A Bimodal Crowdsourcing Platform for Demographic Historical Manuscripts

Alicia Fornés
Computer Vision Center &
Dept. of Computer Science
Ed.O, Campus UAB
08193, Bellaterra, Spain
afornes@cvc.uab.es

Josep Lladós
Computer Vision Center &
Dept. of Computer Science
Ed.O, Campus UAB
08193, Bellaterra, Spain
josep@cvc.uab.es

Joan Mas
Computer Vision Center
Ed.O, Campus UAB
08193, Bellaterra, Spain
jmas@cvc.uab.es

Joana Maria Pujades
Centre for Demographic Studies
Ed.E-2, Campus UAB
08193, Bellaterra, Spain
jpujades@ced.uab.es

Anna Cabré
Centre for Demographic Studies
Ed.E-2, Campus UAB
08193, Bellaterra, Spain
anna.cabre@uab.cat

ABSTRACT

In this paper we present a crowdsourcing web-based application for extracting information from demographic handwritten document images. The proposed application integrates two points of view: the semantic information for demographic research, and the ground-truthing for document analysis research. Concretely, the application has the contents view, where the information is recorded into forms, and the labeling view, with the word labels for evaluating document analysis techniques. The crowdsourcing architecture allows to accelerate the information extraction (many users can work simultaneously), validate the information, and easily provide feedback to the users. We finally show how the proposed application can be extended to other kind of demographic historical manuscripts.

Categories and Subject Descriptors

H [**Information Systems**]: World Wide Web: Web applications: Crowdsourcing: Trust

Keywords

Crowdsourcing; Ground-truth generation; Historical Documents; Document Image Analysis

1. INTRODUCTION

Historical demography scholars use historical documents as source of information. In particular, these documents are usually manuscripts, like birth, marriage, death, or census records. Experts must access to this information physically in situ, unless the documents are digitized. Digitization campaigns are growing worldwide. Digitization allows to preserve documents in a digital format, and to provide universal reading through web portals. It also opens new opportunities and new innovative services. The area of digital humanities bringing together history and technology is an emerging research area related to the analysis of historical digital document images. In the case of historical demographic research, many projects have dedicated important resources to digitize these manuscripts. Examples are FamilySearch, the Ancestry.com World Archives Project [1] [6], or the Mormon Migration project [2], among others.

Once demographic historical manuscripts are digitized, the challenge is the transcription and smart extraction of the contents so the above mentioned services can be provided. These services are aimed to help scholars in the study of genealogies or migratory phenomena with data analytics processes on the transcribed data. But services can be provided also to citizens at large, for example providing tools to search and investigate the lives of their ancestors. Hence, in all these services, the massive transcription of the contents is required. Document analysis research is progressing in handwriting recognition and spotting methods with good performance to massively extract information from document images. However, a fully automatic transcription is still a challenge, and the alternative is the manual transcription of these documents. Although several ground-truthing applications have been developed to make easier the manual transcription and labeling of documents (e.g. DEBORA [5], Pixlabeler [9], ALETHEIA [4], GiDoc [3]), this task is still tedious and time consuming.

An alternative solution to this problem consists in the Crowdsourcing paradigm [2, 10]. Crowdsourcing is an emerging activity consisting in a collaborative transcription or annotation process. The key idea is to split the transcription process in a big amount of small and simple tasks that are distributed among a number of contributor users. There are

.

[1] http://community.ancestry.com/awap
[2] http://mormonmigration.lib.byu.edu/
[3] https://prhlt.iti.upv.es/page/projects/multimodal/idoc/gidoc

different strategies of splitting the work into tasks: in pages, in lines, in words, in semantically significant records, etc. An extreme case can be the use of segmentation algorithms to split the input into snippets of words images so that the users transcribe words individually in a reCaptcha [4] or a game framework such as Digitalkoot [5]. Another strategy is to crowdsource the correction of automatically transcribed handwritten document images [8]. All the crowdsourcing platforms involve document analysis for segmenting blocks or pre-recognize words, and manual tasks for transcribing blocks or correcting the outputs of automatic processes.

In the case of demographic manuscripts, there are several crowdsourcing platforms for transcribing these documents. Examples include the Ancestry.com Keying Tool [6], the Civil War Diaries and Letters Transcription tool [7] [2], the Citizen Archivist Dashboard [8], or the Transcription Project tool [9]. Unfortunately, these platforms are only focused on the needs of demographic scholars, and the output databases are of little use for document analysis research.

For this reason, our goal is to integrate the needs of demographic scholars and document analysis researchers into the same crowdsourcing platform. Concretely, this paper presents an integral bi-modal crowdsourcing platform and the corresponding user experience report in a real scenario of historical demographic documents such as the marriage licenses of the Cathedral of Barcelona. With a user friendly web platform, many volunteers can transcribe and annotate in parallel thousands of documents. The goal is to provide a fully transcribed database for demographic research, and a ground-truth for document image analysis research. From the usage point of view, we distinguish two modalities in the outputs of a crowdsourcing process. First, the mentioned transcription of the contents that we call the *contents or semantic mode*. But a second modality concerns the ground truth data to train or validate the document analysis algorithms required to (semi)automatize the process. Namely, bounding boxes of words or lines, and annotated data related to their properties. We will refer to this second modality of output as *labeling mode*.

The rest of the paper is organized as follows. Section 2 describes the Barcelona marriage licenses. Section 3 describes the crowdsourcing architecture, including the contents and labeling modes, and the output formats. Section 4 describes the running experience from the users point of view. Section 5 is devoted to show that our platform can be easily extended to other kind of demographic documents. Finally, conclusions and future work are described in Section 6.

2. BARCELONA MARRIAGE LICENSES

In 1409, Pope Benedict XIII granted the new Barcelona Cathedral for a tax on marriage licenses. Between 1451 and 1905, a centralized register recorded all the marriages and the fees posed on them according to the social status of the couple. This exceptional documentary treasure (some examples are shown in Figure 1) is still conserved at the Barcelona Cathedral archives, comprising 287 books with information on approximately 600,000 marriages celebrated at 250 parishes in the Barcelona area. The documents were written in Catalan until 1860, and from 1860 until 1905 they were written in Spanish. The documents recorded information about the groom (name, family name, occupation), the bride (name, marital status, and in some cases, her surname) and their parents (name, surname and sometimes, also their occupation). In some cases, the parish and the geographical origin was also included. The amount of the fee depended on their social status, and were fixed in a seven or eight-tiered scale. Each book contains an index with all the groom's surnames (in some cases, the bride's surname is also included).

The information stored in these documents has been used to create the Barcelona Historical Marriage Database, one of the main goals of the European *Five Centuries of Marriages (5CofM)* project (ERC 2010-AdG-20100407). Because of the source's continuity over five centuries, the information extracted from these books will allow demographic and social research, such as genealogical trees, population estimates, migration, survival, intra-generational social mobility, social homogamy, etc.

The structure of the marriage license books (see Figure 2) consists in three blocks: the left column corresponds to the groom's surnames, the main central block corresponds to the license records, and the right column corresponds to the fees. The structure of the license records that are contained at each book is very regular. The information usually appears in the following order: date, husband information, husband's parents, bride information, bride's parents, and place. For more information, the reader is referred to [3, 7].

3. CROWDSOURCING ARCHITECTURE

As stated in the introduction, the proposed bimodal platform integrates the needs from two different research fields. The crowdsourcing architecture that has been proposed (see Figure 3) covers the image space (digitized documents), the transcription space (extraction of information) and the contextual space (where the words have a semantic meaning). In this way, the information extracted from the documents can be used for demographic research. Next we describe the different modules: administration, contents and labeling views.

3.1 Management and Administration

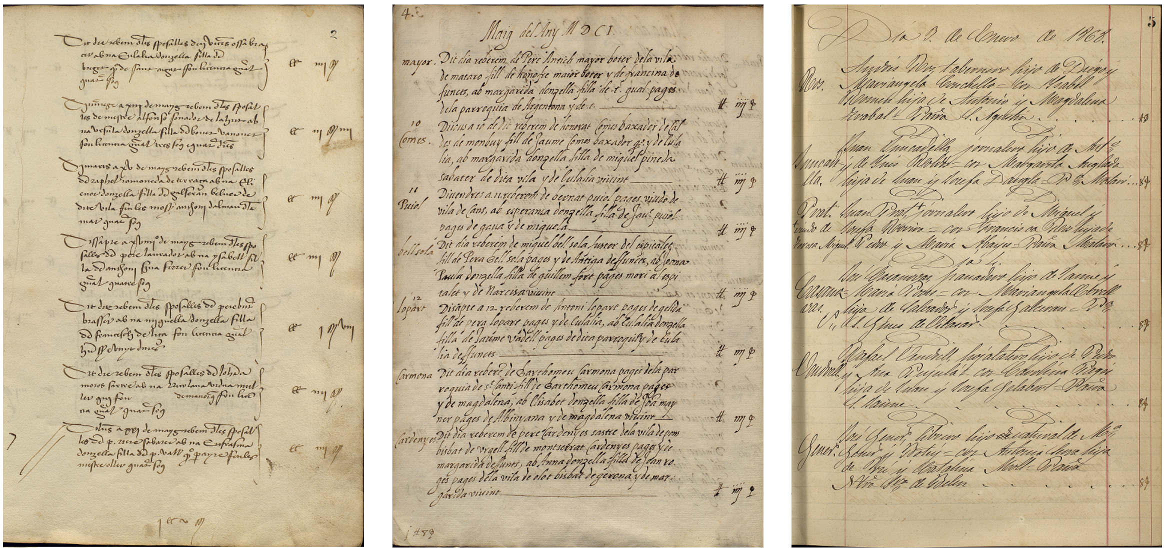
The management of data volumes and users is performed by the administrator, who can add, remove and modify data and users. In addition, the administrator can assign documents to users. Note that the same document can be assigned to several users through different tasks: form filling, transcription, labeling, etc. Whenever the users log in, the welcome screen shows some statistics concerning their finished and remaining work.

3.2 Contents view: Semantics

3.2.1 Form filling: licenses and index

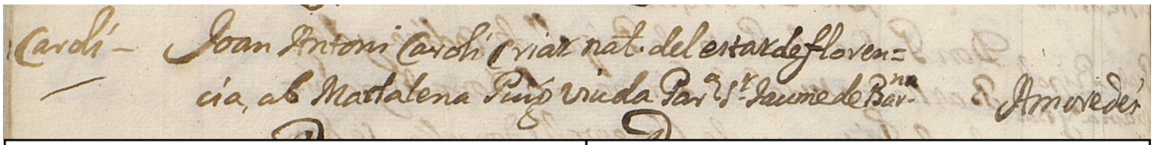
This module is designed for recording the contents of the marriage registers, according to the semantic categories. The generated database is addressed to support the research in historical demography. The transcriber fills in a form with information related to the groom, bride, their parents and the fee. Figure 4 shows an screen-shot of the form filling

[4] http://www.google.com/recaptcha
[5] http://www.digitalkoot.fi/index_en.html
[6] http://www.ancestry.com/wiki/index.php?title=Keying_tool_how-tos
[7] http://diyhistory.lib.uiowa.edu/
[8] http://www.archives.gov/citizen-archivist/
[9] http://transcribe.archives.gov/



License from 1481 License from 1601 License from 1868

Figure 1: Marriage Register Books.



Joan Antoni Caroli criat nat(ural) del estat de florencia, a(m)b Marlalena Puig viuda Par(roquia) s(an)t Jaume de Bar(celo)na Amore dei	Joan Antoni Caroli valet from the state of florence with Marlalena Puig widow from Saint James of Barcelona Parish. Amore dei (no fee)

Figure 2: License register. Left: Catalan. Right: English translation.

module. The screen is divided in several regions: the upper part shows the document image, where the user can zoom in and out for more convenient reading; whereas the middle part shows the different form fields, which have to be filled by the user. The user can also write some comments concerning the marriage register (e.g. the handwriting is not readable, some information is missing or cross-out, etc.). Once all the forms that are contained in a page are completed, the system will mark that page as finished, showing the next pending document page to the user.

An interesting functionality of the application is called the "auto-correction", which is used for detecting spelling mistakes. With this feature, the user can see how many times a specific word (e.g. surname, name, occupation) has been transcribed. Thus, the user can check for possible spelling errors (words that appear only once are more likely to contain spelling mistakes). Finally, the expert user reviews the provided transcriptions and corrects errors if necessary.

3.2.2 *Record Linkage: Genealogical tree*

Since the forms contain the spelling and the semantic meaning of the words (i.e. the word "Montserrat" is the surname of the bride's father in Figure 4), the document can be seen from the knowledge space, and therefore, the contents can be exploited for demographic research. For this purpose, the platform also contains the harmonization and record linkage functionalities. The main idea is that a batch process searches links between individuals described in the books: First, the system searches for the ancestors of the couple: the parent's information has been found as the couple information in a marriage celebrated some years ago. Secondly, the system searches for brothers and sisters of the couple: the parents' information is the same in different registers. The search allows some spelling variations in the names and surnames thanks to the use of the string edit distance, which can be of great help for harmonization. Finally, these links among individuals are shown to the expert who validates or discards them. Note that the system can propose several possibilites for linking two individuals, being the expert the one who decides the correct linkage.

3.3 Labeling view: Annotation

This module is devoted to the literal transcription and labeling of all the words appearing in the document. Note that in the previous module, only the words that contain information concerning the marriage are entered into the database (the markers and stop words are not transcribed). However, for handwriting recognition and word spotting tasks, the ground-truth must contain all these words.

For this purpose, the module developed here is divided into two steps. The first one is designed for transcribing all the words appearing at each text line, so for each register, the user transcribes each text line into a text-box. Note that the transcription has to be done in a literal way, which means

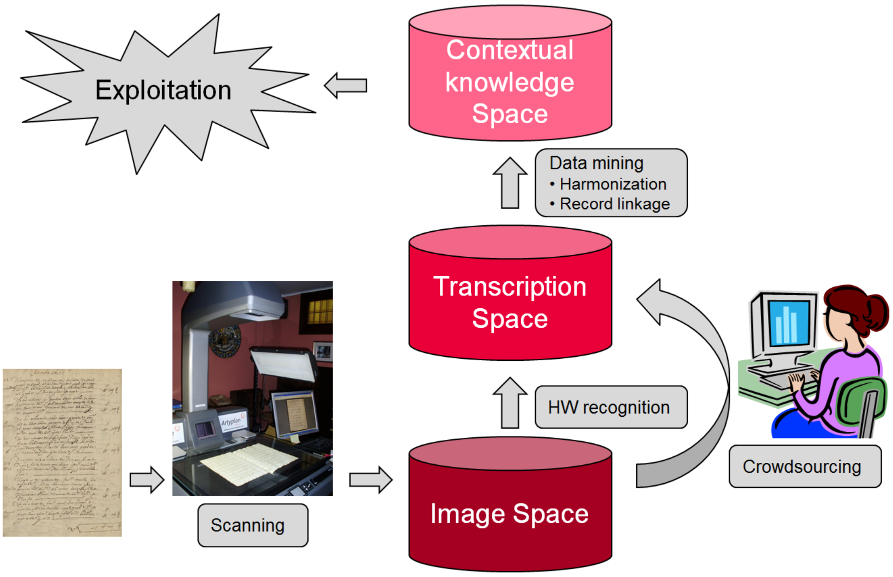


Figure 3: Crowdsourcing architecture.

that spelling mistakes and abbreviations are not corrected.

The second step consists in labeling each word, which means that the user creates the assigned meta-data. The meta-data contains the location and bounding-box of each word in the document, and also the type (semantic category) of word: groom's name and surname, wife's name and surname, occupation, father's name, etc. In case a user had already filled the form corresponding to a marriage record, then the system automatically labels many of these words. Thus, the user only has to validate these automatic correspondences and label the remaining stop words. In this way, our integrated platform allows to put into correspondence the contextual information (contents view) with the words transcription (labeling view).

Figure 4 shows an example of this application module. In the upper part, the document image is shown, with the different bounding-boxes for the words. In the central part, the contextual information form is shown. In the bottom part, the words contained at each text-line are shown. The user clicks every word in the text line and creates the bounding-box that contains it (see the red rectangles in the document image). If the word is related to some field in the form, the user validates the correspondence. Contrary, if the word does not correspond to any field in the form (e.g. it is a determinant), it is automatically labeled as ”other” in the ground-truth.

3.4 Output formats

In this section we describe the output formats.

3.4.1 *Contents mode*

There are two different formats concerning the transcriptions describing record pages in one hand and index pages on the other. Concerning marriage record pages, the CSV file contains the following information: the user who creates it, date of creation, date of modification, volume, page, number of record inside the page. Then, it contains the information about the marriage: day, month, year, fee, parish, information about the groom and his parents or his previous wife in case he is widower, and information about the bride and her parents or her previous husband in case the bride is widow.

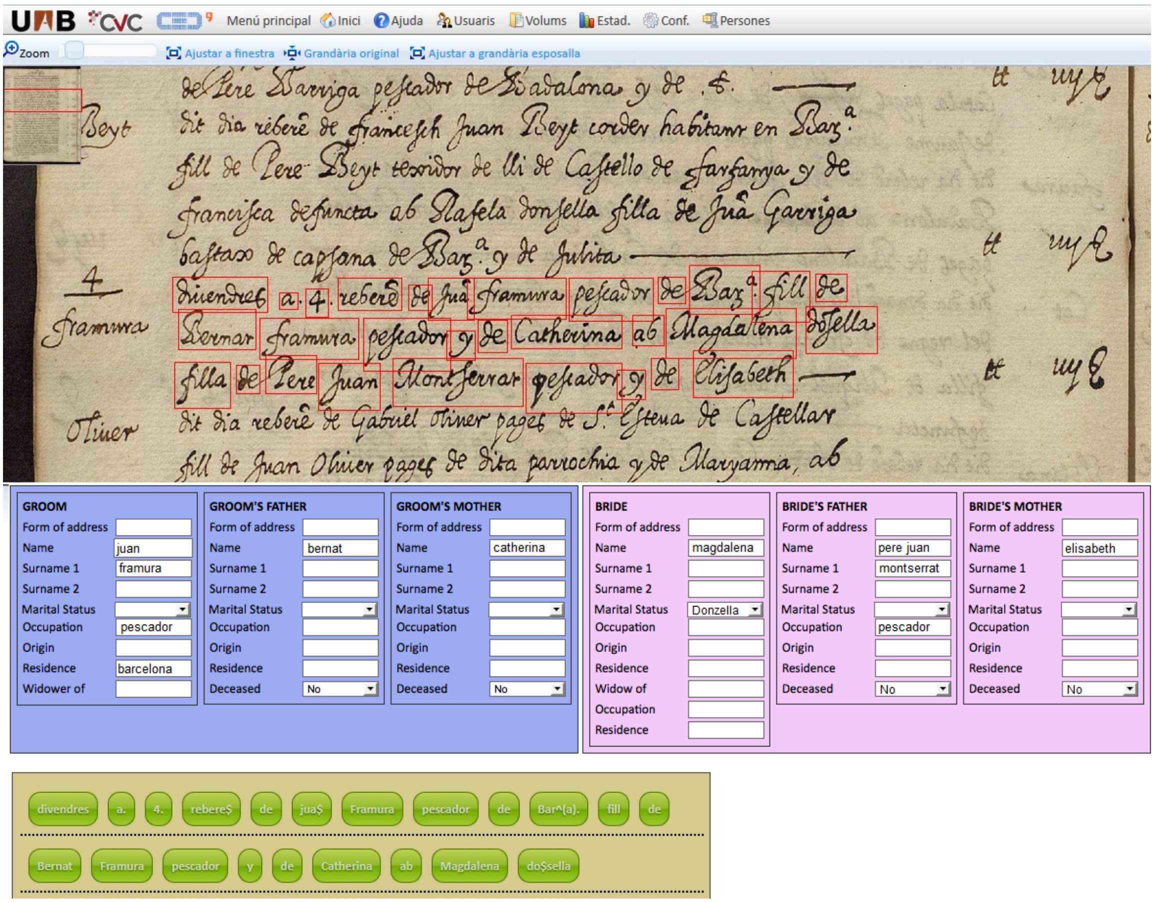
3.4.2 *Labeling mode*

An XML file contains the ground-truth information for evaluating the different document analysis techniques. In this case, we consider a word or graphic element as the atomic unit. We include the following information for each word: location, semantic, and special information. The location defines the physical block where this atomic unit appears, (central, left or right columns in the page), the number of the record in the page, the line number inside a record, the appearing order in the line and, finally, its bounding box. The semantic information includes the reference to a person (e.g. this word is referred to the groom, father-groom, mother-groom, bride, father-bride, mother-bride) and the word type (name, surname, job, place of residence, date, number, etc). Finally, the special information tags define whether the atomic unit is crossed-out, it is a title, a total sum or corresponds to a comment.

4. RUNNING EXPERIENCE

In this section we report on the user experience, according to the volunteers that transcribed and annotated the source documents. The feedback provided by the users was very positive. The main advantages were the following:

- Digital source: the platform containing the digital version of the documents has allowed the users to work from everywhere (it was not necessary to physically go to the Archive, as in the past) and without any timetable limitations (the platform is open 24h a day).
- Parallelization: since more than 150 users were working simultaneously, the extraction of information (contents view) from 287 volumes (with more than 600.000 marriage records) was finished in less than two years.
- Centralization: a centralized platform easies the management of images, users, and the database (forms).



GROOM	
Form of address	
Name	juan
Surname 1	framura
Surname 2	
Marital Status	
Occupation	pescador
Origin	
Residence	barcelona
Widower of	

GROOM'S FATHER	
Form of address	
Name	bernat
Surname 1	
Surname 2	
Marital Status	
Occupation	
Origin	
Residence	
Deceased	No

GROOM'S MOTHER	
Form of address	
Name	catherina
Surname 1	
Surname 2	
Marital Status	
Occupation	
Origin	
Residence	
Deceased	No

BRIDE	
Form of address	
Name	magdalena
Surname 1	
Surname 2	
Marital Status	Donzella
Occupation	
Origin	
Residence	
Widow of	
Occupation	
Residence	

BRIDE'S FATHER	
Form of address	
Name	pere juan
Surname 1	montserrat
Surname 2	
Marital Status	
Occupation	pescador
Origin	
Residence	
Deceased	No

BRIDE'S MOTHER	
Form of address	
Name	elisabeth
Surname 1	
Surname 2	
Marital Status	
Occupation	
Origin	
Residence	
Deceased	No

Figure 4: Crowdsourcing platform. Form filling and word labelling.

The administrator can easily check who is working on each document, assign documents to users, and upload help documents (e.g. dictionaries of surnames, Frequent Asked Questions).

- Automatic control: the system forces the users to fill some required fields (names, surnames), raises warnings whenever the user forgets to fill a register in the page (e.g. the system knows how many marriage registers are written at every page and warns the user if a form is not filled). The system also shows the frequency of the words (such as names, occupations), which helps the user to detect spelling mistakes.
- Security. The system is performing frequent backups (the database and the website). The users can only visualize the documents assigned to them, but not download the images nor the database forms.
- Monitoring. The administrator user can monitor the users' transcriptions, and provide feedback at once whenever it is necessary.
- Visualization and Comfort. The user can drag the page (move), zoom in/out for a better visualization.

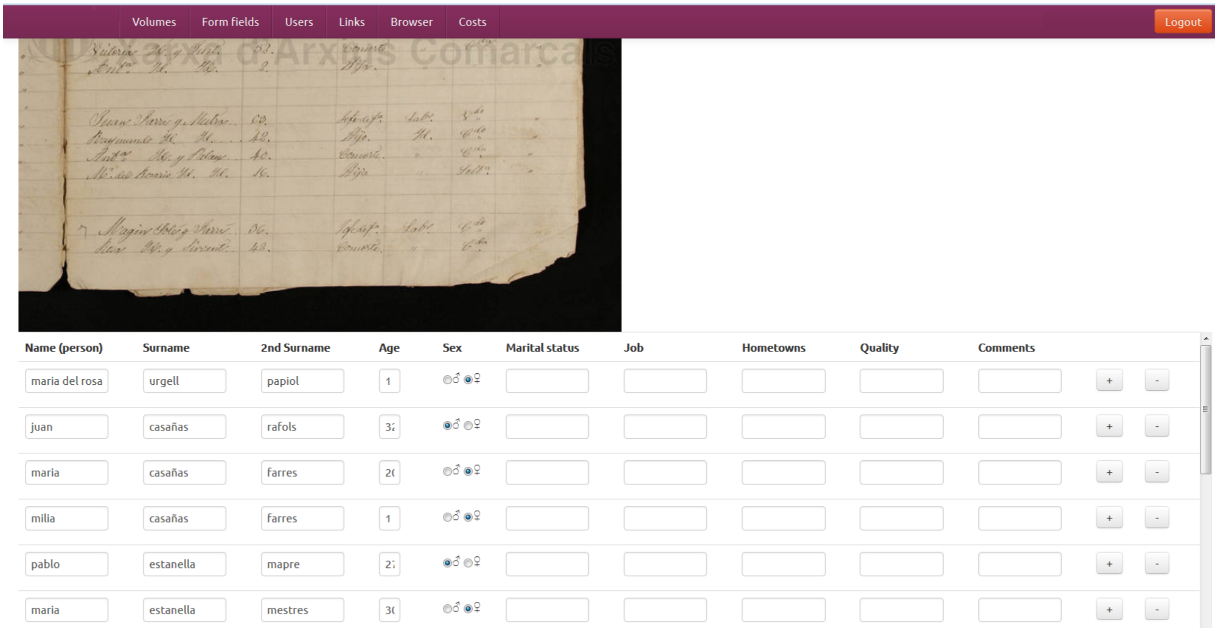
The users also remarked two main disadvantages: there was no possibility to work without internet connection, and secondly, none of the users could work whenever the system was down (internet connection of the server fails, maintenance and upgrading with new functionalities).

5. GENERALIZATION TO OTHER DEMOGRAPHIC MANUSCRIPTS

The platform that has been described for historical marriage register books can be easily adapted to other kind of demographic manuscripts, or even not demographic. Concretely, we have developed a second version of the application (see Fig. 5) for extracting the information contained in census records. The administrator user can upload the different volumes and define the elements that are contained in the census (the information can be different, depending on the kind of census document). Once the volunteer fills the form, there is a batch process that validates the provided information and proposes links among individuals that are living in the same place. Finally, the expert can validate the proposed relations.

6. CONCLUSIONS

In this paper we have described a web-based crowdsourcing platform for demographic manuscripts. The proposed application integrates the needs of demographic researchers as well as computer scientists. The platform has shown to speed up the classical processes of information extraction and ground-truthing. In addition, the centralized applica-



Name (person)	Surname	2nd Surname	Age	Sex	Marital status	Job	Hometowns	Quality	Comments
maria del rosa	urgell	papiol	1	○♂ ●♀					
juan	casañas	rafols	32	●♂ ○♀					
maria	casañas	farres	20	○♂ ●♀					
milia	casañas	farres	1	○♂ ●♀					
pablo	estanella	mapre	27	●♂ ○♀					
maria	estanella	mestres	30	○♂ ●♀					

Figure 5: Crowdsourcing application for census records

tion allows the administrators to monitor and provide feedback to the users in an easy and straight-forward way.

We have recently developed a mobile version [1] for increasing the massive diffusion of crowdsourcing platforms. In addition, mobile devices are user-friendly and more intuitive, and are specially suitable for some kind of tasks (e.g. a touch screen interface is more suitable for obtaining an accurate line and word segmentation).

We are planning to improve the validation of information step. The typical process for detecting errors consists in transcribing the same document by different volunteers and compare the provided information. Since this process is costly, we are planning to use document analysis techniques to extract the information and compare it with the information provided by one single volunteer. In fact, one could even aim for an automatic recognition of these documents, and then the user simply validates the extracted information and corrects errors.

7. ACKNOWLEDGMENTS

This work has been partially supported by the Spanish project TIN2012-37475-C02-02 and the European project ERC-2010-AdG-20100407-269796.

8. REFERENCES

[1] A. Amato, A. Sappa, A. Fornés, F. Lumbreras, and J. Lladós. Divide and conquer: Atomizing and parallelizing a task in a mobile crowdsourcing platform. In *2nd International ACM Workshop on Crowdsourcing for Multimedia (CrowdMM)*, pages 21–22, 2013.

[2] S. Averkamp and M. Butler. The care and feeding of a crowd. In *Code4Lib Conference*, February 2013. http://code4lib.org/conference/2013/averkamp-butler.

[3] N. Cirera, A. Fornés, V. Frinken, and J. Lladós. Hybrid grammar language model for handwritten historical documents recognition. In *Pattern Recognition and Image Analysis*, volume 7887, pages 117–124, 2013.

[4] C. Clausner, S. Pletschacher, and A. Antonacopoulos. Aletheia-an advanced document layout and text ground-truthing system for production environments. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 48–52. IEEE, 2011.

[5] F. Le Bourgeois and H. Emptoz. Debora: Digital access to books of the renaissance. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2-4):193–221, 2007.

[6] A. G. Noll. Crowdsourcing transcriptions of archival materials. In *Graduate History Conference*, pages 1–33, march 2013.

[7] V. Romero, F. A., N. Serrano, J. Sánchez, A. Toselli, V. Frinken, E. Vidal, and J. Lladós. The {ESPOSALLES} database: An ancient marriage license corpus for off-line handwriting recognition. *Pattern Recognition*, 46(6):1658 – 1669, 2013.

[8] V. Romero, A. H. Toselli, and E. Vidal. *Multimodal Interactive Handwritten Text Transcription*. Series in Machine Perception and Artificial Intelligence (MPAI). World Scientific Publishing, 2012. http://www.worldscientific.com/worldscibooks/10.1142/8394.

[9] E. Saund, J. Lin, and P. Sarkar. Pixlabeler: User interface for pixel-level labeling of elements in document images. In *10th International Conference on Document Analysis and Recognition (ICDAR)*, pages 646–650. IEEE, 2009.

[10] M.-C. Yuen, I. King, and K.-S. Leung. A survey of crowdsourcing systems. In *IEEE third International Conference on Privacy, security, risk and trust (PASSAT), and IEEE third International Conference on Social Computing (Socialcom)*, pages 766–773. IEEE, 2011.