INSTITUT FÜR INFORMATIK

Web-based Collaborative System for Transcription of Serial Historic Sources to Structured Data

Dr. Jesper Zedlitz

Bericht Nr. 1604
May 2016
ISSN 2192-6247

CHRISTIAN-ALBRECHTS-UNIVERSITÄT

ZU KIEL

Web-based Collaborative System for Transcription of Serial Historic Sources to Structured Data

Jesper Zedlitz
j.zedlitz@email.uni-kiel.de

Abstract

We present a web-based collaborative system to transcribe serial historic sources to structured data. The web-based system runs completely in the web browser without additional plug-ins. Its 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. The system has been successfully used to transcribe more than 31,000 pages from the German WW1 casualty lists to structured data resulting in more than 8.5 million entries.

1 Introduction

To convert printed historical sources into a structured digital form (i.e., transcribing and indexing) is a very time consuming task. For example [Fur00] notes that the processing of 100,000 entries required a total of 2,400 hours. The authors know a scheduling for a project at a German university, in which the indexing process of the source alone is calculated with 3.5 man-years of work. The processing is obviously not only time- but also cost-intensive. However, structured digital data that is result of such work is important starting material for research—not only for historians. Digital data allows completely novel evaluations. Due to the high effort, archives usually to not perform in-depth-indexing of the archived materials. They content themselves with the development of formal metadata.

Whereas for the transcription of *continuous text* only sequences of characters have to be recognized, the transcription of structured data usually requires knowledge of the language used and the information that is contained in the source. Structuring of data is necessary for example to sort the data by different criteria. To published

the information within the Linked Open Data Cloud [BHIBL08] you also need structured data.

It would be beneficial if it is possible to a) simplify this tedious work of processing through the use of technology and b) to distribute the work via *crowdsourcing* to a large number of people. As a result, enormous savings can be achieve. 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 [How06]. 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 describe the crowdsourcing platform CG-DES ("Computer Genealogy Data Entry System") built in cooperation with the German "Verein für Computergenealogie" (Society for Computer Genealogy).[1] The article is structured as follows: In Section 2 we start with a short description of the source that as been processed with CG-DES as the first project. In Section 3 we present our platform. We discuss our basic considerations during the design of the platform. The next Section 4 describes the five phases of a project in CG-DES. In the following Section 5 we show how the system can be scaled for some larger number of projects and users. We will introduce other approaches for the acquisition of data with crowdsourcing in Section 6. Section 7 concludes this paper with a summary and an outlook.

2 The Source

The characteristics of the source partly influenced the architecture of the CG-DES. Therefore, we want to provide some information about the source: The German casualty lists for the First World War have been published between 1914 and 1919. They are among the most important surviving sources for German soldiers of the First World War. Particular for genealogists they are very helpful to learn something about the whereabouts of their male[2] ancestors during that time period.

The casualty lists include the official notifications from the government on the military losses of the entire armed forces of the German Empire. The lists were published almost daily—sorted by armies, regiments, units, etc. Therefore, it is very difficult to investigate on single persons or groups of persons, e.g. soldiers from one city.

[1] http://www.verlustlisten.de

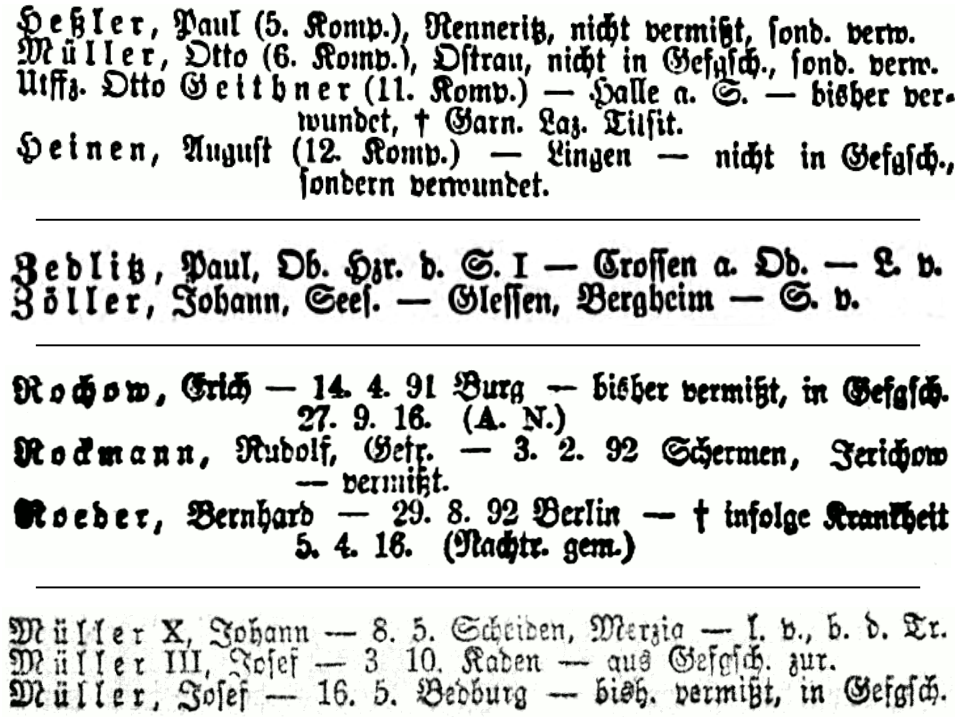
[2] Only very few women are listed in the casualty lists, mostly nurses.

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

Each entry in the list contains information about woundings, MIA, 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 print. One page contains between 200 and 300 individual entries. Figure 1 gives an impression of how a page looks.

Figure 2 shows some excerpts from the casualty list. Family names are quite easy to spot because of the spaced type. The identification of the other data fields usually requires knowledge of the German language. To make matters worse, the entries of the casualty lists are not homogeneous:

- ▻ The order in which the information is listed varies, e.g. order of family name and given names.
- ▻ Abbreviations are often difficult to understand. They also vary—even within a single page.
- ▻ The entries contain varying amounts of information. Some entries contain the complete date of birth, others only day and month. Sometimes the company (military unit) and/or the rank is specified—in many cases it is not.


Heßler, Paul (5. Komp.), Nenneritz, nicht vermißt, sond. verw.
Müller, Otto (6. Komp.), Ostrau, nicht in Gefgsch., sond. verw.
Uffz. Otto Geithner (11. Komp.) — Halle a. S. — bisher verwundet, † Garn. Laz. Tilsit.
Heinen, August (12. Komp.) — Lingen — nicht in Gefgsch., sondern verwundet.

Zedlitz, Paul, Ob. Hzr. d. G. I — Crossen a. Od. — L. v.
Zöller, Johann, Seef. — Glessen, Bergheim — S. v.

Rochow, Erich — 14. 4. 91 Burg — bisher vermißt, in Gefgsch. 27. 9. 16. (A. N.)
Rockmann, Rudolf, Gefr. — 3. 2. 92 Schermen, Jerichow — vermißt.
Roeder, Bernhard — 29. 8. 92 Berlin — † infolge Krankheit 5. 4. 16. (Nachtr. gem.)

Müller X, Johann — 8. 5. Scheiden, Merzig — l. v., b. d. Tr.
Müller III, Josef — 3 10. Kaden — aus Gefgsch. zur.
Müller, Josef — 16. 5. Bedburg — bish. vermißt, in Gefgsch.

Figure 2. Examples from casualty lists. From top to bottom: page 7765 from 22nd July 1915 — two entries from the navy's casualty list #56, page 10184 of the full list from 16th November 1915 — page 20000 from 9th August 1917 — page 23023 from 28th March 1918.

3 Key decisions

Based on our experiences with previous transcription projects with volunteers we decided to try a novel approach for our crowdsourcing platform. According to the classification scheme by Doan, Ramakrishnan und Halevy [DRH11] our platform can be described as such: It is a *standalone system* with *explicit* collaboration in which *volunteers* act as *slaves* to *execute a task*. The *target problem* is digitizing structured text. Currently we have a *manual dispute management* by project supervisors. The project supervisors also do *user evaluation* and are able to ban malicious users.

Four basic decisions characterize the architecture of the system: The data entry is performed entirely web-based. The users are working in their web browser directly "on the scanned image" (c.f. Figure 5)—in contrast to other systems where the display of the scanned image and the data entry is performed in separated areas

or windows. The data entry is carried out in a two-tiered process, “normal” user working on single entries and more experienced “expert” users working on groups of entries. Throughout the entire editing process, instant search in the acquired data is possible. In the following, these four key decisions will be explained and justified.

3.1 Web-based Solution

For offline editing one would have to transport the scans to the users. This could for example be accomplished by DVD shipping or by download. Since the scans of the casualty lists (in JPEG format) have a size of 93GB, such a distribution to a large number of helpers is not practical.

For volunteers working offline a complex management would be necessary to keep track which volunteer is working on which range of pages. To notice that someone has canceled the participation, time-consuming regular requests would be necessary.

In a web-based solution it is easily possible to replace a faulty scan (bend in the page, hand in the field of view, poor exposure, incorrect focus, etc.) with a better one. Already the next retrieval of the image updates it for the user.

Project supervisors can take a look at entries right after the contributor has entered them online. Therefore, if a volunteer does not follow the editorial guidelines, assistance can be given quickly. In a traditional offline data entry such systematic errors are only recognized after hundreds or even thousands of entries have been transcribed incorrectly.

3.2 Work “on the scanned image”

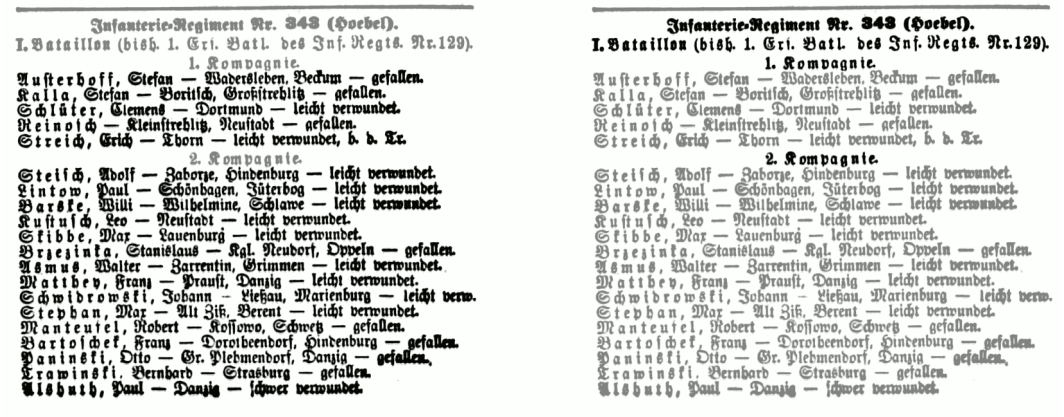
As the entries can be seen directly on the scan, it can easily be verified that each entry of the source has been transcribed. If you accidentally skip one line in a traditional acquisition, this error might never be detected or only by thorough proofreading .

The position of each entry (x- and y pixel coordinates on the scan) is recorded automatically. This makes it possible to add further information by editing entire regions (rectangles) of the page (see description of two-tiered transcription 3 below). During the later search it is not only possible to display on which page the entry you searched for appears, but the user can be led directly to the corresponding position on the page. Especially for the large, three-column pages of the casualty

lists this is an enormous simplification of the search process.

3.3 Two-Tiered Transcription

The transcription is performed in two tiers. Most of the users work in the first tier entering single entries. In the case of the casualty lists a single entry includes the family name, the given names and the place of birth. In the second tier more advanced "expert" users identify regions (rectangles) of a page and enter information about these regions. In the case of the casualty lists such regions contain information about the military units. Figure 3 shows what information is entered in which tier. On the left side the data entered in the first tier is highlighted. Highlighted on the right you can see the information about military units entered in second tier.


Infanterie-Regiment Nr. 343 (Hoebel).
I. Bataillon (bish. 1. Ers. Batl. des Inf. Regts. Nr. 129).
1. Kompagnie.
Austerhoff, Stefan — Wadersleben, Beckum — gefallen.
Kalla, Stefan — Boritsch, Großstrehlitz — gefallen.
Schlüter, Clemens — Dortmund — leicht verwundet.
Reinosch — Kleinstrehlitz, Neustadt — gefallen.
Streich, Erich — Thorn — leicht verwundet, b. d. Tr.
2. Kompagnie.
Steisch, Adolf — Zaborze, Hindenburg — leicht verwundet.
Lintow, Paul — Schönbagen, Jüterbog — leicht verwundet.
Barske, Willi — Wilhelmine, Schlawe — leicht verwundet.
Kustusch, Leo — Neustadt — leicht verwundet.
Skibbe, Max — Lauenburg — leicht verwundet.
Brzezinka, Stanislaus — Kgl. Neudorf, Oppeln — gefallen.
Asmus, Walter — Zarrentin, Grimmen — leicht verwundet.
Matthey, Franz — Praust, Danzig — leicht verwundet.
Schwidrowski, Johann — Ließau, Marienburg — leicht verw.
Steyhan, Max — Alt Zitz, Berent — leicht verwundet.
Manteutel, Robert — Kossowo, Schwetz — gefallen.
Bartoschek, Franz — Dorotheendorf, Hindenburg — gefallen.
Paninski, Otto — Gr. Plehmendorf, Danzig — gefallen.
Trawinski, Bernhard — Strasburg — gefallen.
Aishuth, Paul — Danzig — schwer verwundet.

Infanterie-Regiment Nr. 343 (Hoebel).
I. Bataillon (bish. 1. Ers. Batl. des Inf. Regts. Nr. 129).
1. Kompagnie.
Austerhoff, Stefan — Wadersleben, Beckum — gefallen.
Kalla, Stefan — Boritsch, Großstrehlitz — gefallen.
Schlüter, Clemens — Dortmund — leicht verwundet.
Reinosch — Kleinstrehlitz, Neustadt — gefallen.
Streich, Erich — Thorn — leicht verwundet, b. d. Tr.
2. Kompagnie.
Steisch, Adolf — Zaborze, Hindenburg — leicht verwundet.
Lintow, Paul — Schönbagen, Jüterbog — leicht verwundet.
Barske, Willi — Wilhelmine, Schlawe — leicht verwundet.
Kustusch, Leo — Neustadt — leicht verwundet.
Skibbe, Max — Lauenburg — leicht verwundet.
Brzezinka, Stanislaus — Kgl. Neudorf, Oppeln — gefallen.
Asmus, Walter — Zarrentin, Grimmen — leicht verwundet.
Matthey, Franz — Praust, Danzig — leicht verwundet.
Schwidrowski, Johann — Ließau, Marienburg — leicht verw.
Steyhan, Max — Alt Zitz, Berent — leicht verwundet.
Manteutel, Robert — Kossowo, Schwetz — gefallen.
Bartoschek, Franz — Dorotheendorf, Hindenburg — gefallen.
Paninski, Otto — Gr. Plehmendorf, Danzig — gefallen.
Trawinski, Bernhard — Strasburg — gefallen.
Aishuth, Paul — Danzig — schwer verwundet.

Figure 3. Example from page 9533 of the casualty lists showing the different information captured in the two tiers—single entries on the left and military units on the right.

This division of labor has the advantage that a "normal" user

- ▻ has to do less work. Therefore, he or she can transcribe more entries in the same time.
- ▻ does not have to deal with several complex issues (e.g. the identification of regiments and units in the casualty lists)—and does not have the possibility to make any mistakes in doing so.

3.4 Instant Search

It is known from previous projects that voluntary transcribers want to see “their” data available online quickly.

There is hope that interested visitors that have been attracted via the search will be interested to learn more about the project and might even become actively contributing volunteers.

In the display of results, a possibility is offered to mark typing errors. These error reports are submitted to the project supervisors for reconsideration. Thus there will be a permanent quality control.

4 Phases of a Project

Figure 4 shows an outline of the different phases of a project (e.g. the WW1 casualty lists) in CG-DES.

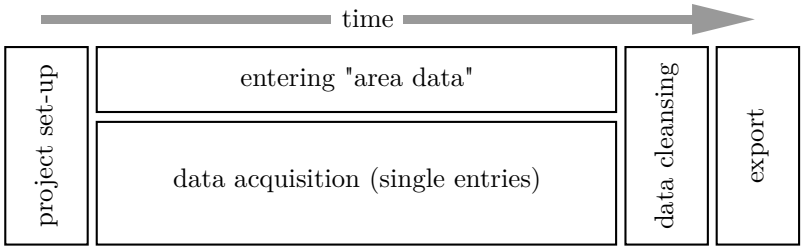


Figure 4. Outline of the different phases of a project in the CG-DES.

4.1 Project Set-Up

The first step is the creation of the project in the system by the project supervisors. In this step the structure of the data to be collected is determined. In many cases it is necessary to evaluate whether information can be summarized in a single input field, or separate fields are required. In general a higher structuring of the data, i.e., more input fields is better. However, it brings some significant problems:

- More time is needed for the transcription of an entry because you have to consider which part of the information belongs into which field.
- The editorial guidelines become more complex. Therefore, they are more difficult to understand and to remember.

▻ The higher complexity of the distribution of information into more input fields offers more possibilities to make errors. These errors have to be spotted and corrected later—causing more time spent.

Therefore, it is often better to use less input fields. In many cases information from one input field can be split in a post-processing step. Some of the information written in the source might not be necessary for subsequent modeling. With respect to the problems of too many data fields mentioned above, in this phase a decision has also to be made if information is not transcribed at all or added in a second transcription pass.

If identical information refers to a multiplicity of entries that are printed close to each other in the source (e.g. inhabitants of one village in an address book), it is smarter to collect this identical information with the help of the "area data" capturing described below. Also that decision falls within this phase of project set-up.

After the data structure has been defined objects representing the pages are created. Pages without any relevant data can already be excluded from the data entry process.

4.2 Data Acquisition

When a user starts with data entry the scan of a free page is presented to him as background image of the browser window. The assignment of a user to a page is memorized so that no one else can edit this page and the user can return to his page later.

The user can add new entries to his page. This is done by clicking on the upper left edge of the text for which the new entry shall be created (c.f. top of Figure 5). An input mask opens at that location as an overlay. It presents the entry fields defined in the previous phase (c.f. center of Figure 5). Since the input mask is an HTML element the project supervisors can be flexible in the design. For example it is easily possible to include tool tips. After the user has entered the data and clicks the save button the data will be sent to the server. The input mask disappears. Instead, the entered data is displayed in a color shaded rectangle at the position of the initial mouse click (c.f. bottom of Figure 5). You can now easily tell which entries have already been processed. Therefore, the user is not required to maintain a particular sequence during the transcription of the entries.

In the case of a two-line entry it might happen that the input mask covers a portion of the text to be transcribed. For such cases, it is possible to move the scan

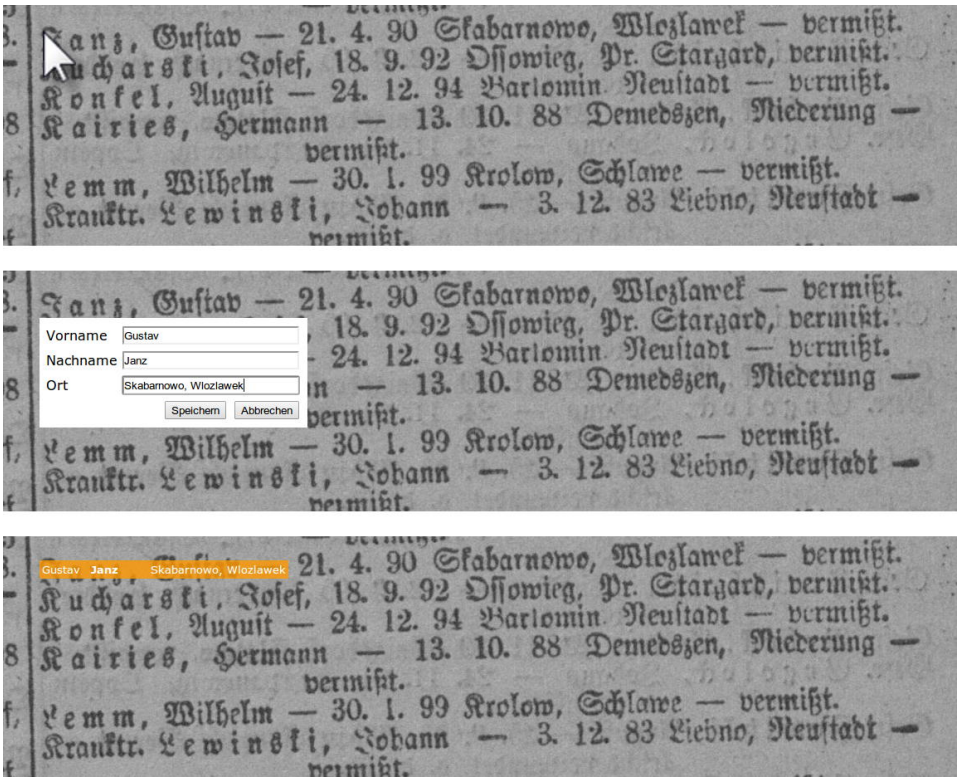


Figure 5. The three steps to enter an entry.

in the browser window with the mouse under the input mask—the position of the initial mouse click is stored as the position of the entry.

4.3 Acquisition of “area data”

In parallel with the transcription by the volunteers, the the so-called acquisition of “area data” is performed. In this process, information is assigned to larger rectangular regions of a scanned page. The type of information depends on the project.

Since in the casualty list entries of a regiment and other units are printed en bloc, one can conveniently draw a frame around these entries. With this one working step one can assign the name of the unit to all entries within this block. Figure 6 shows an example of a page with some already defined areas and the input mask for the data of one new rectangle.

Figure 6. Example for the acquisition of "area data".

4.4 Data cleansing

One action of data cleansing could be the search for systematic deviations from the editing guidelines. For example the editing guidelines for the casualty lists state that noble titles and prefixes are written behind the family name—for example "von Stein" is entered as "Stein, von". It is easy to search the complete dataset for family names starting with "von".

If a normalization of place names is desired, this would also be performed in the phase of data cleansing. By applying a suitable sorting of the complete dataset spelling variations can be spotted easily.

4.5 Export

After a source has been completely transcribed and the data has been cleaned, the data can be exported from CG-DES into a target system. During that export the transcribed data is usually transformed into another data model.

Let us give an example to clarify this step. It is not uncommon that in historic address books one entry contains information about more than one person. For example a single entry lists a widow's name and address but also the first name of the deceased husband. This is transcribed as a single entry. When exporting, two

objects for both persons are created and the transcribed information is assigned appropriately.

5 Scalability

Although so far no performance problems have been encountered, it is likely that a very interesting project attracts a large number of voluntary helpers which will increase the load on the system significantly. Therefore we have been thinking about the scalability of the system, too.

Since the individual projects are independent of each other, a distribution of individual projects to different servers is doable without any problems. The number of projects is therefore not limited by the resources of a single server.

A single project can also be distributed across multiple servers. A central point is the selection and reservation of the next page to be edited. However, this selection and reservation process only accounts for a small fraction of the necessary effort for the processing of a page.

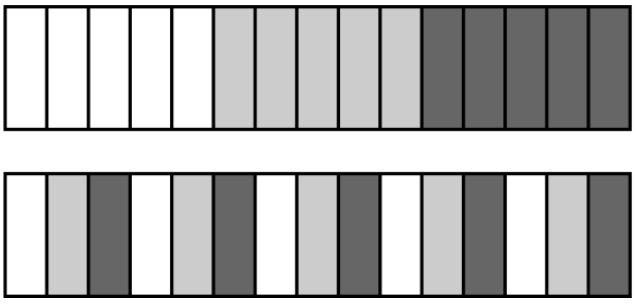


Figure 7. Two possibilities to distribute pages to servers.

After a page has been assigned to a user all communication between the user's browser and the server refers to this one page. Therefore, it is possible to distribute the pages of a project to different servers. Because usually consecutive pages are edited simultaneously, pages should not be allocated block-wise, but in strips as it can be seen in Figure 7. If n servers are available, such a distribution can be achieved by assigning page p to server $(p \mod n) + 1$. A user working on page p only needs resources from this server. The entire communication is performed between the user's browses and this server only.

6 Related Work

There are numerous project to process printed historic sources using crowdsourcing in a web-based environment. In many cases, the scanned image of continuous text is displayed to the user. The user's task is to write the continuous text of the letter in a text box under or next to the scan. These kind of projects include *Transcribe Bentham*[3] [MTW11] [CTW12] [CW12], *FromThePage*[4], *DIY History*[5], and the *Barcelona Marriage Licenses* project [FLM+14].

If the source contains printed text, it can be recognized by machine using optical character recognition (OCR). The result of the OCR process is presented to a human user from the crowd for proofreading and correction. This task can be made—according to the classification of [DRH11]—explicit or implicit. A well-known example of an explicit collaboration is the *Australian Newspapers Digitisation Program* [Hol09]. The approach of an implicit collaboration, in which the improvement of the OCR results is hidden within a computer game was chosen for the National Library of Finland's *Digitalkoot* [CS11].

The aim of the aforementioned projects is the transcription of scanned continuous text. There are also projects for capturing structured data from historic sources using crowdsourcing. As the probably largest project of this kind is *Familysearch Indexing*[6] [HSC+13] should be mentioned here. Initially, it was not a web-based solution. The voluntary helpers needed to install a Java application on their computer. A few years ago, they switched to an online system. The user interface of the application is separated into two areas, the display of the scan and the data entry area.

7 Summary And Outlook

In this article we have presented a novel approach to transcribe structured data from printed historic documents with crowdsourcing. In contrast to previous approaches, the structured data is entered in a web-based system and the data entry is performed directly on the scanned image. The article presents the advantages of a web-based solution and the work "on the image".

[3] http://blogs.ucl.ac.uk/transcribe-bentham/
[4] http://beta.fromthepage.com/
[5] http://diyhistory.lib.uiowa.edu
[6] https://familysearch.org/indexing/

The first project was a success: From January 2011 until August 2014 about 750 volunteers transcribed the German WW1 casualty lists to structured data resulting in more than 8.5 million entries.

We received a lot of positive feedback regarding our crowdsourcing platform. Especially a larger number of request from archives and other organizations that publish historic sources on the internet shows that crowdsouring with a web-based system "on the image" is a good way to transcribe structured data from printed historic sources.

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 a 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 by the users of the data during searches. We have started to investigate how far the data quality can be improved through *double keying* or systematic proofreading—without reducing the processing speed or the motivation of the volunteers too much.

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 on transcription speed and quality.

References

[BHIBL08] Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee. Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, pages 1265–1266. ACM, 2008.

[CS11] Otto Chrons and Sami Sundell. Digitalkoot: Making old archives accessible using crowdsourcing. In *Human Computation*, 2011.

[CTW12] Tim Causer, Justin Tonra, and Valerie Wallace. Transcription maximized; expense minimized? crowdsourcing and editing the collected works of jeremy bentham. *Literary and linguistic computing*, 27(2):119–137, 2012.

[CW12] Tim Causer and Valerie Wallace. Building a volunteer community: re-

sults and findings from transcribe bentham. *Digital Humanities Quarterly*, 6, 2012.

[DRH11] Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4):86–96, 2011.

[FLM+14] Alicia Fornés, Josep Lladós, Joan Mas, Joana Maria Pujades, and Anna Cabré. A bimodal crowdsourcing platform for demographic historical manuscripts. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 103–108. ACM, 2014.

[Fur00] Eli Fure. Interactive record linkage: The cumulative construction of life courses. *Demographic Research*, 3(11):3–11, 2000.

[Hol09] Rose Holley. Many hands make light work: Public collaborative ocr text correction in australian historic newspapers. *National Library of Australia Staff Papers*, 2009.

[How06] Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[HSC+13] Derek L Hansen, Patrick J Schone, Douglas Corey, Matthew Reid, and Jake Gehring. Quality control mechanisms for crowdsourcing: peer review, arbitration, & expertise at familysearch indexing. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 649–660. ACM, 2013.

[MTW11] Martin Moyle, Justin Tonra, and Valerie Wallace. Manuscript transcription by crowdsourcing: Transcribe bentham. *Liber Quarterly*, 20(3/4):347–356, 2011.