Computer-Assisted Transcription by Computer Vision through Citizen-Centered Projects: Crowdsourcing Platforms and Gamesourcing Experiences. The Barcelona Case.

Joana Maria Pujadas-Mora[*], Alícia Fornés[#], Josep Lladós[#], Miquel Valls*, Gabriel Brea[*],

[*] Center for Demographic Studies (CED)

[#] Computer Vision Center (CVC)

Extended abstract

Citizen-centered projects are being common in gathering data for scientific studies in last decades and it has shown to be a productive way of advancing knowledge (Bonney et al., 2009). The engaging of citizens allows to cover larger geographic scale and longer time period for data than it would be possible in more traditional scientific research (Bonney et al., 2009; Curtis, 2015; Wiggins et al., 2018). The expansion of information technologies, the popularization of handheld devices and the integration of internet in everyday life have enabled to develop this specific kind of projects only online (Bonney, 2014; Newman et al., 2012). In spite of pursuing scientific objectives, these projects show to have an important potential of social impact as in terms of literacy or shortening the technological gap (Trumbull et al., 2000; Sunderland et al., 2009; Brossard et al., 2011). Moreover, libraries and archives have devoted great efforts to digitize their historical documentation massively in the last times. Initially, it was oriented towards the preservation of those documentation, however it opens a world of possibilities regarding their access and valorization and it is absolutely beneficial for online citizen projects.

Working with citizens on Historical Demography is hardly new. There is an important precedent on incorporating citizens to collect raw data as The Cambridge Group for the History of Population and Social Structure did for their publication titled *The Population of History of England 1541-1871*. After inventorying the preserved parish registers in England, "*an appeal was then made for local volunteers to help to further the work. They were asked to use standard forms on which monthly totals of baptisms, burials, and marriages were to be recorded from the start of registration until 1837, the year in which civil registration was instituted in England* (Wrigley et al., 1997:6). The aim of this paper is to show the information technology methodologies on engaging citizens used in the research projects CROWDS, TOOLS and NETWORKS in order to build demographic databases using parish and civil registers and census material. Information technologies have been adopted to extract (transcribe) information manually on a crowdsourcing platform (web application) designed ad hoc and by applying computer vision techniques to computer-assisted that manual transcription. At the same time, the extracted information works as a ground-truth for document analysis to develop algorithms of semantic recognition which are useful to train them to perform automatic transcription, which is validated with ad hoc gamesourcing experiences. Besides, the involved citizens in the projects will be monotorized through their sociodemographic profile and their engagement over time using a multivariable analysis.

CROWDS project was titled*: Digital processing of local historical sources. Three experiments in crowdsourcing and user friendly access* (IP: Cabré, A.; Pujadas-Mora, J.M.) and funded by the Spanish Ministry of Economy for the period 2014-2016. TOOLS project had the title: *Tools and procedures for the large scale digitization of historical sources of population* (IP: Lladós, J.; Esteve, A.; Pujadas-Mora, J.M.) and was funded by RecerCaixa program - Obra Social "la Caixa"

for the period 2015-2017. NETWORKS project which is still ongoing is called: *Technology and citizen innovation for building historical social networks to understand the demographic past* (IP: Fornés, A., Pujadas-Mora, J.M.) and funded also by RecerCaixa program. They have been conducted by a team that includes researchers from the Geography Department of Universitat Autònoma de Barcelona, its Center for Demographic Studies (CED), and its Computer Vision Center (CVC). That is to say the team involves demographers, historians, geographers and computer scientists. This team was originally created to respond to the call for Advanced Grant projects by the European Research Council. In this way, the project 'Five Centuries of Marriages' (ERC_2010_Advanced Grant, AdG_269796) was obtained for the period 2011-2016 and directed by Prof. Anna Cabré. This project inspired the abovementioned projects, either in the conjunction of Historical Demography and Computer Sciences and the massive involvement of transcribers through a crowdsourcing web-based application to collect the information of original sources. Even though those transcribers were paid, their massive incorporation (up to 110 transcribers) corresponds with the paradigm of crowdsourcing, which is to split the work into many micro-tasks and ask contributions from a large group of people, especially from the online community (Thorvaldsen et al., 2014; Pujadas-Mora et al. 2016).

Within these projects three databases have been created: The *Baix Llobregat Demographic Database* (*BALL*), which is an ongoing database containing individual census data from the Catalan county of *Baix Llobregat* (Barcelona, Spain), up to now, of 9 different municipalities for the nineteenth and twentieth centuries; the Sant Climent parish register database (Barcelona, Spain) which contains the baptism, the marriages and the death occurred in the parish from 1607 to 1975 and the Montefrío civil register database (Granada, Spain) gathering birth, marriages and deaths for the period 1841 to 1867 (see Table 1).

Table 1: Number of registers in each database.

Database	Registers
The Baix Llobregat (BALL)	221.667
Montefrío	12.395
Sant Climent de Llobregat	26.944
Total	261.006

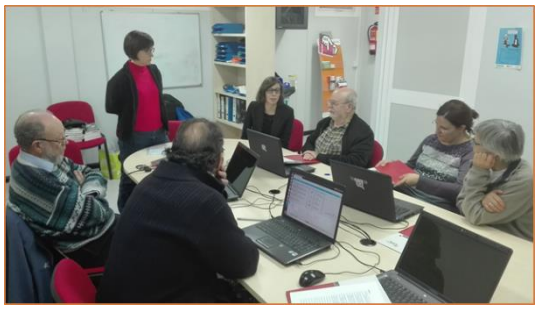
Source: Authors' own elaboration.

Recruiting volunteer citizens and their sociodemographic profile

Thanks to the collaboration with the Regional Archive of the *Baix Llobregat* for the projects TOOLS and NETWORKS and with the Municipal Archive of Montefrío and the local study group "Més de 1000" in Sant Climent for the project CROWDS, groups of volunteer transcribers were founded in every municipality included in the projects. The project CROWDS included the municipalities of Montefrío (Granada, Spain) and Sant Climent (Barcelona, Spain), the project TOOLS, Sant Feliu de Llobregat (Barcelona) and the project NETWORKS, up to February 2019, Castellví de Rosanes, Collbató, Corbera de Llobregat, El Papiol, Molins de Rei, Sant Feliu de Llobregat, Santa Coloma de Cervelló, Torrelles de Llobregat and Begues (Barcelona, Spain).

Figure 1: Pictures of groups of volunteers.

Molins de Rei



Corbera de Llobregat

These volunteers have been properly trained by the researchers of the projects, mainly by the principal investigators, with several sessions on the aims of the projects and the history and features of the demographic sources which they would be transcribed (See Figure 1). Moreover, a specific formation was given to show how the online data entry tool works, which will be presented in the next section, to explain the adopted rules of transcription and to provide some guidelines on paleography. In fact, these sessions succeed to engage the volunteers. Some of the first volunteers signed new other volunteers. Other sessions are carried out regularly to show the obtained results using the data that they are creating. Close to 200 volunteers has been participated in those three projects (see Table 2). Almost all the volunteers turned out to be enthusiasts with local history or/and genealogy, which gave to them a previous knowledge on the source material that they transcribe. Nevertheless, they show diverse cultural backgrounds. In total, 83 men and 77 women have been collaborated. In BALL database the volunteers were mainly people with university degrees with a mean age of 62 years, in Sant Climent database the collaborations were mainly retired showing a mean age of 58 years old and in Montefrío database were mainly unemployed people with graduated studies with a mean age of 43 years. Moreover, we also know by each of those volunteers the number of transcribed records, the number of doubts for each transcribed record and the time devoted to the project. In this way, the sociodemographic variables as well as those variable related to the engagement of the volunteers can be analyzed through multivariable analysis to determine if emerges a determined profile of the ideal transcriber in terms of productivity and dedication, as we will performance by the time of the SSHA conference.

Table 2: Number of transcribers from the projects CROWDS, TOOLS and NETWORKS.

Database	Women	Men	Total
The Baix Llobregat (BALL)	61	59	120
Montefrío	6	9	15
Sant Climent de Llobregat	10	15	25
Total	77	83	160

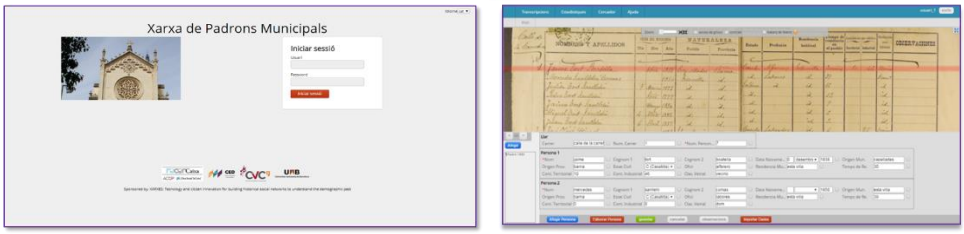
Source: Authors' own elaboration.

Engaging citizens through crowdsourcing platforms & videogames to build demographic databases.

All those projects have in common that the citizens have been involved in the construction of demographic databases through a crowdsourcing web-based application and gamesourcing experiences. The crowdsourcing app integrates two research perspectives to extract information from demographic handwritten document images: the semantic information for demographic research, and the ground-truthing for document analysis research. Concretely, the

application has the contents view as a data entry tool, where the information is typed into forms by the volunteers (see Figure 2), and the labelling view, also made by the volunteers, which corresponds to the annotation (location) over the image of each line and word. In this way, the typed data is also used to train the *ad hoc* designed computer vision algorithms based on deep learning to read the original sources (manuscripts) in order to improve the efficiency of the automatic transcription and also to adapt itself to every new handwriting style. As a result, these algorithms can be used to speed-up the transcription in two different scenarios: first, to assist the transcription via information transfer, and second, to perform an automatic transcription with manual validation through the use of videogames. Both scenarios look for facilitating the manual transcription done by the volunteers.

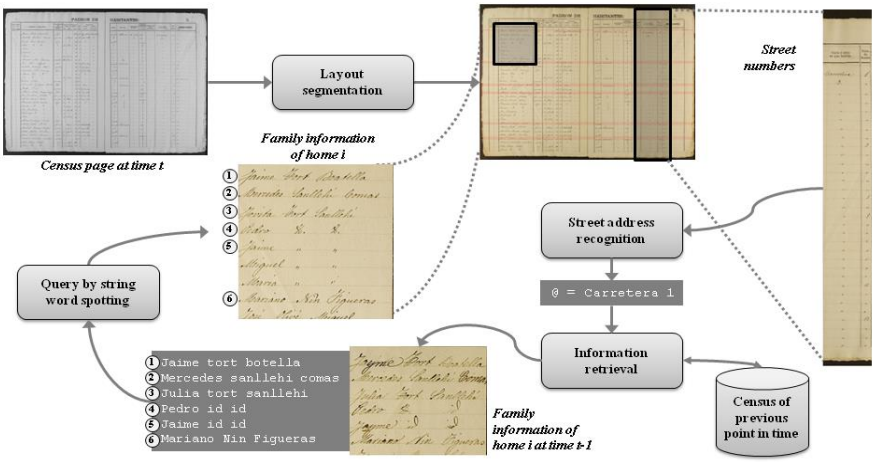
Figure 2. Crowdsourcing platform: http://dagapp.cvc.uab.es/PadronsXarxes/



Source: Authors' own elaboration

The first scenario is designed for building databases using censuses. In fact, the redundancy in censuses, due to they were carried out in intervals of few years, can be used to automatically transfer the repeated information. to assist the transcription. Once a census has been manually transcribed, the redundant information (names, surnames and address) is transferred to the next one, so it is only necessary to manually update the changes: adding new members or deleting those who leave or died for each household. For this purpose, household records of consecutive censuses are automatically aligned using the street address. Then, the individuals are automatically located using word image search (namely, word spotting). Concretely, the individual names and surnames from a census are searched in the corresponding home record in the next census (see Figure 3). Since the process is based on a focused search, the accuracy is very high. In this way, a 70% of reduction in the transcription time of a new census is obtained (Mas et al., 2016).

Figure 3. Architecture of the proposed system in first scenario.



Source: Mas et al., 2016

In the second scenario, the transcription is performed fully automatic, followed by a manual validation. This scenario is applied to transcribe any of the mentioned demographic sources. First, snippets corresponding to word images are segmented from the manuscripts. Then, these word images are clustered to find high frequency words that can be jointly transcribed using a small percentage of representative instances. Afterwards, the words in each cluster have been sent to the deep learning-based transcription system, which provides, the most plausible transcriptions for each word image (see Figure 4). Finally, clusters that contain words with the same transcription, can be transcribed at once. In this way, we can avoid the validation of every single word, speeding up the transcription while maintaining the performance. The validation has been done through game-sourcing (a serious game based strategy for crowdsourcing). Concretely, two Android games have been developed (Chen et al., 2018). The first game is designed to validate the word clustering algorithm. The second game is designed to validate the output of the transcription algorithm (see Figure 4). Up to now, these videogames have been played by selected users showing that the transcription effort can be significantly reduced, and also, the user engagement is higher than with the crowdsourcing web-based application. The performance of both games can be watched at: https://goo.gl/4PVupn.

Figure 4. Android games developed in second scenario.

Source: Chen et al., 2018

Quoted references

Bonney, R.; Cooper, C.; Dickinson, J.; Kelling, S.; Phillips, T.; Rosenberg, K.; Shirk, J. (2009). "Citizen science: a developing tool for expanding science knowledge and scientific literacy." *BioScience* 59(11): 977-984.

Bonney, R.; Shirk, J.; Phillips, T.; Wiggins, A.; Ballard, H.; Miller-Rushing, A.; Parrish, J. (2014) "Next steps for citizen science." *Science.* 343.(6178): 1436-1437.

Brossard, D.; Lewenstein, B.; Bonney, R. (2011). "Scientific knowledge and attitude change: The impact of a citizen science project". *International Journal of a Science Education*, 27(9):1099-1121. https://doi.org/10.1080/09500690500069483

Chen, J.; Fornés, A.; Mas, J.; Lladós, J.; Pujadas-Mora, J. M. (2017). "Word-Hunter: Speeding up the transcription of historical documents through gamesourcing". *16th International Conference on Document Analysis and Recognition (ICDAR)*:528-533

Curtis, V. (2015). "Online citizen science projects: an exploration of motivation, contribution and participation". Ph.D. thesis. The Open University. URL: http://oro.open.ac.uk/42239/.

Fornés, A.; Lladós, J.; Mas, J.; Pujadas-Mora, J.M.; Cabré, A. (2014) "A bimodal crowdsourcing platform for demographic historical manuscripts". *Digital Acces to Textual Culture Heritage 2014 DATeCH '14 Proceedings ofthe first international conference on digital access to textual cultural heritage*, 103-108. DOI: 10.1145/2595188.2595199

Mas, J.; Fornés, A.; Lladós, J. (2016). “An interactive transcription system of census records using wordspotting based information transfer”. *Proceedings of the International Workshop on Document Analysis Systems (DAS) Santorini : 54-59. doi: 10.1109/DAS.2016.47*

Newman, Greg, et al. (2012) "The future of citizen science: emerging technologies and shifting paradigms". *Frontiers in Ecology and the Environment*. 10 (6): 298-304.

Pujadas-Mora, J.M.; Fornés, A.; Lladós, J.; Cabré, A. (2016). “Bridging the gap between Historical Demography and Computing: Tools for computer-assisted transcription and analysis of demographic sources”.*The Future of Historical Demography: upside down and inside out*. Leuven, Acco.: 222 - 226.

Sunderland, T.; Sunderland-Groves, J.; Shanley, P.; Campbell, B. (2009). “Bridging the gap: how can information access and exchange between conservation biologists and field practitioners be improved for better conservation outcomes?”. *Biotropica*, 41(5), 549-554.

Thorvaldsen, G., Pujadas-Mora, J., Andersen, T., Eikvil, L., Lladós, J., Fornés, A. & Cabré, A. (2015). “A Tale of Two Transcriptions. Machine-Assisted Transcription of Historical Sources”. *Historical Life Course Studies*, 2: 1-19.

Trumbull, D., Bonney, R., Bascom, D., & Cabral, A. (2000). “Thinking scientifically during participation in a citizen-science project”. *Science Education*. 84: 265–275.

Wiggins, A.; Bonney, R.; Lebuhn, G.; Parrish, J.; Weltzin, J. (2018). “A Science products inventory for citizen-science planning and evaluation”. *Biosciecne*, 68(8); 436-444. doi: 10.1093/bioscience/biy028.

Wrigley, E.A.; Davis, R.S.; Oeppen, J.E.; Schofield, R.S. (1997). *English Population History from Family Reconstitution.* Cambridge, Cambridge University Press.