Artículo de revista:

PUJADAS-MORA, Joana M.; FORNÉS, Alícia; RAMOS TERRADES, Oriol; LLADÓS, Josep; CHEN, Jialuo; VALLS-FÍGOLS, Miquel; CABRÉ, Anna (2022) "The Barcelona Historical Marriage Database and the Baix Llobregat Demographic Database. From Algorithms for Handwriting Recognition to Individual-Level Demographic and Socioeconomic Data". *Historical Life Course Studies*, 12 (5): 99-132 (eISSN: 2352-6343)
https://doi.org/10.51964/hlcs11971

The Barcelona Historical Marriage Database and the Baix Llobregat Demographic Database. From Algorithms for Handwriting Recognition to Individual-Level Demographic and Socioeconomic Data

By Joana Maria Pujadas-Mora, Alícia Fornés, Oriol Ramos Terrades, Josep Lladós, Jialuo Chen, Miquel Valls-Fígols and Anna Cabré

To cite this article: Pujadas-Mora, J. M., Fornés, A., Ramos Terrades, O., Lladós, J., Chen, J., Valls-Figols, M., & Cabré, A. (2022). The Barcelona Historical Marriage Database and the Baix Llobregat Demographic Database. From Algorithms for Handwriting Recognition to Individual-Level Demographic and Socioeconomic Data. *Historical Life Course Studies, 12*, 99–132. https://doi.org/10.51964/hlcs11971

HISTORICAL LIFE COURSE STUDIES

Content, Design and Structure of Major Databases with Historical Longitudinal Population Data

VOLUME 12, SPECIAL ISSUE 5, 2020

GUEST EDITORS
George Alter
Kees Mandemakers
Hélène Vézina

EHPS NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies is the electronic journal of the *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles
This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent, on the EHPS-Net website.

Research articles
This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation (ESF, http://www.esf.org), the Scientific Research Network of Historical Demography (FWO Flanders, http://www.historicaldemography.be) and the International Institute of Social History Amsterdam (IISH, http://socialhistory.org/). Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at hlcs.nl.

Co-Editors-In-Chief:
Paul Puschmann (Radboud University) & Luciana Quaranta (Lund University)
hislives@kuleuven.be

The European Science Foundation (ESF) provides a platform for its Member Organisations to advance science and explore new directions for research at the European level. Established in 1974 as an independent non-governmental organisation, the ESF currently serves 78 Member Organisations across 30 countries. EHPS-Net is an ESF Research Networking Programme.

The European Historical Population Samples Network (EHPS-net) brings together scholars to create a common format for databases containing non-aggregated information on persons, families and households. The aim is to form an integrated and joint interface between many European and non-European databases to stimulate comparative research on the micro-level.
Visit: http://www.ehps-net.eu.

HISTORICAL LIFE COURSE STUDIES
VOLUME 12 (2022), published 23-06-2022

The Barcelona Historical Marriage Database and the Baix Llobregat Demographic Database

From Algorithms for Handwriting Recognition to Individual-Level Demographic and Socioeconomic Data

Joana Maria Pujadas-Mora — Open University of Catalonia & Center for Demographic Studies, Autonomous University of Barcelona
Alícia Fornés — Computer Vision Center, Autonomous University of Barcelona
Oriol Ramos Terrades — Computer Vision Center, Autonomous University of Barcelona
Josep Lladós — Computer Vision Center, Autonomous University of Barcelona
Jialuo Chen — Computer Vision Center, Autonomous University of Barcelona
Miquel Valls-Fígols — Center for Demographic Studies, Autonomous University of Barcelona
Anna Cabré — Center for Demographic Studies, Autonomous University of Barcelona

ABSTRACT

The Barcelona Historical Marriage Database (BHMD) gathers records of the more than 600,000 marriages celebrated in the Diocese of Barcelona and their taxation registered in Barcelona Cathedral's so-called Marriage Licenses Books for the long period 1451–1905 and the BALL Demographic Database brings together the individual information recorded in the population registers, censuses and fiscal censuses of the main municipalities of the county of Baix Llobregat (Barcelona). In this ongoing collection 263,786 individual observations have been assembled, dating from the period between 1828 and 1965 by December 2020. The two databases started as part of different interdisciplinary research projects at the crossroads of Historical Demography and Computer Vision. Their construction uses artificial intelligence and computer vision methods as Handwriting Recognition to reduce the time of execution. However, its current state still requires some human intervention which explains the implemented crowdsourcing and game sourcing experiences. Moreover, knowledge graph techniques have allowed the application of advanced record linkage to link the same individuals and families across time and space. Moreover, we will discuss the main research lines using both databases developed so far in historical demography.

Keywords: Individual demographic databases, Computer vision, Record linkage, Social mobility, Inequality, Migration, Word spotting, Handwriting recognition, Local censuses, Marriage licences

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs11971

1 INTRODUCTION

The Barcelona Historical Marriage Database (BHMD) gathers records of the more than 600,000 marriages celebrated in the Diocese of Barcelona and their taxation registered in Barcelona Cathedral's so-called Marriage Licenses Books for the long period 1451–1905. The Baix Llobregat Demographic Database (BALL) brings together the individual information recorded in the population registers, censuses and fiscal censuses of the main municipalities of the county of Baix Llobregat (Barcelona). In this ongoing collection 263,786 individual observations have been assembled, dating from the period between 1828 and 1965 by December 2020. The two databases started as part of different interdisciplinary research projects at the crossroads of Historical Demography and Computer Vision, within the well-known Digital Humanities, by a team of demographers, historians, geographers, mathematicians, and computer scientists from the Geography Department of Universitat Autònoma de Barcelona, the Center for Demographic Studies (CED), and the Computer Vision Center (CVC). To this end, these databases are conceived not only as demographic databases, but also as interdisciplinary research infrastructures.

Artificial intelligence methods are used to reduce the time needed for the construction of the database execution and to cope with the expansive volume of data, since gathering original information from primary sources is a time-consuming process, even when using a computer. In particular, we use computer vision technologies to extract information from the scanned document images. Computer Vision is the subfield of Artificial Intelligence that designs software to give machines the ability to see. In this work, Computer Vision aims to automatically read the text of original handwritten sources. The current development of optical character recognition, including the successful adoption of deep learning and machine learning in handwritten text recognition (Lladós, Rusinol, Fornés, Fernández, & Dutta, 2012; Toledo, Dey, Fornés, & Lladós, 2017), opens up the possibility of their integration into data collection to shorten the process. It also contributes to expand the volume of information in line with the big data revolution. However, the present approaches are specific for each data source considering language, structure, vocabulary, etc. Notably, these techniques are moving towards 'Document Understanding' to narrow the semantic gap between interpreting the original contents and automatically extracting and filling a database (Toledo, Carbonell, Fornes, & Lladós, 2019). Integrating these techniques brings Historical Demography into the realm of Big Data Sciences and allows the study of new topics and revisitation of old ones from a life course perspective, and for taking a multigenerational approach.

The research group responsible for constructing these databases was originally created to respond to the European Research Council's call for Advanced Grant projects, obtaining the 'Five Centuries of Marriages — 5CofM' project for the period 2011–2016 directed by Prof. Anna Cabré, one of the main objectives of which was the creation of the BHMD. This project was followed by others such as 'Tools and procedures for the large scale digitization of historical sources of population — *Eines*' (PI: Josep Lladós and Albert Esteve) (2015–2017) and 'Technology and citizen innovation for building historical social networks to understand the demographic past — *Xarxes*' (PI: Alícia Fornés and Joana Maria Pujadas-Mora) (2017–2019), both supported by the RecerCaixa program, and 'Demographic determinants of economic inequality, a historical approach (18th–20th centuries) — *Demodesigual*' funded by the Spanish Ministry of Sciences (PI: Joana Maria Pujadas-Mora) (2019–2021), under which the BALL database was built.

This article is organized in seven sections. Section 2 describes the source materials on which the BHMD and the BALL are based. In the following section, the data structures and data harmonization for the two databases are discussed. In section 4, we describe how the databases were constructed, presenting the use of web-based crowdsourcing platforms to transcribe the original sources, which were validated using game sourcing mobile applications, both assisted by computer vision techniques. In section 5, data linkage is presented to show how string distances, visual word search, and knowledge graph-based methods are useful to build individual life courses and multigenerational families. Section 6 is devoted to the research fields addressed, involving both databases, in historical demography, as the study of intergenerational transmission of social status, intermarriage and kinship marriages within the social reproduction process, the evolution of migratory flows in a long-term view and the estimation of economic inequality within preindustrial and industrial periods (15th–20th centuries).The paper closes with some final comments on the future agenda to expand the two databases with other textual sources and images.

2 SOURCE MATERIALS

2.1 MARRIAGE LICENSES BOOKS

The BHMD brings together the 612,487 marriages recorded in the so-called Marriage Licenses Books of the *Ius Tabulae* of the Cathedral of Barcelona, covering the Diocese of Barcelona between 1451 and 1905. This is a unique data source dating back to 1409 when Pope Benedict XIII (1328–1423) visited Barcelona and granted the new cathedral the power to impose a tax on marriage in accordance with the socio-economic status of the couple to fund the cathedral's construction and maintenance (Baucells, 2002; Carreras Candi, 1913). The names and surnames of the grooms were registered (one surname up to 1876 and two surnames thereafter), while the names of the brides started to be registered from 1481 onwards. Previously unmarried brides were related to their fathers and widows to their late husbands. The occupations of the grooms, the marital status of both spouses, and the tax paid were also recorded throughout the entire life of the source, as were the first and surnames of the spouses' parents, except for the period 1645–1715. In fact, the availability of occupational and fiscal information for so many centuries is one of the main strengths of this database, enabling us to adopt a long-term perspective approach in our studies. However, the occupations of the fathers were not systematically recorded except for the period 1545–1643 and the occupations of the wives and mothers were never recorded. The original books were written in Catalan until 1860, after which Spanish was used (see Table 1).

From 1575 onwards, marriage taxes were organized on a seven- or eight-tiered scale), ranging from the highest tax paid by the nobility to exempt from tax for those declared poor. From 1451 to 1565, there were up to 27 different payment levels, while seven levels were set from 1583 onwards. In this seven-level system the first level corresponded to the titled nobility and the next two to the knights and honest citizens, or those who could hold public office. The fourth and fifth levels of payment corresponded to the commercial bourgeoisie, liberal professionals, and masters of guilds. The sixth was paid by farmers and artisans, and the last level was the exempt from tax category. A new level was already added in 1649 and continued till 1857, corresponding exclusively to the merchants of Barcelona. The highest tax level was 120 times higher than the lowest one and 40 times more than the average tax.

As mentioned previously, the territorial coverage of the BHMD is the Diocese of Barcelona (see Map 1), which was made up of four deanships: Barcelona, the main one, which is why it was called Oficiality, Piera, Vallès and Penedès. This territory covered the main population centers of the time including Barcelona, Mataró, Sabadell, and Terrassa, and a conglomerate of rural towns located in the current counties of Baix Llobregat, Barcelonès, Maresme, and Vallès Occidental (see Map 1). In 1900, the diocese was comprised of 250 parishes. This source is not only extraordinary for its territorial coverage and chronological amplitude, starting a few centuries before the parish marriage books, but also for its state of preservation compared with the low conservation of the parish archives in Catalonia, especially in the study area.

Map 1 *Geographical coverage of BHMD and BALL database. Area of the Barcelona Diocese*

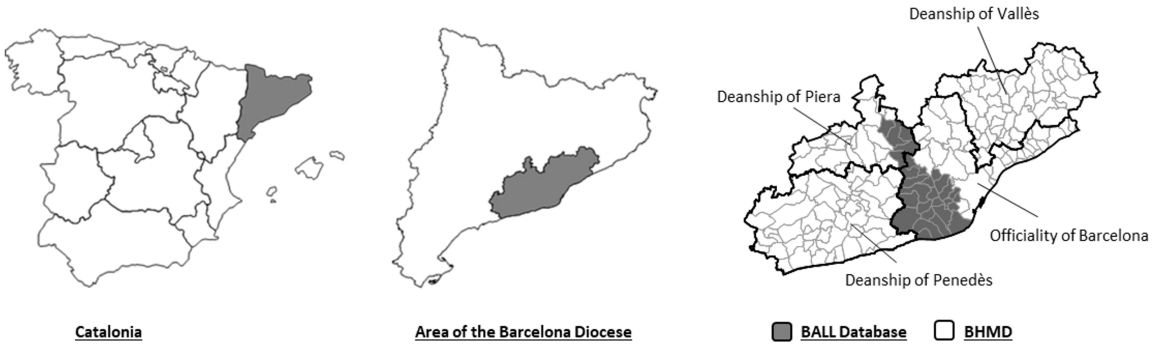


Table 1 *Number of marriages and their compiled information in the marriage licenses books, 1451–1905*

		Marriage Licenses								
		1450–1499(a)	1500–1549	1550–1599	1600–1649	1650–1699(b)	1700–1749	1750–1799(c)	1800–1849	1850–1905
Number of marriages		7,002	23,766	44,567	57,050	41,718	69,744	82,608	96,579	189,453
General	Priest first name and surname	X								
	Date of the payment	X	X	X	X	X	X	X	X	X
	Parish of celebration						X	X	X	X
	Fee	X	X	X	X	X	X	X	X	X
Groom	First name	X	X	X	X	X	X	X	X	X
	Surname 1	X	X	X	X	X	X	X	X	X
	Surname 2									X
	Marital status				X	X	X	X	X	X
	Occupation	X	X	X	X	X	X	X	X	X
	Geographical origin		X	X	X	X	X			
	Residence	X	X	X	X					
Bride	First name		X	X	X	X	X	X	X	X
	Surname 1					X	X	X	X	X
	Surname 2									X
	Marital status		X	X	X	X	X	X	X	X
	Occupation									
	Geographical origin									
	Residence									
Father of the groom	First name			X	X			X	X	X
	Surname 1			X	X			X		
	Surname 2									
	Marital status									
	Occupation			X	X					
	Geographical origin									
	Residence			X	X					
Mother of the groom	First name			X	X			X	X	X
	Surname 1							X	X	X
	Surname 2									
	Marital status									
	Occupation									
	Geographical origin									
	Residence									
Father of the bride	First name	X	X	X	X			X	X	X
	Surname 1	X	X	X	X			X		
	Surname 2									
	Marital status									
	Occupation		X	X	X					
	Geographic origin			X	X					
	Residence	X	X	X	X					
Mother of the bride	First name			X	X			X	X	X
	Surname 1					X		X	X	X
	Surname 2									
	Marital status									
	Occupation									
	Geographic origin									
	Residence									

Notes: (a) Except 1456–1480 and 1491–1499; (b) Except 1653–1659; (c) Except 1782–1784.

2.2 POPULATION REGISTERS, NATIONAL CENSUSES AND PERSONAL TAXES

The BALL database of which the construction is still ongoing, compiles the population registers — called padrones in Spanish — and censuses of 14 municipalities in the county of Baix Llobregat (province of Barcelona) and the personal taxes of nine of these municipalities collected between the 19th and 20th centuries (see Map 1). In Spain, the population registers were introduced with the Decree issued on 3 February 1823 being household registers. Initially, there was no simultaneity or fixed periodicity in the population registers until a five-year interval was enacted by the Municipal Law of August 20th, 1870, which was followed mainly from the beginning of the 20th century. The population registers contain information about individuals and households for urban centers, small villages or hamlets, and isolated houses. The first population registers in the 1820s and 1830s were simply lists of inhabitants, becoming complete registers from the end of the 1880s onwards, recording the first name and surnames, age or date of birth, marital status, and occupation of each individual, in addition to the family or the working relationship with the head of the household and the complete address. For some periods and municipalities, there is also information about the individuals' literacy skills, migratory status, and income (see Table 2). The variability of information among the registers of the different municipalities was because the number and type of variables were not standardized for the whole country until the Municipal Statute of March 8th, 1924 (García Pérez, 2007; García Ruipérez, 2012). All this information may originally have been used for statistical purposes, as a way of counting the number of inhabitants or their main socioeconomic and demographic features in national censuses, but unlike the national censuses their main purpose was in fact to meet the military and fiscal requirements of the State and to provide proof of residence in the municipality.

The population registers contain almost the only census data preserved at the individual level in Spain because national census manuscripts were generally destroyed once the population count had been estimated and the main variables aggregated, as established by law. Therefore, only the census tabulations are preserved in the current National Institute of Statistics. However, the individual registrations of the national censuses were found for different municipalities in the Baix Llobregat county. Wherever the national censuses had been preserved locally they were also included in BALL even though the information they provided was very similar to that contained in the population registers. Moreover, the national censuses tended to provide more information of a socioeconomic nature or about literacy. And last, this meant an increased number of observations in the BALL mainly from the 20th century onwards because the censuses were carried out in the years ended in 0, while the population registers were compiled in the ones ending in 5 (Reher & Valero Lobo, 1995). In this way, the population registers throughout the 19th century are quite similar to the usual population registers in the north of Europe given their high cadence in time — especially in certain municipalities — although the consecutive years have not necessarily been collected in the BALL database, because of the higher priority of expanding the geographical and time coverage. On the other hand, in the 20th century they are more similar to the national censuses being snapshots in time.

The first modern national census in Spain was carried out in 1857, although there are direct precedents such as the censuses of Aranda (1768), Floridablanca (1787), and Godoy (1797). While we did not find the individual information, we did find the aggregated data. From 1857 onwards, national censuses began to take on the distinguishing features of modern population censuses sponsored by the State with appropriate administrative procedures and legislation, national boundaries, universality, individual enumeration, simultaneity in data collection, periodicity, publication, and diffusion (Goyer & Draaijer, 1992; Reher & Valero Lobo, 1995). National censuses were also taken in 1860, 1877, 1887, 1897, 1900, 1910, 1920, 1930, 1940, and so on until 2011 when the census started to be based on the population registers combined with a 10% survey. From 2021 the national census will only be constructed using administrative data as in the Netherlands, Denmark, Sweden and other European countries. Notably, when the state started to compile the national censuses in the mid-19th century, the public administration with the most experience in compiling demographic data in population registers and in which the largest number of civil servants or public employees were involved was the municipal administration (Salas-Vives & Pujadas-Mora, 2021).

In 1854, the personal taxes, first called *cédulas de vecindad*, later *cédulas de empadronamiento,* and finally, *cédulas personales*, were established as capitation taxes and were a type of personal identity document up until 1943 (Martín Niño, 1972). It is a source that is hardly used in Spain and even less linked with the population registers, which is one of the novelties of our *Demodesigual* project. In 1870, three personal tax classes were created but in 1873 the tax was abolished completely for one year (Melis Maynar, 2019). It was reinstated in 1874 with nine tax classes and in 1884 11 classes were instituted that lasted for the rest of its duration. From 1874 onwards, the tax classes were based on information about direct taxes paid on real estate or industrial activities, annual salaries (only workers with annual contracts, not day laborers), and

annual housing rents (Pérez Hernández, 2020) (see Table 2). Its form was similar to that of the population registers. The register was organized on a personal basis, gathering personal, occupational, address and fiscal information for all citizens — including married women — from the age of 14 years on. All residents of a municipality were considered taxable persons.

These taxes were modest in terms of yields, but they were important in terms of the creation of a liberal tax system. They were more progressive and inclusive than previous taxes as they taxed all social strata (Marín Corbera, 2010). Preparation of the collection lists and the tax collection itself was a municipal responsibility, like the population registers, which is another indication of the quality of these sources. Its broad coverage in terms of measures of wealth and income for much of the population is also worth noting. However, active and inactive women alike fall into a single fiscal category, making its use as an economic measure unadvisable, whereas the men are represented in all the tax categories.

In mid-December 2020, the BALL contained 263,786 individual observations from the population registers and censuses from 13 different municipalities and 38,103 from the personal taxes' material from eight of them. Their population size of these cities and industrial and rural towns went from 500 to 12,000 inhabitants due to cities as Sant Feliu which showed an important urbanization as a result of industrialization and their role as head of the county (see Table Appendix 1). Other towns, as Santa Coloma de Cervelló, included an important industrial colony, where Molins de Rei and Martorell presented a relevant textile industry. The rest of the towns were mainly rural with an economy based on a commercial agriculture as a result, among other reasons, of its proximity to the city of Barcelona. The average size of households ranges from almost five to three as a clear consequence of the demographic transition, being Catalonia one of the forerunners in Spain mainly because of an early fertility decline. Actually, the whole territory covered by the BALL database — Baix Llobregat county — is part of the territory of the Diocese of Barcelona included in the BHMD, which makes both databases compatible, although they only coincide in time in the 19th century and the first years of the 20th century. At the same time, the BALL database will continue growing by adding new counties. In fact, the Maresme county is already under construction and in September 2021 began the collection of the Selva county. All these counties share a border.

Table 2 *Compiled information in the population registers, national censuses and personal taxes, 19th–20th centuries*

	Population registers/Censuses												
	1820–1829	1830–1849	1850–1859	1860–1869	1870–1879	1880–1889	1890–1899	1900–1909	1910–1919	1920–1929	1930–1939	1940–1949	1950–1965
First name	X	X	X	X	X	X	X	X	X	X	X	X	X
Surname 1	X	X	X	X	X	X	X	X	X	X	X	X	X
Surname 2			X	X	X	X	X	X	X	X	X	X	X
Occupation			X			X	X	X	X	X	X	X	X
Sex										X	X	X	X
Number of men in household	X												
Number of women in household	X												
Relation to the head of household						X	X	X	X	X	X	X	X
Marital status		X	X	X		X	X	X	X	X	X	X	X
Age		X	X	X	X	X	X	X	X	X	X	X	X
Street name		X			X	X	X	X	X	X	X	X	X
Citizenship										X	X	X	X
Municipality of birth						X	X	X	X	X	X	X	X
Province of birth						X	X	X	X	X	X	X	X
Date of birth						X	X	X	X	X	X	X	X
Municipality of usual residence						X	X	X	X	X	X	X	X
Province of usual residence										X	X	X	X
Years living in the municipality						X	X	X	X	X	X	X	X
Months living in the municipality						X	X	X	X	X	X	X	X
Property tax						X	X	X	X				
Industrial tax						X	X	X	X				
Literacy						X	X	X	X	X	X	X	X
Wages and salary										X		X	X

Table 2 *Continued*

	Personal taxes		
	1868–1879	1880–1889	1890–1943
First name	X	X	X
Surname 1	X	X	X
Surname 2	X	X	X
Age		X	X
Marital status		X	X
Occupation		X	X
Geographical origin (municipality)		X	X
Geographical origin (province)		X	X
Street	X	X	X
House number		X	X
Floor			X
Tax class	X	X	X
Number of family members	X		
Salary (annual contract)		X	X
Annual housing rent		X	X
Property and industrial tax		X	X

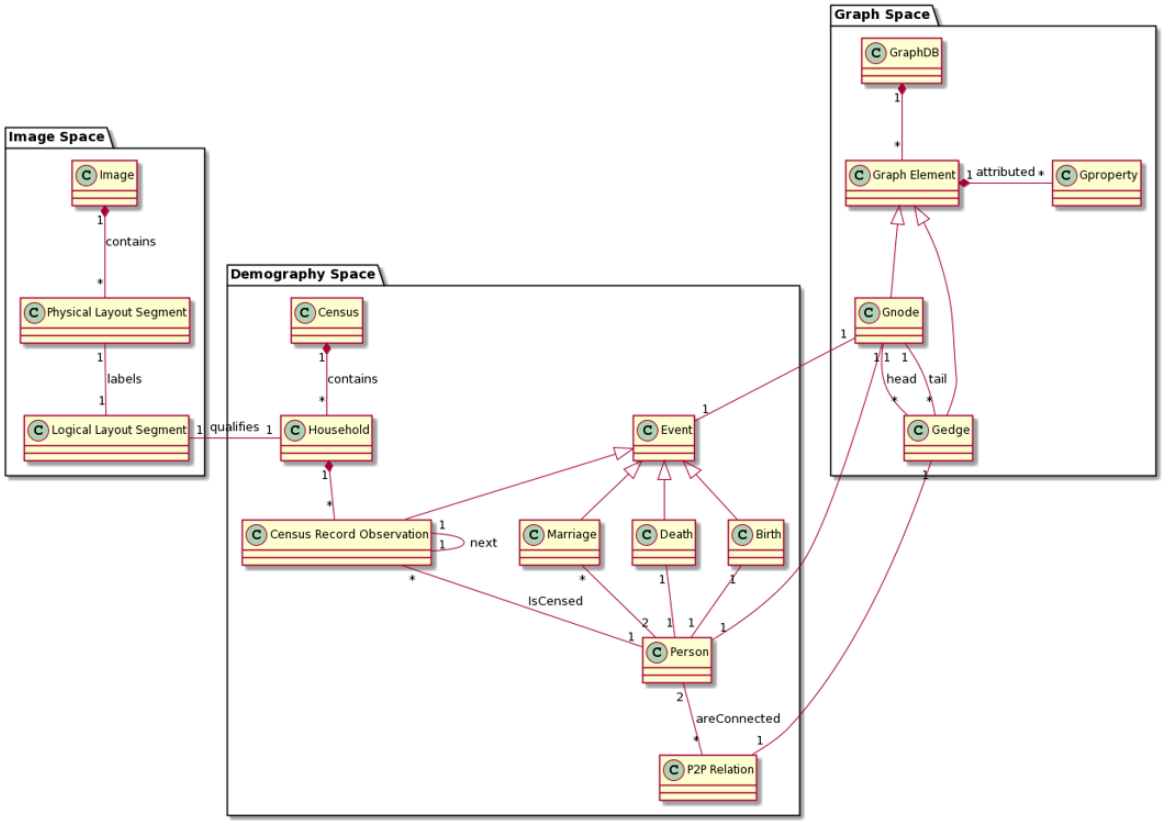
3 DATA STRUCTURE AND HARMONIZATION

3.1 DATA STRUCTURE

The information extracted from the sources (marriage license books and the population and fiscal census materials) are stored in relational databases. The demographic data of the BHMD and the BALL are complementary so the two databases share the same storage structure, allowing the information to be merged and record linkage operations to be performed to organize the data more succinctly. Record linkage operations result in discovered relations between information items. This suggests a second representational layer based on linked data (graph-like) representation. Thus, there are two levels of information using two kinds of database management systems (DBMS), namely a relational database and a graph database (see Figure 1). The relational database is the core DBMS, storing the information extracted directly from the digitized documents using the different data entry procedures: automatic information extraction using the computer vision techniques described in section 4, and the crowdsourced data collection through web-based platforms and game-sourcing mobile applications. The second level extends the original database with semantic information, implemented as a knowledge graph. A knowledge graph is a typical representation in artificial intelligence of a collection of interlinked entities. In our case, the entities are individuals, events, or places. This representation puts the data into context via semantic linking, providing a framework for data analytics and reasoning (e.g., finding communities, genealogies, life courses). Although the relational database satisfies the functional requirements, the use of a graph database, with native linked-data analytics operations, allows more efficiency and flexibility.

The complete information architecture is graphically described in Figure 1 using an object-oriented notation (a class diagram). It is made up of three abstraction layers: the image space, the demography space and the graph space. First, the image space stores the information directly related to the digitized documents and the physical/visual and logical structure of the relevant regions. The document images in a raster format are stored together with the bounding box coordinates corresponding to the segmented regions, along with the corresponding labels (logical segmentation). The image-related information is the input to the computer vision algorithms for information extraction.

Figure 1 *Class diagram of the data model architecture of the BHMD and the BALL*



All the information ingested from the original documents (user transcription/validation, automatic reading) is stored according to the schema presented in Figure 1 in the demography space. Its central class, *Person,* contains information that is permanent over time. For each *Person* there are different *Observations,* obtained from the analyzed censuses/marriage licenses. These observations are grouped into *Households* for the census. Each demographic document is represented as an *Event*, which can be specified as a *census record* or a *marriage record*. The proposed relational architecture has a flexible design to be able to store the information contained in the different census records as primary data, coping with the heterogeneity of different sources, time periods, and locations, and easily scaled to integrate other demographic sources like marriage, birth, or death records. This representation is compatible with the *Intermediate Data Structure* (IDS) format proposed by Alter and Mandemakers (2014). IDS is a simplified data schema that aims to map the diversity of data in European historical micro databases into a common format.

The graph space uses a graph-based model to represent a knowledge graph that infers semantics in terms of interlinked data. This knowledge graph is a conceptualization of the "historical social network" constructed from demographic sources. Using a graph representation, nodes correspond mainly to persons (observations) and events, according to the structure described in Figure 1. Graph edges represent relations between individuals (both directly extracted from the source documents, like genealogy relations, or semantic relations deduced by artificial intelligence reasoning tasks). The graph is dynamic because it evolves over time and is composed of different subgraphs. Each individual subgraph corresponds to a particular census, providing a static picture of the population at a specific time, or a marriage record. A knowledge graph also allows graph-based data analytics techniques for querying the database to be applied, including record linkage and community detection. Thus, life courses of individuals can be constructed by linking the corresponding record of the observation at time *t* with the observation at time *t+1*. This structure also allows future users to add specific events for individuals to achieve a more complete reconstruction of their life course. We provide further details of the knowledge graph techniques used in these databases in Section 6.

From a technical point of view, the main difference between a relational database and a graph database is the way relationships between entities are stored. Relational databases are good for highly structured data that can be organized in different tables. Relations between individual records must be defined at table level. Graph databases store relationships at individual record level. It seems an ironic, but relational databases are not good with relationships, especially when they are dynamically created. This is why we use a graph model

on top of a relational model. Relational databases are faster when handling large numbers of records because the structure of the data is known a priori. But graph databases are more efficient when handling linked data, and answering queries that involve a navigation through this linked data. We can roughly differentiate the levels saying that the relational level (demography space) stores the demographic information of persons, and the graph level (graph space) stores the contexts as linked data, which facilitates the subsequent analysis and browsing.

A side information that is stored is the metadata associated with the entities collected from the web-based crowdsourcing platform and game sourcing mobile application. This is basically the support metadata for these platforms (golden tasks, number of transcriptions per image, forms presented to the transcribers, etc.).

3.2 DATA HARMONIZATION

The BHMD has already been harmonized and the ongoing BALL database is in the process of being harmonized to make the data comparable through time and compatible with other data from similar or different national or international sources. This procedure consists of two steps: first, a linguistic standardization of names, surnames, occupations, and geographic locations, and second, a codification of the last two variables following international classifications, tasks shared among different researchers and technicians.

The name and surname standardization was a key step because the same individual could appear with seemingly different names and surnames because they were recorded using different spellings or abbreviations (Goiser & Christen, 2006; Herzog, Scheuren, & Winkler, 2007; Schürer, 2007). These issues needed to be addressed mostly for the entries in Catalan, which was not standardized until 1913 having important dialectal differences and at the same time influenced by Spanish, French, and Occitan, all of which favored the appearance of many written variations of the same name or surname (Peytaví Deixona, 2010; Rubio Vela, & Rodrigo Lizondo, 1997). To this end, we compile dictionaries with the different variations of each string variable name present in our databases (names, surnames, occupations, and geographical location) (Bloothooft, 1998).

The dictionary of first names contains 22,951 different entries, 10,667 of which are female and 12,284 male first names, all Catalan and Spanish. The one for the surnames gathers 141,129 entries in different languages, including Catalan, Spanish, French, Occitan, Italian, and English, among others. The number of spelling variations ranges from 1 till 96, with the surname *Vall* as the entry with the most variations (96). There are 24,399 items of Catalan and Spanish vocabulary relating to occupations, some of which are variations of the same ones. The dictionary of the geographical locations contains 47,559 items divided into levels corresponding to the geographical divisions of parishes/municipalities, regions, and countries. While this is an intensive task, the record linkage success rates using string distances are high (Villavicencio, Jordà, & Pujadas-Mora, 2015).

The information regarding occupations was codified according to the Historical International Classification of Occupations (HISCO), based on the International Labor Organization's (ILO) ISCO 68 classification (van Leeuwen, Maas, & Miles, 2002). Its application to the BHMD needed some adjustments due to most occupations coded belonging to the preindustrial period (Pujadas-Mora, Romeo-Marín, & Villar, 2014). Special attention was paid to the nobility titles and political institutions and a new subsidiary classification created to maintain their nature and diversity. There are hardly any women's occupations in the BHMD, although their conditions were frequently reported as slave, captive, freed slave, or prostitute, which the HISCO classification did not originally contemplate. Moreover, new codes had to be created for some controversial occupations such as gambler or bandit. The adaptation of the HISCO classification to the BALL database, however, was smoother thanks to the experience gained building the BHMD and the time span reference period of the database, the 19th and 20th centuries.

The HISCO classification also facilitates the application of other stratification schemes such as HISCLASS, HISCAM, and SOCPO (Lambert, Zijdeman, van Leeuwen, Maas, & Prandy, 2013; Van de Putte, & Miles, 2005; van Leeuwen, & Maas, 2011). The 12 categories gathered in HISCLASS were regrouped into eight classes in line with the socioeconomic characteristics of the Barcelona area. Groups 3, 4, and 6 (lower manager, lower professional and clerical sales, and foremen) and 7 and 8 (low-skilled and unskilled workers, and low-skilled farm worker and unskilled farm workers) were combined. Moreover, an extra level for the nobility was included in all three classifications due to its social class significance (score 100 on HISCAM, group 0 on HISCLASS, and group 6 on SOCPO), with the two databases totaling 565 different HISCO codes.

Last, the geographical locations were coded following the 1860 and 1991 municipal bases of the Spanish National Institute of Statistics and all the parishes were geo-referenced within the corresponding municipal

borders. For those locations outside Spain a specific country code was created. Altogether, 1,458 codes were assigned.

4 TRANSCRIBING THE DATA BY WAY OF CROWDSOURCING AND GAME SOURCING

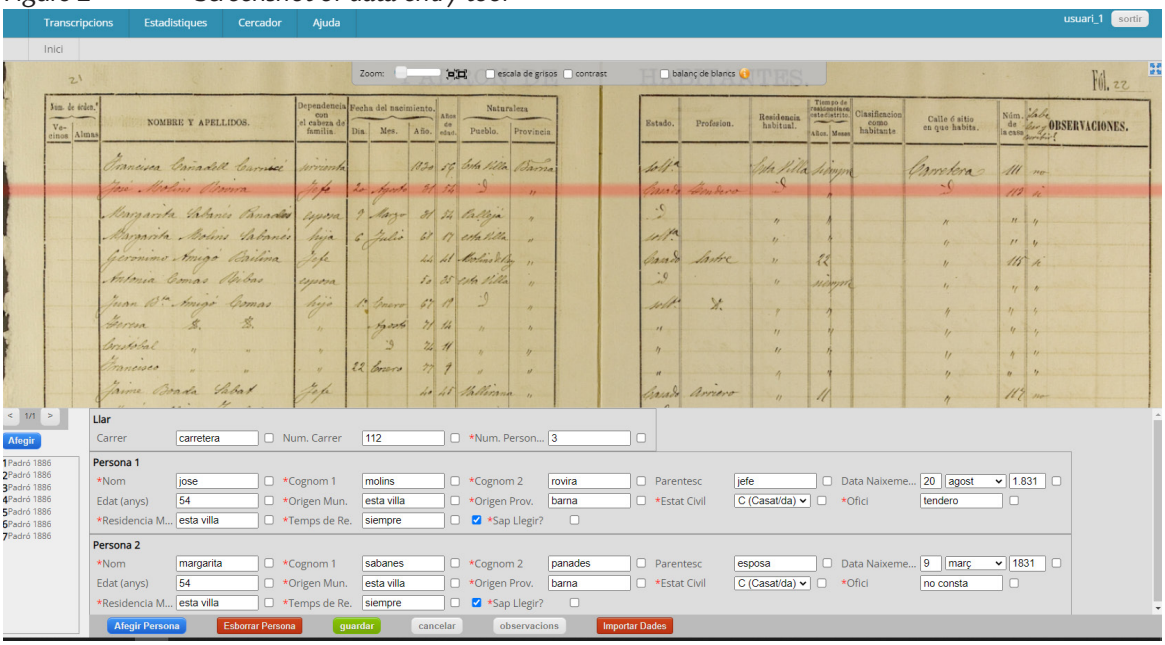
Manual transcription from original sources using a computer or not has been the system traditionally used since the mid-20th century. This way of constructing demographic databases has proven to be tedious and time-consuming. To speed up and to facilitate the data entry process of the BHMD and the BALL database, we have developed several computing alternatives based on crowdsourcing and game sourcing paradigms. Secondly, we used computer vision algorithms to extract the information contained in the marriage licenses, population registers, national censuses, and personal taxes.

4.1 MANUAL TRANSCRIPTION USING CROWDSOURCING PLATFORMS

For this first approach, we developed several web-based platforms (Fornés, Llados, Mas, Pujadas-Mora, & Cabré, 2014) to extract the information contained in the demographic-like sources. The BHMD and the BALL database have their own platforms, which are used by many users in parallel, following the crowd-based outsourcing concept. This consists in dividing a long task into many smaller tasks and distributing them among a large group of people, particularly from the online community, thereby shortening the time required and also allowing a higher volume of data covering larger geographical areas to be gathered. The user interface of our web-based platform (Fornés et al., 2014) consists in a data entry tool with a user-friendly environment. This means that both the scan of the source and the data entry form are combined in one view (see Figure 2). The web incorporated several tools to improve legibility such as zoom and image contrast. This is particularly useful in the case of paper degradation, as has happened with one of our sources. The iron gall ink used having corroded some parts of the original documents, mainly from the 16th century. The BHMD has been compiled entirely using a web-platform. For its part, the BALL platform incorporates a computer assisted transcription system, as explained in the following subsection.

A total of 153 paid transcribers (69 men and 84 women) participated in building the BHMD from backgrounds as diverse as secondary school pupils to university students, consultants in history, local historians, retired people, and PhD holders in history. Volunteers totaling 175 (81 men and 94 women) collaborated in constructing the BALL, almost all of which were local history and/or genealogy enthusiasts with previous knowledge of the source materials they were transcribing. They were from diverse cultural backgrounds, were mainly university graduates, and had a mean age of 59 years.

Figure 2 *Screenshot of data entry tool*



4.2 SEMI-INTERACTIVE TRANSCRIPTION SYSTEM

Since population registers and census records are recorded in intervals of a few years, the individuals in each household are quite stable. Hence, once the records for a specific year have been transcribed, this redundancy can be exploited to assist the transcription. The main idea behind the semi-interactive transcription system we developed (Mas, Fornés, & Lladós, 2016) is to transfer the individuals in one previous population register/census to the next one (see Figure 3). The computer vision algorithm first detects and transcribes the street address in the original manuscript page. The system then accesses the database and retrieves the list of individuals that were living in that household in the previous population register/census. In case the transcribed street address is not found in the previous census, it means that no previous information is available on that particular household, so the data must be manually introduced.

In case the household is found in the previous census, the system proceeds to check which individuals still appear in the current one. The search for each individual (name and surnames) in the new population register/census is performed using a word spotting algorithm. Word spotting consists in finding a particular word in a manuscript that has not been transcribed (using OCR or similar). This means that the search is performed directly on the digitized image. Since the task consists in searching concrete words (e.g., names and surnames), the method is more accurate than automatic transcription.

So, when the word spotting finds the name and surnames of the individual, it means that this particular individual appears in both the previous and the current census (so, no change in address). So, the individual's information from the previous census is copied on the data entry form of the current census. The user simply needs to update the still living individuals with their age or marital status, complete the new fields (where appropriate), transcribe the new members of the households (birth, marriage, or immigration), and delete the individuals who emigrated or died between two consecutive rounds.

4.3 AUTOMATIC TRANSCRIPTION USING COMPUTER VISION

4.3.1 THE HANDWRITTEN TEXT SYSTEM

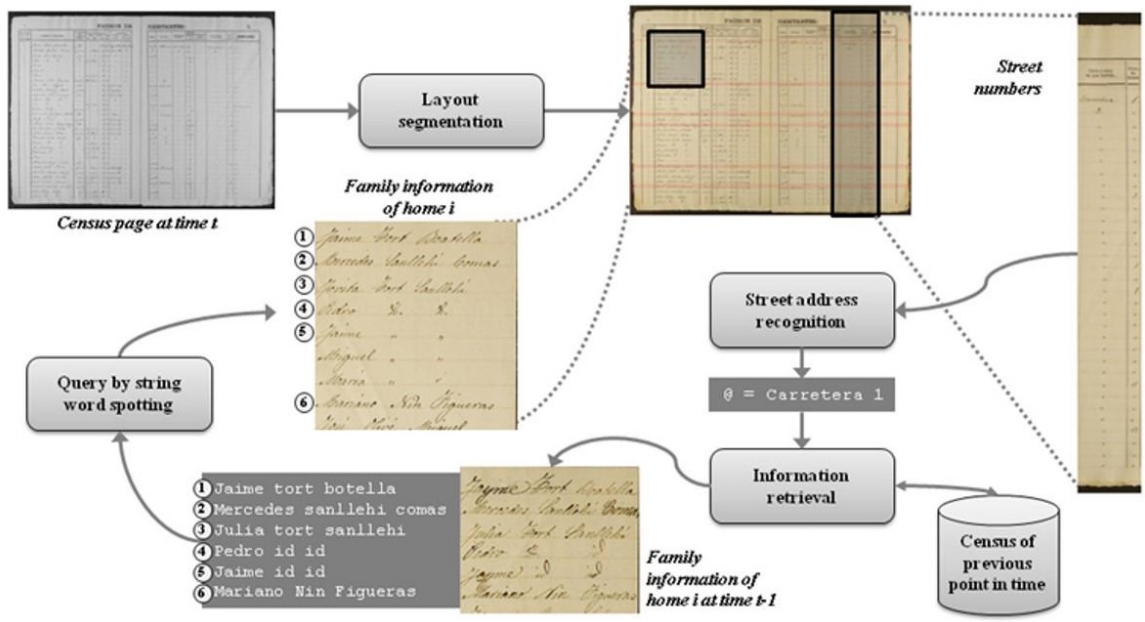
In the second approach, we assume that there is enough labelled (transcribed) data from the sources to be able to use computer vision and document image recognition techniques. We used a Handwritten Text Recognition (HTR) system (Toledo et al., 2017) trained on these kinds of documents (e.g., similar structure, same time period, similar handwriting style) and based on deep learning architectures, or more specifically on recurrent neural networks.

Since the marriage license records are written in blocks, once the different blocks and text lines have been automatically segmented and transcribed, a named entity recognition algorithm (Toledo et al., 2019) is used to extract names, surnames, occupations, places, etc.

For the population register/census documents, the steps of the automatic recognition system are:

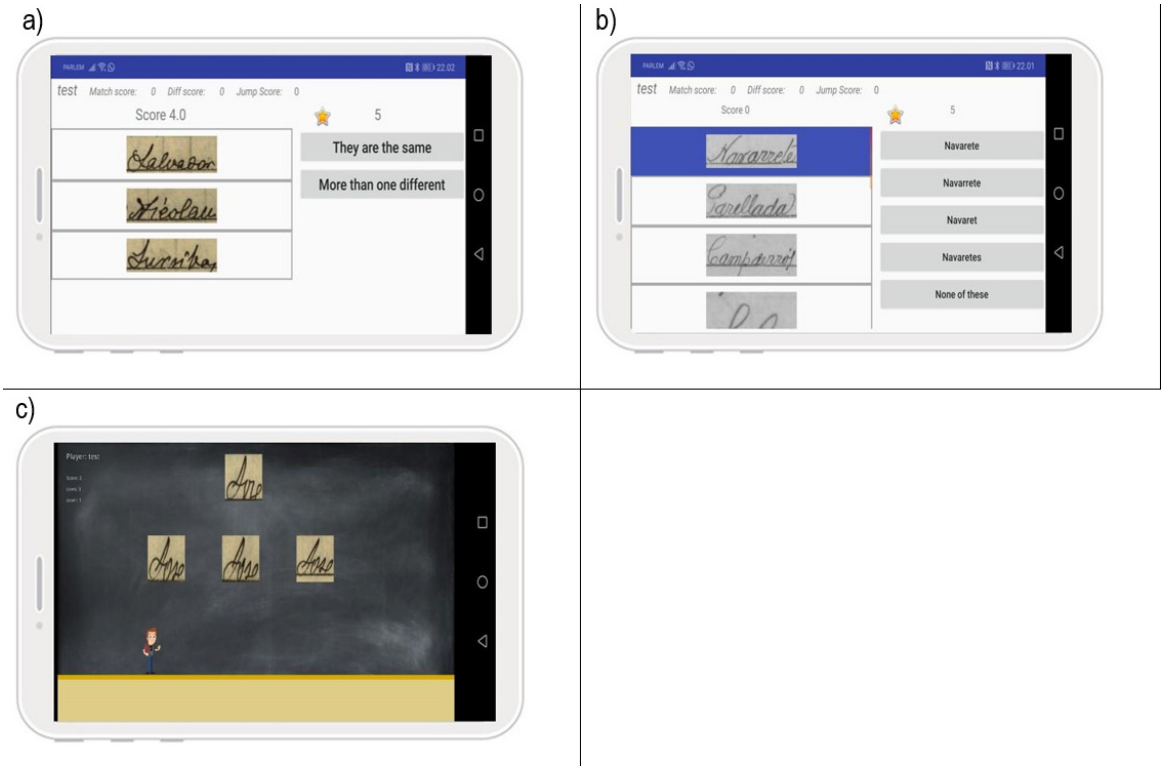
1. A layout analysis algorithm detects the different columns and rows in the tabular census documents. Each cell in the table is then segmented into words.
2. For each column, a word image clustering method groups words by visual similarity. Thus, the groupings will likely contain instances of the same name, surname, etc.
3. The most populated clusters are selected, which correspond to the most common names, surnames, etc.
4. The words that belong to each one of the populated clusters and transcribed using a Handwritten Text Recognizer algorithm.
5. Since the system knows the exact location in the page (this means the column and row in the table) for each word in the clusters, the transcription of each word is introduced in the corresponding field in the form (e.g. if a cluster contains many instances of a common surname, all individuals with this specific surname will be filled in the form, in one single step).

Figure 3 *Examples of the semi-interactive transcription system*



Source: Mas, Fornés and Lladós (2016).

Figure 4 *Example of text confirmation by way of videogaming*



Given that the accuracy and performance of all these methods depends highly on the amount of labelled data to train the models, differences in handwriting styles, and degradation of the paper, the transcriptions must be validated manually. Instead of using a crowdsourcing web-based platform, we opted to incorporate gamification principles to engage users in the validation, commonly referred as game sourcing. Game sourcing consists in the application of game design elements into non-game contexts, like crowdsourcing. More specifically, we have developed several video games for mobile devices (Chen et al., 2018). These mini games (see Figure 4), grouped together in a video game named *WordHunter,* are designed to engage users

in validating the transcriptions and confirming the computer vision algorithm output applied to population registers/censuses or personal taxes. The three games are:

a) *WordHunter-Difference:* A game to validate the word clustering algorithm, consisting in word images contained in a given cluster, which the user must identify as being the same or not.

b) *WordHunter-Match:* A game to validate the transcription algorithm. A set of word images are shown on the screen along with the most probable automatic transcriptions provided by the HTR system, from which the user must select the correct answer.

c) *WordHunter-Jump:* A third game, also used to validate the transcriptions, which emulates a platformer game. The user makes the videogame character jump to catch the correctly transcribed words by the HTR system.

More than 40 volunteers participated in the validation using videogames. Of these, 28 were men and 14 were women, 32 were experts (with a background in history and/or paleography) and 8 had no previous knowledge of the material. The videogames have been available for download from Google Play for the last 6 months and almost 100 users have already played them.

4.3.2 STEPS OF THE AUTOMATIC TEXT RECOGNITION SYSTEM

Layout Analysis

Layout analysis is crucial for the semantic interpretation of a document. However, layout usually depends on signals in the image (e.g., lines that define the rows, columns, and cells of a table), which serve to define the limits between the different parts of the document. In some cases, such as the marriage records of the BHMD, the entries are recorded in blocks rather than tables. In others, the signals are not sufficiently clear due to the quality of the image or degradation of the paper. For these reasons, the layout analysis module must use this incomplete information to understand the logical structure and extract the page layout. Consequently, it is important to study how the maximum layout information can be obtained from the parts of the document where there are clear image signals, using automatic inference to determine the coherent parts of the document according to these clues (Kleber, Diem, Dejean, Meunier, & Lang, 2018). In addition, the recent success of graph neural networks for semantic segmentation in scene images makes them promising architectures for the analysis of complex population documents (Carbonell, Riba, Villegas, Fornés, & Lladós, 2021).

Handwritten Text Recognition

Once the layout of the document has been recognized, the next step in the extraction of information from digitized documents is the recognition of text. Although Optical Character Recognition is quite accurate for recognizing printed text, the recognition of handwritten text is still a challenge. Even though Handwritten Text Recognition (HTR) systems based on deep learning models perform better than traditional approaches based on Hidden Markov Models or Recurrent Neural Networks, the accuracy is still unsatisfactory regarding historical manuscripts. The difficulties in processing historical handwritten documents are mainly due to paper degradation, different handwriting styles, old vocabulary, etc., causing frequent errors in the transcription.

Research on more robust and generic HTR models is therefore required. Some recent deep learning approaches based on sequence-to-sequence models (Kang, Riba, Villegas, Fornés, & Rusiñol, 2020a) and transformer networks (Kang, Riba, Rusiñol, Fornés, & Villegas, 2020b) have been proposed. These models have significantly improved on the HTR performance of previous architectures, but the large differences in the handwriting styles in documents from different regions or centuries still pose a problem. Therefore, there is a need for research into models that can adapt to unseen handwriting styles, meaning that techniques for domain adaptation and transfer learning should be explored. Some attempts have already been made towards unsupervised writer style adaptation (Kang, Rusiñol, Fornés, Riba, & Villegas, 2020c) to recognize segmented handwritten words, but these techniques need to be further explored and extended to deal with full sentences.

Named Entity Recognition

Analysis of the document layout and the transcription of text is not enough to fill the corresponding knowledge database, the reason being that the transcribed text needs to be semantically labelled, providing a semantic meaning. In other words, the named entities (e.g., names, surnames, locations, dates) must be

detected so that the information contained in the documents can be properly entered in the way required by the database.

The typical procedure consists in applying Named Entity Recognition models from the Natural Language Processing field to the transcribed text. However, the transcribed text contains errors, which will affect the performance of the subsequent Named Entity Recognition module. Consequently, there have been some attempts in recent years to perform text detection, text recognition, and named entity categorization in an integrated neural network model (Carbonell, Fornés, Villegas, & Lladós, 2020), commonly referred as multi-task learning. This joint architecture must therefore learn to perform several tasks at the same time, which can potentially improve the overall performance and reduce errors. However, this kind of approach consists in large deep learning architectures that need many labelled documents to train. Since labelled data is costly to produce and barely available, the generation of "realistic" synthetic documents (Das et al., 2020) to increase the amount of training data is worth exploring. All in all, the automatic extraction of information from images of population documents have shown to speed up the data entry process, although the performance of such techniques is not perfect, so a manual validation is still needed. In general, the performance of our developed methods vary depending on the paper degradation and the handwriting style. For example, the transcription accuracy varies from 60–90%, whereas the word spotting method reaches a performance between 80–90%. These values demonstrate that, although beneficial, automatic tools need human supervision for quality check. For this reason, further research will be focused on improving the automatic analysis of the document layout, the recognition of handwritten text, and the recognition of named entities.

5 DATA LINKAGE

Record linkage as the linkage of different appearances of the same persons is necessary to reconstruct the life course of individuals. This happened with both databases, in the BALL database to reconstruct families from a multigenerational perspective and to reconstruct a succession of marriages forming lineages, as is the case with the BHMD. The patrilineality of the transmission of surnames, which prevailed from the medieval period onwards in Catalonia mostly for men, and not changing female surnames throughout their life from modern times onwards, facilitated enormously the reconstruction of families and the application of different approaches for the data linkage. The varying number of reconstructed generations can be explained not only by the social and biological success of their descendants but also by migration. However, loss of migration is not that problematic, given the size of the territory and the continuity of the sources. At the same time, the areas covered by the two databases have historically been poles of attraction. We describe the three different types of methods we developed for linking the data and creating the social network, including constructing pedigrees and life courses. These three methods cover from the most basic statistical methods, which are based on string distances, to more sophisticated methods which build intermediate representations that allow record linkage between heterogeneous sources of information. The first proposed method, based on string distances, and the second one based on visual word search (namely word spotting) have been applied to both databases. The third method, which is based on knowledge graphs, has been applied only to the BALL database.

5.1 STRING DISTANCES

The first method is based on the comparison of strings (e.g., names and surnames). In the case of the marriage license records contained in the BHMD, the objective was to recursively link the married couples with their corresponding parents following a backward approach, thus creating lineages as pedigrees. For this purpose, we developed a software application that runs three different similarity measures between strings. These algorithms consider the length of each string and the position of each letter. The string comparison is based on the "Levenshtein distance" and a combination of the so-called "bag distance" and the "longest common substring distances". These algorithms are adapted to Catalan phonetics, allowing spelling variations. This meant that we could define which letters of the alphabet were usually used in the same place (similar phonetics) so that the similarity algorithm could take them into account when computing the final string distance. Besides the string distance, we also imposed several restrictions for linking the data for the BHMD when lineages were created. These restrictions consider data plausibility between marriages, setting a minimum difference of 15 years between the marriages of parents and children, and a 50-year maximum difference in extreme cases. To minimize the overlinks due to the similarity of onomastics, we carried out

a complete cross check of the links following historical and probabilistic criteria and considering both the number of kilometers between the geographical location and the social distance between the marriages of parents and children. The spatial distance between the place where the marriages of parents and children took place should follow a logical sequence according to the mobility patterns of the study period. The occupations of parents and their descendants and the amount of tax paid following each marriage should not be more than two tiers different in the fee scale described.

By applying the software and the aforementioned rules, family trees have been created for two different periods, since until 1481 the filiation of the bride and for the period 1645–1753 the name of the parents were not recorded, being key for the linkage of the generations. The first period ranges from 1500 to 1643 and the second period, from 1715 to 1880. Among the links generated by our rules and software (with a threshold of 85%), 33% of the total were validated. This reflects the fact that the Barcelona area was (and still is) a dynamic and mobile place and the restrictive nature of the method. The maximum number of linked generations for the 16th–17th centuries is four and for the 18th–19th centuries it is six, which could be explained both by a higher survival and greater stability of the surnaming system. Nevertheless, in both periods, two-level generations (parents–children) are the most frequent (see Table 3).

In the case of the BALL database, individuals were linked according to the string edit distance, also with a threshold of 85%. The procedure and restrictions differed slightly from those mentioned above for the BHMD, with the individuals in each household linked according to the relationship with the head of the household (e.g., sisters/brothers, sons). When a new census is introduced in the database, the algorithm searches whether these individuals also appear in the previous census. The string edit distance is computed together with some restrictions, namely that the street address should be the same and the individual's age should increase in a coherent way. If these constraints are satisfied, the unification is fully automatic and the new relatives are linked (e.g., newborns). By December 2020, all individuals present in the national censuses and censuses of five of the 14 municipalities (Sant Feliu de Llobregat, Collbató, Castellví de Rosanes, Santa Coloma de Cervelló and Sant Vicenç dels Horts) comprising the BALL database have been linked both across time and space. Thus, on average, more than 80% of the individuals have been linked to create a unique person. This implies that 80% of the individuals present in these municipalities have at least two observations over time either in the same municipality or in different municipalities (see Table 4). The maximum number of observations for the same individual is 12. In addition, tax information from personal taxes has also been linked, with a total linkage ratio of 95%.

In the coming months, new municipalities will continue to be linked to the BALL database. But it is also planned to interconnect the BHMD and BALL database for the overlapping years — 19th century and early 20th century — as the territory is totally coincident since the municipalities of the Baix Llobregat county were part of the Diocese of Barcelona, ecclesiastical demarcation of the BHMD, as explained above.

Table 3 *Number of marriage licenses, linkage validation and number of generations linked in the Barcelona Historical Marriage Database*

BHMD	**XVI–XVII (1500–1643)**	**XVIII–XIX (1715–1880)**
Marriage licenses	125,140	334,011
Linked marriages	21,200	109,224
	Genealogies	
2 generations	19,425	95
3 generations	172	12
4 generations	55	2
5 generations		255
6 generations		11

Table 4 *Number of inhabitants and linked individuals between two consecutive population register/census in BALL database*

Sant Feliu de Llobregat											
	1878	**1881**	**1889**	**1906**	**1910**	**1915**	**1920**	**1924**	**1930**	**1936**	**1940**
Total inhabitants	2,747	3,002	3,118	3,606	3,807	4,329	4,352	5,569	6,383	7,020	6,720
Total households	673	641	645	804	867	937	923	1,048	1,162	1,458	1,678
Linked individuals	-	2,573	2,627	3,169	3,411	3,788	3,888	4,544	5,295	5,940	5,109
% Linked individuals	-	85.7%	84.3%	87.9%	89.6%	87.5%	89.3%	81.6%	83.0%	84.6%	76.0%
Persons by household	4.08	4.68	4.83	4.49	4.39	4.62	4.72	5.31	5.49	4.81	4.00

Santa Coloma de Cervelló						
	1901	**1924**	**1936**	**1940**	**1945**	**1950**
Total inhabitants	542	1,132	1,311	1,218	1,167	1,222
Total households	127	260	325	329	307	308
Linked individuals	-	805	1,122	1,124	1,063	984
% Linked individuals	-	71.1%	85.6%	92.3%	91.1%	80.5%
Persons by household	4.27	4.35	4.03	3.70	3.80	3.97

Sant Vicenç dels Horts								
	1921	**1925**	**1930**	**1935**	**1940**	**1946**	**1950**	**1955**
Total inhabitants	2,097	1,985	2,950	3,129	2,980	3,014	3,323	3,717
Total households	535	409	783	716	763	791	1,030	1,168
Linked individuals	-	1,844	2,560	2,809	2,721	2,713	2,911	2,676
% Linked individuals	-	92.9%	86.8%	89.8%	91.3%	90.0%	87.6%	72.0%
Persons by household	3.92	4.85	3.77	4.37	3.91	3.81	3.23	3.18

Callbató														
	1887	**1889**	**1896**	**1900**	**1905**	**1910**	**1916**	**1920**	**1924**	**1930**	**1936**	**1940**	**1945**	**1950**
Total inhabitants	781	812	686	703	665	809	796	560	754	531	525	474	425	418
Total households	172	170	147	182	144	162	166	141	170	138	138	162	125	147
Linked individuals	-	651	647	607	634	716	748	544	650	516	503	440	399	352
% Linked individuals	-	80.2%	94.3%	86.3%	95.3%	88.5%	94.0%	97.1%	86.2%	97.2%	95.8%	92.8%	93.9%	84.2%
Persons by household	4.54	4.78	4.67	3.86	4.62	4.99	4.80	3.97	4.44	3.85	3.80	2.93	3.40	2.84

Castellví de Rosanes						
	1924	**1930**	**1936**	**1940**	**1945**	**1950**
Total inhabitants	283	279	273	281	242	268
Total households	61	64	62	73	50	60
Linked individuals	-	260	252	236	222	195
% Linked individuals	-	93.2%	92.3%	84.0%	91.7%	72.8%
Persons by household	4.64	4.36	4.40	3.85	4.84	4.47

5.2 VISUAL WORD SEARCH

The visual word search corresponds to the word spotting procedure described in the semi-interactive transcription system section. As explained previously, the key is to benefit from the redundant information between consecutive census records, which can be used to speed up the transcription while at the same time unifying the individuals that appear in the population register/census, and involves linking the same individual across different censuses, thus creating the life course. For each household in a previous census, the word spotting algorithm searches each individual living in that household in the current census. This search is performed by visual similarity, which means that the algorithm internally creates a visual representation of each name and surname. The individual is considered found if the system finds the same combination of name and surname/s, and the age is feasible. In such cases, the linkage (i.e., unification) is automatically performed. This procedure has proven to reduce the time of transcription of a new population register or census by 70% (Mas et al., 2016).

5.3 KNOWLEDGE GRAPH-BASED METHODS

A knowledge graph (KG) is a data representation that uses graphs to emphasize the links between data elements (Schneider, 1973). A KG formally represents object semantics by describing their relationships. Moreover, they can also make use of ontologies in an abstraction layer to allow logical inference for implicit knowledge retrieval (Baader, Calvanese, McGuiness, Nardi, & Patel-Schneider, 2003). The KG has become more popular since Google introduced their KG to complement their string-based search engine (Singhal, 2012), and they are currently used by many multinational firms to provide their customers with tailored services. There are two major KG representations: entity-relation (ER) graphs and entity-attribute-relation (EAR) graphs (see Figure 5). In an ER graph, nodes are the world objects (entities), which contain a set of attributes that describe each node, and edges (relations), which link nodes between them. This is the simplest KG representation and any AI algorithm reasoning on it must navigate through the graph edges to infer a result. However, databases like the BHMD and the BALL database, contain slightly different information from individuals that cannot be modelled by a single ER KG. EAR graphs model this data better than ER graphs because there are two kinds of nodes: entities and attributes. The node attributes in an ER KG are extracted from the node and are "promoted" to nodes in the EAR KG and, consequently, each attribute node is connected to its original node in the equivalent ER KG representation. In Figure 5 we show how this procedure works. For instance, individual information like "first_name", "surname", "marital_status", etc. that are defined as attributes of individual nodes in Figure 5a are defined as nodes in the EAR KG representation which are linked to individuals in Figure 5b. Thus, an individual is not represented only by a set of attributes but by a subgraph, in which the central node is the Individual. This representation provides more flexibility as kin relations and individual information are represented homogenously and all of them contribute to build the numerical representation that will be used to represent individuals for record linkage tasks.

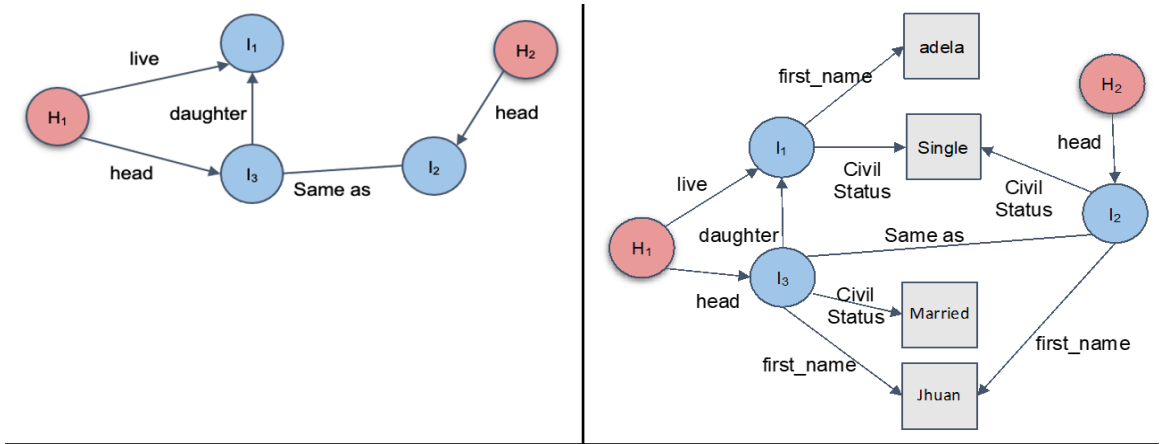
As motivated above, we use EAR KG to model data extracted from the BALL database. Following this representation, entities mainly correspond to individuals and households, while attribute nodes are the field values (name, surname, age, occupation, household address, etc.) describing each individual or household. The KG also has two kinds of edges. On the one hand, there are the edges linking entity nodes representing family relationships between individuals living in the same household, the edges linking individuals and households to denote the household in which each individual lived, and the edges between two or more individual nodes representing the same person through their records over years of censuses. On the other hand, there are also edges between entity nodes and attribute nodes. As mentioned above, attribute nodes contain the values of the fields extracted from the BALL dataset. All the attribute nodes are therefore essentially composed of the word vocabulary appearing in this database. The edges between entities and attributes represent the semantics, i.e., name, surname, occupation, and so on, assigning the item to each attribute node when it is linked to an entity.

Record linkage on an EAR KG is an inference method to discover links between individual entities representing the same person over time. Latest advances in this field compute numerical representations of each node in the KG. Each numerical representation is a set of N real values, where N is an arbitrary value set beforehand by an expert and usually takes a value ranging from 32 to 64, that encodes each node. Since each node in this database represents, either an individual or a household, each numerical representation is built from the attribute nodes that are linked to each individual or household. The main differences between these methods are the way that these numerical representations are computed for each individual, but all of them essentially seek the same purpose, for the two nodes that represent the same individual, or household, to

have their numerical representation close to each other. For instance in Figure 5b, we can see two individuals, I2 and I3, that should be identified as being the same person by any record linkage method. The proposed EAR KG methods will learn numerical representations that will be encoded in a N-dimensional space. These representations will consider the edges that connect nodes I2 and I3, respectively. If both nodes had the same neighbors, i.e. they live in the same household, had the same first name, marital status, etc. they would exactly have the same numerical representations and hence the record linkage method would provide a perfect match. In the given example, I2 lives in a different household and his marital status has also change from "single" to "married". Consequently, the numerical representations will differ but should be close enough to provide a match of the record linkage method.

We have applied a representative set of these techniques to a subset of the BALL, choosing records from the municipality of Sant Feliu de Llobregat and splitting the data into the following partitions. We have imported the data into a local Neo4j Database server, which is a native graph database system. We used two pairs of datasets for training, (A,B) as (1889, 1906) and (1930, 1936), (1906, 1910) and (1936, 1940) for validation, and (1910, 1924) and (1924, 1930) for testing. The selected methods are TransE (Bordes, Usunier, Garcia-Duran, Weston, & Yakhnenko, 2013) and TransH (Wang, Zhang, Feng, & Chen, 2014), which are applied to ER KG; SEEA (Guan et al., 2017) and KR-EAR (Lin, Liu, & Sun, 2016), which are applied to EAR KG; and the WERL method, which is applied to an extension of the EAR KG that we recently developed in (Gautam, Ramos-Terrades, Pujadas-Mora, & Valls-Fígols, 2020).

Figure 5 *Types of KG representations*



(a) ER KG. All nodes are of the same kind (entities). Individual information (name, surname, marital status, ...) are attributes of each node.

(b) EAR KG. Circled nodes represent entities and the squared, attributes.

Source: Gautam, Ramos-Terrades, Pujadas-Mora, and Valls (2020).

Table 5 *Record linkage results for the BALL database only for the town of Sant Feliu de Llobregat*

Method	Accuracy	Precision	Recall	F-Score
TransE (ER)	0.87	0.19	0.25	0.21
TransH (ER)	0.81	0.12	0.26	0.17
KR-EAR (EAR)	0.79	0.13	0.30	0.18
SEEA (EAR	0.91	0.07	0.10	0.08
WERL (ER)	0.99	0.99	0.67	0.80
WERL (EKG)	**0.99**	**0.98**	**0.89**	**0.93**

Source: Gautam, Ramos-Terrades, Pujadas-Mora, and Valls (2020).

Comparing benchmark methods, the reported results show a significant performance improvement on the WERL-based method in terms of the F-Score (see Table 5). The F-Score is a performance measure, ranging between 0 to 1, that combines the precision and recall of a system like a record linkage method. It is formally defined as the harmonic mean of the Precision and the Recall measures and is easily computed as:

$$F - Score = 2\,\frac{Precision\ x\ Recall}{Precision + Recall}$$

The Precision and Recall, aka sensitivity, are two measures used to evaluate the impact of false positives (Type I error) and false negatives (Type II error) in the system performance:

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

where TP, FP and FN stand for true positive pairs, false positive pairs and false negative pairs, respectively. This improvement is weakening when we observe the accuracy of the methods. A few comments are needed to better understand these figures. First, record linkage is a highly unbalanced classification problem, meaning that there are many more pairs of not linked records (true negative pairs) than records to be linked (true positive pairs). Thus, automatic record linkage methods biased to not link candidates of pairs are likely to increase their accuracy. They therefore achieve low precision and recall values and, consequently, low F-score values. Conversely, the proposed WERL methods have been shown to have good accuracy and precision scores, meaning that linked pairs are likely to be correct by relatively low recall scores, and indicating that some record links are missed. Given the amount of data to be processed, the challenge is to increase the recall score while losing minimal precision.

6 CONTRIBUTIONS TO HISTORICAL DEMOGRAPHY

The historical lines of research on demography using the BHMD and the BALL deal with the intergenerational transmission of social status, intermarriage and kinship marriages within the social reproduction process (15th–20th centuries), the evolution of migratory flows in a long-term view (15th–20th centuries), and the estimation of economic inequality within preindustrial and industrial periods.

6.1 SOCIAL REPRODUCTION

Social reproduction is a classic research topic in historical demography, referring to the economic, institutional, legal, political and/or cultural mechanisms used by individuals, families, or social groups to maintain, improve and/or transmit their social position (Bourdieu, 2013). However, dimensions such as the intergenerational transmission of social outcomes (mostly of occupations or social status) and intermarriage or kin marriages, which appear to be fundamental to social dynamics and change, have been scarcely studied for preindustrial periods or from a long-term perspective. This is, in part, owing to the non-availability of sources and data for earlier periods and different historiographical traditions in every country.

Families used intergenerational transmission, intermarriage, or kin marriages, which were subject to demography, legal systems, and economy expansion and contraction (Berkner & Mendels, 1976; Ganzeboom, de Graaf, & Treiman, 1991), to place their members in better-off positions and to perpetuate their lineages biologically and socially over time. This phenomenon led preindustrial society to appear static, although it is true that levels of social transmission between generations or social endogamy, and even territorial endogamy, remained high in these societies. However, this may have varied among different social/migratory groups since they might have adapted their strategies depending on their specific circumstances and the general context (Bourdieu, 1976; Laslett, 1983), which ultimately involves group identities and consequently class formation (Kocka, 1984).

The socio-occupational destinations of sons and daughters of peasants and artisans in the area of Barcelona in the period 1547 to 1643 were analyzed using the BHMD to explore how intergenerational transmission worked in preindustrial periods (Pujadas-Mora, Brea-Martínez, Jordà, & Cabré, 2018). Destinations are understood as a combination of both ascription and achievement, i.e., the influence of parents' socioeconomic

positions and individual improvements in relation to family background. Notably, the legal context, and specifically inheritance systems, can strengthen the intergenerational transmission of occupations or status, particularly when the system is based on the principle of impartibility, which usually granted eldest sons the privilege of being heir, as was the case in Catalonia. At the same time, this research aims to shed light on the competitive or cooperative strategies adopted by families which could prevent or facilitate parent-offspring conflict (Schlomer, Ellis, & Garber, 2010; Trivers, 1974) and resource dilution (Low & Clarke, 1991; Shenk et al., 2010). A multilevel approach was applied using hierarchical linear models and logistic regressions. The benefit of this kind of approach lies in the possibility of analyzing how similar siblings were in status attainment in view of the relatively high level of ascription in pre-industrial periods.

According to the findings, high levels of intergenerational status inheritance and ascription were closely linked to the effect of the parents' social class, a plausible fact in an ordered society with an inheritance system based on the impartibility of property. The pattern among farmers, whose children also tended to be farmers, accounts in part for the high levels of intergenerational transmission. Artisans also contributed to raising the total intergenerational transmission, given that sons of artisans mainly inherited the condition but not the occupation. First-married children were the main inheritors of parental status in all social groups, and especially among farmers and artisans. However, some second- or third-married farmers' children were able to follow artisan careers. Likewise, the farmers (mainly non-inheritors) joined the then-existing waged rural labor market, which could sometimes be used to complement earnings from their own land. However, low availability of cropland (Alvarez Nogal & Prados de la Escosura, 2007; Ferrer-Alós, 1983, 2014; Vilar, 1986), accompanied by the expansion of the textile sector (García Espuche, 1998; Torras, 1998) in early modern Catalonia could have prompted families to encourage some of their children (usually the non-first-married) to take up artisan activities. Hence, rather than a model of sibling competition for family resources, we see a pattern of family economic diversification, which is one of the most relevant and novel conclusions. The most paradigmatic example of this is probably the significant number of farmers with children who became *pelaires*, or fiber preparers.

Using the BHMD, intermarriage is approached through the study of immigrant marriage due to the important migratory flow to Catalonia from France that took place mainly in modern centuries (Amengual-Bibiloni & Pujadas-Mora, 2020). The analysis of this flow has an important tradition in Spanish historiography, starting with the seminal works of Moreu Rey (1959) and Nadal and Giralt (1960), although it is little known in the international literature. The volume and length of this flow makes it one of the most important in modern Europe, one result of which is the important number of these migrants who got married in Catalonia, a neglected topic in both national and international historiography. Up to 25% of the marriages that took place in the Diocese of Barcelona in the period 1568–1639 were between a French groom and a native wife. Multivariate logistic models confirm that French grooms were significantly less likely to get married to a bride from an equivalent social group than natives, with high odds of marrying a widow bride. Nevertheless, the interaction between a specific period (the whole period, 1568–1639, was split into sub-periods of ten years) and geographical origin of the groom (native and French) presents a decrease over time of the odds of finding a partner of the same social status for French and native grooms. On the one hand, the increase in the relative weight of French grooms modified the structure of the whole marriage market, and on the other hand, native brides, mainly widows, became less reluctant to get married with a French groom. Notably, group size is a key factor to better understand the social homogamy: the larger the size, the greater the probability of homogamy (except for the nobility), even when the marriage market is considered as a segmented market by social group, geographical origin, and period. The important presence of French migrants on the marriage market shows that there was an important marriage squeeze due to more male candidates, which would explain why native widows were remarrying more than in previous or subsequent periods.

Last, the practice of kin marriages as a mode of social reproduction was explored through the BHMD. The high increase in the number of consanguineous and affine marriages throughout the 19th century was a common phenomenon in different European countries (Bittles & Black, 2010; Bras, Kok, & Mandemakers, 2010; Chacón Jiménez & Hernández Franco, 1992; Delille, 1994; Gouesse, 1984; Sutter, 1968). Two types of kinship-affine marriages, simultaneous marriages of two or more couples of siblings and marriages of widowers or widows with the sister or brother of their deceased spouses, phenomena known as levirate and sororate marriages, are considered (Pujadas-Mora, Brea-Martínez, Jordà, & Cabré, 2018). The aim of this study was to establish the social profile of these types of marriages and to try to find the determinants of choosing a partner among kin in the Barcelona area in the period 1780–1880. The farmers and skilled and lower skilled workers, mainly artisans, and the day laborers were the groups with the highest percentage of simultaneous marriages, and particularly notable was the weight of farmers in simultaneous marriages of brothers and sisters. This could be because the eldest son, the universal inheritor, was marrying at the same

time as one of his sisters, who was marrying another universal inheritor, meaning that she was marrying upwards. The high proportion of simultaneous marriages among day laborers may be due to the broad application of that occupational title. Men were classified as day laborers even when their father or brother was a land owner.

Logistic regressions were performed to disentangle the weight of the factors that determined sororate and levirate marriages, with the novelty of a control variable for the frequency (distribution) of surnames in the population. The widowers from the highest social groups, mostly elites and professionals, showed higher odds of marrying their sisters-in-law than the day laborers. Furthermore, there was no difference in this practice between urban and rural areas. An important concern in this study is the fact that these kinds of marriages were completely dependent on the demographic circumstances of the death of a spouse, the existence of an available sister or brother for sororate or levirate marriage, or the survival of at least two siblings of a marriageable age. The demographic transition helped the occurrence of these kinds of marriages due to an increased number of relatives on the marriage market following an increase in survival at marriageable ages.

6.2 MIGRATION

The analysis of historical migrations is a difficult issue due to the scarcity or even lack of specific sources that compile information about migratory events. This is the reason why onomastics, and mainly the exploration of family names, has been used in demography, history, and biology to assess the mobility of historical populations. In medieval and modern Europe, onomastics in general and particularly surnames were subject to various changes resulting from the continuous flow of the rural population to the urban centers, in addition to the homeless, pilgrims, and traders, and in modern Catalonia to a lesser extent, immigrants from the Kingdom of France. The combination of all these processes meant that between medieval and modern times there were important variations in the number, prevalence, and distribution of Catalan surnames. A total of 47% of the surnames appear for the first time in the Barcelona Diocese's Marriage Licenses Books, compiled in the BHMD, between the periods 1451–1487 and 1487–1500, which could be a sign that these surnames were brought to the area of Barcelona by recent immigrants. Conversely, 64% of surnames disappeared in the same period. These results suggest that approximately half of the surnames recorded in the BHMD during these years could have a migrant origin (Jordà, Ameijerias-Alonso, & Pujadas-Mora, 2018; Jordà, Pujadas-Mora, & Cabré, 2016).

When an isonomic analysis of onomastics is carried out for the whole of Catalonia using the household counts of 1497 and 1553, three different internal migratory patterns are observed. The first is a north-south population flow, seemingly because of a steady migration from the Catalan Pyrenees (north) to the Tarragona and Montblanc areas (south). The second is an inland migration trend. And the third is a continuous settlement pattern in the Barcelona hinterland. Moreover, it should be noted that a significant increase in the number of households is observed from 1497 to 1553, which would have led to a dramatic reduction in the distribution of ancient Catalan family names (Jordà et al., 2016). These results could support the notion that during the first half of the 16th century the territory underwent several internal migratory flows, favored by the economic growth of the time and the resulting increase in the demand for labor, which was also in part covered by migrants from France (Amengual-Bibiloni & Pujadas-Mora, 2016).

Moreover, it should be noted that Catalan onomastics are the result of the interrelations between Catalan, Occitan, French, and Castilian surnames. Immigration from France and then from the rest of the different Spanish kingdoms enriched Catalan onomastics by introducing new forms of surnames, not always distinguishable from a linguistic criterion given a relatively high number of homonymic surnames among the languages and the absence of standardized languages. Cluster analysis allowed for organizing the surnames into 3 different groups by similarities of growth, independent of linguistic and historical origins. The first group shows the highest number of Catalan surnames; the second has the highest percentage of surnames borne by foreigners; and the third, which is the most heterogeneous of all the groups, comprises the largest number of Castilian surnames and the smallest number, in relative terms, of surnames influenced by French immigration (Jordà et al., 2016).

All these studies contribute to confirming that there was much geographical mobility throughout the preindustrial period, both internally and internationally, allowing a better understanding of how migrants contributed to revitalizing the cities from a demographic point of view, which had suffered a chronic negative natural growth rate due to poor living conditions, as had the Catalan coastal villages. At the same time, they help to demystify the ethnic origin of surnames.

6.3 ECONOMIC INEQUALITY

Increased economic inequality is one of the most worrying problems nowadays and has evolved into a central topic of historiographical debate. A historical perspective shows the roots of current inequality and is key in the understanding of social mobility trends (Álvarez Nogal & Prados de la Escosura, 2013; Milanovic, 2016). Social mobility is a major topic in Historical Demography, proving its interaction with demographic behaviour (Dribe, Van Bavel, & Campbell, 2012; Schnore, 1961; Song, 2021). In this way, knowing the general context of inequality, and its evolution, appears to be determinant in any study of social stratification. However, the evolution of economic inequality in a long perspective is difficult due to the lack of direct sources (Alfani, 2015; Milanovic, Lindert, & Williamson, 2011; Piketty, 2015). The link between inequality and social mobility has been little explored from a historical view, and this will be one of our next research topics.

A long view analysis of the estimation of economic inequality was made possible thanks to the BHMD, which offers continuity over time (1451–1905) and space (the Barcelona area, the most populated area of Catalonia, also including the city of Barcelona). The fact that it is a fiscal source showing what each couple paid to get their marriage license meant that it was a measure of wealth. These taxes were imposed on the whole of society, which is not always the case with fiscal sources, one example being the tax exemption of the nobility throughout preindustrial periods (Pujadas-Mora & Brea-Martínez, 2020). Inequality in the Barcelona area was higher in preindustrial societies than in industrial ones, as recently stated by other authors (Alfani, 2015; Álvarez-Nogal & Prados de la Escosura, 2007; Milanovic, 2012, some of them already quoted) who disagree with the conclusions of the seminal work *Economic Growth and Income Inequality* (1955) by Simon Kuznets.

Three periods in the evolution of inequality in the Barcelona area are described. In the first, between 1481 and 1649, there was low inequality, which stabilized with the economic stagnation after the Catalan Civil War (1462–1472) and a whole series of mortality crises caused by different episodes of the plague. In the second, between 1650 and 1749, the levels of disparity increased significantly, coinciding with the transformation of the rural economy during the 17th century when new sharecropping contracts (*Rabassa Morta*) were established and the export of wine products increased. Inequality peaked around 1740 with the resurgence of the Catalan economy within the framework of an early proto-industrialization that took place in many rural areas, along with considerable population growth and the introduction of the first industrial establishments in Barcelona. In the third, between 1750 and 1880, there was a stabilization and subsequent decline in the disparity because of the War of Independence (1808–1814), followed by an upward trend that concurs with the date that traditionally marks the beginning of industrialization in Catalonia (1833). Inequality almost doubled between 1830 and 1880, driven by a significant increase in the levels of tax exemption (number of marriages receiving the marriage license by Amore Dei) among day laborers and workers.

As a result of decomposing the trends of economic inequality by social groups, it was observed that the nobility, liberal professionals, and merchants contributed positively to inequality, or in other words these groups had a proportionally greater weight in terms of wealth than population numbers. For their part, artisans, day laborer, and peasants contributed negatively by presenting a greater demographic weight than wealth. In fact, inequality in the Old Regime can be explained by a hierarchical society where the nobility concentrated much of the wealth, as expected, while from the second half of the 18th century onwards the groups that contributed to increasing inequality were those that worked in the factory system (day laborers, weavers, etc.), showing a clear proletarian effect. On analyzing the accumulation of wealth/income among the top 1% of the population along the five centuries covered by the BHMD, it was seen that by the end of the 17th century this group had gone from holding 10% of the total wealth to almost 17%, which was the highest point, decreasing again to 12% by 1880 (Pujadas-Mora & Brea-Martínez, 2020). In fact, the richest members of society were much richer in the preindustrial period, even though the development of capitalism allowed for greater levels of concentration. Socially, the nobles constituted more than 50% of the richest 1%, with their presence declining from the mid-17th century onwards to reach a figure of 10% at the end of the 18th century and during the 19th century, in a situation resembling what occurred in other European countries. The nobility was replaced by the large merchants and traders and the liberal professionals who by the end of the 19th century made up more than 80% of the top 1% in the Barcelona area.

Focusing the research on industrialization, it has been shown that labor market transformation and inequality were fundamental aspects in its consolidation (Brea-Martínez & Pujadas-Mora, 2018a). There was a massive incorporation of unskilled workers into industries from agriculture, the rural exodus caused by technical improvements in agricultural production and the development of the factory system (Martínez-Galarraga & Prat, 2016; Nadal, 1975; Sánchez Suárez, 1993). Using the BHMD, we show how all these processes

contributed to both the increased levels of economic inequality in the area of Barcelona along the 18th and 19th centuries and the transformation of the economic sectors. From 1780 onwards, there was an important decrease in the primary sector and a simultaneous increase in the secondary and tertiary sectors, which would advance the traditional starting date of industrialization in Catalonia. Moreover, between-sector economic inequality rose, as well as within the secondary sector (textile), due to a process of polarization. Conversely, the primary sector levelled-off in terms of disparity levels and the tertiary sector showed a declining pattern.

In general, the use of the BHMD for this specific research topic allowed us to introduce three fundamental innovations. First, using a marriage tax as a measure of wealth implies that the individuals are measured at the same time in the life cycle (in the same age range), making them more comparable to each other since the accumulation of wealth happens over time. Second, these tax records included everyone who married, even the nobles, who were otherwise almost universally tax exempt, and those exempt due to poverty. Thus, we have a cross-sectional view not usually provided by regular tax sources in Old Regime societies, which were subject tax fraud and tax evasion. And third, by merging the occupation or social status of the grooms together with the level of the marriage tax the economic concept of 'ability to pay' can be applied. The combination of the two magnitudes provides a more precise approach to the human capital of the individuals analyzed.

7 FINAL CONSIDERATIONS AND FUTURE AGENDA ON EXPANDING BHMD AND BALL DATABASES

The BHMD and the BALL provide proof of how artificial intelligence can facilitate the application of Handwritten Text Recognition (HTR) techniques to data collection of primary sources, which is key to building individual-level databases and reducing the time required for their construction. However, the current state-of-the-art of Handwriting Recognition means that some human intervention is still required, which explains the crowdsourcing and game-sourcing experiences implemented. Moreover, knowledge graph techniques have allowed the application of advanced record linkage, meaning that data complexity can be increased and its entire volume analyzed. In short, these approaches undoubtedly contribute to the so-called big data in terms of historical data.

Our future agenda is to continue building the BALL with the individual data of population registers/censuses and personal taxes for the 19th and 20th century from other Catalan counties. In December 2020, we initiated a crowdsourcing experience in the county of Maresme in the north of Barcelona. The forecast is that the same data for five new counties will be added in 2021. We estimate having 1.5 million individual observations by the end of the year. Besides, we plan to merge new textual sources to extend the individual socioeconomic information for the already compiled demographic data and document images to extract meaningful information to complement these demographic and economic data.

Future research in historical demography using the Barcelona Historical Marriage Database will examine:

- Assortative mating among French immigrant women in the great migratory wave received during the Hispanic Monarchy between the 16th and 18th centuries;
- Sibling similarities in social and occupational attainment in a long-term perspective from the 16th to the 19th century;
- Structural and strategic determinants of the formation of kin (isonymic) marriages;
- The association of residential segregation and socioeconomic inequality in 18th and 19th century Barcelona;
- The roles of politics and industrialization in the abandonment of the traditional marriage calendar;
- The diminishing influence of families on labor careers during the 19th and 20th centuries and the transition from stem families to nuclear families.

From the technological perspective, databases with the corresponding ground truth (images, transcriptions, named entity annotation) are being used to organize challenges and competitions in computer vision and document analysis and recognition. The databases will continue to be used for tasks like document object detection (layout segmentation), handwriting recognition, word spotting and information extraction in order

to advance in the automatic transcription of manuscript demographic sources. At the highest level, the representational models based on linked data structures will be a valid benchmark for testing algorithms on record linkage/link discovery. Another future line of research will be the transfer of the current algorithms to other types of archival data that contain complementary information for the databases, such as marriage advertisements in historical newspapers).

ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish Science Ministry projects RTI2018-095533-B-I00, RTI2018-095645-B-C21, RYC-2014-16831, the Catalan project 2017-SGR-1783, the Catalan Ministry of Culture, the CERCA Program/Generalitat de Catalunya and the Culture Department of the Generalitat de Catalunya.

REFERENCES

Alfani, G. (2015). Economic inequality in northwestern Italy: A long-term view (fourteenth to eighteenth centuries). *The Journal of Economic History, 75*(4), 1058–1096. Retrieved from https://EconPapers.repec.org/RePEc:cup:jechis:v:75:y:2015:i:04:p:1058-1096_00

Almazán, J., Gordo, A., Fornés, A., & Valveny, E. (2014). Word spotting and recognition with embedded attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(12), 2552–2566. doi: 10.1109/TPAMI.2014.2339814

Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies, 1*, 1–26. doi: 10.51964/hlcs9290

Álvarez-Nogal, C., & Prados de la Escosura, L. (2007). The decline of Spain (1500–1850): Conjectural estimates. *European Review of Economic History, 11*(3), 319–366. Retrieved from https://EconPapers.repec.org/RePEc:cup:ereveh:v:11:y:2007:i:03:p:319-366_00

Álvarez-Nogal, C., & Prados de la Escosura, L. (2013). The rise and fall of Spain (1270–1850). *The Economic History Review, 66*(1), 1–37. doi: 10.1111/j.1468-0289.2012.00656.x

Amengual-Bibiloni, M., & Pujadas-Mora, J. M. (2016). Orígens i destins de la immigració francesa a l'àrea de Barcelona (1481–1643). Aportacions a partir de la Barcelona Historical Marriage Database [Origins and destinations of French immigration in the area of Barcelona (1481–1643). Contributions from the Barcelona Historical Marriage Database]. *Manuscrits. Revista d'Història Moderna, 34*, 35–61. Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo=5974891

Amengual-Bibiloni, M., & Pujadas-Mora, J. M. (2020). Cruce de caminos: Matrimonios de viudas y franceses en el área de Barcelona en los siglos XVI y XVII [Crossroads: Marriages between local widows and French migrants in the Barcelona area in the 16th and 17th centuries]. In Tovar Pulido, R. (Ed.), *De humilde e ilustre cuna: Retratos familiares de la España moderna (siglos XV-XIX)* (pp. 86–123). Évora : Publicações do Cidehus. doi: 10.4000/books.cidehus.10726

Baader, F., Calvanese, D., McGuiness, D. L., Nardi, D., & Patel-Schneider, P. F. (2003). *The description logic handbook: Theory, implementation and applications*. Cambridge: Cambridge University Press.

Baucells, J. (2002). 'Esposalles' de l'arxiu de la catedral de Barcelona: Un fons documental únic (1451–1905) ['Marriage licenses' of the Barcelona cathedral's archive: A unique documentary collection (1451–1905)]. *Butlletí del Servei d'Arxius, 35*, 1–2.

Berkner, L. K., & Mendels, F. F. (1976). Inheritance systems, family structure, and demographic patterns in Western Europe, 1700–1900. In C. Tilly, *Historical studies of changing fertility* (pp. 209–223). Princeton: Princeton University Press. doi: 10.1515/9781400871452-006

Bittles, A. H., & Black, M. L. (2010). Consanguineous marriage and human evolution. *Annual Review of Anthropology, 39*, 193–207. doi: 10.1146/annurev.anthro.012809.105051

Bloothooft, G. (1998). Assessment of systems for nominal retrieval and historical record linkage. *Computers and the Humanities, 32*(1), 39–56. Retrieved from https://www.gerritbloothooft.nl/Publications/CHUM36.htm

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26 (NIPS 2013)* (pp. 2787–2795). Retrieved from https://papers.nips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf

Bourdieu, P. (1976). Marriage strategies as strategies of social reproduction. In R. Forster & O. Ranum (Eds.), *Family and society: Selections from the annales, économies, sociétés, civilisations* (pp. 117–144). Baltimore: Johns Hopkins University Press.

Bourdieu, P. (2013). *Distinction: A social critique of the judgment of taste*. Abingdon: Routledge. (Original work published 1984)

Bras, H., Kok, J., & Mandemakers, K. (2010). Sibship size and status attainment across contexts: Evidence from the Netherlands, 1840–1925. *Demographic Research, 23*(4), 73–104. doi: 10.4054/DemRes.2010.23.4

Brea-Martínez, G., & Pujadas-Mora, J. M. (2018a). Transformación y desigualdad económica en la industrialización en el área de Barcelona, 1715–1860 [Transformation and economic inequality in Industrialization in the area of Barcelona, 1715–1860]. *Revista de Historia Económica/ Journal of Iberian and Latin American Economic History, 36*(2), 241–273. doi: 10.1017/S0212610917000234

Brea-Martínez, G., & Pujadas-Mora, J. M. (2018b). Estimating long-term socioeconomic inequality in southern Europe: The Barcelona area, 1481–1880. *European Review of Economic History, 23*(4), 397–420. doi: 10.1093/ereh/hey017

Carbonell, M, Fornés, A., Villegas, M., & Lladós, J. (2020). A neural model for text localization, transcription and named entity recognition in full pages. *Pattern Recognition Letters, 136*, 219–227. doi: 10.1016/j.patrec.2020.05.001

Carbonell, M., Riba, P., Villegas, M., Fornés, A., & Lladós, J. (2021). Named entity recognition and relation extraction with graph neural networks in semi structured documents. *2020 25th International Conference on Pattern Recognition (ICPR)*, 9622–9627. doi: 10.1109/ICPR48806.2021.9412669

Carreras Candi, F. (1913). Les obres de la Catedral de Barcelona 1298–1445 [The works of the Cathedral of Barcelona 1298–1445]. *Butlletí de la Reial Acadèmia de Bones Lletres de Barcelona, 7*(49), 22–30. Retrieved from https://raco.cat/index.php/BoletinRABL/article/view/201731

Chacón Jiménez, F., & Hernández Franco, J. (Eds.) (1992). *Poder, familia y consanguinidad en la España del antiguo régimen* [Power, family and consanguinity in ancien régime Spain]. Madrid: Anthropos.

Chen, J., Riba, P., Fornés, A., Mas, J., Lladós, J., & Pujadas-Mora, J. M. (2018). Word-Hunter: A gamesourcing experience to validate the transcription of historical manuscripts. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 528–533. doi: 10.1109/ICFHR-2018.2018.00098

Das, S., Sial, H. A., Ma, K., Baldrich, R., Vanrell, M., & Samaras, D. (2020). Intrinsic decomposition of document images in-the-wild. *Proceedings of the British Machine Vision Conference (BMVC)*, 1–14. doi: 10.48550/arXiv.2011.14447

Delille, G. (1994). Consanguinité proche en Italie du XVIe au XIXe siècle [Consanguinity in Italy from the sixteenth to the nineteenth century]. In Bonté, P. (Ed.), *Épouser au plus proche. Inceste, prohibitions et stretégies matrimoniales autour de la Méditerranée* (pp. 323–340). Paris: EHESS.

Dribe, M., Van Bavel, J., & Campbell, C. (2012). Social mobility and demographic behavior: Long term perspectives. *Demographic Research, 26*,173–190. doi: 10.4054/DemRes.2012.26.8

Ferrer-Alòs, Ll. (1983). Censals, vendes a carta de gràcia i endeutament pagès al Bages (s. XVIII) ['Censals', sales by letter of grace and peasant indebtedness in the Bages (18th century)]. *Estudis d'història agrària, 4*, 101–128. Retrieved from https://raco.cat/index.php/EHA/article/view/99535

Ferrer-Alòs, Ll. (2014). Derechos de propiedad y mercado de la tierra en la Cataluña Vieja (s. XV-XIX). El caso de Artés (Bages) [Property rights and land market in Old Catalonia (15th–19th century). The case of Artés (Bages)]. *Historia agraria: Revista de agricultura e historia rural, 62*, 47–82. Retrieved from https://ideas.repec.org/a/seh/journl/y2014i62maprilp47-82.html

Fornés, A., Llados, J., Mas, J., Pujadas-Mora, J. M., & Cabré, A. (2014). A bimodal crowdsourcing platform for demographic historical manuscripts. *DATeCH '14: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 103–108. doi: 10.1145/2595188.2595199

Ganzeboom, H. B. G., de Graaf, P. M., & Treiman, D. J. (1991). A standard international socio-economic index of occupational status. *Social science research, 21*(1), 1–56. doi: 10.1016/0049-089X(92)90017-B

García Espuche, A. (1998). *Un siglo decisivo: Barcelona y Cataluña, 1550–1640* [A decisive century: Barcelona and Catalonia, 1550–1640]. Barcelona: Alianza Editorial.

García Pérez, M. S. (2007). El padrón municipal de habitantes: Origen, evolución y significado [Population register: Origin, evolution and significance]. *Hispania Nova: Revista de historia contemporánea, 7*, 1–5. Retrieved from http://hispanianova.rediris.es/7/articulos/7a005.pdf

García Ruipérez, M. (2012). El empadronamiento municipal en España: Evolución legislativa y tipología documental [The population register in Spain: Legislative evolution and documentary typology]. *Documenta & Instrumenta, 10*, 45–86. doi: 10.5209/rev_DOCU.2012.v10.40485

Gautam, B., Ramos-Terrades, O., Pujadas-Mora, J. M., & Valls, M. (2020). Knowledge graph based methods for record linkage. *Pattern Recognition Letters, 136*, 127–133. doi: 10.1016/j.patrec.2020.05.025

Goiser, K., & Christen, P. (2006). Towards automated record linkage. *AusDM '06: Proceedings of the fifth Australasian conference on Data mining and analytics, 61*, 23–31.

Gouesse, J. M. (1984). Mariages de proches parents (XVIe–XXe siècle). Esquisse d'une conjoncture [Marriages of close relatives (16th–20th century). Sketch of a conjuncture]. *Publications de l'École française de Rome, 90: Le modèle familial européen. Normes, déviances, contrôle du pouvoir. Actes des séminaires organisés par l'École française de Rome et l'Università di Roma* (pp. 31–61). Rome: École Française de Rome.

Goyer, D. S., & Draaijer, G. E. (1992). *The handbook of national population censuses: Europe*. Santa Barbara: Greenwood.

Guan, S., Jin, X., Jia, Y., Wang, Y., Shen, H., & Cheng, X. (2017). Self-learning and embedding based entity alignment. *IEEE International Conference on Big Knowledge (ICBK)*, 33–40. doi: 10.1109/ICBK.2017.15

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). Record linkage — Methodology. In T. N. Herzog, F. J. Scheuren & W. E. Winkler, *Data quality and record linkage techniques* (pp. 81–92). New York: Springer. doi: 10.1007/0-387-69505-2_8

Jordà, J. P., Ameijerias-Alonso, J., & Pujadas-Mora, J. M. (2018). Chronicle of an early demise, surname extinction in the fifteenth and the seventeenth centuries. *Historical Methods, 51*(3), 190–201. doi: 10.1080/01615440.2018.1462747

Jordà, J. P., Pujadas-Mora, J. M., & Cabré, A. (2014). The footprint of migrations on surnames: Onomastic changes in the Barcelona area at the late Middle Ages (1451–1500). *Onoma, 49*, 105–136. doi: 10.2143/ONO.49.0.3285500

Jordà, J. P., Pujadas-Mora, J.,M., & Cabré, A. (2016). Surnames and migrations: The Barcelona area (1451-1900). *Names and their Environment. Proceedings of the 25th International Congress of Onomastic*, 131–143. Retrieved from https://ddd.uab.cat/record/174338

Kang, L., Riba, P., Rusiñol, P., Fornés, A., & Villegas, M. (2020b). Pay attention to what you read: Non-recurrent handwritten text-line recognition. *arXiv:2005,13044*. Retrieved from https://arxiv.org/abs/2005.13044

Kang, L., Riba, P., Villegas, M., Fornés, A., & Rusiñol, M. (2020a). Candidate fusion: Integrating language modelling into a sequence-to-sequence handwritten word recognition architecture. *Pattern Recognition, 112*, 107790. doi: 10.1016/J.PATCOG.2020.107790

Kang, L., Rusiñol, M., Fornés, A., Riba, P., & Villegas, M. (2020c). Unsupervised adaptation for synthetic-to-real handwritten word recognition. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3491–3500. doi: 10.1109/WACV45572.2020.909339

Kleber, F., Diem, M., Dejean, H., Meunier, J.-L., & Lang, E. (2018). Matching table structures of historical register books using association graphs. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 217–222. doi: 10.1109/ICFHR-2018.2018.00046

Kocka, J. (1984). Family and class formation: Intergenerational mobility and marriage patterns in nineteenth-century Westphalian towns. *Journal of Social History, 17*(3), 411–433. Retrieved from https://www.jstor.org/stable/3787212

Kuznets, S. (1955). Economic growth and income inequality. *The American Economic Review, 45*(1), 1-28. Retrieved from https://www.jstor.org/stable/1811581

Lambert, P. S., Zijdeman, R. L., van Leeuwen, M. H. D., Maas, I., & Prandy, K. (2013). The construction of HISCAM: A stratification scale based on social interactions for historical comparative research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 46*(2), 77–89. doi: 10.1080/01615440.2012.715569

Laslett, P. (1983). Family and household as work group and kin group: Areas of traditional Europe compared. In R. Wall (Ed.), *Family forms in historic Europe* (pp. 513–564). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511897535.018

Lin, Y., Liu, Z., & Sun, M. (2016). Knowledge representation learning with entities, attributes and relations. In S. Kambhampati (Ed.), *Proceedings of the twenty-fifth International Joint Conference on Artificial Intelligence (IJCAI-16)* (pp. 2866–2872). Palo Alto: AAAI Press/International Joint Conferences on Artificial Intelligence. Retrieved from https://www.ijcai.org/Proceedings/16/Papers/407.pdf

Lladós, J., Rusinol, M., Fornés, A., Fernández, D., & Dutta, A. (2012). On the influence of word representations for handwritten word spotting in historical documents. *International Journal of Pattern Recognition and Artificial Intelligence, 26*(5),1263002. doi: 10.1142/S0218001412630025

Low, B. S., & Clarke, A. L. (1991). Family patterns in nineteenth-century Sweden: Impact of occupational status and landownership. *Journal of Family History, 16*(2), 117–138. doi: 10.1177/036319909101600202

Lucassen, J., & Lucassen, L. (2009). The mobility transition revisited, 1500-1900: What the case of Europe can offer to global history. *Journal of Global History, 4*(3), 347–377. doi: 10.1017/S174002280999012X

Marín Corbera, M. (2010). La gestación del Documento Nacional de Identidad: Un proyecto de control totalitario para la España Franquista [The gestation of the National Identity Card: A totalitarian control project for Francoist Spain]. In C. Navajas Zubeldia & D. Iturriaga Barco (Eds.), *Novísima: II Congreso Internacional de Historia de Nuestro Tiempo* (pp. 323–338). Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo=3313002

Martín Niño, J. (1972). *La hacienda española y la revolución de 1868* [The Spanish treasury and the 1868 revolution]. Madrid: Instituto de Estudios Fiscales.

Martínez-Galarraga, J., & Prat, M. (2016). Wages, prices, and technology in early Catalan industrialization. *The Economic History Review, 69*(2), 548–574. doi: 10.1111/ehr.12127

Mas, J. Fornés, A., & Lladós, J. (2016). An interactive transcription system of census records using word-spotting based information transfer. *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 54–59. doi: 10.1109/DAS.2016.47

Melis Maynar, F. (2019). Distribución personal y provincial de la renta en 1926 según el impuesto de cédulas personales [Personal and provincial distribution of income in 1926 according to the personal taxes]. *Papeles de trabajo del Instituto de Estudios Fiscales. Serie economía, 3*, 1–44. Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo=7206051

Milanovic, B. (2012). *Global income inequality by the numbers: In history and now* (Policy Research Working Paper; No. 6259). Washington, DC: World Bank. Retrieved from https://openknowledge.worldbank.org/handle/10986/12117

Milanovic, B. (2016). *Global inequality: A new approach for the age of globalization*. Cambridge: Harvard University Press.

Milanovic, B., Lindert, P. H., & Williamson, J. G. (2011). Pre-industrial inequality. *The Economic Journal, 121*(551), 255–272. doi: 10.1111/j.1468-0297.2010.02403.x

Moreu Rey, E. (1959). *Els immigrants francesos a Barcelona (segles XVI al XVIII)* [French immigrants in Barcelona (16th to 18th centuries)] (Vol. 20). Barcelona: Institut d'Estudis Catalans.

Nadal, J. (1975). *El fracaso de la revolución industrial en España, 1814–1913* [The failure of the industrial revolution in Spain, 1814–1913]. Barcelona: Ariel.

Nadal, J., & Giralt, E. (1960). *La population catalane de 1553 à 1717: L'immigration française et les autres facteurs de son développement* [The Catalan population from 1553 to 1717: The French emigration and the others factors of its development] (Vol. 3). Paris: S.E.V.P.E.N.

Pérez Hernández, E. (2020). *The distribution of income in Madrid over the first half of the 20th century: An estimation using rental values* (Master thesis). Universidad Carlos III, Madrid.

Peytaví Deixona, J. (2010). *Antroponimia, poblament i immigració a la Catalunya moderna. L'exemple dels comptats de Rosselló i Cerdanya (segles XVI-XVIII)* [Anthroponymy, settlement and immigration in modern Catalonia. The example of the counties of Rosselló and Cerdanya (16th–18th centuries)]. Barcelona: IEC. Retrieved from https://publicacions.iec.cat/repository/pdf/00000233/00000077.pdf

Piketty, T. (2015). *The economics of inequality*. Cambridge: Harvard University Press.

Pujadas-Mora, J. M., & Brea-Martínez, G. (2020). Five centuries of inequality and socioeconomic transformation in the Barcelona area, 1451–1880. *Perspectives Demogràfiques, 18*, 1–4. doi: 10.46710/ced.pd.eng.18

Pujadas-Mora, J. M., Brea-Martínez, G., Jordà, J.P., & Cabré, A. (2018). The apple never falls far from the tree: Siblings and intergenerational transmission among farmers and artisans in the Barcelona area in the sixteenth and seventeenth centuries. *The History of the Family, 23*(4), 533–567. doi: 10.1080/1081602X.2018.1426483

Pujadas-Mora, J. M., Fornés, A., Lladós, J., Brea-Martínez, G., & Valls-Fígols, M. (2019). The Baix Llobregat (BALL) demographic database, between historical demography and computer vision (nineteenth–twentieth centuries). In E. Glavatskaya, G. Thorvaldsen, Georg F., & M. Szoltysek (Eds.), *Nominative data in demographic research in the East and the West* (pp. 29–61). Ekaterinburg: Ural University Press. doi: 10.15826/B978-5-7996-2656-3.03

Pujadas-Mora, J. M., Fornés Bosquerra, A., Lladós, J., & Cabré, A. (2016). Bridging the gap between historical demography and computing: Tools for computer-assisted transcription and analysis of demographic sources. In. K. Matthijs, S. Hin, J. Kok, & H. Matsuo, *The future of historical demography: Upside down and inside out* (pp. 222–226). Leuven/Den Haag: Acco. Retrieved from https://soc.kuleuven.be/ceso/fapos/publications/the-future-of-historical-demography-upside-down-and-inside-out

Pujadas-Mora, J. M., González-Murciano, C., & Cabré, A. (2018). Fenomen rodstvennykh brakov v Katalonii v XIX v. (po materialam istoricheskoy bazy dannykh Barselony) [Kin marriages in the 19th century Catalonia. With reference to findings from the Barcelona historical marriage database]. *Izvestia. Ural Federal University Journal. Series 2. Humanities and Arts, 20*(4(181)), 27–45. doi: 10.15826/izv2.2018.20.4.064

Pujadas-Mora, J. M., Romeo-Marín, J., & Villar, C. (2014). Propuestas metodológicas para la aplicación de HISCO en el caso de Cataluña, siglos XV-XX [Methodological proposals for the application of HISCO to the case of Catalonia, 15th–20th centuries]. *Revista de Demografía Histórica, 32*(1), 181–219. Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo=5161081

Reher, D. S., & Valero Lobo, A. (1995). *Fuentes de información demográfica en España* [Sources of demographic information in Spain] (Vol. 13). Madrid: CIS.

Rubio Vela, A., & Rodrigo Lizondo, M. (1997). *Antroponímia valenciana del segle XIV: Nòmines de la ciutat de València (1368–69 i 1373)* [Valencian anthroponymy of the 14th century: Name list of the city of Valencia (1368–69 and 1373)]. Valencia: Universitat de València.

Salas-Vives, P., & Pujadas-Mora, J. M. (2021). Bottom-up nation-building: National censuses and local administration in nineteenth-century Spain. *Journal of Historical Sociology, 34*(2), 287–304. doi: 10.1111/johs.12323

Sánchez Suárez, A. (1993). 1993 Les activitats econòmiques a Barcelona (1717–1833) [Economic activities in Barcelona (1717–1833)]. In J. Sobrequés i Callicó (Ed.), *Història de Barcelona. Vol. 5: El desplegament de la ciutat manufacturera (1714–1833)* (pp. 217–265). Barcelona: Enciclopedia Catalana.

Schlomer, G. L., Ellis, B. J., & Garber, J. (2010). Mother–child conflict and sibling relatedness: A test of hypotheses from parent–offspring conflict theory. *Journal of Research on Adolescence, 20*(2), 287–306. doi: 10.1111/j.1532-7795.2010.00641.x

Schneider, E. W. (1973). Course modularization applied: The interface system and its implications for sequence control and data analysis. *Association for the Development of Instructional Systems (ADIS)*, 1–17. Retrieved from https://files.eric.ed.gov/fulltext/ED088424.pdf

Schnore, L. F. (1961). Social mobility in demographic perspective. *American Sociological Review, 26*(3), 407–423. doi: 10.2307/2090668

Schürer, K. (2007). Focus: Creating a nationally representative individual and household sample for Great Britain, 1851 to 1901: the Victorian Panel Study (VPS). *Historical Social Research/ Historische Sozialforschung, 32*((2), 211–331. doi: 10.12759/hsr.32.2007.2.211-331

Shenk, M. K., Borgerhoff Mulder, M., Beise, J., Clark, G., Irons, W. G, Leonetti, D., … Piraino, P. (2010). Intergenerational wealth transmission among agriculturalists: Foundations of agrarian inequality. *Current Anthropology, 51*(1), 65–83. doi: 10.1086/648658

Singhal, A. (2012). Introducing the Knowledge Graph: Things, not strings. *Official Google Blog*. Retrieved from https://blog.google/products/search/introducing-knowledge-graph-things-not/

Song, X. (2021). Multigenerational social mobility: A demographic approach. *Sociological Methodology, 51*(1), 1–43. doi: 10.1177/0081175020973054

Sutter, J. (1968). Fréquence de l'endogamie et ses facteurs au XIXe siècle [Frequency of endogamy and its factors in 19th century]. *Population, 23*(2), 303–324. doi: 10.2307/1527490

Toledo, J. I., Carbonell, M., Fornés, A., & Lladós, J. (2019). Information extraction from historical handwritten document images with a context-aware neural model. *Pattern Recognition, 86*, 27–36. doi: 10.1016/j.patcog.2018.08.020

Toledo, J. I., Dey, S., Fornés, A., & Lladós, J. (2017). Handwriting recognition by attribute embedding and recurrent neural networks. *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, 1038–1043. doi: 10.1109/ICDAR.2017.172

Toledo, J. I., Sudholt, S., Fornés, A., Cucurull, J., Fink, G. A., & Lladós, J. (2016). Handwritten word image categorization with convolutional neural networks and spatial pyramid pooling. In A. Robles-Kelly, M. Loog, B. Biggio, F. Escolano & R. Wilson (Eds.) *Structural, Syntactic, and Statistical Pattern Recognition. S+SSPR 2016. Lecture Notes in Computer Science* (Vol. 10029) (pp. 543–552). Cham: Springer. doi: 10.1007/978-3-319-49055-7_48

Torras, J. (1998). Small towns, craft guilds and proto-industry in Spain. *Jahrbuch für Wirtschaftsgeschichte/ Economic History Yearbook, 39*(2), 79–96. doi: 10.1524/jbwg.1998.39.2.79

Trivers, R. L. (1974). Parent-offspring conflict. *American Zoologist, 14*(1), 249–264. doi: 10.1093/icb/14.1.249

Van de Putte, B., & Miles, A. (2005). A social classification scheme for historical occupational data. *Historical Methods, 38*(2), 61–94. doi: 10.3200/HMTS.38.2.61-94

van Leeuwen, M. H. D., & Maas, I. (2011). *HISCLASS: A historical International Social Class Scheme*. Leuven: University Press.

van Leeuwen, M. H. D., Maas, I., & Miles, A. (2002). *HISCO: Historical International Standard Classification of Occupations*. Leuven: University Press.

Vilar, P. (1986). *Catalunya dins l'Espanya moderna: III. Les transformacions agràries del segle XVIII català* [Catalonia in modern Spain: III. The agrarian transformations of the Catalan 18th century]. Barcelona: Edicions 62.

Villavicencio, F., Jordà, J. P., & Pujadas-Mora, J. M. (2015). Reconstructing lifespans through historical marriage records of Barcelona from the sixteenth and seventeenth centuries. In G. Bloothooft, P. Christen, K. Mandemakers & M. Schraagen (Eds.), *Population Reconstruction* (pp. 199–216). Cham: Springer. doi: 10.1007/978-3-319-19884-2_10

Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014). Knowledge graph embedding by translating on hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence, 28*(1), 1112–1119. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/8870

APPENDIX

Appendix Table 1 *Number of inhabitants and households by town included in BALL database, 19th–20th centuries (1822–1965)*

Population registers and national censuses

Sant Feliu de Llobregat																			
	1828	**1839**	**1857**	**1878**	**1881**	**1889**	**1906**	**1910**	**1915**	**1920**	**1924**	**1930**	**1936**	**1940**	**1945**	**1950**	**1955**	**1960**	**1965**
Total inhabitants	2,209	2,233	2,471	2,747	3,002	3,118	3,606	3,807	4,329	4,352	5,569	6,383	7,020	6,720	6,977	7,327	7,529	10,201	12,945
Total households	426	432	533	673	641	645	804	867	937	923	1,048	1,162	1,458	1,678	1,702	1,812	2,153	2,784	3,742
Persons by household	5.19	5.17	4.64	4.08	4.68	4.83	4.49	4.39	4.62	4.72	5.31	5.49	4.81	4.00	4.10	4.04	3.50	3.66	3.46

Santa Coloma de Cervelló						
	1901	**1924**	**1936**	**1940**	**1945**	**1950**
Total inhabitants	542	1,132	1,311	1,218	1,167	1,222
Total households	127	260	325	329	307	308
Persons by household	4.27	4.35	4.03	3.70	3.80	3.97

Sant Vicenç dels Horts										
	1857	**1860**	**1921**	**1925**	**1930**	**1935**	**1940**	**1946**	**1950**	**1955**
Total inhabitants	1,772	1,731	2,097	1,985	2,950	3,129	2,980	3,014	3,323	3,717
Total households	357	350	535	409	783	716	763	791	1,030	1,168
Persons by household	4.96	4.95	3.92	4.85	3.77	4.37	3.91	3.81	3.23	3.18

Collbató																			
	1842	**1845**	**1857**	**1860**	**1863**	**1866**	**1867**	**1868**	**1869**	**1870**	**1871**	**1872**	**1875**	**1880**	**1887**	**1889**	**1892**	**1896**	**1897**
Total inhabitants	794	746	865	859	840	828	790	765	785	802	798	828	795	791	781	812	743	686	679
Total households	109	115	140	152	154	145	139	142	148	153	160	157	149	168	172	170	154	147	148
Persons by household	7.28	6.49	6.18	5.65	5,45	5.71	5.68	5.39	5.30	5.24	4.00	5.27	5.34	4.71	4.54	4.78	4.82	4.67	4.59

Collbató (continued)											
	1900	**1905**	**1910**	**1916**	**1920**	**1924**	**1930**	**1936**	**1940**	**1945**	**1950**
Total inhabitants	703	665	809	796	560	754	531	525	474	425	418
Total households	182	144	162	166	141	170	138	138	162	125	147
Persons by household	3.86	4.62	4.99	4.80	3.97	4.44	3.85	3.80	2.93	3.40	2.84

Population registers and national censuses (continued)

Castellví de Rosanes							
	1866	**1924**	**1930**	**1936**	**1940**	**1945**	**1950**
Total inhabitants	353	283	279	273	281	242	268
Total households	63	61	64	62	73	50	60
Persons by household	5.60	4.64	4.36	4.40	3.85	4.84	4.47

Molins de Rei														
	1852	**1871**	**1872**	**1899**	**1905**	**1911**	**1915**	**1920**	**1924**	**1925**	**1930**	**1940**	**1950**	**1955**
Total inhabitants	1,339	3,387	**	2,896	3.003	3,354	3,383	3,842	4,917	**	6,469	7,622	**	**
Total households	259	502		856	717	763	817	982	1,093		1,845	2,212		
Persons by household	5.17	6.75		3.38	4.19	4.40	4.14	3.91	4.50		3.51	3.45		

El Papiol														
	1857	**1875**	**1885**	**1889**	**1890**	**1895**	**1900**	**1916**	**1921**	**1924**	**1936**	**1940**	**1945**	**1950**
Total inhabitants	1,100	**	**	1,022	**	**	912	**	**	**	**	**	**	**
Total households	198			190			183							
Persons by household	5.56			5.38			4.98							

Martorell							
	1842	**1916**	**1924**	**1940**	**1945**	**1950**	**1955**
Total inhabitants	**	3,822	**	6,113	6,214	**	**
Total households		906		1,940	1,936		
Persons by household		4.22		3.15	3.21		

Olesa de Montserrat								
	1860	**1872**	**1875**	**1910**	**1915**	**1920**	**1924**	**1930**
Total inhabitants	3,176	3,148	2,810	3,850	3,734	4,060	4,369	5,647
Total households	621	690	620	927	824	951	949	1,330
Persons by household	5.11	4.56	4.53	4.15	4.53	4.27	4.60	4.25

Population registers and national censuses (continued)

Pallejà							
	1827	**1857**	**1889**	**1910**	**1924**	**1945**	**1965**
Total inhabitants	642	838	718	653	800	998	3,127
Total households	111	178	166	169	178	237	705
Persons by household	5.78	4.71	4.33	3.86	4.49	4.21	4.44

Abrera													
	1850	**1855**	**1857**	**1865**	**1875**	**1889**	**1906**	**1912**	**1920**	**1930**	**1936**	**1940**	**1945**
Total inhabitants	**	431	**	**	**	875	838	**	835	**	**	**	**
Total households		103				233	175		226				
Persons by household		4.18				3.76	4.79		3.69				

Torrelles de Llobregat																		
	1842	**1845**	**1850**	**1852**	**1857**	**1860**	**1862**	**1863**	**1864**	**1865**	**1866**	**1867**	**1887**	**1889**	**1890**	**1895**	**1897**	**1900**	**1906**
Total inhabitants	465	388	533	**	494	496	457	**	**	**	**	**	675	**	705	697	683	693	715
Total households	72	74	103		96	105	98						140		129	120	122	153	160
Persons by household	6.46	5.24	5.17		5.15	4.72	4.66						4.82		5.47	5.81	5.60	4.53	4.47

Torrelles de Llobregat (continued)									
	1910	**1917**	**1920**	**1924**	**1930**	**1936**	**1940**	**1945**	**1950**
Total inhabitants	726	**	755	767	817	1,573	762	733	780
Total households	173		180	156	192	326	198	191	187
Persons by household	4.20		4.19	4.92	4.26	4.83	3.85	3.84	4.17

Begues										
	1857	**1860**	**1871**	**1881**	**1900**	**1905**	**1910**	**1915**	**1924**	**1928**
Total inhabitants	**	858	902	955	**	**	**	**	**	**
Total households		158	176	193						
Persons by household		5.43	5.13	4.95						

Population registers and national censuses (continued)

Corbera de Llobregat																		
	1857	**1860**	**1877**	**1880**	**1888**	**1892**	**1897–98**	**1900**	**1916**	**1920**	**1921**	**1923**	**1924**	**1930**	**1936**	**1940**	**1945**	**1950**
Total inhabitants	**	869	**	981	1,022	981	**	964	1,092	1,197	**	**	1,301	1,247	1,271	1,175	**	**
Total households		180		276	234	274		230	240	232			259	272	282	278		
Persons by household		4.83		3.55	4.37	3.58		4.19	4.55	5.16			5.02	4.58	4.51	4.23		

** Under construction

Personal taxes

San Feliu de Llobregat																		
	1868	**1869**	**1882**	**1883**	**1884**	**1885**	**1886–89**	**1890**	**1891**	**1892**	**1907**	**1909**	**1911**	**1912**	**1913–15**	**1916**	**1917**	**1918**	**1919**
Personal taxes	438	**	1,185	426	352	393	**	986	903	**	1,891	1,849	**	2,057	**	2,861	**	2,822	**

San Feliu de Llobregat (continued)						
	1920	**1921–23**	**1924**	**1925–28**	**1929**	**1930–41**
Personal taxes	2,894	**	2,673	**	3,848	**

Santa Coloma de Cervelló														
	1910	**1913–15**	**1916**	**1917–19**	**1920**	**1921–23**	**1924**	**1925**	**1927–28**	**1929**	**1931–34**	**1935**	**1937**	**1941**
Personal taxes	221	**	298	**	308	**	612	**	**	644	**	934	**	**

Sant Vicenç dels Horts			
	1879	**1880**	**1881**
Personal taxes	325	303	332

Collbató							
	1878–79	**1879–80**	**1880–81**	**1881–82**	**1883–84**	**1884–85**	**1885–86**
Personal taxes	202	168	176	218	488	486	482

Personal taxes (continued)

Torrelles de Llobregat												
	1882–86	**1887**	**1888–91**	**1906**	**1907–08**	**1911**	**1912–19**	**1920**	**1921–23**	**1924**	**1925–29**	**1930**
Personal taxes	**	327	**	460	**	492	**	523	**	530	**	569

Corbera de Llobregat							
	1881–84	**1888**	**1892–93**	**1920–21**	**1924**	**1930**	**1936**
Personal taxes	**	674	**	**	792	**	**

Martorell								
	1911	**1916**	**1919**	**1923–24**	**1927**	**1931**	**1935**	**1940**
Personal taxes	**	2,780	**	**	**	**	**	**

Abrera				
	1921	**1922–24**	**1928–32**	**1934**
Personal taxes	626	**	**	**

** Under construction