750 Volunteers Transcribing 31,000 Pages with 8.5 million Entries Online—an Evaluation

Jesper Zedlitz
Dept. of Computer Science
Christian-Albrechts-Universität
Kiel, Germany
j.zedlitz@email.uni-kiel.de

Norbert Luttenberger
Dept. of Computer Science
Christian-Albrechts-Universität
Kiel, Germany
n.luttenberger@email.uni-kiel.de

ACM Reference format:
Jesper Zedlitz and Norbert Luttenberger. 2017. 750 Volunteers Transcribing 31,000 Pages with 8.5 million Entries Online—an Evaluation. In *Proceedings of DATeCH2017, Göttingen, Germany, June 01-02, 2017,* 6 pages.
DOI: http://dx.doi.org/10.1145/3078081.3078086

1 INTRODUCTION

Serial sources (lists, tables, registers, ...) are an important starting material for any kind of social research, including research in history. If the amount of data contained in these sources is beyond some threshold, digital data structures are preferred. Structured digital data is a prerequisite to almost all kinds of statistical evaluations, from simple counting to elaborate computations. It is also a prerequisite for publishing the information to the Linked Open Data Cloud [2].

Today, many serial sources—especially from the historic domain—are available only in print format, i.e. as printed books, sheets, loose-leaf-collections and other kinds of "papers". For transcribing serial sources into structured data it is necessary to understand the language as well as the data model of the source. Unfortunately, the conversion of a document from printed to structured digital format (i.e., transcribing, tagging, encoding, and indexing; "in-depth-indexing" in the remainder of this paper) is a very time-consuming and cost-intensive task. For example, the author of [8] notes that the in-depth-indexing of 100,000 entries required a total of 2,400 hours. Because of this high effort, archives usually don't perform in-depth-indexing of the archived materials; in most cases, sources are enriched with some formal metadata (provenience, publication information, etc.) only.

It would be beneficial a) to simplify the tedious work of in-depth-indexing through the use of technology and b) to distribute the work via *crowdsourcing* to a large number of people. Projects become possible that seemed to be impossible before due to limited budgets. The term "crowdsourcing" has first been used by Jeff Howe in 2006 in [10]. Howe later summaries crowdsourcing as "the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call."

In this paper we present our experience with the voluntary contributors during transcription of the German WW1 casualty lists using the—at the time of its launch—novel crowdsourcing platform CG-DES ("Computer Genealogy Data Entry System") built in cooperation with the German "Verein für Computergenealogie" (Society for Computer Genealogy).

The article is structured as follows: In Section 2 we start with a short description of the source, the German WW1 casualty lists. In this section we also present key features of CG-DES. Section 3 presents the results of the analysis of the statistical data collected during the project time. We present seven hypotheses about the behavior of volunteers and verify/falsify these. We show related projects in Section 4. Section 5 concludes the paper with a summary and an outlook on future work.

2 BACKGROUND

2.1 Historic Source

The German casualty lists for the First World War have been published between 1914 and 1919. They include the official notifications from the Prussian government on the military losses of the entire armed forces of the German Empire. Each entry in the list contains information about woundings, missing persons, captures, deaths, as well as corrections to previous messages. In total there are 31,200 pages in Tabloid Extra format (305 x 455 mm) written in Gothic script. One page contains between 200 and 300 individual entries.

Figure 1 shows a page from the casualty list. The order in which the information is listed as well as the abbreviations vary—even within a single page. In contrast to family names which are quite easy to spot because of the spaced type the identification of the other data fields usually requires knowledge of the German language.

2.2 Software

CG-DES is described in detail in [13]. Therefore, we just mention its main features here. CG-DES is web-based and can be used with any web-browser without additional plug-ins. CG-DES' key feature is the fact that data entry is performed directly on scanned images: The scan is used as background image of the browser window; it is overlaid by the entry mask as well as text boxes with already transcribed data.

3 EVALUATION

In this section we report on the growth of the number of volunteers and the number of entered records, the age structure of the

DATeCH2017, Göttingen, Germany

978-1-4503-5265-9/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078081.3078086

Figure 1: Page 12135 of the German WW1 casualty list from 20th April 1916.

volunteers, the typing error rate, and the migration of volunteers between projects. To characterize volunteer behavior patterns, we set up a number of hypotheses. Using these hypotheses, it should be possible to optimize the labor put into the supervision of the crowdsourcing volunteers. The hypotheses might also give hints how to better motivate certain groups of volunteers.

3.1 Project Progress

Figure 2 shows how many volunteers have joined the project and how many entries have been transcribed. As one can clearly see there is a strong increase in the number of volunteers after the initial announcement of the project. As of mid 2012, the increase is slowed and settled around 14 new people joining the project per month.

The number of transcribed records increases approximately linearly (8,550 entries per day). Since the number of volunteers also increased, a quadratic growth would have been expected. However, not every volunteer who joined the project contributed daily. Some volunteers also eventually left the project. Figure 3 shows the details of this behavior. The upper line shows the number of active volunteers per day. Within the first two months each day 80 volunteers were active – probably the enthusiasm about the novel data entry approach. After this first phase the number levels of to 50 active volunteers per day.

The bars in the lower part of Figure 3 indicate the number of volunteers joining and leaving the project. A volunteer is considered as leaving if he does not enter data anymore. These numbers are aggregated per week. At the start the crowdsourcing project could attract many volunteers. During most of the project duration, we see a coming and going of volunteers. The numbers of leaving volunteers at the end of the project become distorted due to the fact that occasionally no free pages were available for editing anymore.

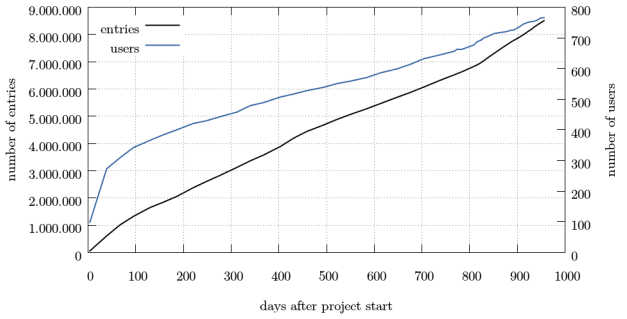

Figure 2: Number of volunteers and number of entered records.

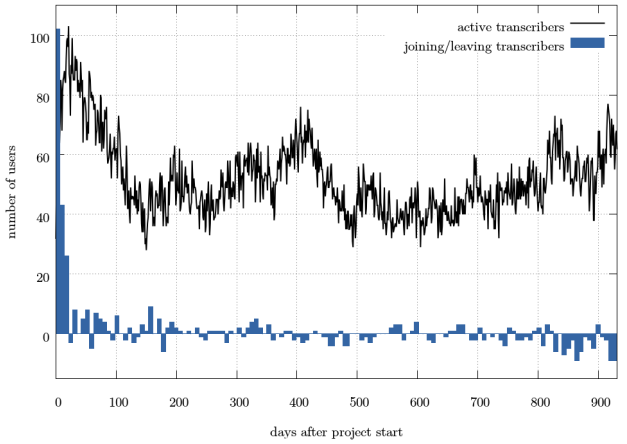

Figure 3: The diagram shows the number of active transcribers per day as well as the number of volunteers joining (positive number) and leaving (negative number) the project aggregated per week.

3.2 Age structure

Figure 4 shows the age distribution of the volunteers. 262 of 767 volunteers have told us their birthday. The smaller bars indicate the expected number of volunteers based on the internet usage in Germany.[1] As you can see, for a web application the participants are relatively old.

3.3 Error Rate

An important question when transcribing sources is the rate of typing errors. For the transcription of the casualty lists a single-keying approach was used. Users searching for entries were able to mark transcription errors. There was no systematic software-supported proofreading process. However, a couple of volunteers have specialized in typo finding. They developed there own search strategies to spot errors. As one can easily imagine, such a non-systematic proofreading is not effective. We selected a sample of

[1]The expected numbers have been calculated based on the data from [1] and [11]: The number of persons for each birth cohort is multiplied with the proportion of online users in this age group according to [1]. Proportional to the total number of online users, the 240 users are distributed to the age cohorts.

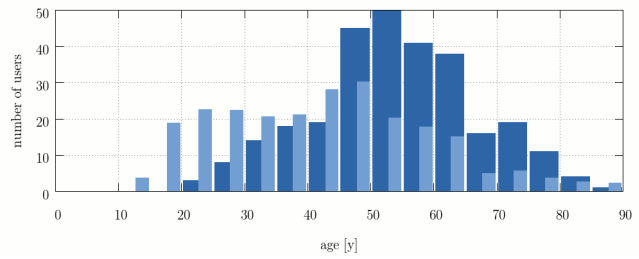

Figure 4: Age distribution of the volunteers. The smaller bars indicate the expected number of users based on the internet usage in Germany.

5,000 entries from the transcribed entries and checked them using double-keying with n-keyed run-off votes [3]. This check revealed transcription errors in 15.6% of the entries. In contrast, the users only reported 349,012 (=4.1% of 8.5 million) incorrect entries. That means that 74% of the transcription errors remained unnoticed with non-systematic proofreading. However, for comparing error rates of volunteers the coarse numbers of proofreading are sufficient. But one has to keep in mind that the numbers can not be used to infer absolute error rates for single volunteers or groups.

3.4 Volunteer Migration

After the processing of the first source started very successfully and cooperation requests from archives arrived, we decided to use CG-DES for further projects. We were concerned that only few additional volunteers would join. Instead, the new projects might drain volunteers from the first project and the time to complete the first project would increase significantly.

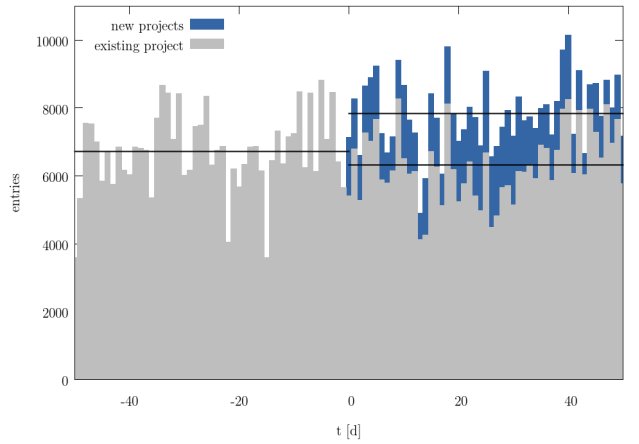

Figure 5: Number of entries entered per day before and after new projects have been added. The horizontal lines show the average number of entries for the existing project and the total number after new projects have been added.

Figure 5 shows the number of entries entered per day 50 days before and 50 days after the new additional projects have been added and announced. It can be seen that the performance in the existing project only slightly decreases (on average 6,717 entries per day before to 6,332 entries after). In contrast, an average of 1,500 entries are added to the new projects per day.

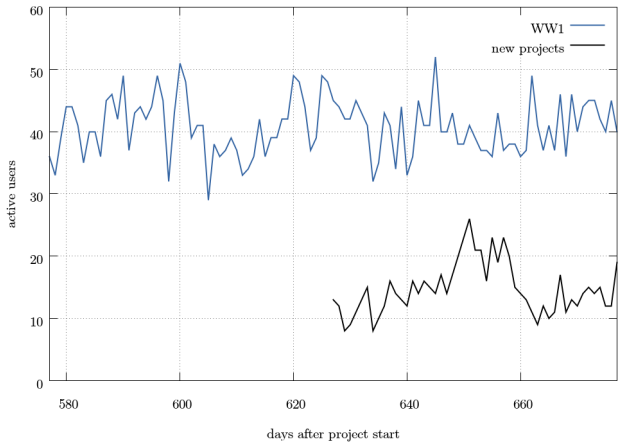

Figure 6: Number of active volunteers per day before and after new projects have been added.

However, one has to keep in mind that the new projects had other topics (civil records and address books) than the first project (military). It is conceivable that there would have been greater migration of volunteers if the projects would have been more similar.

3.5 Hypotheses

We set up seven hypotheses on volunteer behavior within the project. In this section we are going to evaluate these hypotheses using the data collected during the project runtime.

Hypothesis 1. *80% of the data is entered by 20% of the volunteers. (Pareto principle)*

According to the Pareto principle (i.e. 80% of the effects come from 20% of the causes) only one fifth of the volunteers enter 80% of the data. During the observation period 767 volunteers have contributed data. Therefore, it is to expect that 153 volunteers have entered 80% of the data. However, the data shows that 80% of the data entered in the observation period has been contributed by only 75 volunteers (~10% of all volunteers).

Figure 7 shows the cumulative contributions of the volunteers, sorted in descending order according to the activity of the volunteer. The y-value at x = 100 thus gives the number of data that has been contributed by the 100 most productive volunteers. The grey vertical line indicates the expected number of 137 volunteers. Highlighted is the point at which 80% of the data has been acquired.

The principle of the unequal distribution of work can therefore also be observed in CG-DES. The ratio is even more unequal than 80:20.

Hypothesis 2. *Most of the work is in the evening hours and on weekends.*

This hypothesis is mainly useful for planning of resource distribution and maintenance windows for the web application. The

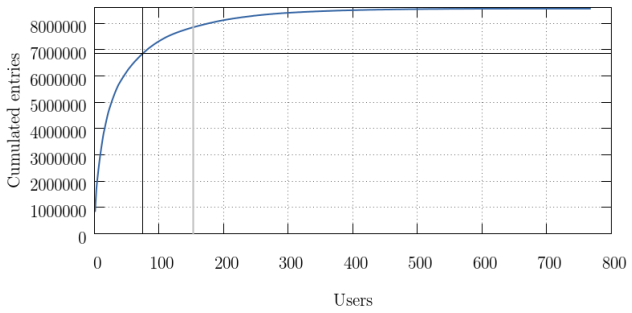

Figure 7: Cumulated number of entries entered by volunteers (in descending order according to the activity of the volunteer).

time from 5 pm until midnight is considered as "evening hours". Saturday and Sunday from 10 am until midnight is considered as "weekend".

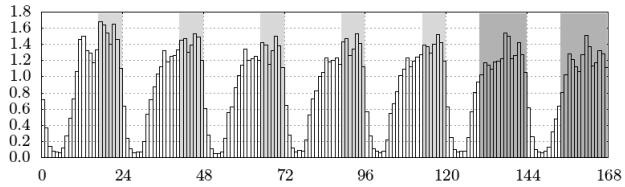

Figure 8: Percentage of entries entered per hour of the day and week.

Figure 8 shows what proportion of the total entries were entered in a certain hour of a week. For example Monday during 5 pm and 6 pm 0.903% of all entries were made. The evening hours and weekends are highlighted.

By averaging over all days of the week (weekdays and weekends together) and all hours of the day, an average of 0.595% of all entries is entered in one hour. Within an hour of a weekend 0.890% of all entries are contributed. In an evening hour on weekdays even more entries are made: 0.939% of all entries.

The hypothesis has thus been found to be correct.

Hypothesis 3. *"Power users" leave the project less than other volunteers.*

Two definitions are required to check this hypothesis:

- A "power user" is one of the volunteers that contribute more than 80% of the data (see Hypothesis 1).
- A volunteer has left the project after he has not contributed an entry for more than one month.

Within the last month of the project 122 volunteers have contributed data. This is 15.9% of all volunteers that have ever worked on the project. In contrast, 53 of the 75 volunteers determined as "power users" in Hypothesis 1 were still active in the last month. This is 70% of all "power users".

The hypothesis has thus been found to be correct.

Hypothesis 4. *After a volunteer joins the project the number of edits per day increases strongly, than drops to a long-term average value.*

We expected a trend of the number of entries contributed by a volunteer on his x'th day working in the project as shown in Figure 9. Two variants seemed plausible:

(1) In the beginning the volunteer is strongly motivated for a few days and therefore contributes a high number of entries. Subsequently, the motivation and the number of entries per day decreases and settles at a lower value.
(2) In the beginning, a volunteer is not so fast in entering data because he or she has to learn how to use the system and get familiar with the Gothic print. By learning success and motivation, the acquisition performance increases day by day. Subsequently, the motivation and the number of entries per day decreases and settles at a lower value.

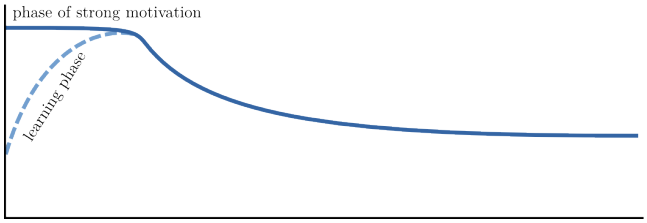

Figure 9: Expected number of entries per day.

Figure 10 shows the average number of entries a volunteer has entered on this x'th day working in the project. Plotted in gray is the raw value. Since this value varies significantly, a smoothed curve is shown in black.

As you can see the number of entries per day increases up to 315 entries per day within the first two weeks. Until the 81th day in the project (twelve weeks) it decreases down to 212 entries per day. Subsequently, it settles down to an average of 212 entries per day.

The hypothesis has thus been found to be correct.

Hypothesis 5. *The longer a volunteer works on the project the less errors he makes.*

Figure 11 shows the average relative error of all volunteers for each week of project participation within the first year. First, it was determined how many entries a volunteer i has made in his x'th week of project participation: n_x^i. It was additionally counted how many of these entries had been marked as erroneous: e_x^i. These values are summed up for all volunteers per week:

$$N_x = \sum_{i=1}^{767} n_x^i \qquad E_x = \sum_{i=1}^{767} e_x^i$$

Now, the average relative error of the x'th week can easily be computed:

$$r_x = \frac{E_x}{N_x}$$

It can be seen in Figure 11 that the relative error decreases rapidly within the first weeks of project participation from approximately

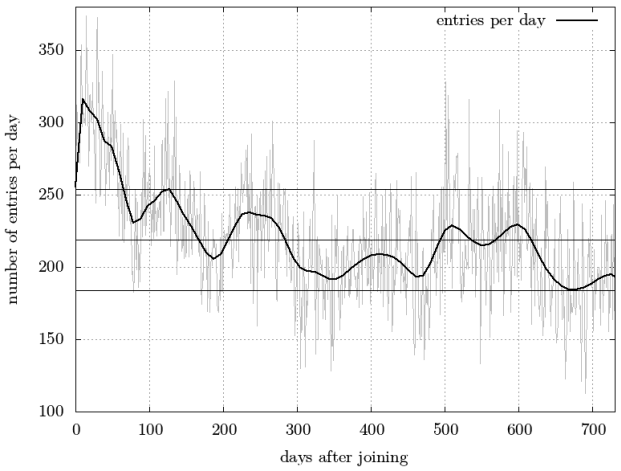

Figure 10: Number of entries entered by a volunteer per day after she joined the project.

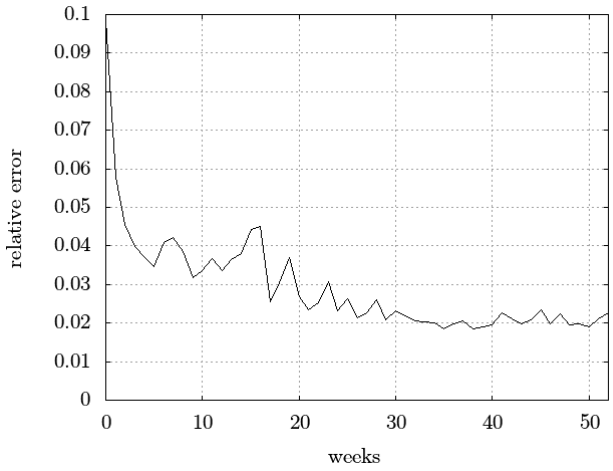

Figure 11: Average relative error of all volunteers within the first year of participation.

9% to a value around 2%. The error is decreasing in the following time, but only to a minor extent.

The hypothesis has thus been found to be correct.

Hypothesis 6. *"Power users" make less errors.*

Since "power users" are more engaged in the project and constantly keep in practice, it is expected that they make fewer mistakes than other volunteers. Therefore, we take a look at the total number of entries N_i entered by a volunteer i and the number of reported errors within these entries E_i. With these two values the relative error r_i for one volunteer can be easily calculated:

$$r_i = \frac{E_i}{N_i}$$

Figure 12 shows the frequency distribution of the relative error per volunteer, for both the "power users" (highlighted) as well as for the remaining volunteers. While the majority of volunteers contributes about 3% incorrect entries, there are few volunteers where the error rate in some cases is significantly higher. For "power users", however, error rates around 1.5% dominate.

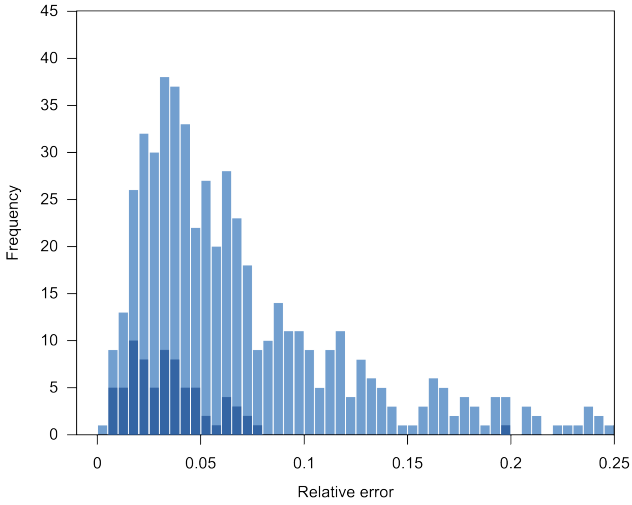

Figure 12: Distribution of relative errors per volunteer.

The hypothesis has thus been found to be correct.

Hypothesis 7. *Volunteers that leave the project after a very short time contribute only little and cause a lot of work.*

As previously seen in Hypothesis 1, a large portion of volunteers contribute only little to the project. This contrast is even more dramatic when you look at those volunteers who leave the project after only one week of participation. This table shows the amount of data entered and the average error rate for "power users", short-term volunteers and other volunteers.[2]

	Count	Contributed entries	Error rate
Short-term v.	211	0.32%	19.74%
"Power users"	74	78.28%	2.68%
Other volunteers	481	19.55%	6.26%

With 211 persons (27.5%) this is a significant part of the users. On the one hand, these users contribute only a very small proportion (0.32%) of the data. On the other hand, these short-term users make significantly more errors than other volunteers. Because these errors require a manual check by the project supervisors, the short-term volunteers cause a lot of work.

The hypothesis has thus been found to be correct.

[2]One outlier has been removed from the group of "power users": After two years with a low error rate the volunteer entered a lot of incorrect entries with an error rate up to 50%. Since we use anonymized data we can only speculate about the reasons.

4 RELATED WORK

Apparently micro-level data of user behavior in a crowdsourcing project is hardly published. Hirth, Hoßfeld, and Tran-Gia provide a more detailed insight in [9]. However, their object of investigation are commercial crowdsourcing platforms where the users get paid for their contribution. Therefore, their observations are complementary to ours. For example, the most active times of the weeks reported in this study are working hours from Monday to Friday whereas it could be observed that CG-DES' volunteers were active mostly in the evenings and on weekends.

In [7] Franzoni and Sauermann note generally that in voluntary crowdsourcing projects "the majority of participants make only small and infrequent contributions, often stopping quickly after joining." However, no micro-level data is provided to underpin this observation.

Some data for a project on a voluntary basis is provided by Causer, Tonra, and Wallace in [4]. Their report covers six month of the *Transcribe Bentham* project and gives information on the number of volunteers and completed documents. A somewhat superficial evaluation of "user experience" is given by Fornés et al. in [6].

The question of how people are motivated to take part in a crowdsourcing project, is investigated, inter alia, by Doan et al. [5] and Zarndt [12].

5 SUMMARY AND OUTLOOK

In this article we have presented our experiences with volunteers behavior in a crowdsourcing project transcribing structured data from printed historic documents. We set up seven hypotheses how volunteers behave and underpin these hypotheses with the statistic data. It is likely that these hypotheses are valid for other crowdsourcing projects, too. Hypothesis 5 and Hypothesis 7 suggest a provocative proposal: Data that has been entered by a volunteer in his/her first week is generally discarded silently. The entries show up during the search and also typing errors within the entries can be reported. However, these reports are not presented to the project supervisors. Without loosing too much data the work of the project supervisors is reduced significantly. Entries with a high probability of error do not remain in the system.

Especially new volunteers make lots of mistakes. Therefore, a mandatory learning program seems to be useful. With increasing difficulty, such a learning program would present to the newcomer parts of the sources of which the content is already known. The data entered by the new volunteer can be compared to the known content and the volunteer will be given feedback on misread characters or positioning errors.

So far, the detection of typos is done only through the users of the data during searches. We are working on a method to improve the data quality through *double keying* or systematic proofreading—without reducing the processing speed or the motivation of the volunteers too much.

While the casualty lists requires only three data input fields, other types of sources (e.g., historical address books) require more data input fields. It should be investigated what effect a greater number of input fields has on motivation, processing speed and error rate.

For tabulated sources a different type of the input window might be more useful, which reflects the columns of the table better. Once again, evaluation is necessary whether another kind of input window has a positive effect.

The seven hypotheses presented in this article can be used to specifically address volunteers. An early exit of a volunteer from the project might be prevented and permanent motivation as well as high input quality of the voluntary contributed can be ensured. The results of such influence could also be the subject of further investigation.

REFERENCES

[1] BITKOM. 2011. Netzgesellschaft – Eine repräsentative Untersuchung zur Mediennutzung und dem Informationsverhalten der Gesellschaft in Deutschland. (2011). http://www.bitkom.org/files/documents/BITKOM_Publikation_Netzgesellschaft.pdf

[2] Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee. 2008. Linked data on the web (LDOW2008). In *Proceedings of the 17th international conference on World Wide Web*. ACM, 1265–1266.

[3] Ben Brumfield. 2012. Quality control for crowdsourced transcription. *Collaborative Manuscript Transcription Blog* (2012).

[4] Tim Causer, Justin Tonra, and Valerie Wallace. 2012. Transcription maximized; expense minimized? crowdsourcing and editing The Collected Works of Jeremy Bentham. *Literary and linguistic computing* 27, 2 (2012), 119–137.

[5] Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy. 2011. Crowdsourcing systems on the world-wide web. *Commun. ACM* 54, 4 (2011), 86–96.

[6] Alicia Fornés, Josep Lladós, Joan Mas, Joana Maria Pujades, and Anna Cabré. 2014. A bimodal crowdsourcing platform for demographic historical manuscripts. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*. ACM, 103–108.

[7] Chiara Franzoni and Henry Sauermann. 2014. Crowd science: The organization of scientific research in open collaborative projects. *Research Policy* 43, 1 (2014), 1–20.

[8] Eli Fure. 2000. Interactive record linkage: The cumulative construction of life courses. *Demographic Research* 3, 11 (2000), 3–11.

[9] Matthias Hirth, Tobias Hoßfeld, and Phuoc Tran-Gia. 2011. Anatomy of a crowdsourcing platform-using the example of microworkers. com. In *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on*. IEEE, 322–329.

[10] Jeff Howe. 2006. The rise of crowdsourcing. *Wired magazine* 14, 6 (2006), 1–4.

[11] Statistisches Bundesamt. 2013. Fortschreibung des Bevölkerungsstandes – Bevölkerung: Deutschland, Stichtag, Altersjahre. (2013). https://www-genesis.destatis.de/genesis/online?sequenz=tabelleAufbau&selectionname=12411-0005

[12] Frederick Zarndt. 2012. Putting the world's cultural heritage online with crowdsourcing. (2012).

[13] Jesper Zedlitz. 2016. Web-based Collaborative System for Transcription of Serial Historic Sources to Structured Data. *Technische Berichte des Instituts für Informatik der CAU Kiel* TR_1604 (2016).