Bridging the gap between historical demography and computing: tools for computer-assisted transcription and the analysis of demographic sources

Joana Maria Pujadas-Mora, Alicia Fornés, Josep Lladós & Anna Cabré

The construction of large-scale databases in the field of historical demography has proliferated in recent decades as a result of the development of information technology and the specific financing of research projects. Indeed, it is said that we are in the midst of a Big Data Revolution (Ruggles 2012). However, data building is still a time-consuming process, mainly due to manual data entry and the lack of digital copies of the original sources.

Massive digitization of historical sources has become customary nowadays. As a result, digital copies can now be accessed on-line, but generally through platforms that only offer visualization functionalities. Knowledge becomes dematerialized as information is extracted, organized semantically into large databases, stored and valorised. Another advantage of this trend is that it offers possibilities to standardize the data as part of the process, which eases subsequent or future nominative record-linkage. There is an emerging trend in the development of web-based crowdsourcing platforms that allows people to type in data online, thus splitting this task among large numbers of transcribers.

The advances in the field of computer vision, and in particular in the sub-field of document image analysis and recognition, make the automation of some tasks feasible. During the last decade, scholars in the humanities and computer scientists have, to their mutual benefit, started to work together in the emerging discipline of Digital Humanities. Computational algorithms and services arising from this research activity are gaining relevance, as they start to be integrated into crowdsourcing platforms as assisting tools for scholars and transcribers at large. We can identify two major categories of tools; namely, tools for recognition and tools facilitating the understanding of image contents.

Since sources are usually manuscripts, handwriting recognition techniques (Romero *et al*. 2013) are at the heart of the first category. Handwritten text-recognition consists in automatically transcribing the content of an image into a text; in other words, to convert an image (in pixels) into its textual representation (typically ASCII), which can be later managed using a text-processing application. For this purpose, most existing technologies consist of an optical model (such as hidden Markov-models or neural networks) for modelling the appearance of the characters. These are integrated with dictionaries and language models for lexical, syntactical and structural validation.

A particular scenario of handwriting recognition is called *word spotting* (Mas *et al*. 2016). There are situations where the recognition of images is very difficult and/ or suffers from a high error rate. This occurs when images are ‘noisy’, due to the physical degradation of a document, or the use of old

scripts and languages, or the compilation of the document by multiple writers, etc. In such cases, the strategy of word spotting proposes a holistic approach in which words are treated as visual patterns. Instead of splitting the input into small units (letters, graphemes, etc.), words are recognized on the basis of their shape, using some visual features. Word spotting can be used to directly retrieve the pages where a given query appears; in the transcription process; to find links between registers; or to cluster named entities that frequently reappear.

However, a single literal transcription of the documents is useless for the purpose of analysis (for example, to generate genealogies with these data, to establish individual and family life spans, and to spatially locate family networks). To understand a document, we need to be able to semantically analyze and categorize its content. It does not suffice to merely recognize a word; we also need to be able to tag it as being a name, an occupation, a date, etc. A key concept in the activity of document understanding is the use of contextual knowledge. Document sources are highly heterogeneous. Generally used tools for document recognition (for example, line or word segmentation, writer identification, word recognition or word spotting) are not generic enough to perform well on different types of documents from different periods that are written in different scripts on different topics. This is why we need to use contextual knowledge. Two categories of contextual knowledge can be defined. The first is intrinsic contextual knowledge, which refers to contextual information that can be derived from the document itself. Such information may concern the relationships between and the frequency of use of terms in the document; for example, the presence of one term increases the probability of another one. The second type of contextual knowledge is extrinsic and concerns the correlation and cross-linkage between data on separate pages or in different sources, as well as the knowledge provided by an expert (for example, the socio-economic or temporal context in which the document was written). The use of contextual knowledge allows us to adapt recognition and interpretation tasks to the domain of the processed documents.

We have implemented the architecture described above in historical demographic settings (see Figure 1). In particular, we have constructed the Barcelona Historical Marriage Database in the context of the EU-ERC Advanced Grant project ‘Five Centuries of Marriages’. Currently, we are constructing databases on the basis of census records, such as the 19 censuses held between 1828 and 1955 in the Catalan town of Sant Feliu de Llobregat. In all our research projects, both past and present, researchers from both historical demography and computer vision are brought together to share their insights. All the projects have also included a crowdsourcing task (see Thorvaldsen *et al*. 2015 for more information). More than 200 transcribers, some of them volunteers, participated in the projects. They not only helped with the crowdsource-based transcription, but were early adopters of the services arising from the research. Their valuable feedback helps our interdisciplinary research team to solve new research challenges. The implementation of handwriting recognition and word spotting techniques can identify frequent words, thereby speeding up the transcription made by users (they only need to type once some of the names that occur very frequently).

At the most sophisticated level of automatization, we take advantage of the extrinsic and intrinsic contextual knowledge offered by the documentation to automate processes. For example, the censuses from Sant Feliu de Llobregat were recorded in intervals of just a few years and the information on individuals in each household was quite stable from one point in time to the next. This seeming redundancy of information is used to assist the transcription. The redundant information is transferred from the census already transcribed to the next one – a process that is facilitated by automated searches for family member names that correspond to the same household record, using word spotting procedures.

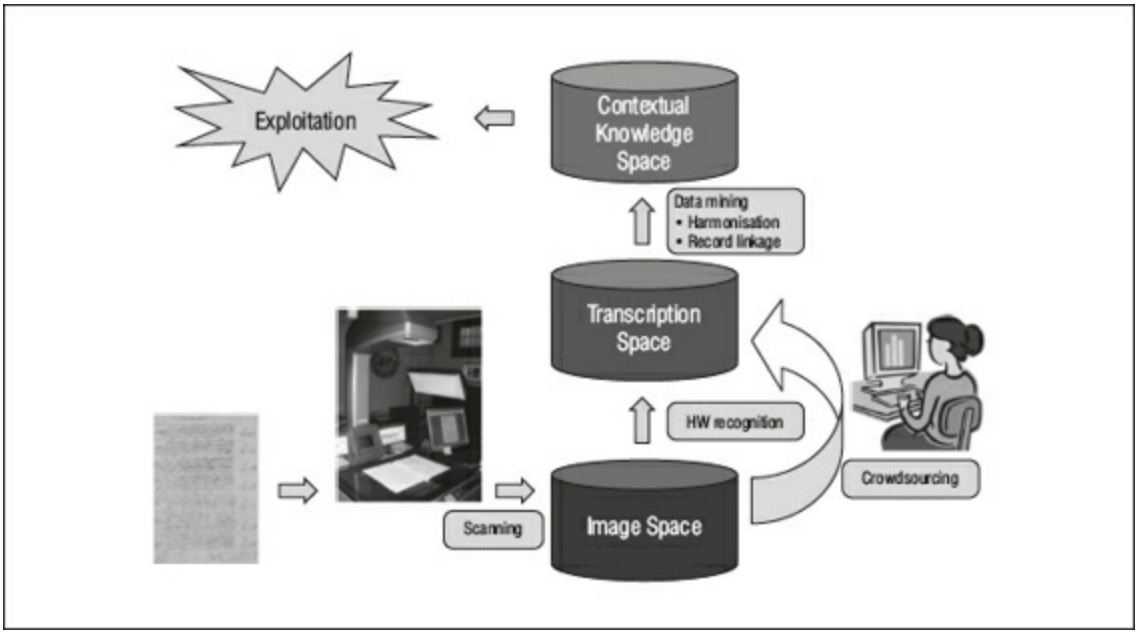

Figure 1. Technical architecture

The semi-automatization of data entry and the definition of standard formats for demographic databases, such as the Intermediate Data Structure (IDS) proposed by Mandemakers and others (see Alter & Mandemakers 2014), lead us to envision the creation of a kind of 'social network' of the past, similar to the way that people today are connected to each other through Facebook and other platforms. The use of artificial intelligence techniques permits analyses of their habits, their preferences and their social behaviour. We could achieve the same for historical populations by applying similar data analysis strategies to demographic data extracted from historical sources. Three key factors will be required to achieve this goal. First, the massive processing of scanned sources. Automatic reading software, adapted to different writers, languages and scripts, will be essential. Second, we need interoperable databases. This means more than standard formats and connectors across different platforms. The architectures underlying the database systems must also be flexible and dynamic enough to adapt themselves to the increases in and the enrichment of the data they contain. Third, the interpretation of the data contained in the databases will require that they be integrated with knowledge provided by people. This knowledge is the so-called 'natural archives', maintained by humans as memories of their societies.

Against this optimistic scenario, it should be remembered that fully automatic reading systems, which can operate on any source, are not a realistic expectation. Human intervention will always be needed. But this raises another challenge: how to place the user in the transcription loop in an efficient and effective way. From past experience of crowdsourcing the transcription process (Fornés *et al*. 2014), we can conclude that humans tend to introduce errors. Moreover, the transcription task tends to become tedious for the people involved. Redundancy in the transcription of some critical sources is necessary. However, this redundancy should be designed in a smart way; for example, with human and machine transcribers working in parallel. It also is vital, according to us, that the transcription activity is integrated into engaging platforms. Gamesourcing is an emerging paradigm that is worth considering.

In conclusion, in the mid-term future citizens of many countries will be able to navigate through networks of knowledge constructed from large-scale and cross-community demographic databases. This will generate new services for the interpretation of the past, not only for scholars but also for wider groups of the public. The incorporation of powerful image recognition tools will be at the heart

of data entry software. It will provide the computational power for semi-automatically processing large document collections, creating databases in a faster and more effective way. At the same time, interdisciplinary and cooperative work is needed to drive the construction of these databases. This interdisciplinarity should consist of a symbiosis between historical demography and computer science. In this mutually beneficial relationship, the demographers provide the historical, social and linguistic knowledge that allows the computer scientists to design algorithms adapted to the document sources.

References

Alter, G. & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies*, 1, 1-26.
DOI: http://hdl.handle.net/10622/23526343-2014-0001?locatt=view:master

Fornés, A., Llados, J., Mas, J., Pujadas-Mora, J. M. & Cabré, A. (2014). A bimodal crowdsourcing platform for demographic historical manuscripts. In the Digital Access to Textual Cultural Heritage Conference (DATeCH), pp. 103-108. Online publication, see http://dl.acm.org/citation.cfm?id=2595199.

Mas, J., Fornés, A. & Lladós, J. (2016). An interactive transcription system of census records using word-spotting based information transfer. *Conference presentation* at the 12th IAPR International Workshop on Document Analysis Systems (DAS).

Romero, V., Fornés, A., Serrano, N., Sánchez, J. A., Toselli, A.H., Frinken, V., Vidal, E. & Lladós, J. (2013). The ESPOSALLES database: an ancient marriage license corpus for offline handwriting recognition, *Pattern Recognition*, 46 (6), 1658-1669.
DOI: http://dx.doi.org/10.1016/j.patcog.2012.11.024

Ruggles, S. (2012). The future of historical family demography. *Annual Review of Sociology*, 38, 423-441.
DOI: http://dx.doi.org/10.1146/annurev-soc-071811-145533

Thorvaldsen, G., Pujadas-Mora, J. M., Andersen, T., Eikvil, L., Lladós, J., Fornés, A. & Cabré, A. (2015). A tale of two transcriptions. *Historical Life Course Studies*, 2, 1-19.
DOI: http://hdl.handle.net/10622/23526343-2015-0001?locatt=view:master

Acknowledgments

This work is part of the Advanced Grant Project Five Centuries of Marriages (IP. Anna Cabré, 2011-2016) funded by the European Research Council (ERC 2010-AdG-269796).

Biographies

Joana Maria Pujadas-Mora is a researcher at the Department of Geography of the Universitat Autònoma de Barcelona and coordinator of the Advanced Grant Project 'Five Centuries of Marriages'. Her main research interests are mortality, migration and marriage during the Ancien Régime and the Demographic Transition. She also engages in the construction and standardization of nominative databases, and in record linkage.

Alicia Fornés is a researcher at the Universitat Autònoma de Barcelona and the Computer Vision Centre. Her research interests include document image analysis, historical documents, handwriting recognition, symbol recognition, optical music recognition and writer identification. Since 2011, she has been a member of the leadership team of the IAPR TC-10 (Technical Committee 10 on Graphics Recognition).

Josep Lladós is an Associate Professor at the Computer Sciences Department of the Universitat Autònoma de Barcelona and director of its Computer Vision Centre. His current research fields are document analysis, structural and syntactic pattern recognition, and computer vision.

Anna Cabré is Emeritus Professor of Demography and Geography at the Universitat Autònoma de Barcelona and honorary director of

the Centre d'Estudis Demographics (CED). She is the principal investigator of the ERC-Advanced Grant project 'Five Centuries of Marriages', which has created the Barcelona Historical Marriage Database, containing information about 615,000 marriages celebrated in the Barcelona diocese between 1451 and 1905.

Historical population databases and the Intermediate Data Structure, 1980-2050

Kees Mandemakers

Datasets with historical demographical data were limited around 1980. But up in the North, a number of scientists with a 'cool' mind had already started ambitious databases with longitudinal data (Umea, Chicoutimi). Thirty-five years later, there has been an enormous expansion of datasets with historical population data of this kind.

In the 1990s, IPUMS (*Integrated Public Use Microdata Series Project*) started building systematic structured samples of the American census from 1850 onwards, which had already been further expanded before the turn of the millennium with the advent of two other major projects: a) The North Atlantic Population Project, with 100% population coverage for the USA, the UK and other Anglo-Saxon countries, and b) the start of IPUMS International, which rescues census data from all over the world. Taken together, all the IPUMS projects will number over 2 billion personal records by 2018 (see weblinks and Ruggles 2014). In Europe, the MOSAIC Project was started some ten years later. It collects census data from all over Europe and now includes almost 1 million persons (see weblinks).

IPUMS and MOSAIC are collections of static data, which refer to the moment of census-taking. From 1980 onwards, datasets with dynamic longitudinal data have grown enormously in all directions. Currently, there are now about 40 to 50 serious databases with longitudinal data worldwide, of which 30 are systematically described on the website of the European Historical Population Samples network (see weblinks). Until now, no-one has made the effort to count the number of persons in all these datasets, but a rough calculation suggests that the total must be at least 20 million. In addition to this, there are a lot of small-scale datasets that were mostly built by individual researchers. Their expansion generally halted when the researcher lost his or her interest in the subject.

What will the situation be like in 2050? For the Netherlands, given the enormous flow of activity in indexing by genealogists and archivists over the last decade, it is expected that all genealogical sources will be scanned and indexed – at least the information needed for identification. I expect the same development for most of the countries in northwestern Europe, the USA and Canada. The UK has already made the complete censuses of 1850-1910 available for scientific research. Their next step will be the inclusion of church registers and the linking of all person appearances (Schürer2007). We may call such a database semi-longitudinal, since it links different points in time without following persons day by day. There is no doubt that by 2050 IPUMS will have expanded its 100% count for 1850, 1880 and 1940 to all the censuses between 1850 and 1960, and will have linked them as well. The existing longitudinal databases will have expanded in many different directions: covering longer periods; covering more and larger regions; linking with modern

population registrations; and creating long family trees.

During the last ten years, we have seen the introduction and development of the so-called Intermediate Data Structure (IDS). This is an open data structure that provides a technical solution for disseminating data from historical population databases in a harmonized way (Alter, Mandemakers & Gutmann 2009; Alter & Mandemakers 2014). Figure 1 shows the various stages of the processing of person data. Each database uses several sources that differ in detail, but overall are more or less the same in terms of their basic structure. By converting these data into a common data structure, it becomes possible to use generic software for building datasets for analysis – so-called extraction software. The IDS presupposes that important integrating tasks, such as standardizing, dating and linking persons, will be performed by the database owners or creators themselves, since they have the best knowledge of their own sources. Of course, they can learn from each other and adopt techniques from other databases.

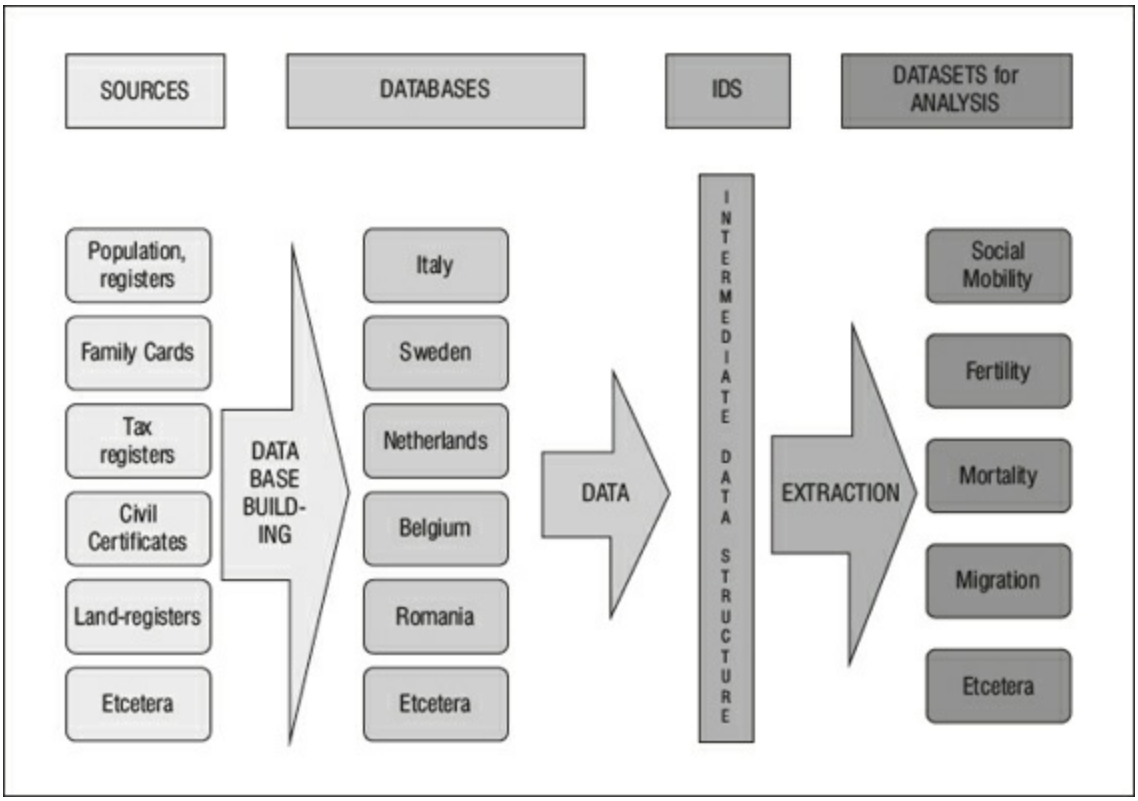

Figure 1. From source to harmonized datasets for analysis via the IDS structure

Since its official introduction in 2009, the IDS has been adopted by more than ten databases, including some very large ones, like DDB Umea. The production of software has taken off as well, both for small and bigger extraction components (a.o. Quaranta 2015). As historical demographers, we can therefore be very satisfied, especially when we realize that most of the work has been (and is being) done on a voluntary basis by professionals, who all adhere to the idea of a single collective data structure. In this sense, the demographic world has fully entered the 'open access' movement.

What are the challenges for the IDS during the next ten to fifteen years? I can see three key issues: a) outreach, b) maintaining standardization and c) integration.

Outreach is hampered by an old problem: lack of education. Historical demographers with a humanities background do not always have the skills to work with these big datasets. Teaching in statistics, database handling, etc. is very poor in most history faculties, especially at the bachelor level. This situation will not improve easily, since many history students are not fond of these more

technical subjects. In the very first article on IDS, Alter *et al.* (2009) suggested a three-step structure to overcome this problem: 1) courses on methodology; 2) easy data files; and 3) extraction software. The EHPS network has started to take up these issues by organizing a summer school system and by organizing the process of building and disseminating extraction software. In comparison, the second remedy – the construction of easy data files – is lagging behind, but this is only a question of time. However, maintaining the current network is highly dependent on the willingness of scholars to organize courses and to develop software. Another solution for the 'skill' problem is greater interdisciplinary cooperation, since only a few scientists are able to cover all the necessary aspects of research (theory, statistics and data handling).

Maintaining standardization and preventing 'dialectization' of the IDS is another continuing challenge. There is a risk of databases choosing new values or variables without consulting the community, which can lead to different variables and values for the same content. The solution here is to continue the already existing authority that decides on the IDS system. As far as software is concerned, there is the risk that all kinds of languages and packages will be used, which are not very durable or are too expensive for non-Western countries. For the moment, the best solution is to make a resolute choice for the open software community R when developing programmes.

The integration of all kinds of datasets will become an important issue; not the technical integration as such, which is realized by the IPUMS and IDS systems, but the integration of different data from the same realm. Historical demographers are interested in the micro world of specific villages and cities, simply because they need more variables for their analyses than can be offered by the big databases. However, this poses an enormous problem: that of generalization. Here, large datasets may offer firm ground for the selection of persons and data. This is where I see a future for a Historical Person Identifier or HPI, which will make the linkage of these multiple sources easier and more secure. The kernel of such a system is a register that contains the identifying information of persons (and their HPI number), comparable to the modern national population administration in continental Europe, but limited to deceased persons only. This HPI register will constitute the authoritative national reference source on historical persons. The users will be institutes with data collections, archives, individual researchers or research groups, and the genealogical community. The HPI can only be successfully introduced if the HPI register is grounded in a central, stable and transparent institutional setting.

To conclude, I see a bright future for our kinds of datasets, with more and more researchers other than historians using these data for many different types of research, including epigenetics, societal change (migration, mobility) and demography. It is important that historical demographers should keep up with these developments. At the same time, the existing historical databases need to integrate systems like the IDS, standard person identifiers and modern registrations into their data.

References

Alter, G., Mandemakers, K. & Gutmann, M. (2009). Defining and distributing longitudinal historical data in a general way through an intermediate structure. *Historical Social Research*, 34(3), 78-114.
DOI: http://www.jstor.org/stable/20762377

Alter, G. & Mandemakers, K. (2014).The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies*, 1, 1-26.
DOI: http://hdl.handle.net/10622/23526343-2014-0001?locatt=view:master

Quaranta, L. (2015). Using the Intermediate Data Structure (IDS) to construct files for statistical analysis. *Historical Life Course*

Studies, 2, 86-107.
DOI: http://hdl.handle.net/10622/23526343-2015-0007?locatt=view:master
Ruggles, S. (2014). Big microdata for population research. *Demography*, 51(1), 287-297.
DOI: http://dx.doi.org/10.1007/s13524-013-0240-2
Schürer, K. (2007). Creating a national and representative individual and household sample for Great Britain 1850-1901 – the Victorian Panel Study (VPS). *Historical Social Research*, 32(2), 211-331.
DOI: http://www.jstor.org/stable/20762213

Weblinks

EHPS: http://www.ehps-net.eu
IPUMS: https://www.ipums.org
MOSAIC: http://www.censusmosaic.org

Biography

Kees Mandemakers is a senior research fellow at the International Institute for Social History (IISG). He is Head of the Historical Sample of the Netherlands (HSN) and Professor of Large Historical Databases at the Erasmus School of History, Culture and Communication (ESHCC) of the Erasmus University, Rotterdam. His main research interests are the methodology of large historical databases, family and demography, social stratification and mobility, and the social history of education. For publications, see https://socialhistory.org/en/staff/kees-mandemakers.