Task Design and Crowd Sentiment in Biocollections Information Extraction

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

***Abstract*—Citizen science projects have successfully taken advantage of volunteers to unlock scientific information contained in images. Crowds extract scientific data by completing different types of activities: transcribing text, selecting values from pre-defined options, reading data aloud, or pointing and clicking at graphical elements. While designing crowdsourcing tasks, selecting the best form of input and task granularity is essential for keeping the volunteers engaged and maximizing the quality of the results. In the context of biocollections information extraction, this study compares three interface actions (transcribe, select, and crop) and tasks of different levels of granularity (single field vs. compound tasks). Using 30 crowdsourcing experiments and two different populations, these interface alternatives are evaluated in terms of speed, quality, perceived difficulty and enjoyability. The results show that Selection and Transcription tasks generate high quality output, but they are perceived as boring. Conversely, Cropping tasks, and arguably graphical tasks in general, are more enjoyable, but their output quality depend on additional machine-oriented processing. When the text to be extracted is longer than two or three words, Transcription is slower than Selection and Cropping. When using compound tasks, the overall time required for the crowdsourcing experiment is considerably shorter than using single field tasks, but they are perceived as more difficult. When using single field tasks, both the quality of the output and the amount of identified data are slightly higher compared to compound tasks, but they are perceived by the crowd as less entertaining.**

Keywords—crowdsourcing, crowdsourcing interface, task complexity, crowd sentiment.

I. INTRODUCTION

Di-Palantino and Vojnovic define "crowdsourcing" as "[a set of] methods of soliciting solutions to tasks via open calls to large communities" [8]. Utilizing computing as instrument, crowdsourcing has been used as a form of collaborative computing that enables large numbers of individuals to use computer-based interfaces and networks to accomplish challenging tasks. Crowdsourcing has long been recognized as an enabler of massive Information Extraction (IE) for scientific use. Ten years ago, one of the pioneers and most visible scientific crowdsourcing projects – Galaxy Zoo [15] – used a custom web site to ask volunteers to classify galaxies based on their morphology.

Similar crowdsourcing projects are ongoing in important institutions such as the Smithsonian [23] and NASA [19]. Initiatives as Zooniverse [32] and SciStarter [21] serve as citizen science portals and platforms for tens of scientific crowdsourcing projects. Crowdsourcing applications, specifically developed for biocollections' data extraction, include Symbiota [24], Atlas of Living Australia's DigiVol [3] and Les Herbonauts [20].

In comparison to general purpose crowdsourcing, which typically consists of straightforward tasks, scientific crowdsourcing usually requires domain knowledge, and expertise that needs to be given to volunteers through training (e.g., understanding what qualifies as a scientific name, knowledge about the shape of neurons and dendrites, and ability to distinguish various types of galaxies).

There are four main factors that contribute to the quality of the crowdsourcing results: task definition, user interface, granularity, and compensation policy [1]. This paper studies how user interface input options and task granularity affect the duration of crowdsourced IE, the quality of its result, and the sentiment of the participants.

This study considers three types of interface interactions for scientific information extraction from images, according to their associated actions:

- *Transcription*: typing via the keyboard the exact value (or an implicit value) found in images. An implicit value is not explicitly present in the image, but it can be derived from associated information present in the image.
- *Selection*: pre-defined values are selected with the use of radio buttons, check-boxes or drop-down lists. The need to define the values available to the user limits the use of this interface to data for which curated dictionaries exist. The advantage of this interface is that output is implicitly validated.
- *Cropping*: using a point and click device, the user graphically selects an area of the image containing the expected value. This region of the image is further processed by an Optical Character Recognition (OCR) application to generate the final result.

This paper studies the impact of crowdsourcing task design on the output quality, processing time, and crowd engagement in the context of biocollections information extraction. To this end, crowdsourcing tasks that support the above described three types of interactions were developed. Some of them require the

0-7695-6303-1/17/31.00 ©2017 IEEE
DOI 10.1109/CIC.2017.00056

extraction of one single value, while others are compound, asking for several pieces of information in the same task. Among these interface alternatives, we wondered: what do crowds prefer? and, do any of these options produce results of better quality or at a higher rate than the others?

Towards answering the above questions, a set of 30 crowdsourcing tasks was prepared, covering three common user actions utilized in scientific data extraction: typing, selecting from drop-down lists, and cropping parts of an image. Users were presented images, exemplified in Figure 1, coming from three biological collections, and were asked to extract Darwin Core terms found in these images. The number of terms to extract was also varied among different crowdsourcing tasks. The information extracted by the users was compared to experts' transcriptions for the same dataset and utilized to study the effects of different interface options and task complexities in the processing time, quality of the results, and user sentiment (i.e. enjoyability and difficulty as perceived by the users).

Figure 1. Specimen EMEC 609,651 of the insects (Entomology) collection. Image extracted from the data set[1] prepared by the A-OCR group of iDigBio.

The crowdsourcing tasks were implemented in a website [25] that users utilized to access the images and insert their correspondent information. The *Transcription* tasks were also implemented on the Zooniverse platform, for which a project called HuMaIN was created [7].

On average, obtained results show that users perceive these scientific crowdsourcing activities as "Slightly easy", but also "Slightly boring" (terms used in the surveys). The study demonstrates that compound tasks, which are considered by the crowd as more complex than single field tasks, require less time per extracted value, but generate results of slightly lower quality.

Transcription is relatively quick for small values, but *Cropping* could be a better alternative for long values, especially when the image quality allows the OCR software to get an acceptable recognition rate. *Selection* is the fastest interface and provides the highest output quality among the three interface options, but it is not always an alternative as previously discussed.

II. Related Work

Cheng et al. [5] compare microtasks and macrotasks in crowdsourcing, and their results are consistent with the related results in this paper: transcription macrotasks save time but produce results with lower quality than microtasks. Tasks utilized in their experiments (adding receipt costs, sorting numbers, and transcribing audio) are common tasks in general purpose crowdsourcing platforms, and typing is the only tested interface. Experiments in this paper use three types of interfaces and add sentiment analysis, covering two types of crowds (on-site paid participants and Zooniverse volunteers), and consider the learning process and their effects in extraction time and quality.

The work of Finnerty [10] was one of the inspirations to make this research. It made very clear the importance of motivating workers to get high quality results, but we believe sustaining a whole study on a single phrase experiment is insufficient. Moreover, the notion of complexity utilized was not realistic and made artificially complex the IE task.

This paper builds on the results presented in [2], which studied the advantages of hybrid (Human- and Machine-Intelligent) workflows for scientific data extraction. Results demonstrated that the mix of human and machine processes has advantages in data extraction time and quality over a machine-only workflow. This paper focuses on improvements to human processes by comparing different interfaces, levels of task complexity, and their effect on crowd sentiment.

Vaish [28] presented benefits of microtasks, and argued that it is more productive when users participate in very short periods of time. However, scientific and biocollections crowdsourcing may require the understanding of complex concepts, which has an associated learning curve that makes user participation in short periods of time ineffective. Utilization of crowd engagement techniques, such as gratitude messages [14], can be important in scientific crowdsourcing to keep volunteers working the maximum amount of time possible.

The seek for better engagement of the crowd and improvements to enjoyability of tasks led to the development of products such as Tomnod [27] and fold.it [11]; studies on gamification [18]; and financial reward. While good solutions to engage the public are present in these researches, the goal of this paper is to evaluate the most common biocollections crowdsourcing interface options, regarding the impact on task processing time, output quality, and the volunteers' sentiment.

The effect of different interfaces and task complexity on information extraction performance is illustrated in [29], which studied the impact of pairwise and multi-item tasks in the context of Crowd Entity Resolution. The study shows that clusters of similar images can be grouped efficiently when using multi-item tasks. However, some pairs with high level of difficulty are best resolved with pairwise tasks. The proposed hybrid approach, called Waldo, selects an optimal set of multi-item and pairwise tasks to best utilize the available resources. In this paper, a different set of interfaces in the context of biocollections IE is the subject of study.

[1] https://github.com/idigbio-aocr/label-data

In Human-Computer Interaction (HCI), [9] and [17] study geometrical factors and interface objects in task efficiency, but their notion of task is more minimalistic: pointing, selecting, or any other graphical activity to interact with the application. It is not the same task efficiency that we cover in this paper, where we deal with the quality and duration of the IE process. Given the type of activity to register the information, the IE process may require several clicks, dragging, typing, and other activities that are otherwise studied in isolation.

One of the topics covered in this paper is Crowd Sentiment from the perspective of how users feel about different IE interfaces and tasks, and seeking to find what users like and dislike in order to improve interface designs. We are distant from common researches about encouraging participation by gamification of tasks [18], competitiveness, rewards [30], and other incentives [22], which could even create a biased population in the experiments [13]. There are studies that focus on improving output's quality, with solutions based on increasing the amount of crowd work [31] [16], but the fact that it is not always easy to convince a crowd to participate in IE (and probably for free or receiving a low payment) is neglected. In a human-oriented activity as crowdsourcing, we want to know how people feel about tasks to convince them to participate, keeping them for the longest time possible, and getting from them the best attitude to complete the experiments with a high quality in the results.

III. Experimental Design

A. Dataset

Between 2011 and 2014, the Augmenting OCR Working Group (A-OCR) [4] of the iDigBio project[2], organized several events to generate content and tools for the scientific digitization community. Among their results, there is a data set of 400 images of cataloged specimens[3], which have textual content transcribed by domain experts and parsed into individual fields. These images are divided in three collections: Entomology, Herbs, and Lichens, of 100, 100, and 200 images respectively.

For this study, we randomly selected 100 of these images: 34 insects (entomology), 33 herbs, and 33 lichens; and improved the experts' transcription by adding some verbatim (literal) columns and completing some missing values. The resulting "gold" data, along with scripts utilized in this study, are available at [26].

B. Fields

The metadata to be extracted from the images consists of twelve (12) fields: *Event date*, *Scientific name*, *Identified by*, *Country*, *State*, *County*, *Latitude*, *Longitude*, *Elevation*, *Locality*, *Habitat*, and *Recorded by*. These are Darwin-Core terms [6] commonly used in biocollections projects. Some of these terms – e.g., *Event date* and *Country* – are easy to understand by untrained crowd; but others can be confusing – e.g., *Locality*, *Habitat*, and *Identified by* – and need a description of what they are and what to transcribe. A brief explanation of each field is found in the GitHub site of this paper [26].

C. Information Extraction Platforms and Tasks

Two web platforms were utilized by the participants:

- HuMaIN (on-site) platform [25]: developed specifically for collecting the data for this research.
- Zooniverse platform [7]: The objective with this platform was to evaluate the effects of an increased number of participants and to compare the results generated by the two populations – i.e., users with on-site support and training and open platform users.

30 tasks were used throughout this study:

- *Transcription* of:
  - 12 fields: *Event date*, *Scientific name*, *Identified by*, *Country*, *State*, *County*, *Latitude*, *Longitude*, *Elevation*, *Locality*, *Habitat*, and *Recorded by*.
  - 8 fields (textual): *Scientific name*, *Identified by*, *Country*, *State*, *County*, *Locality*, *Habitat*, and *Recorded by*.
  - 4 fields (numerical): *Event date*, *Latitude*, *Longitude*, *Elevation*.
  - Each of the 12 fields, independently.
- *Selection* of:
  - *Event date*.
  - *Identified by*.
  - *Country*, *State*, and *County*.
- *Cropping* of:
  - Each of the 12 fields, independently.

In *Cropping*, the user selects the areas of the image where the value to extract is located. In a posterior offline process, the image fragments are processed by an Optical Character Recognition (OCR) software tool to generate the final output. In this study, the utilized OCR software was OCRopus, freely available at https://github.com/tmbdev/ocropy.

In the Zooniverse platform, due to its limited functionality, only the 15 transcription tasks could be implemented; while in the HuMaIN platform all the 30 tasks were available to the participants.

D. Work Sessions

For the tasks completed using the HuMaIN platform, participants were asked to follow three steps: 5-15 minutes of training, 40 minutes of work (or less, when a user is able to process all the 100 images in less time), and 5 minutes to answer the task's survey. Participants could complete a maximum of three tasks, but these were carefully distributed to (1) ensure that they do not repeat action (*Transcription*, *Selection*, *Cropping*) or field; and (2) minimize the experience or learning factor from one session to another. Each task was completed by at least three different participants.

Figure 2 presents excerpts of the *Cropping*, *Selection* and *Transcription* task interfaces. The system displays a photo of a

[2] https://www.idigbio.org/

[3] https://github.com/idigbio-aocr/label-data

specimen which contain labels with metadata. Each user has to locate and transfer, using one of these interfaces, the values correspondent to the fields requested in the task.

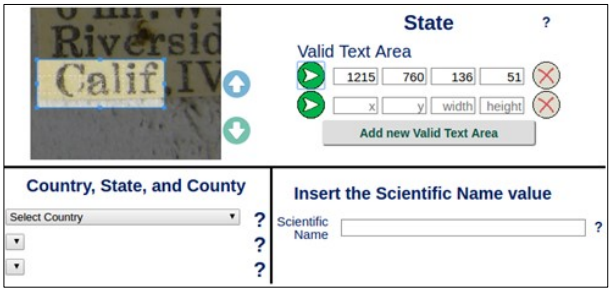

Figure 2. Excerpts of the *Cropping* (upper), *Selection* (lower left), and *Transcription* (lower right) interfaces.

Training was verbal and included general information about the project, utilization of the interface, and how to recognize the metadata in the images. A member of our team was always available to clarify questions and misunderstandings.

HuMaIN and Zooniverse sites provided the same online help, which describes the fields to be digitized and some examples. In Zooniverse, users were asked to complete a tutorial before starting to work on transcription tasks.

E. Computation of Quality

Quality estimation of *Transcription*, *Selection*, and *Cropping* tasks was obtained computing the similarity of the extracted values to the experts' transcription. In order to get the similarity, the normalized Damerau-Levenshtein (DL) algorithm was used. DL algorithm computes the distance between two strings as the minimum amount of insertions, deletions, substitutions, and transpositions of two adjacent characters, required to convert one string into the other [12].

Symbols were included in the similarity computation since, for several fields (e.g., *Elevation* and *Locality*), they are an important part of the transcription meaning. Strings were converted to lowercase before computing the similarity, which is defined as the complement of the normalized DL distance:

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

F. Categories of Extracted Values

Extracted values are categorized using the confusion matrix terminology:

- True Positive (TP) case: correctly identified value. This is a desirable case, in which experts transcribed a value for the correspondent field and the user found a value for it. The user's transcribed value may or may not exactly match the experts' (gold) value. The quality of the results is estimated using the DL algorithm.
- False Negative (FN) case: incorrectly omitted value. The user did not find in the image a value for that field, but experts did. This is considered a user miss. The quality of the results in this category is zero.
- False Positive (FP) case: incorrectly identified value. This is considered a mistake made by the user, who finds a value for a field when experts said there was no value for it. It can reflect some confusion by the user or lack of training. The quality of the results in this category is zero.
- True Negative (TN) case: correctly omitted value. Neither the user nor the expert found a value for the field. The quality of the results in this category is one.

G. Populations

Due to the utilized platforms, we can divide the participants in two populations:

Zooniverse crowd: They are volunteers from around the world who collaborate with Zooniverse in the evaluation and completion of crowdsourcing experiments. In total, 436 users participated, 284 of them (62%) were registered users. For this population, no demographic information is available.

On-site Participants: 41 people who were paid at a rate of $10 per hour to complete *Transcription*, *Selection*, and *Cropping* tasks using the HuMaIN website. Unless otherwise specified, the results provided by this group are the ones used in our analysis. This is the "by default" population because not all the tasks could be implemented using the Zooniverse project builder tool and for these users we collected demographic data and their opinion about the different tasks. The interaction and collection of the data from these participants required the approval of an Institutional Board Review (IRB) project.

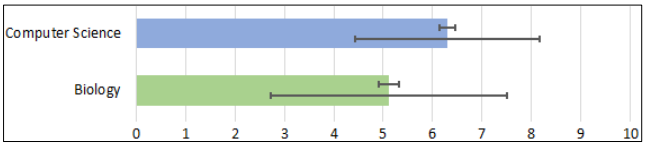

Figure 3. Average volunteers' perceived knowledge. Standard Error of the Mean (upper bar) and Standard deviation (lower bar).

When asked about their Computer Science and Biology knowledge, participants considered that their background was average or slightly over the average, see Figure 3. In general, they believe to have a better understanding on Computer Science than Biology.

IV. Results

A. Information Extraction Quality

1) Quality by Interface Type and Field:

Figure 4 shows the similarity of the values extracted by the participants to the experts' transcription. For each field, the resulting similarity of the different interface types are plotted. Similarity values go from 0 to 1, one (1) represents the similarity of two identical strings.

Only five of the twelve fields were extracted using a *Selection* interface. In four of the five fields, *Selection* generated a result of higher quality than *Transcription*. With exception of *Country* field, *Transcription* interface generated better quality than the *Cropping* (and OCR) option. This is expected since *Cropping* highly depends on the quality of the images and the OCR software. The output quality generated by *Cropping* can be improved using training data tailored to the dataset and cleaning up the images before the OCR processing.

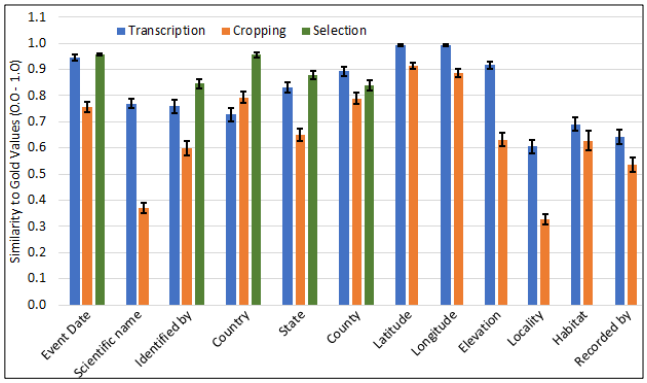

Figure 4. Average similarity of values extracted by volunteers to the golden values for each term and interface type.

Specifically for *Country*, the quality of the *Transcription* interface was negatively affected by two users who inferred the country value by other geographical information available in the image. When compared to the non-existent country value of the experts, the assigned similarity was 0 - i.e., FP error.

2) Compound vs. Simple Transcription Tasks:

Figure 5 shows the similarity of the transcribed values to the gold (experts) data, for the compound 12 Fields transcription task compared to the twelve tasks of 1 single field. In most of the fields, the similarity is higher when values are obtained from single field tasks.

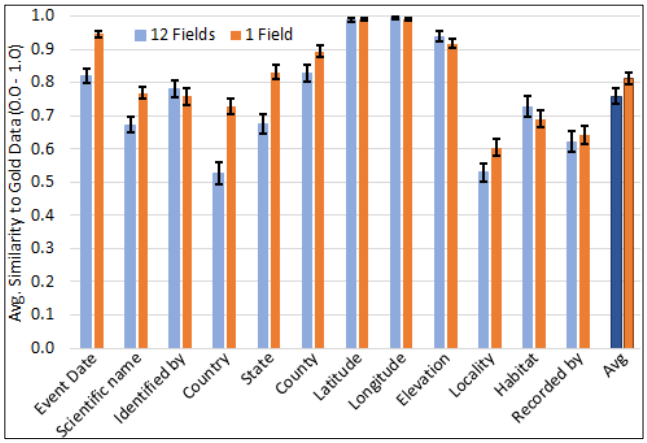

Figure 5. Damerau-Levenshtein similarity for compound and single field tasks.

For half of the twelve fields, single field tasks clearly generated results of better quality, while for the other half, compound and single field tasks obtained results of similar quality. On average, considering all the fields, the similarity for the compound task was 0.759, while the single field tasks generated an overall similarity of 0.814. This indicates that dividing the information extraction process in smaller tasks improved the overall quality of the result by 7.25%.

Figure 6 shows the breakdown of extracted values, organized in categories using the confusion matrix terminology. Single field tasks, when compared to the compound 12 fields task, generated results with less negative cases (FN and FP), more positive cases (TP), higher similarity values, and higher number of output identical to golden values.

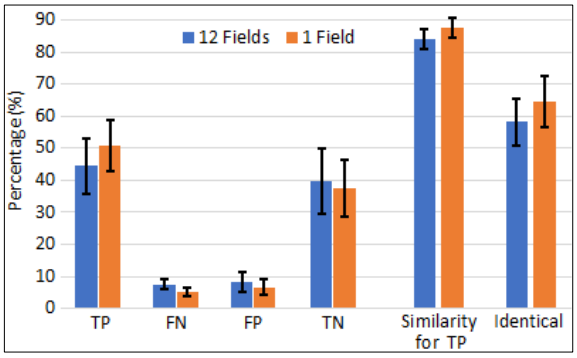

Figure 6. Breakdown of extracted values subdivided into confusion matrix categories obtained using the HuMaIN platform. Blue bars represent results for the compound transcription task (12 Fields) while orange bars representresults for the simple transcription tasks.

The results of graphs 5 and 6 reflect that, when extracting information from biocollections, single field tasks can generate results with quality that is slightly better than multiple field (compound) tasks.

3) Textual vs. Numerical Fields:

Figure 5 indicates that numerical fields (*Event date*, *Latitude*, *Longitude*, and *Elevation*), which tend to have short values, generate results with the highest quality. On the other hand, longer value fields, e.g., *Habitat* and *Locality*, generated results with quality lower than the overall average.

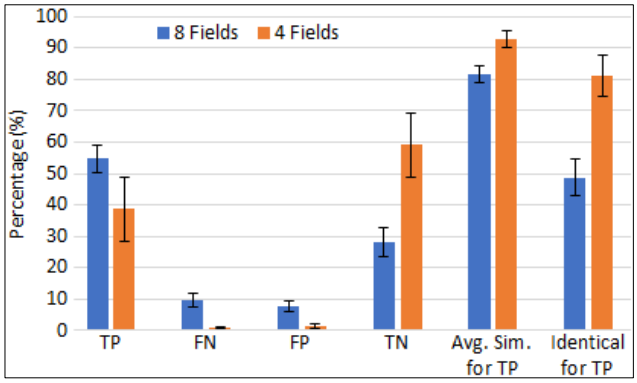

Figure 7. Breakdown of extracted values categorized into confusion matrix categories. Blue bars represent results for textual (8 fields) tasks while orange bars represent results for numerical (4 fields) tasks.

Two additional transcription tasks were completed using the HuMaIN and Zooniverse (detailed in Section IV.A.6) platforms. In the first one, eight (8) textual fields: *Scientific name*, *Identified by*, *Country*, *State*, *County*, *Locality*, *Habitat*, and *Recorded by* were transcribed by participants. The remaining four (4) fields (numerical), were transcribed in another crowdsourcing task.

Figure 7 shows the quality of the output of these two information extraction processes. Textual fields generated 15% more errors (i.e., number of FNs and FPs) than numerical fields.

Numerical fields generated results with 11% higher similarity and 33% more identical values than textual fields. These facts indicate that the information extraction of numerical fields generate results with better quality than the information extraction of textual fields, and suggest that, whenever possible, programmers should break information extraction workflows into short or numerical value extraction tasks in order to get higher quality extraction results.

4) Cropping + OCR as an alternative to transcribing:

Figure 8 shows the distribution of the Damerau-Levenshtein similarity of the extracted information for *Cropping* tasks, where the selected area is processed by an OCR software to generate the correspondent text.

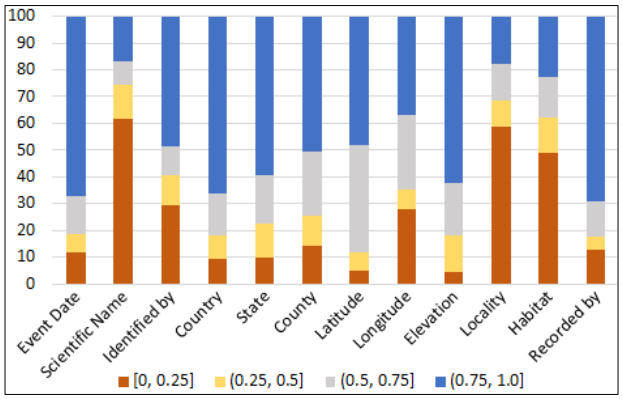

Figure 8. Similarity of the Cropping tasks by range

Values with a similarity greater than 0.75 are considered acceptable and they can usually be recognized or matched to the true value. Similarities lower than 0.5 are unnaceptable and tipically correspond to unrecoverable values. Fields with handwritten values in the images, such as *Scientific name*, *Locality*, and *Habitat*, got the worst quality in their results.

The output quality for this type of task is affected by the image quality, handwritten text, background color, text with overlapped graphical objects, skewed text, and many other conditions, which are common in scientific and historical records. Therefore, *Cropping* demands an extra effort to train the OCR and preprocessing the images in order to get a quality similar to *Transcription*.

5) Selection Interface:

Selecting a value from a dropdown list requires the normalization of the values to be presented as options. This has intrinsic advantages:

- Users can learn from the list and can verify if a candidate string is an acceptable value.
- Transcription typos and OCR errors are avoided.
- The output to be processed is already normalized, saving data cleaning and processing time.

Table I shows similarity of the extracted values, using drow-down lists for 5 terms: *Event date*, *Identified by*, *Country*, *State*, and *County*.

For the two most universally understood fields: *Event date* and *Country*, the average similarity is very high: about 0.96. The difference of about 10% when compared to other fields suggests a lack of training or understanding by the participants. Particularly for *State*, several users raised concerns about countries for which they did not know the geography and therefore, they were not able to identify the *State* in the text. The definition of *Identified by* field is domain specific and not always clearly identifiable in images, especially for non-experts.

TABLE I. QUALITY OF THE SELECTING TASKS

	Event Date	*Identified by*	*Country*	*State*	*County*
Mean	0.96	0.85	0.96	0.88	0.84
SEM	0.01	0.02	0.01	0.02	0.02
Std. Dev.	0.20	0.36	0.20	0.32	0.37

Nevertheless, as seen in Figure 4, *Selection* interface generates the highest quality among the information extraction methods studied in this paper.

6) Zooniverse:

Results of *Transcription* tasks completed using the Zooniverse platform are summarized in Figure 9.

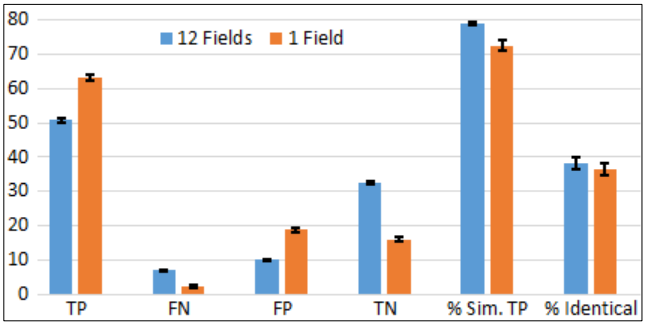

Figure 9. Breakdown of extracted values subdivided into confusion matrix categories obtained using the Zooniverse platform. Blue bars represent results for the compound transcription task (12 Fields), while orange bars represent results for the 12 simple transcription tasks.

In terms of output quality, results obtained the Zooniverse platform was not as conclusive as in the HuMaIN platform (see Figure 6), where the values generated by single field tasks showed a better quality than the data generated by the compoud task. In the Zooniverse platform, while the number of TPs were higher than observed in the HuMaIN platform, the output quality was not as good, especially for single field tasks. Moreover, the error conditions (FN and FP) show mixed results for single field and compound tasks.

Overall, compared to the Zooniverse platform, the HuMaIN platform generated results with similarity for both the compound and single field tasks that is 10% higher; the

percentage of identical values is about 22% higher (see Figures 6 and 9); lower total negative cases (FN and FP); and higher total positive cases (TP and TN).

These differences can be potentially explained by lack of training of Zooniverse volunteers. Although the offered tutorial and online help were similar in both platforms, several volunteers commented in the Zooniverse' log that they started the transcription work without completing the tutorial because it looked sufficiently simple. Reading the directions was crucial for understanding the expected output format, which was verbatim (literal).

This lack of training could also have affected the comparison between the compound and the single field tasks. It is highly probable that volunteers who decided to complete the 12 Fields task, followed the tutorial and had a better understanding in comparison to users who directly started to complete single Field tasks. Single field tasks, such as *Country* or *State*, are commonly understood, but have special extraction requirements that can be only known by completing the tutorial.

Another possible cause of the difference in quality is the monetary incentive (payment) received by on-site participants, in comparison to Zooniverse users, who participated for free.

B. Information Extraction Duration

In this section, the duration of the IE process is examined by field, interface type, task complexity, and platform.

1) Duration by Interface Type and Field:

Figure 10 shows, by field, the duration of the IE process for the *Transcription* and *Cropping* interfaces.

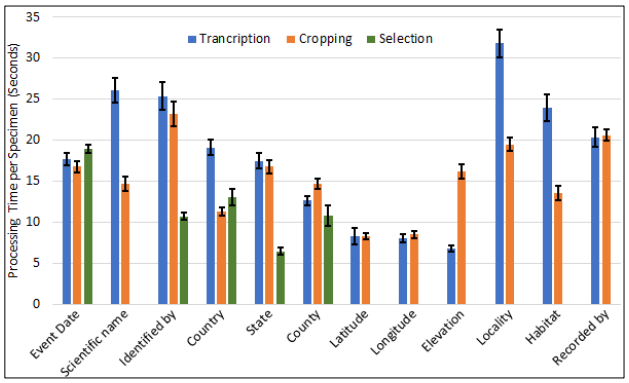

Figure 10. Average duration and error bars, in seconds, for *Transcription*, *Cropping*, and *Selection* interfaces.

Selection was faster than *Transcription* and *Cropping* in 3 of the 5 fields where it was tested, which seems to indicate that it is the fastest option. For the *Event date* field, it is understandable that *Selection* is not the fastest option because users have to make, for the common case, three selections: month, day, and year. In the case of a range of dates, there are 6 values to select.

In fields that require long text, including *Scientific name*, *Locality*, and *Habitat*; *Cropping* is faster than *Transcription*. In the rest of fields, results are mixed. *Cropping* has less variability than *Transcription* because the size of the values does not affect much the completion time.

For the field *Identified by*, *Selection* takes almost one third of the time of *Transcription* or *Cropping*, which is understandable considering that when users type the first letter of a name, the interface lists the names with that initial, thus reducing the amount of scrolling time.

Selecting from drop down lists has the advantage that, in confusing fields, a user can validate the candidate value with the listed ones. Moreover, typos are eliminated. However, *Selection* interface requires that all possible values are known, which it is not always the case. Another drawback of this interface type is that, for very large lists of values, the amount of scrolling can make the process very tedious.

The IE duration of the task for selecting the *Country*/*State*/*County* fields does not have a simple analysis. When the three values exist, users required on average 22.9 seconds per image, which could be considered better than *Transcribing* and *Cropping* (Figure 10). But when the *State* value or the *County* were not present, the duration time were higher: 52.4 and 32.3 seconds respectively. Observing the behavior of some participants, we found that when a user did not have knowledge of a country, s/he searched in the dropdown list for the candidate values to be sure that no *State* was in the image, which consumed most of the measured time.

2) Compound vs. Simple Tasks:

Figure 11 presents the average duration, per image, of the transcription tasks. In the fourth column of the graph, the average duration of the 12 single field tasks was added. The scale to the left corresponds to the total duration of the tasks (columns), while the scale to the right shows the per field duration (line). All the durations are expressed in seconds.

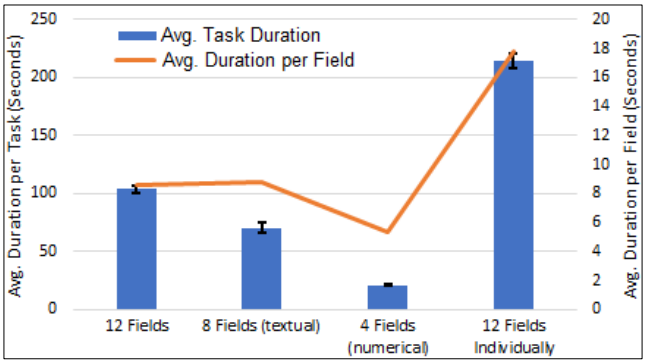

Figure 11. Average IE duration for different transcription tasks.

Separately executing the 12 single field tasks takes twice the time taken to process the 12 fields compound task (104 vs. 208 seconds). Textual fields, usually larger than numerical fields, tend to take more time to be transcribed.

3) Learning Process:

Figure 12 shows the difference between the number of correctly processed images (TP and TN) during the first and last 6 minutes of crowdsourcing.

With the exception of *Habitat*, users have a higher rate of processed images towards the end of their work session. This implies a learning process, i.e., users require some time or practice to internalize the concept, learn how to identify the value in the image and use the interface.

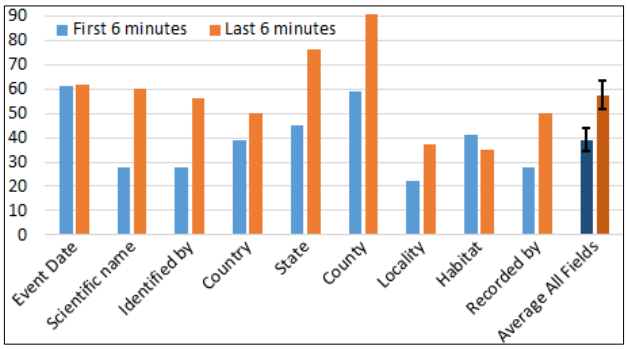

Figure 12. Number of processed images during the first and last 6 minutes of crowdsourcing (3 participants per Field).

This suggests that, in scientific crowdsourcing, platforms where users spend few minutes before quitting or changing task (microtasks) will require more time to complete the same experiment than in platforms where users stay longer working on a similar set of tasks.

However, this does not hold true for the output quality, which basically stays the same at the beginning and towards the end of the experiments, as illustrated in Figure 13.

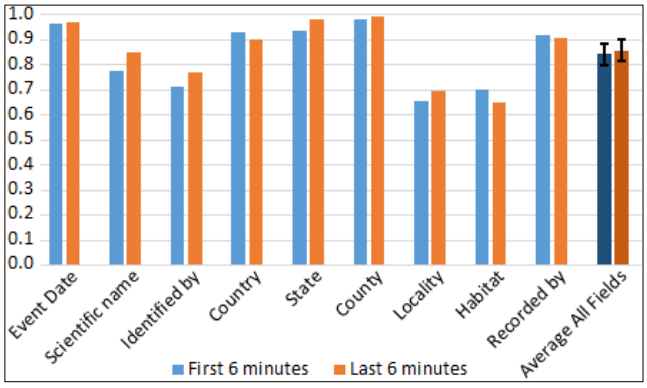

Figure 13. Average similarity during the first and last 6 minutes of crowdsourcing (3 participants per Field).

4) Zooniverse:

Figure 11 was created from the data generated by the HuMaIN platform. The same graph was computed for the Zooniverse platform, see Figure 14. Similar tendencies observed in Figure 11 are present in Figure 14: transcription of the 12 fields using one compound task takes less time than using 12 single field tasks. Two observations to highlight are:

- On average, Zooniverse users took 3 times (3x) longer compared to HuMaIN users to complete the same tasks on the same images.
- The duration per field seems to indicate that it is positive to group the fields in small groups, but not in single field tasks.

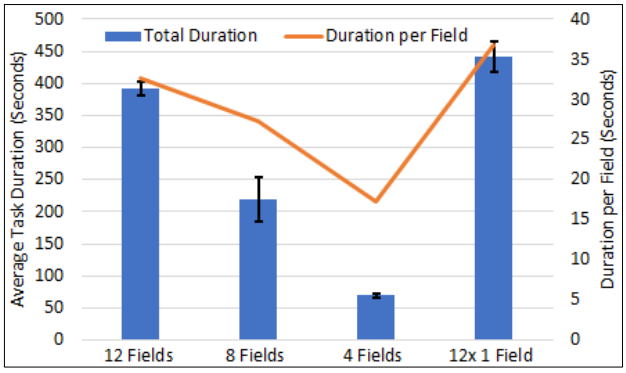

Figure 14. Average duration of transcription tasks on the Zooniverse platform.

C. Crowd Sentiment

1) About the Crowdsourcing Activity:

In general, the results of the surveys about complexity and enjoyability suggest that the experiment was considered "slightly easy" and "slightly boring" (see Table II). The complete scales and survey models can be found in the GitHub site of the study [26].

Participants considered that the received training was sufficient and they had an average biology knowledge and a slightly over the average computer science background.

TABLE II. CROWD SENTIMENT

	Easy-Difficult	Fun-Boring	Not prepared-Trained	Biology Knowledge	CS Knowledge
Median	Slightly easy 3.333	Slightly boring 6.199	Probably yes 7.500	5.500	6.500
Avg.	Slightly easy 3.624	Slightly boring 6.667	Probably yes 8.321	5.117	6.303
SEM	0.226	0.221	0.165	0.209	0.163
Std. Dev.	2.593	2.542	1.897	2.403	1.870

The feedback provided by 74 volunteers of the Zooniverse project was similar to the HuMaIN crowdsourcing experiment, with some differences in the perception of adequate training (see Table III).

TABLE III. CROWD'S OPINION ABOUT THE RECEIVED TRAINING

	Positive (~yes)	Negative (~no)	Standard Error
HuMaIN	96.67 %	3,33 %	1.64 %
Zooniverse	67.12 %	32.88 %	5.50 %

The perception of the participants whether the provided help (tutorial, directions, and online help) were adequate to complete the tasks was lower in Zooniverse than in the HuMaIN platform.

2) Perceived Easiness/Difficulty.

The perceived easiness decreases when the number of fields to transcribe grows (the task is perceived as more difficult when it includes more fields), see Figure 15.

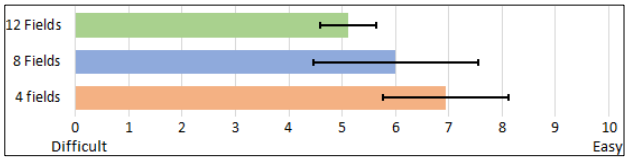

Figure 15. Average appraised easiness (Difficult) 0 – 10 (Easy), by number of fields per task

Numeric fields tend to be perceived as easier to complete than textual fields. This can be justified by the fact that, in numeric fields, the explanations about what to transcribe are simpler and values are short. *State*, an a priori simple field, turned to be the most difficult because states from other countries were involved.

Identified by and *Recorded by* are similar fields (names of people), but *Recorded by* is usually better specified or easier to identify in the image, which can be the main reason of higher appraised easiness (see Figure 16).

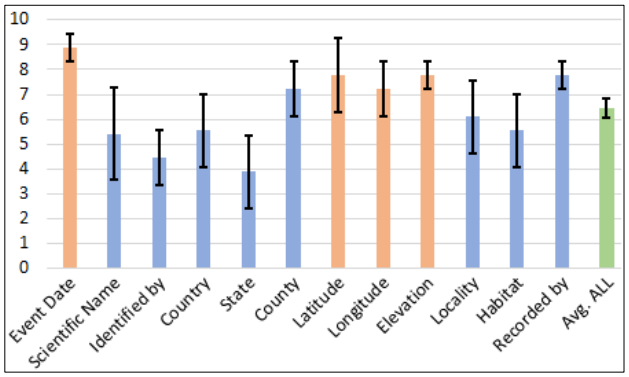

Figure 16. Average appraised easiness (Difficult) 0 – 10 (Easy).

The average perceived easiness when the 12 fields are transcribed in independent tasks (last column of Figure 16) is higher than the average perceived easiness when the 12 fields are extracted in a single compound task (top bar of Figure 15). Hence, quality is improved and perceived easiness is increased when using single field tasks.

3) Perceived Enjoyability:

It was also asked to the participants how they qualified each experiment between Boring and Fun.

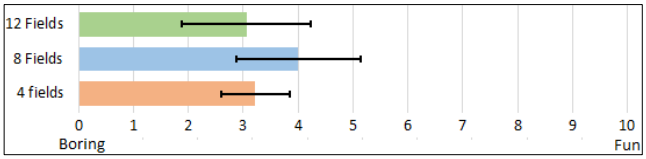

Figure 17. Average appraised enjoyability (Boring) 0 – 10 (Fun), by number of fields per task.

In general, transcription is considered boring (see Figure 17): independently of the number of fields to process, the three averages are lower than 5.

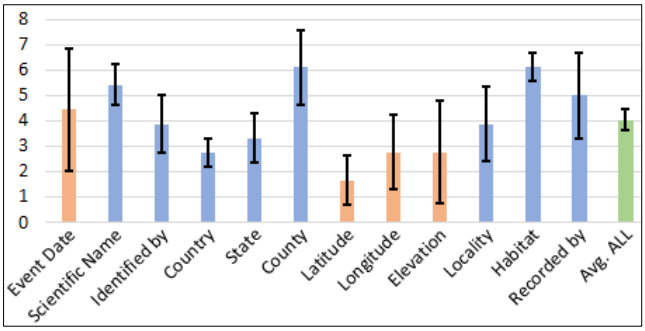

Figure 18. Average appraised enjoyability by field, (Boring) 0 – 10 (Fun).

Figure 18 shows the perceived enjoyability per field in *Transcription* tasks of one single field. Only the transcription of *County* and *Habitat* were, on average, perceived as slightly fun. Interestingly, although the transcription of numeric fields is perceived as an easier task than the transcription of textual fields, it tends to be considered more boring.

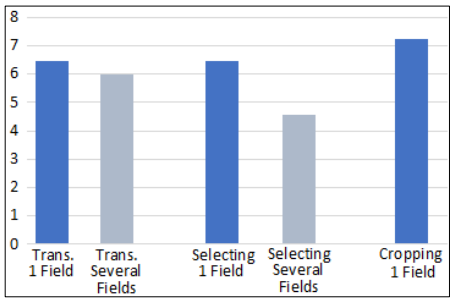

Figure 19. Average appraised easiness per task granularity (0.0-1.0)

Figure 19 compares the perceived easiness for each of the three interface types and the granularity of the tasks. *Cropping* was considered slightly easier than *Transcription* or *Selection*, independently of the skills required to control the interface. For *Transcription* and *Selection*, participants considered that it was easier to work on single field tasks than compound tasks.

CONCLUSIONS

Comparing *Transcription*, *Selection* and *Cropping* as information extraction interface alternatives:

- *Transcription* is fast for small fields. But for large text fields, it can be slow and prone to errors; alternatives as *Cropping* (and then OCR) could be considered.
- *Selecting* from dropdown lists, when it can be implemented, generates the highest quality results and is the fastest option. It is perceived as boring, hence user incentives are recommended.
- *Cropping*, and probably similar graphical activities, are perceived as fun. *Cropping* is fast for fields with large values, but its quality will depend on the image characteristics and the OCR process.

Multiple fields (compound) tasks are perceived as more enjoyable than single field tasks.

When possible, it is recommended to design the crowdsourcing tasks using short and numerical formats for the values instead of long or textual fields. This will improve the quality of the information extraction process.

Compound tasks save experimental time but can generate a lower quality result in comparison to single field tasks.

In scientific information extraction, there is a domain specific learning curve. Microtasks, where many users process few subjects, may take more time per subject than tasks where users stay a long time performing similar work. Nevertheless, the generated crowdsourced information seems to have equivalent quality in both cases.

ACKNOWLEDGMENT

We would like to express our deep gratitude to the volunteers of the Florida Natural History Museum, Zooniverse, and on-site participants for their kindness participation.

We want to especially recognize the work done by Grace Hong understanding and completing the Institutional Review Board (IRB) and payment processes in our University.

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] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. R. Motahari-Nezhad, E. Bertino, and S. Dustdar, "Quality control in crowdsourcing systems: Issues and directions," IEEE Internet Computing, vol. 17, pp. 76-81, 2013.

[2] 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.

[3] "Atlas of Living Australia," [Online]. http://www.ala.org.au/. [Accessed: July 30, 2017].

[4] "Augmenting OCR Working Group," [Online]. https://www.idigbio.org/wiki/index.php/Augmenting_OCR. [Accessed: July 30, 2017].

[5] J. Cheng, J. Teevan, S. T. Iqbal, and M. S. Bernstein, "Break it down: A comparison of macro-and microtasks," in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 4061-4064, 2015.

[6] "Darwin-Core terms," [Online]. http://tdwg.github.io/dwc/terms/. [Accessed: May 2, 2017].

[7] "Digitization of Scientific Data with Human and Machine Collaboration," [Online]. https://www.zooniverse.org/projects/ialzuru/humain. [Accessed: July 30, 2017].

[8] E. Estellés-Arolas and F. González-Ladrón-de-Guevara, "Towards an integrated crowdsourcing definition," Journal of Information Science, vol. 38, pp. 189-200, 2012.

[9] A. Feizi and C. Y. Wong, "Usability of user interface styles for learning a graphical software application," in Computer & Information Science (ICCIS), 2012 International Conference on, pp. 1089-1094, 2012.

[10] A. Finnerty, P. Kucherbaev, S. Tranquillini, and G. Convertino, "Keep it simple: Reward and task design in crowdsourcing," in Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI, p. 14, 2013.

[11] "fold it. Solve Puzzles for Science," [Online]. https://fold.it/portal/. [Accessed: July 30, 2017].

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

[13] G. Hsieh and R. Kocielnik, "You get who you pay for: The impact of incentives on participation bias," in Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pp. 823-835, 2016.

[14] A. Segal, Y. Gal, E. Kamar, E. Horvitz, A. Bowyer and G. Miller, "Intervention strategies for increasing engagement in crowdsourcing: Platform, predictions, and experiments," in Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp. 3861-3867, 2016.

[15] K. Land, A. Slosar, C. Lintott, D. Andreescu, S. Bamford, P. Murray, R. Nichol, M.J. Raddick, K. Schawinski and A. Szalay, "Galaxy Zoo: the large-scale spin statistics of spiral galaxies in the Sloan Digital Sky Survey," Monthly Notices of the Royal Astronomical Society, vol. 388, pp. 1686-1692, 2008.

[16] T. McDonnell, M. Lease, T. Elsayad and M. Kutlu, "Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments," in Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP), pp. 10, 2016.

[17] R. Michalski, J. Grobelny and W. Karwowski, "The effects of graphical interface design characteristics on human–computer interaction task efficiency," Int.J.Ind.Ergonomics, vol. 36, pp. 959-977, 2006.

[18] B. Morschheuser, J. Hamari and J. Koivisto, "Gamification in crowdsourcing: a review," in System Sciences (HICSS), 2016 49th Hawaii International Conference on, pp. 4375-4384, 2016.

[19] "NASA Mars Clickworkers," [Online]. http://nasaclickworkers.com/classic/crater-marking.html. [Accessed: May 2, 2017].

[20] G. Rouhan, S. Chagnoux, B. Dennetière, and M. Pignal, "The herbonauts website: recruiting the general public to acquire the data from herbarium labels," in Botanists of the twenty first century: roles, challenges and opportunities, UNESCO International conference, 2014.

[21] "SciStarter," [Online]. https://scistarter.com/. [Accessed: May 2, 2017].

[22] A. Singla and A. Krause, "Truthful incentives in crowdsourcing tasks using regret minimization mechanisms," in Proceedings of the 22nd international conference on World Wide Web, pp. 1167-1178, 2013.

[23] "Smithsonian Digital Volunteers: Transcription Center," [Online]. https://transcription.si.edu. [Accessed: May 2, 2017].

[24] "Symbiota," [Online]. https://sourceforge.net/projects/symbiota/. [Accessed: May 2, 2017].

[25] "Task Complexity in Crowdsouring," [Online]. http://humain.acis.ufl.edu/complexity/. [Accessed: July 31, 2017].

[26] "Task Design and Crowd Sentiment in Biocollections Information Extraction," [Online]. https://github.com/acislab/HuMaIN_Crowdsourcing_Complexity. [Accessed: July 30, 2017].

[27] "Tomnod," [Online]. http://www.tomnod.com. [Accessed: May 2, 2017].

[28] R. Vaish, K. Wyngarden, J. Chen, B. Cheung and M. S. Bernstein, "Twitch crowdsourcing: crowd contributions in short bursts of time," in Proceedings of the 32nd annual ACM conference on Human factors in computing systems, pp. 3645-3654, 2014.

[29] V. Verroios, H. Garcia-Molina and Y. Papakonstantinou, "Waldo: An Adaptive Human Interface for Crowd Entity Resolution," in Proceedings of the 2017 ACM International Conference on Management of Data, pp. 1133-1148, 2017.

[30] Y. Yang and R. Saremi, "Award vs. Worker Behaviors in Competitive Crowdsourcing Tasks," in Empirical Software Engineering and Measurement (ESEM), 2015 ACM/IEEE International Symposium on, pp. 1-10, 2015.

[31] O. F. Zaidan and C. Callison-Burch, "Crowdsourcing translation: Professional quality from non-professionals," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 1220-1229, 2011.

[32] "Zooniverse," [Online]. https://www.zooniverse.org/. [Accessed: May 2, 2017