experimental Setup*

The machine used to run OCRopus and Tesseract has the following characteristics:

Hardware:
- System: IBM BladeCenter HS22 7870-AC1
- CPU: 2x Intel Xeon, E5540 (8 cores/16 HyperThreads).
- Storage: HGST HTS725050A7 (HD, 2.5 Inch, 500 GB)
- Memory: 48 GB, 12 x 4GB PC3-10600R DDR3 RAM

Software:
- CentOS Linux release 7.2.1511
- Python 2.7.5 (default, Nov 20 2015, 02:00:19)
- gcc (GCC) 4.8.5 20150623 (Red Hat 4.8.5-4)
- Tesseract 3.04.00, with leptonica-1.72
- OCRopy v1.0 (OCRopus)

### B. *Approach 1 (Machine-only – OCR whole image)*

The OCR process was executed on the 400 images (divided in Entomology, Bryophyte, and Lichen specimens) using OCRopus and Tesseract; and evaluated considering the Damerau-Levenshtein (DL), Jaro-Winkler (JW), and matching words (mw) similarity metrics. The average similarity, with respect to the experts' transcription, is shown in Figure 3 and summarized, for the DL metric, in table 2.

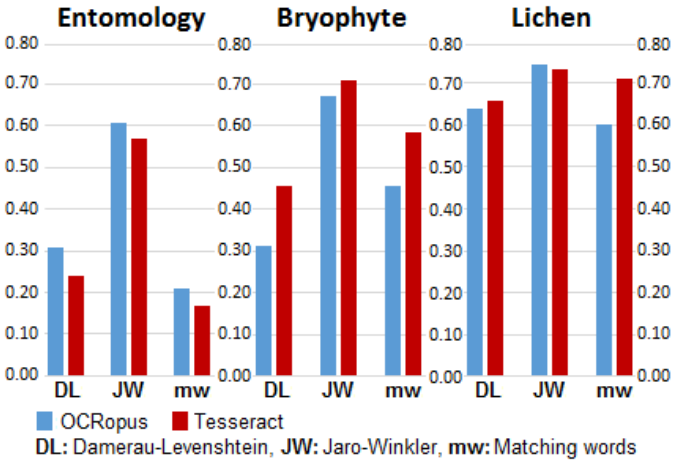

Figure 3. Average similarity per specimen type, OCR tool, and metric

The character recognition process worked better for lichens, followed by bryophyte and entomology images, considering the values obtained for the Damerau-Levenshtein (DL) metric.

Table 2. Average Damerau-Levenshtein (DL) similarity

| | Entomology | Bryophyte | Lichen |
|---|---|---|---|
| **OCRopus** | 0.31 | 0.31 | 0.64 |
| **Tesseract** | 0.24 | 0.46 | 0.66 |

Despite having the lowest resolution (only 96dpi), the text in lichen images is the easiest to be interpreted by the OCR: the three similarity metrics used were better for lichen images than for entomology and bryophyte images. In lichen images more than 60% of the words are recognized in both OCR technologies. These images are basically the label of the specimen, with few stamps, bar codes, or graphical elements in them, and a mostly white background.

Entomology images have text with non-white background, many lines (boxes and underlined text), and dark or shadowed regions around the text labels, that create additional borders or lines which mix with the text. Even though bryophyte images have more graphical elements than entomology images, their clean background, better resolution and contrast, make them get more accurate OCR results.

In Figure 3, the matching words (mw) metric, a string similarity measure which can be thought as more restrictive than the DL and JW metrics, got better or similar results (for Bryophyte and Lichen) than the DL metric. This shows that many of the words in the image are recognized, but added noise characters make the DL algorithm getting a worse similarity.

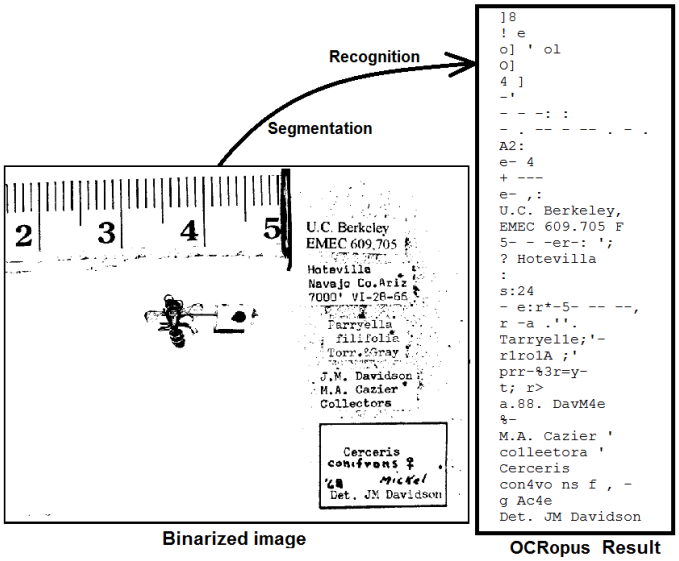


Figure 4. Result of the binarization (left) and interpreted text (right) of specimen EMEC 609,705

The result of the OCRopus binarization process for the entomology image presented in Figure 1 is illustrated in Figure 4. After the binarization, OCRopus executes the segmentation and recognition processes, to get the final result (right).

If the background is not completely white, the binarization process removes some pixels from the characters, making these letters more difficult to recognize. The dark areas around the small labels, create figures that the OCR tries to interpret. Boxes and underlined text also confuse the recognition algorithms.

For the specimen in Figure 4 (EMEC 609,705), the following similarity results were obtained:

Table 3. Similarity values obtained for specimen EMEC 609,705

| Similarity \ OCR software | OCRopus | Tesseract |
|---|---|---|
| Damerau-Levenshtein | 0.3627 | 0.3941 |
| Jaro–Winkler | 0.5943 | 0.6514 |
| Matching words | 0.4 | 0.4 |

In general, Jaro-Winkler metric shows more "optimistic" results, returning a higher similarity value than the Damerau-Levenshtein and matching words similarity metrics.

The execution time of the OCR process is shown in table 4. Tesseract was in average 18.5 times faster than OCRopus. During the study, none of the tuning features these tools provide were utilized. The main reason for the execution time difference is that OCRopus generates intermediate on-disk results during the binarization and segmentation steps, while Tesseract works entirely in memory, as a single process.

Table 4. Approach 1 – 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 |

In Figure 5, it is shown the similarity box chart for approach 1. We observe the Jaro-Winkler metric offers less variability in the results. We also notice high maximums (images where almost every word was identified) and very low similarity results (images for which the OCRs were not able to correctly identify a single word).

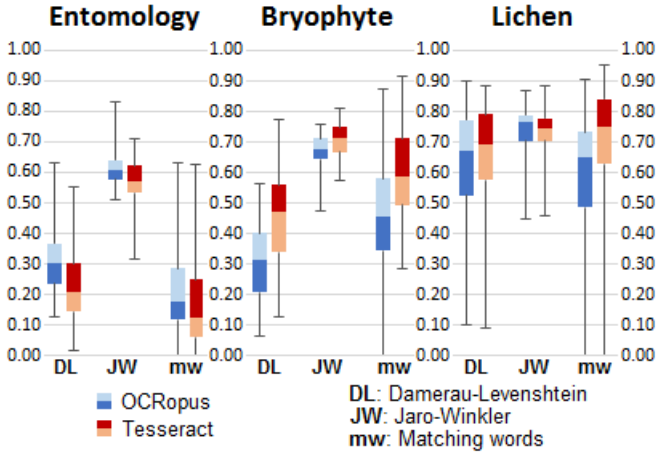


Figure 5. Box and whisker representation of the obtained similarity per specimen type, metric, and OCR tool.

Lichen images have some attributes which make their text easier to be recognized than the other specimen types, but not all the images have the same quality. One of the images with poor conditions is lichen TENN-L-0000003, see Figure 6, for which the DL similarity was 0.1 in OCRopus and 0.15 in Tesseract; and none OCR tool was able to match a single word. The low contrast of the image makes it difficult for the OCR tools generate good results. Increasing the contrast would improve these cases, but another machine process would be needed to auto-select the level of contrast, as elevating contrast for all will cause other labels to decrease their recognition rate.

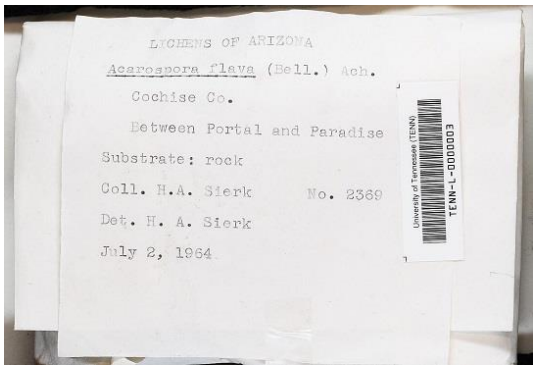


Figure 6. Lichen TEN-L0000003

### C. *Approach 2 (Cooperative – Crop and OCR label)*

Using the HuMaIN interface [32], the label of the 400 images were cropped by two volunteers. The result of this process is a rectangular image with the text in it. In the case of entomology images, there are several pieces of paper with data, hence the rectangle includes all these areas. In the case of bryophyte images, the final cropped label may include pieces of other elements, which reverts to Approach 1 in these cases. Figure 7 shows the cropped version of the images in Figure 1.


Figure 7. Cropped labels of images in Figure 1

The HuMaIN's app, randomly picks, the next image to be cropped. The coordinates of each cropped label are stored in a database, as well as the time the user spent in the process. Every label was cropped at least three times by the volunteers. The consensus criteria to select the coordinates (image) to process with the OCR was choosing the image with the largest area.

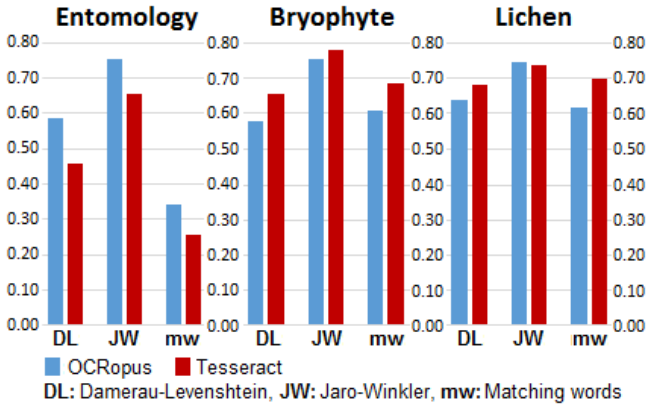


Figure 8. Average similarity for cropped label images

Figure 8 shows the DL, JW, and mw similarity values for the cropped entomology, bryophyte, and lichen labels, OCRed with OCRopus and Tesseract.

Figure 9 exhibits the absolute similarity variation when executing the OCR on the original images (Approach 1) vs. the cropped labels. For lichen images, there was no significant variation because the cropped labels are very similar to the original images. Lichen images are basically a cropped label.

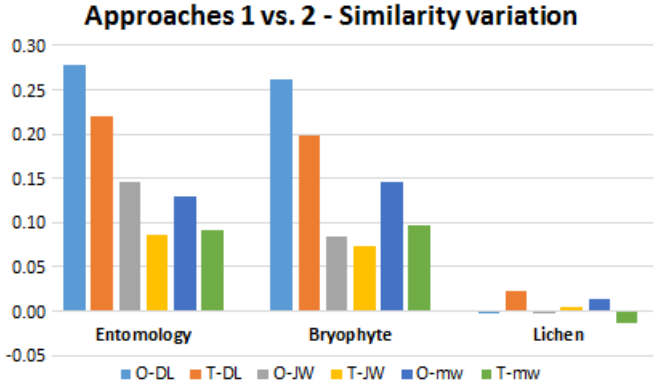


Figure 9. Average similarity difference between approaches 1 and 2

In the case of entomology and bryophyte images, there was a clear improvement. The similarity of the OCR generated text with respect to the experts' transcription improved about 0.22 considering Damerau-Levenshtein, 0.08 for Jaro-Winkler, and 0.12 considering the matching words similarity metric. The DL metric showed a higher improvement than the JW metric, likely because JW similarity values were already high for Approach 1.

Figure 10 shows the cropped text area of specimen EMEC 609,705 after being binarized, and the final OCRopus result. In comparison to Figure 4, we can observe that the initial undesired characters were eliminated and a pair of zones (around "Navajo" and the first appearance of "J.M. Davidson") were better interpreted.

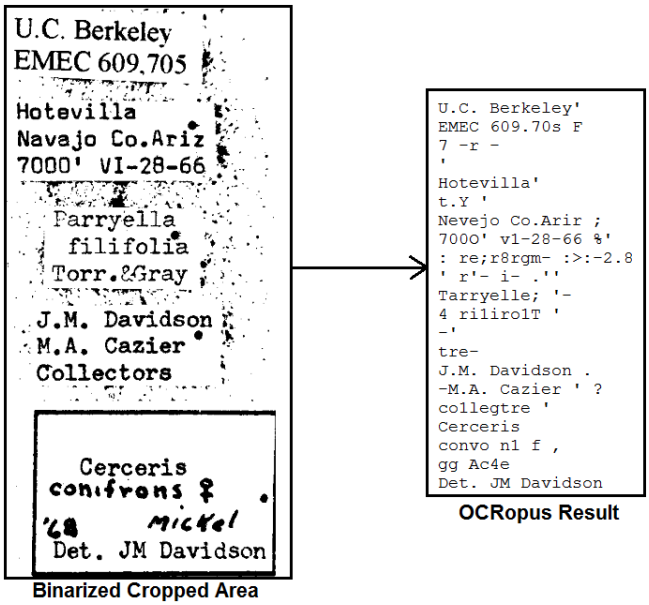


Figure 10. Result of the binarization (left) and interpreted text (right) of the cropped area of specimen EMEC 609,705

In the box chart of Approach 2 (Figure 11) it is remarkable how the matching word metric for some bryophyte images raised to 100%. This increment together with the average results for approach 2 shown in Figure 8, confirm that cropping the text area (made by humans) before running the OCR, improved the quality of the OCR output (a machine-intelligent process).

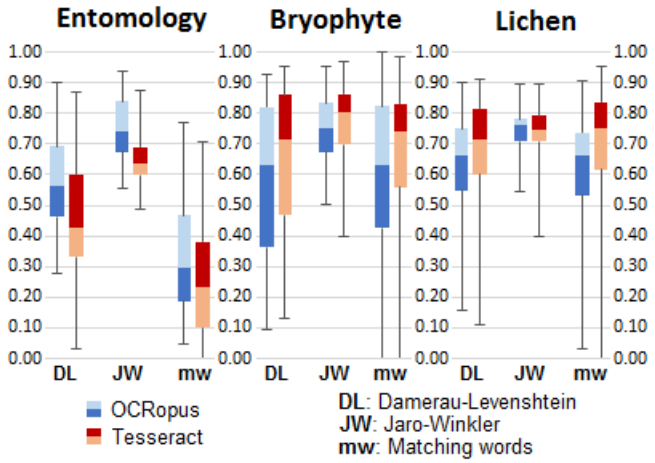


Figure 11. Similarity of the cropped labels per specimen type, metric, and OCR tool.

Comparing figures 11 and 5, we also observe that the variability increased. Although the average OCR's recognition rate improved, there are images which OCR process did not benefit from processing a smaller version of them.

In Approach 1 the total time is equal to the OCR execution time. For Approach 2 (see Table 5), we include the time users spent cropping the image (O: OCRopus, T: Tesseract). This cropping time was measured as the interval between the web page loads and the user pressing the "Save and Next" button. In-between these events, the image loads, the user interprets the image, marks the area to crop, and clicks the "take coordinates" (green arrow) button, see Figure 3.

Table 5. Approach 2 - Average execution time (s)

| | Execution time (s) | | | | |
|---|---|---|---|---|---|
| **Type \ Tool** | Cropping | Ocropus | Tesser. | Tot. O | Tot. T. |
| **Entomology** | 15.36 | 15.65 | 2.47 | 31.01 | 17.83 |
| **Bryophyte** | 24.56 | 32.74 | 1.68 | 57.30 | 26.24 |
| **Lichen** | 15.13 | 25.52 | 1.82 | 40.65 | 16.95 |

For entomology and lichen images the average cropping time was 15 seconds, but for bryophytes the process was more complex, and took on average 25 seconds. Considering only the OCR execution time, Approach 2 was on average 2.62 and 1.74 times faster than Approach 1, for OCRopus and Tesseract respectively. Considering the total time, Approach 2 was 1.48 times faster with OCRopus and 6.48 times slower with Tesseract, with respect to Approach 1.

### *D. Approach 3 (Cooperative – Crop and OCR fields)*

Using the HuMaIN web interface, 8 Darwin Core fields: scientific name, event date, latitude, longitude, identified by, country, county, and state/province were cropped and then OCRed. It is important to highlight that only 6 of these fields are not modified or interpreted. The fields Scientific name, Event date, Latitude, and Longitude were collected as dwc:verbatim [26], which means they are the original values present in the image. Additionally, the fields Identified by, and County are usually not modified or completed by the expert. On the other hand, State/Province and Country fields, use to be abbreviations which are completed or interpreted. For example, in the country field, "Mex." is interpreted as "Mexico"; while "US", "U.S.A.", and "USA" are interpreted as "United States". This affects the comparison because we are not doing these interpretations of the OCR output. As mentioned before, the objective of this study is not to develop a data extraction product as SALIX [29] or LabelX [30], but showing the benefits of the human-machine cooperation in data extraction. Some random examples of cropped fields are shown in Figure 12.


Figure 12. Examples of cropped fields

Every cropped field image was resized to 600 x 600 pixels because this is the minimum image size permitted by OCRopus. During this enlargement process, the cropped area was not changed, but the surrounding area was filled with "silver" (light gray) color. In preliminary tests, we filled the image with white background; but we found, after trying with white, silver, gray, and black colors, that silver gave the best result. In 20 test images, silver background increased the character recognition rate by about 12% (in OCRopus and Tesseract) with respect to white background. The reason for this difference is the artificial border around the cropped area created by the binarization process when white background is used. In the case of silver background, the contrast is reduced and the border disappears or is minimized, see Figure 13. The border shown around the scientific name at the left side of Figure 13 does not exist in the original image, see Figure 1 – entomology specimen. When the border is present, the recognition rate decreases.


Figure 13. Binarization result of the same cropped image, filled with white (left) and filled with silver (right)

The amount of cropped fields, per specimen type and field, is shown in the following table:

Table 6. Numbers of cropped fields per specimen type and field

| | Entomology | Bryophyte | Lichen |
|---|---|---|---|
| dwc:country | 7 | 75 | 63 |
| dwc:county | 55 | 0 | 52 |
| dwc:verbatimEventDate | 89 | 98 | 191 |
| dwc:identifiedBy | 51 | 55 | 89 |
| dwc:verbatimLatitude | 2 | 6 | 89 |
| dwc:verbatimLongitude | 2 | 8 | 97 |
| aocr:verbatimScientificName | 52 | 97 | 196 |
| dwc:stateProvince | 61 | 26 | 157 |

The "matching words" (mw) metric, used in the first 2 approaches and defined as the percentage of exact words recognized by the OCR, was not used in this case because most of the fields have only one or a very low number of words.

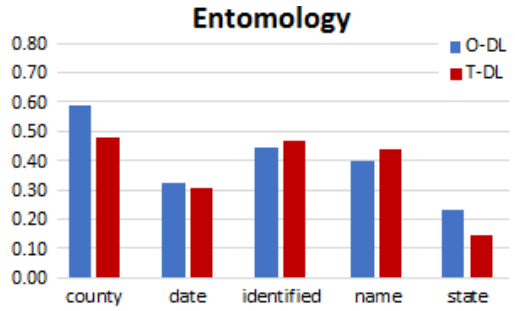


Figure 14. Average similarity for entomology fields

Figures 14, 15, and 16 present the similarity value for the entomology, bryophyte, and lichen fields. For simplicity purposes, we only show the results for the Damerau-Levenshtein similarity metric in OCRopus (O-DL) and Tesseract (T-DL).

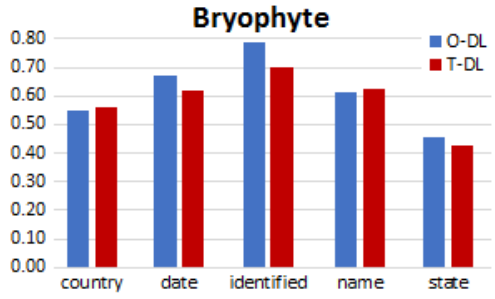


Figure 15. Average similarity for bryophyte fields

Figures 14 and 15 omit the fields for which we collected less than 10 values per specimen type, see Table 6. Entomology images show the worst results. Each field has its own challenges:

- The degree symbol in Latitude and Longitude is usually not recognized. The slash and hyphen symbols of the Event date also confuse the interpreters.
- Some Scientific names are underlined, which mixes with the letters and confuses the OCRs.
- State/Province and Country are sometimes abbreviated in the images and completed in the experts' results. Hence, the value for these fields do not really represent the quality of the OCR's output.

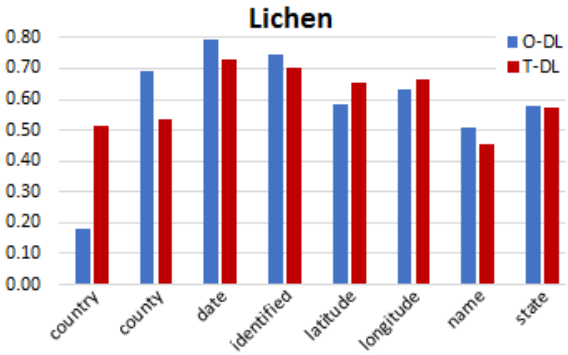


Figure 16. Average similarity for lichen fields

On average, "Identified by" was the easiest field to interpret; while Country and State fields, because they can be abbreviated, present the worst similarity. Lichen fields obtained a higher similarity than entomology and bryophyte fields. Table 7 shows the average Damerau-Levenshtein similarity per specimen type in each of the 3 approaches. Fields State/Province and Country were not considered to calculate the averages for Approach 3.

Table 7. Average DL similarity by approach

| | Entomology | Bryophyte | Lichen |
|---|---|---|---|
| **A1: whole image** | 0.31 | 0.31 | 0.64 |
| **A2: cropped label** | **0.59** | 0.58 | 0.64 |
| **A3: cropped field** | 0.44 | **0.69** | **0.66** |

Bryophyte and Lichen information was better interpreted in Approach 3 (Cropped fields), while entomology text was better recognized in Approach 2 (Cropped label). The result was not absolute in terms of better similarity for Approach 3, but data consistently show that when unnecessary areas are discarded, the OCR generates a better result.

Considering only entomology and bryophyte images, (since lichens images are already cropped labels), the similarity improvement when humans cropped the label or the fields, with respect to approach 1, was on average 0.27. Table 8 shows the percentage improvement with respect to the machine-only approach, obtained by the two cooperative approaches.

Table 8. Avg. DL improvement of A2 and A3 with respect to A1

| | Entomology | Bryophyte |
|---|---|---|
| **A2 vs. A1.** | 90% | 87% |
| **A3 vs. A1** | 42% | 123% |

The machine-intelligent process performed by the OCR was improved when a human-intelligent process (cropping the text areas), was added to the processing workflow.

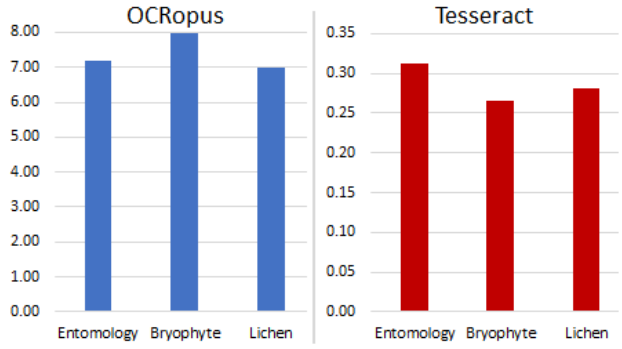


Figure 17. OCR Execution time (s) by tool and specimen type

Figure 17 shows that in OCRopus, bryophyte images are the slowest in being processed, while for Tesseract those are precisely the fastest to process. It is important to note that the scales are different. On average, Tesseract spends 0.29 sec processing a field image, while OCRopus takes 7.39 sec, which is about 25 times slower than Tesseract.

### *E. Approach 4 (Machine-only - Data cleaning)*

In Approach 1, we executed the OCR process and observed that the results include unexpected characters, which could affect the quality of the similarity values. We developed a very simple and fast filtering script which omits the words (sequence of characters) that contain symbols which are uncommon punctuation characters [22].

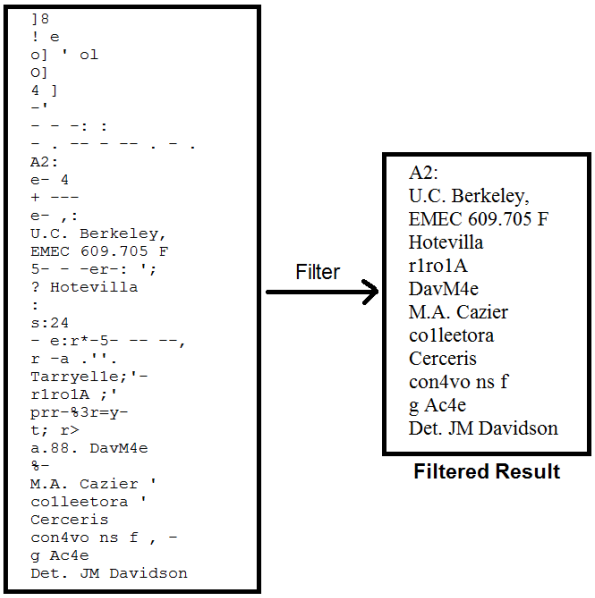


Figure 18. Data cleaning process: source (left), result (right)

Our objective was to simplify the output to improve the readability and verifying if the similarity values get better. Figure 18 shows the output after cleaning the OCRopus result of our example entomology specimen 609,705.

Table 9. Similarity variation when cleaning the Approach 1's output

| | Damerau-L. | | Jaro-Winkler | | Matching w. | |
|---|---|---|---|---|---|---|
| | Ocro. | Tess. | Ocro. | Tess. | Ocro. | Tess. |
| **Entomology** | 0.00 | 0.00 | 0.01 | 0.00 | -0.01 | -0.01 |
| **Bryophyte** | 0.08 | 0.01 | 0.02 | -0.01 | -0.04 | -0.02 |
| **Lichen** | -0.04 | -0.02 | -0.01 | 0.01 | -0.03 | -0.05 |

The cleaning script was executed on OCRopus and Tesseract results of Approach 1. The similarity values were recomputed for these outputs and we compared the obtained results with the Approach 1's similarity values. Table 9 shows the variation in similarity for each OCR tool, metric, and specimen type. The cleaning process did not improve the similarity, therefore the comparison of Approaches 2 and 3 with respect to Approach 1 is fair even after cleaning its output. Despite not improving similarity, the cleaning process reduced the output size and simplified the result, which is a positive behavior.

### *F. Time, Cost, and Scalability*

Consider 1 billion of scientific images to be processed, a single server with a Total Cost of Ownership (TCO) of $3000 per year [35], a single person with a base salary of $10/ hour, an OCR with the average behavior between Tesseract and OCRopus, and the following 4 approaches:

0) Human-only: where the average transcription time is about 9 minutes [7]. Instead of DL similarity, the percentage of tasks where consensus is reached was used [7].

1) Machine-only: where the average OCR time is 37.9 sec (OCRopus and Tesseract average), see Table 4.

2) Cooperative-Crop Label: where the average crowdsourcing (cropping) time is 18.3 sec (see Table 5), and the average OCR time is 13.3 sec.

3) Cooperative–Crop fields: where 10 fields are transcribed per specimen, the average cropping time is 20 sec per field (according to our experiments with volunteers), and the average OCR time is 3.84 sec per field.

Time and cost can be estimated as summarized in Table 10. In general, adding human effort increases execution time and cost, but improves the quality of the result. Note that even though the total time in years is provided, both machine and human efforts are fully parallelizable.

Table 10. Time, Cost, and DL Similarity per Approach

| Appr. | Human+Machine (Time in years) | Cost ($ in Millions) | DL similarity |
|---|---|---|---|
| **0** | 17123 – 0 (17123) | 1500.00 | 0.79 [7] |
| **1** | 0 – 1202 (1202) | 3.61 | 0.42 (Table 7) |
| **2** | 580 – 422 (1002) | 52.10 | 0.60 (Table 7) |
| **3** | 6342 – 1218 (7560) | 559.21 | 0.60 (Table 7) |

The machine-only option is the cheapest option, but generates the worst output quality. On the other hand, the human-only option is the most expensive and the most accurate. The hybrid approaches balance these two extreme cases. When adding the most trivial human work (cropping whole labels), the cost increases, the required overall time is actually reduced and the quality improves. Cropping fields, requires detecting the different fields, increasing the time and cost while maintaining quality when compared to the label-cropping hybrid approach.

In this cost evaluation, we considered that workers are compensated. However, if the crowdsourcing task can be made entertaining (e.g., as an application that museum visitors could use while interacting with the items in displays, or as an online competition game), where volunteers would be willing to contribute their work, then the cost would be drastically reduced.

## V. CONCLUSIONS

In this work, we demonstrated that a single workflow with cooperation of human- and machine-intelligent processes led to improved quality, while not sacrificing significant time, when compared to a machine-only workflow. Even though we did not explicitly compare to human-only workflows, related work [7] has shown that human-only workflows demand user training, are time intensive (require multiple users to perform the same task), quality is not perfect, and require solutions to deal with variations in human opinion, bias, and error.

Improvements in output quality were assessed for two workflows with machine and human processes that require minimal user training to generate segments with text from the image: whole labels and individual parsed fields. These workflows were compared to a typical machine-only workflow and an improved machine-only workflow that implicitly removes noise from the output. The quality of the hybrid workflow was at least 42% superior. A secondary future goal of collecting text region information, is to investigate the ability to train a machine-learning model to look for and find regions with this characteristic. During segmentation, OCRopus uses different heuristics to find region of text and eliminate images, but these heuristics assume publication type layout, and specimen images do not follow such a constrained format. We also experimented with other tools that can detect text on photographs. While those could detect text within an image with

texture, they failed on biological specimen images. Collecting training data for tools that make use of supervised machine-learning algorithms is time demanding, and a hybrid workflow as presented in this work can also facilitate the tuning of such machine-only workflows.

In addition to these main findings, we also provided detailed insights into the performance of OCRopus and Tesseract. Because Tesseract was on average 25 times faster than OCRopus while maintaining the quality of output, cropping the label accelerated the OCRopus execution time, but decreased Tesseract execution performance. Similarly, when considering the crowdsourcing time, the hybrid approach resulted in time efficiency gains with OCRopus, and time efficiency loss with Tesseract. Factors such as yellowed paper, underlined text, low contrast text and graphical elements that touch characters, had a higher negative impact in the character recognition rate than the resolution.

## ACKNOWLEDGEMENT

This work is supported in part by the National Science Foundation (NSF) grants No. ACI-1535086, No. EF-1115210, DBI-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] *"A Matter of Life and Death: Natural science collections: why keep them and why fund them?,"* 1st ed. UK: NatSCA, 2005.

[2] A.C. Bentley, *"Scientific Collections: Mission-Critical Infrastructure for Federal Service Agencies,"* Interagency Working Group on Scientific Collections (IW/GSC), 2009.

[3] R.S. Beaman and N. Cellinese, *"Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science,"* Zookeys, pp. 7-17, 07/09. 2012.

[4] A.H. Ariño, *"Approaches to estimating the universe of natural history collections data,"* Biodiversity Informatics; Vol 7, no 2 (2010) DO - 10.17161/bi.v7i2.3991, 10/09. 2010.

[5] V. Blagoderov, , I. Kitching, , L. Livermore, , T. Simonsen, , V. Smith. *"No specimen left behind: industrial scale digitization of natural history collections,"* ZooKeys, 209. 133–146, 2012.

[6] P. Flemons and P. Berents, *"Image based Digitisation of Entomology Collections: Leveraging volunteers to increase digitization capacity,"* No Specimen Left Behind: Mass Digitization of Natural History Collections. ZooKeys, vol. 209, pp. 203-217, 2012.

[7] A. Matsunaga, A. Mast, J.A.B. Fortes, "*Workforce-efficient consensus in crowdsourced transcription of biocollections information*," Future Generation Computer Systems, Vol 56, March 2016, 526-536, ISSN 0167-739X, http://dx.doi.org/10.1016/j.future.2015.07.004.

[8] Notes From Nature. Available: http://notesfromnature.org. [Accessed: 23- May- 2016].

[9] D. Wagner, *"Store First & Ask Questions Later,"* allanalytics.com, 2014. Available: http://www.allanalytics.com/author.asp?doc_id=275286. [Accessed: 23- May- 2016].

[10] *"Augmenting OCR,"* iDigBio Wiki, 2014. Available: https://www.idigbio.org/wiki/index.php/Augmenting_OCR. [Accessed: 22- May- 2016].

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

[12] 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.

[13] A. Ul-Hasan, S. Bukhari, and A. Dengel, "*OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters,*" 12th IAPR International Workshop on Document Analysis Systems, 2016.

[14] W.E. Moen, J. Huang, M. McCotter, A. Neill, and J. Best, *"Extraction and Parsing of Herbarium Specimen Data: Exploring the Use of the Dublin Core Application Profile Framework,"* 2010. Available: http://hdl.handle.net/2142/14920.

[15] S. Dhiman, A.J. Singh, *"Tesseract Vs GOCR A Comparative Study),"* International Journal of Recent Technology and Engineering (IJRTE); Vol 2, Issue 4, 2013.

[16] T. Breuel, A. Ul-Hasan, M. Al Azawi, and F. Safait, *"High-Performance OCR for Printed English and Fraktur using LSTM Networks,"* 12th International Conference on Document Analysis and Recognition, 2013.

[17] T. Breuel, *"ocropy"*, *GitHub*, 2010. Available: https://github.com/tmbdev/ocropy. [Accessed: 23- May- 2016].

[18] R. Smith and Z. Podobny, *"Tesseract"*, GitHub, 2007. Available: https://github.com/tesseract-ocr/tesseract. [Accessed: 23- May- 2016].

[19] R. Smith, *"An Overview of the Tesseract OCR Engine,"* in Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02, pp. 629-633, 2007.

[20] R. Minerich, *"Record Linkage Algorithms in F# – Extensions to Jaro-Winkler Distance (Part 3) « Inviting Epiphany,"* richardminerich.com, 2011. Available: http://richardminerich.com/2011/09/record-linkage-algorithms-in-f-extensions-to-jaro-winkler-distance-part-3/. [Accessed: 05- May- 2016].

[21] *"Optical character recognition,"* Wikipedia, 2016. Available: https://en.wikipedia.org/wiki/Optical_character_recognition. [Accessed: 04- May- 2016].

[22] "Collaborative Data Extraction Scripts," Available: https://github.com/acislab/HuMaIN_Collaborative_Data_Extraction. [Accessed: 19- August- 2016].

[23] J. Turk and M. Stephens, *"jellyfish 0.5.3,"* pypi.python.org, 2016. Available: https://pypi.python.org/pypi/jellyfish. [Accessed: 04- May- 2016].

[24] *"Jaro–Winkler distance"*, Wikipedia, 2016. Available: https://en.wikipedia.org/wiki/Jaro–Winkler_distance. [Accessed: 04- May- 2016].

[25] *"Darwin Core"*, rs.tdwg.org, 2015. Available: http://rs.tdwg.org/dwc/. [Accessed: 24- May- 2016].

[26] *"Darwin Core Terms: A quick reference guide,"* rs.tdwg.org, 2015. Available: http://rs.tdwg.org/dwc/terms/. [Accessed: 24- May- 2016].

[27] *"label-data"*, https://github.com/idigbio-aocr/label-data. [Accessed: 24- May- 2016].

[28] J. Leskovec, A. Rajaraman, and J. D. Ullman, *"Mining of Massive Datasets"*, 2nd. ed., 2014, p. 74.

[29] L. Lafferty, *"SALIX 2,"* 2013. Available: https://www.idigbio.org/content/idigbio-hackathon-salix-2. [Accessed: 24- May- 2016].

[30] B. Heidorn. *"LABELX. Label Annotation Through Biodiversity Enhanced Learning,"* 2014. Available: http://github.com/BryanHeidorn/LABELX. [Accessed: 24- May- 2016].

[31] W. Gomaa and A. Fahmy, *"A survey of text similarity approaches. International Journal of Computer Applications,"* 2013, 68(13) doi:http://dx.doi.org/10.5120/11638-7118.

[32] *"HuMaIN: Human and Machine Intelligent Network,"* Advanced Computing and Information Systems (ACIS) laboratory. Available: http://humain.acis.ufl.edu. [Accessed: 24- May- 2016].

[33] G. Fairchild, *"pyxDamerauLevenshtein,"* GitHub, 2015. Available: https://github.com/gfairchild/pyxDamerauLevenshtein. [Accessed: 04- May- 2016].

[34] "Symbiota Introduction," Available: http://symbiota.org/docs/. [Accessed: 24- May- 2016].

[35] Barroso, L. A., Clidaras, J., & Hölzle, U., "*The datacenter as a computer: An introduction to the design of warehouse-scale machines*". Synthesis lectures on computer architecture, 8(3), pp. 1-154, 2013.