# Application of Crowdsourcing to Generate Datasets of Public Procurement Processes

*Andrés* Rey[1*] *Luis Gomez*[1], and Angel Lozada[1]

[1]Informatics Engineering, Faculty of Engineering, Central Unit of the Valley of Cauca, Tuluá, Colombia.

**Abstract.** In order to fight against Corruption, there are currently laws of free access to public information that favor transparency in many countries. In this opening scenario, information and communication technologies play an important role in facilitating the delivery of information that is made available to citizens seeking to enable greater control over the actions of government and civil servants. Nevertheless, in many cases, the information provided is aimed at human reading and interpretation and it is not published in machine-processable formats, a condition that hinders automatic processing using state-of-the-art computational techniques. This article presents a new method of data collection based on Crowdsourcing that allows generating data sets of public procurement processes from the unstructured information published on the Web. The results obtained show that our method allows the generation of reliable public procurement data sets in a scalable way whose scope is limited by the availability of data and the transparency of the country.

## 1 Introduction

The World Bank defines corruption as the abuse of public power for private benefit, pointing out that this problem is the greatest obstacle to the economic and social development of the countries [1]. Corruption is a problem that is highly complex and difficult to deal with because its causes are related to different social, economic, legal, cultural, and psychological problems, among others. However, it is believed that regardless of the complexity of the problem of corruption, transparency is the key to face it [2]. For this reason, in many countries in the framework of what is known as Electronic Government (e-Government) and thanks to the laws of free access to public information (Freedom of Information, FOL), the information made available to the public has increased in order to improve qualitatively and quantitatively the possibilities of control over the actions of the government [3].

For effectively confronting corruption, it is required that citizens could be able to supervise government work, but for citizens to demand accountability from the government and officials for their actions they must first know what those actions are [4]. In this work, the corruption that occurs in public procurement is focused, where currently one of the difficulties in knowing or extracting knowledge about what happens in the procurement processes does not occur mainly due to lack of information, but rather because of large volumes of information on public procurement processes [5] that are provided in many cases in non-processable formats by machine, a situation that makes it difficult to process them automatically with state-of-art computational techniques.

Considering that in recent years, information and communication technologies have begun to be used to help citizens fight and prevent the phenomenon of corruption in public procurement [6] and also in the field of the Web 2.0 has provided a model for problem solving called Crowdsourcing that is used in data collection and information generation processes (establishing a relatively new approach in the process of knowledge acquisition) [7]. This article presents a new method of data collection based on Crowdsourcing that allows generating data sets of public procurement processes from the unstructured information published on the Web.

The use of crowdsourcing as an alternative to the programming of specialized software code for the extraction of data is because currently there are tasks that are much easier for humans to perform than for machines. These tasks are known as human intelligence tasks (Human Intelligence Task, HIT) and the Information Extraction, (IE) required to generate data sets is much more viable and easier to perform using Crowdsourcing than with Natural Language Processing (NLP) techniques and Optical Character Recognition (OCR).

The structure of the document is as follows: section 2, deals with the fundamentals; in section 3, the proposed method for data sets of public procurement processes is presented; Section 4 shows the experimental results; Section 5 contains the discussion of results; and section 6, presents the conclusions.

\* Corresponding author: arey@uceva.edu.co

## 2 Background

This section briefly addresses two fundamental issues for the work. The first of them deals with the importance of transparency in the fight against corruption and the other corresponds to Crowdsourcing, the section ends with mention of part of the related work.

### 2.1 Importance of Transparency in the fight against Corruption

The Transparency International defines corruption as the misuse of the public power entrusted to obtain private benefits. Although this definition is very general, it gives a little direction to the meaning of this concept, but in order to have a framework for analyzing the problem and explaining the relationship between governments and citizens, the agency theory that applies to any situation in where one or more people have delegated authority, at the time of making decisions, to another/other person/s [8].

The agency theory illustrates the relationship between two important actors: the principal, which represents those who delegate authority, which for this case will be citizens; and the agent that represents those who receive the delegated authority, that is, the elected civil servants. In this relationship, citizens are the directors and, in contrast, governments are agents that work for citizens. By delegating certain decision-making power to the agent, the principal expects the agent to defend the interests and maximize the profits, appropriately managing the discretionary capacity of action that has been delivered to this one [8].

The agency theory allows analyzing the problem of corruption that is facilitated when the agent has more information than the principal, which is called information asymmetry [9]. It is indisputable that the principal has the right to know exactly all the actions of the agent (or all the information this one has) so that he does not act against his interests. Nevertheless, when information asymmetries are created in the principal-agent relationship, the agent's corruption is facilitated, which can act to favor his personal interests to the detriment of the interests of the principal [9].

The agency theory provides a framework to analyze the problem of corruption in which transparency is fundamental, because the lack of it does not allow society to know the actions of state entities and their officials creating the conditions that favor corrupt practices by making them less risky and more attractive [2].

Although there is no unanimity in the literature regarding the definition of transparency, [10] this can be defined as the degree to which information is made available to people outside the government administration looking for people to influence decisions [11]. Transparency involves the opening and flow of information from public entities to the public domain, allowing prospective stakeholders to review and analyze the information in order to detect possible anomalies [12]. Finally, it is relevant to indicate, in relation to the problem of corruption, that transparency is fundamental to facilitate what is known as accountability, which can be seen as the opportunity that citizens have to analyze and judge the act of the government entities and their officials making them responsible for their actions and punishing or rewarding them for their actions [13].

Given the importance of transparency in recent years, many countries have evolved in the direction of adjusting the legal system to favor transparency by incorporating access to public information as a right explicitly guaranteed to citizens. For this reason, laws of free access to public information (Freedom of Information, FoI) are beginning to be available around the world, which increase the public information made available to citizens seeking to improve qualitatively and quantitatively the possibilities of control over the government's actions. [2].

### 2.2 Crowdsourcing

The concept of Crowdsourcing is widely disseminated nowadays, which refers, according to its creator Jeff Howe, to the "act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call" [14]. The previous definition, although quite concrete, provides elements to suggest that Crowdsourcing can be seen as a new business model that is considered particularly useful in tasks that require a large number of points of view or different solutions to problems [15]. Furthermore, a key aspect of Crowdsourcing is that it aims to take advantage of the perceptive and cognitive abilities of a large group of individuals to solve a certain task [16].

Crowdsourcing is generally carried out through the Web and over the years has proven to be an effective and scalable approach to solve different problems in various fields [15], even to be used to solve problems that are computationally expensive or impossible to solve for machines, but which are rather simple for human beings; hence, sometimes the tasks proposed to the crowd are called Human Intelligence Task (HITs).

Regarding this study, it should be highlighted that Crowdsourcing can be used in data acquisition and information generation processes establishing a relatively new approach in the knowledge acquisition process [7]. In this sense, it is proposed that the Information Extraction, (IE) required to generate public procurement data sets from the non-structured information available on the Web can be much more viable and easier to perform using Crowdsourcing than with Natural Language Processing (NLP) techniques and Optical Character Recognition (OCR).

With regard to the implementation at the time of carrying out a Crowdsourcing initiative, it is required to have 4 main elements, these are: (1) The crowd, which is the multitude made up of people who participate in a crowdsourcing activity, the workers are called Crowdworker; (2) The crowdsourcer, which is the entity (a person, a for-profit organization, a non-profit organization, etc.) that seeks the power and wisdom of the crowd for a task in question; (3) The task, assigment

or activity in which people participate; (4) The platform, the crowdsourcing platform that is the system (software or non-software) within which a crowdsourcing task is carried out [17].

## 2.3 Related work

In the conducted review of the scientific literature, no works were found where crowdsourcing is used as a method of data collection of public procurement processes. Similarly, despite the fact that there are numerous works that use crowdsourcing for different tasks related to data collection, there were no data collection methods based on Crowdsourcing that could serve as a general reference.

With regard to public procurement data available on the Web, the situation is similar in many countries of the world where data is published and made available to citizens in formats that can not be processed by machine, a condition that hinders automatic processing with state-of-art computational techniques. In addition, another drawback is presented in the lack of standardization in aspects such as the information to be published, finding that the procurement information is not published in full, although for several years thanks to the World Bank, an open data standard is available for the publication of structured information covering all stages of a procurement process, from planning to execution, called the Open Contracting Data Standard (OCDS).

However, in spite of the different problems presented by the public procurement information available on the Web, if important studies are being carried out such as those carried out by Fazekas in [18], where taking advantage of the available data of public procurement in Hungary and helping by a small group of collaborators, he generated a set of data that he later used to formulate an indicator that measures the risk of corruption (Corruption Risk Index, CRI). Fazekas' work shows that despite not having complete first class data, the public information available on the Web together with the use of models has great potential to address problems such as corruption [19].

## 3 Data collection method based on crowdsourcing

### 3.1 General description of the data collection method

The method consists of two stages: the first called "Initial generation of the data set", has the purpose of allowing the interested parties to generate an initial data set that will contain basic data of the contracting processes of interest; and the second one called "Enrichment of the data set", seeks to complement the initial data set by adding more intermediate level data of the processes of interest selected in the previous stage. Each one of these stages is in turn made up of works proposed with a special objective that is achieved by performing specific tasks that are performed by the Crowdsourcer and others by the workers of the crowd

(Crowdworker). Figure 1 and Figure 2 presents the stages, works and tasks that make up the proposed method.
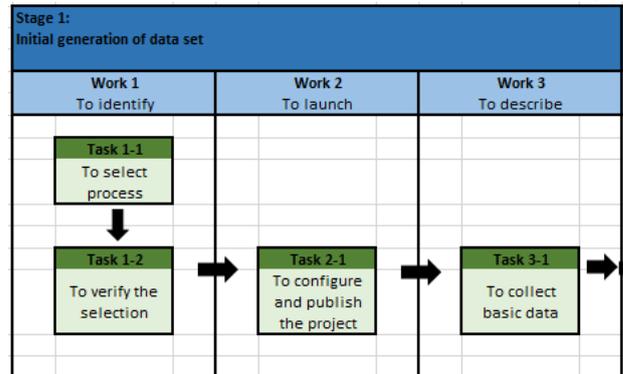


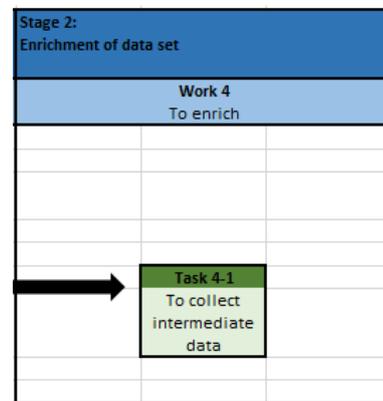**Fig. 1.** Works and tasks of the proposed method.



**Fig. 2.** Works and tasks of the proposed method.

Basically, the process defined by the method for generating the data sets is as follows. The first work of stage 1 that is called "to identify" has the objective of allowing those interested in generating the data set to indicate the transparency portal where procurement data are available and define which procurement processes of interest are. To carry out the above, two tasks are defined: the first task (task 1-1) called "to select" is precisely what allows stakeholders to define the transparency portal together with the criteria to be used to select the hiring processes of interest; the second task (task 1-2) called "to verify" allows confirming that the data provided in the previous task is correct and can proceed to create the crowdsourcing project to advance the data collection.

The second work of stage 1 is called "to launch" and aims to allow the stakeholders or those responsible for the data generation project to configure the characteristics of the crowdsourcing project and launch or publish the project to the multitude of possible workers. As an example of characteristics that must be established in this point, we have the number of answers (or contributions) required for the crowdsourcing tasks and the incentives to deliver among others. The above actions are performed in task 2-1 called "to configure

and publish". The last work of stage 1 (work 3) is called "to describe" and it is what allows the collection of basic data to the workers of the crowd selected to perform this task that is carried out in task 3-1 called "to collect basic data".

On the other hand, in stage 2 it is found work 4 called "to enrich" that seeks to complement the initial data set by adding more intermediate level data of the selected processes, for which task 4-1 called "to collect intermediate data" where the workers of the crowd selected to carry out this task collect the indicated intermediate data.

## 4 Experimental results

To evaluate the data collection method based on Crowdsourcing designed to generate public procurement data sets, it was necessary to develop a web application prototype that allowed testing the method in the most important points.

The results presented in this section correspond to an initial test carried out with the objective of allowing evaluation. First, the process defined in the proposed method for data collection that is made up of a series of steps that when carried out they must allow generating the desired data set. Second, to examine how close the final data generated are with respect to the exact values that should be collected.

The materials used to perform the test on the proposed method and the results obtained are described below.

### 4.1 Web application developed to test the proposed data collection method

The developed prototype Web application served as a Crowdsourcing platform facilitating the completion of some steps and others, which are not yet supported by this prototype, were carried out manually. The prototype implemented to perform the tests allowed registering users (type Administrator, Crowdsourcer, Crowdworker and Citizen), creating Crowdsourcing data collection projects, advance the collection of basic data and performing the aggregation of results to generate the data set.

Figure 3 shows one of the GUI that allowed the Crowdsourcer user to create a data collection project, indicating, among other things, the transparency portal and the selection criteria for procurement processes as established by the method in the steps 1 to 13 that make up the tasks "To select processes", "To verify selection" and "To configure and publish project"; the authors of this work carried out the role of Crowdsourcer creating and configuring the data collection project. The criteria defined for the selection of the interest processes were carefully selected so that the number of processes found was low and it would be feasible to perform the test with the available resources. The search revealed 6 procurement processes carried out in Colombia by the institution called UCEVA (*Unidad Central del Valle del Cauca*) under the public bidding modality in 2016,

although the number may seem low, these processes total $ 3,096,848,898,

The Data collected and results collected in the test was obtained thanks to the participation of students from the Systems Engineering program of the UCEVA, which were called to attend and carried out the role of Crowdwokers, users basically responsible for collecting the data. data that was carried out by copying and/or transcribing, from the transparency portal, the data requested in the GUI. Figure 3 presents one of the GUI that allowed Crowdworker users to enter the basic data of the hiring processes of interest assigned in the data collection tasks.

The data collection started properly with step 14 of the "to collect basic level data" task where the Crowdsourcer assigned the data collection tasks to the participating students or Crowdworkers. Each Crowdworker was assigned 6 tasks of data collection in different processes, Figure 3 shows the form used in this task, the Crowdworkers were able to carry out some preliminary tests to prepare before carrying out the final exercise and subsequently it was carried out the data collection (step 15). In addition, in the prototype application, two treatments were programmed for the input chains: elimination of the tildes and conversion of the strings to a capital letter.



**Fig. 3.** Graphical user interface of the project creation

**Fig. 3.** Graphical user interface used for the collection of basic data



**Fig. 4.** Data associated with the procurement process identified with No.113 (Part1)



**Fig. 5.** Data associated with the procurement process identified with No.113 (Part 2)



**Fig. 6.** Data associated with the procurement process identified with No.170 (Part 1)



**Fig. 7.** Data associated with the procurement process identified with No.170 (Part 2)

## 4.2 We Data collected and results

Once the collected data was available, the final data set was formed by examining how close the final data generated is with respect to the exact values to be collected. To make this comparison, it was used a reliable data set that was previously generated, reviewed and adjusted to avoid having errors. Figures 4 to 7 show the following data related to one of the procurement processes studied, as follows: the first row (blue background) corresponds to the reliable data that do not contain errors, the value of each of these fields co-sign with the original value of the field published in the transparency portal; the second row (green background) contains the data called endings that are generated from processing the raw data corresponding to the four final rows (pink background), where each field was collected by four Crowdworkers to be able to implement quality control and correction of errors through the aggregation of results approach.

The set of data generated has 23 fields in total. The columns correspond to the fields of the data set and in red the fields that present some error are indicated. In the first column, it is the process number that is an internal

identifier assigned by the procurement public entity. For the process 181 that is taken as an example, the final data set presents error in 2 of the 23 fields, this figure is striking because the objective would be to have no error. But analyzing the raw data set, which is the input to generate the final data, it is observed that it had 15 fields that had at least one error in any of its entries. This means that by applying quality control and error correction through the mechanism of aggregation of results that implements the evolutionary algorithm that aggregates the raw inputs and generates the final data, it was possible to generate correct outputs for 13 of the 15 fields that contained errors.

Similar behaviors and variations were observed with the data sets of the remaining processes.

Table 1 summarizes the results obtained for the test performed where, all the final data generated for the procurement processes have fewer fields with errors than the sets in raw data, which always presented errors.

**Table 1.** Final results obtained

| Process | No. Fields in raw with errors | No. Final fields with errors | No. Corrected fields |
|---|---|---|---|
| Process No.181 | 15 of 23 | 2 of 23 | 13 |
| Process No.33 | 17 of 23 | 7 of 23 | 10 |
| Process No.113 | 14 of 23 | 6 of 23 | 8 |
| Process No.159 | 9 of 15 | 3 of 15 | 6 |
| Process No.160 | 3 of 8 | 0 of 8 | 3 |

## 5 Discussion of results

To The data collection method based on proposed Crowdsourcing was tested using a prototype web application developed to support it. The results obtained in this initial test show that when carrying out, in the Web application, the different steps specified by the method, the data set of the procurement processes of interest is generated.

When examining how close are the final data generated with respect to the exact values to be collected, it was found that the incorporation of the human element in the data collection process introduced some errors that in certain cases could be corrected when performing the aggregation of results that was carried out by means of the evolutionary algorithm that implemented the correction of errors, by means of the approach of aggregation of results that is one of the most used in Crowwdsourcing for quality control.

Within the identified errors, it was found that some are generated by the Crowdworkers: when transcribing the

requested data, this type of error introduces wrong characters in any position of the input string. When the requested data is copied from the source page, this error can eliminate characters at the beginning or end of the chain or even incorporate characters that are not part of the original data that it is intended to copy; copy or transcribe a totally wrong value. Likewise, a failure is generated when trying to calculate a final data by adding fields that contain chains whose size exceeds 12 characters, this is due to a restriction of the evolutionary algorithm that performs the aggregation implemented in C language that overflows the memory in these situations. It was also identified that fewer errors occur when the GUI uses inputs with closed controls, such as a selection list, but this type of controls can only be used when the different possible input values are known. However, there are several techniques such as "grid adjustment" that can be tested to facilitate the capture of structured public procurement data [20].

The results obtained are consistent with results presented in similar works such as [21] where Crowdsourcing was used to generate a set of biological sample data. However, between the two works, the way in which the aggregation of results mechanisms are implemented is differentiated, finding that for the proposed method the aggregation was carried out by applying the concept of Media Chain (String Media) [22] that they face a computationally complex problem associated with chain treatment [23] and in the cited work a different algorithm was developed. In this part, it should be noted that the application of the concept of Medium Chain to perform the aggregation of results in Crowdsourcing and its implementation with evolutionary algorithms presents a novel character for the developments in this field.

It should also be noted that Crowdsourcing has great potential to be used in processes of data acquisition and information generation favoring a new approach that can be used in various fields. Furthermore, we believe that in order to optimally solve different problems such as the one addressed in this research, it is necessary to combine the automatic processing, which can be programmed by software, with human intelligence tasks that can be solved by means of Crowdsourcing.

## 5 Conclusions

To Thinking about facilitating the fight against corruption, many countries now have laws of free access to public information that allow citizens to access public information, including procurement information. In this opening scenario, information and communication technologies play an important role in facilitating the delivery of information that is made available to citizens seeking to enable greater control over the actions of government and officials. Bearing in mind that in some countries public procurement information is provided in unstructured formats, this paper presented a data collection method based on Crowdsourcing that allows the generation of public procurement data sets based on the information published in the Web.

The initial tests performed on the data collection method presented allowed obtaining results that show that by carrying out the different steps specified, it is possible to generate the data set of the procurement processes of interest. Similarly, when examining how close the final data generated is to the exact values to be collected, it was found that the incorporation of the human element in the data collection process introduced some errors that in certain cases could be corrected by the evolutionary algorithm that implemented the correction of errors, through the aggregation of results approach that is one of the most used in Crowwdsourcing for quality control.

Since the presented method was formulated following a scheme that includes different key elements found in the literature of similar works, we believe that it can serve as a guide for further works that intend to use Crowdsourcing to generate data sets in different fields of application. This is an improvement with respect to previous work.

Similarly, it can be noted that the innovative idea of the Application of Crowdsourcing to Generate Datasets of Public Procurement Processes, provides a simple but upgradeable alternative to address the problem of generating datasets from sources that provide little re-useability.

It can be pointed out that the different data published in the transparency portals of governments are a valuable source of information that can be used for accountability and there is a need to continue researching as new information and communication technologies, they can be used effectively to make it easier for citizens to fight and prevent the phenomenon of corruption.

Finally, within the future work to be carried out, it can be mentioned that experimentation is required, since the results presented in this work are preliminary and require tests that are more exhaustive. Similarly, in relation to what is mentioned in the previous paragraph, there is the possibility of developing a new method of procurement data collection that adequately combines information recovery and extraction techniques with crowdsourcing, seeking to take advantage of the advantages each approach presents. More robust quality control mechanisms can also be developed that combine several approaches used in Crowdsourcing such as pre-selection, gold data, multilevel reviews, reputation, etc. [24]. And to perform a new implementation of the evolutionary algorithm of aggregation of results implementing the concept of Medium Chain using a non-typed language such as Racket (Scheme) that does not impose limitations on the size of memory that is assigned to the variables, so the length could be extended of the input chains.

## References

1. B. Mundial, Helping countries combat corruption: the role of the World Bank PREM, (1997).
2. Y. Fan, Z. Zhang, & Q. Yue, E-government, Transparency and Anti-corruption, *Management of e-Commerce and e-Government,* 101-104 (2009).
3. Y. Welp & others, ¿Hacia el fin de la administración pública burocrática?*, Universitat Pompeu Fabra*, (2007).
4. J. I. Criado y A. Corojan, ¿Pueden las TIC cambiar la transparencia, lucha contra la corrupción y rendición de cuentas en los gobiernos latinoamericanos?, *XIV ELE*, 2037-2072 (2010).
5. D. Dietrich, J. Gray, T. McNamara, A. Poikola, P. Pollock, J. Tait, & T. Zijlstra, *Open data handbook*, (2009).
6. M. Fazekas, I. Tóth., & L. King., An objective corruption risk index using public procurement data, *European Journal on Criminal Policy and Research*, **22**(3), 369-397 (2016).
7. C. Papadopoulou & M. Giaoutzi, Crowdsourcing as a tool for knowledge acquisition in spatial planning, *Future. Internet*, **6**(1), 109–125 (2014).
8. R. W. Smith & M. Bertozzi, Principals and agents: An explanatory model for public budgeting, *Public Budgeting, Account. Financ. Manag.*, **10**(3), 325–353 (1996).
9. A. Schedler, L. Diamond, & M. Plattner, The self-restraining state: power and accountability in new democracies. *Lynne Rienner Publishers*, (1999).
10. J. Relly & M. Sabharwal, Perceptions of transparency of government policymaking: A cross-national study, *Gov. Inf. Q.*, **26**(1), 148–157 (2009).
11. A. Florini, The right to know: transparency for an open world, *Columbia University Press*, (2007).
12. M. Valverde, Transparencia, acceso a la información y rendición de cuentas: experiencias en la Unión Europea y México. *ITESM*, (2004).
13. C. der Eijk & M. Franklin, Elections and voters, *Palgrave Macmillan*, (2009).
14. J. Howe, The rise of crowdsourcing, *Wired Mag.*, **14**(6), 1–4 (2006).
15. D. Brabham, Crowdsourcing as a model for problem solving: leveraging the collective intelligence of online communities for public good. *The University of Utah*, (2010).
16. T. Erickson, Geocentric crowdsourcing and smarter cities: Enabling urban intelligence in cities and regions, *1st Ubiquitous Crowdsourcing Workshop at UbiComp*, (2010).
17. M. Hosseini, K. Phalp, J. Taylor, & R. Ali, The four pillars of crowdsourcing: A reference model, *IEEE (RCIS)*, 1–12 (2014).
18. M. Fazekas, I. Tóth, & L. King, Anatomy of grand corruption: A composite corruption risk index based on objective data, (2013).
19. J. Barr, E. Pinilla, & J. Finke, A Legal Perspective on the Use of Models in the Fight against Corruption, *SCJ Int'l L. Bus.*, **8**, 267 (2011).
20. T. Gruber, Collective knowledge systems: Where the social web meets the semantic web, *Web Semant. Sci. Serv. agents World Wide Web*, **6**(1), 4–13 (2008).
21. A. Matsunaga, A. Mast, & J. Fortes, Workforce-efficient consensus in crowdsourced transcription of biocollections information, *Futur. Gener. Comput. Syst.*, **56**, 526–536 (2016).
22. T. Kohonen, Median strings, *Pattern Recognit. Lett.*,

     **3**(5), 309–313 (1985).
23. C. de la Higuera & F. Casacuberta, Topology of strings: Median string is NP-complete, *Theor. Comput. Sci.*, **230**(1), 39–48 (2000).
24. A. Rey & P. Vélez, Control de calidad en sistemas crowdsourcing: un mapeo sistemático, *Sci. Tech.*, **22**(1), 73–82 (2017).