POPP. An OCR-Generated Database of the Population Censuses of Paris (1926–1936)

By Sandra Brée, Victor Gay, Marion Leturcq, Baptiste Coulmont, Yoann Doignon, Thomas Constum, Thierry Paquet and Pierrick Tranouez

To cite this article: Brée, S., Gay, V., Leturcq, M., Coulmont, B., Doignon, Y., Constum, T., Paquet, T., & Tranouez, P. (2026). POPP. An OCR-Generated Database of the Population Censuses of Paris (1926–1936). *Historical Life Course Studies, 16*, 3–28. https://doi.org/10.52024/hlcs18627

HISTORICAL LIFE COURSE STUDIES

VOLUME 16

2026

EHPS NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies was established within *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.

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, the International Institute of Social History, the European Society of Historical Demography, Radboud University Press, Lund University and HiDO Scientific Research Network Historical Demography. 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.

HISTORICAL LIFE COURSE STUDIES
VOLUME 16 (2026), 3–28, published 12-02-2026

POPP

An OCR-Generated Database of the Population Censuses of Paris (1926–1936)

Sandra Brée	French National Centre for Scientific Research (CNRS)
Victor Gay	Toulouse School of Economics
Marion Leturcq	French National Institute for Demographic Studies (INED)
Baptiste Coulmont	École Normale Supérieure Paris-Saclay
Yoann Doignon	French National Centre for Scientific Research (CNRS)
Thomas Constum	LITIS, University of Rouen
Thierry Paquet	LITIS, University of Rouen
Pierrick Tranouez	LITIS, University of Rouen

ABSTRACT

Empirical research in historical demography is usually time-consuming and labour-intensive. Recent developments in machine learning offer new possibilities for building very large databases with reduced time and costs, though these new methods raise new challenges as well. This article describes the process of constructing the POPP database, a data collection project based on the exploitation of the nominative lists of the Parisian population censuses of 1926, 1931, and 1936. This database provides a host of information for almost 9 million individuals: their name and surname, year and location of birth, nationality, relation to the household head, and occupation. The article discusses the digitisation of archival sources — several hundred thousand handwritten pages — their transformation into a database by computer scientists using machine learning techniques, and the work required on the part of social scientists to correct and adapt the resulting data for statistical purposes. Beyond its methodological contribution, this article also discusses the various ways in which the POPP database will improve our knowledge of the economic, social, and demographic evolution of an important European urban population.

Keywords: Database, Census, Machine learning, Artificial Intelligence, Paris, France, Interwar

e-ISSN: 2352-6343
DOI article: https://doi.org/10.52024/hlcs18627

1 INTRODUCTION

From the very beginning of the field of historical demography, researchers have recognised the value of computers not only for computing statistics, but also for the creation and processing of source material — particularly in a field where collecting census or civil registration data is both highly time-consuming and repetitive. But recent advances in artificial intelligence and deep learning have now opened new possibilities to historical demographers for automated data collection. In particular, optical character recognition (OCR) technology enables computers to interpret the shapes of digitised images and translate them into alphanumeric characters. However, handwritten text recognition (HTR) remains particularly complex. This is one of the key challenges faced by the POPP project — the first historical demographic database in France to rely on artificial intelligence.[1] Indeed, the aim of this project is to construct a database of nearly 9 million entries derived from the handwritten nominative lists of the Parisian population censuses of 1926, 1931 and 1936. A further challenge is to correct and adapt the resulting database for statistical analysis.

Although the nominative lists of the census are generally straightforward to interpret when read — often allowing identification of individual and household characteristics — they are far more difficult to analyse globally and automatically, especially in an urban setting. Indeed, in large cities, the size of the population often required multiple enumerators — sometimes even within a single district — and individual census forms were filled without the proper application of (limited) instructions. As a result, an HTR algorithm — however efficient — is insufficient on its own. We therefore had to adapt the database specifically for statistical analysis. Tasks that are manageable by hand when processing individuals one at a time — such as standardising the spelling of birthplaces or inferring relationships within households — become infeasible when working with nearly 9 million records. The advantages of this type of large-scale data collection — such as full population coverage and substantial time savings — thus come with trade-offs inherent to processing data at scale. This requires automating not only the transcription, but also the interpretation, correction, and structuring of the source material, some tasks complicated by the fact that historical documents were never intended for use in a modern database.

The technical steps involved in creating the POPP database using artificial intelligence have already been detailed in a dedicated data paper (Constum et al., 2022). The aim of this article is to demonstrate that the work of IT specialists alone is not sufficient to transform digitised source images into a structured database amenable to statistical analysis. Its primary objective is to describe the additional steps required to achieve this transformation in the case of the Paris censuses of 1926, 1931 and 1936. Following a brief overview of the relationship between computing and historical demography (Section 2), and an introduction to the value of nominative census records (Section 3), we turn to the creation and the post-processing of the database (Section 4). We then outline future developments for the POPP database and show how it will serve both academic research and civil society (Section 5). Finally, we describe the availability of the POPP database (Section 6).

2 HISTORICAL DEMOGRAPHY AND COMPUTERS

Almost from the inception of historical demography — if not immediately — historical demographers recognised the value of computers for the creation, processing, and statistical analysis of databases (Schofield, 1972). Indeed, anyone who has ever collected census or civil registration data knows how time-consuming and repetitive this task can be. In France, as early as 1966, Marcel Couturier introduced a "new mechanographic methodology" (which would become Forcod B) that enabled data to be collected using a tape recorder (Couturier, 1966, p. 61). At the time, he noted that researchers were then "approaching the time when [they] will be able to write instructions for the machine themselves, as 'scientists' (sic) have been doing for a long time." Marcel Couturier emphasised the central role of the researcher in defining their analytical objectives. Pierre Goubert — one of the founders of historical demography — expressed a similar view: "what machines give you is largely what you give them, [and that requires] a very thorough intellectual analysis." This sentiment is echoed by Antoinette Fauve-

1 POPP stands for *Project for the OCR-ization of the Parisian Population census* (see https://popp.hypotheses.org/).

Chamoux, who later remarked, in the introduction to an article on family reconstruction, that "[t]he miracle will not come from the computer, but from the researcher himself" (Fauve-Chamoux, 1972, p. 1083). The discussion following Couturier's article is especially interesting because it highlights the tensions between time savings, cost, data quality, and the scale of the populations analysed.[2] One of the primary challenges identified even then — tracking the same individual across multiple sources — remains a fundamental difficulty today, as we will return to later in this article.

Over the past 65 years, numerous databases in historical demography have been developed (Edvinsson et al., 2023a; Kesztenbaum, 2021; Mandemakers, 2025). The earliest efforts relied on the Louis Henry method of family reconstitution (Fleury & Henry, 1956, 1985) and typically covered only a single village or small town. Mandemakers (2025) distinguishes two main types of longitudinal databases emerging around 1990: (i) event databases based on baptism/birth, marriage, and burial/death registers, and (ii) life-course databases that follow individuals over longer periods using sources such as church records or population registers. In some countries, population registers in particular enabled the construction of richer longitudinal histories by recording changes over an individual's lifetime. More recently, researchers have augmented event databases by linking census records, thereby creating "semi-longitudinal" databases (Mandemakers, 2025). Other resources rely primarily on census microdata, such as the (I)PUMS datasets (Ruggles, 2014).[3] However, the ultimate objective of these data construction efforts is to trace individuals' lives by linking data from different sources (Edvinsson et al., 2023b; Mandemakers et al., 2023).

Until very recently, these databases were assembled manually, which helps explain the widespread use of sampling. In the French context, a common approach has been to select individuals with last names beginning with the letter "B", which yields about one-tenth of the population across social strata. This strategy was used, for instance, in the Geneva (Perrenoud, 1979) and Charleville surveys (Boudjaaba et al., 2010). Other projects have relied on different letters, notably the TRA survey (Bourdieu et al., 2013; Bourdieu et al., 2014; Dupâquier, 1984) and the Antwerp COR*-database (Puschmann et al., 2022).

Recent advances in artificial intelligence have opened up new possibilities for database creation and thereby new relationships between historical demographers and computational tools. Several research teams are actively exploring the most effective techniques for automating data extraction and processing. A significant milestone in this effort was the first European workshop on automatic registration, organised in 2019 by the International Union for the Scientific Study of Population (IUSSP). Among the four studies presented, two focused on the digitisation of historical records in Danish and Spanish (Pujadas-Mora, 2019; Sandholt Jensen & Nørmark Sørensen, 2019). Longstanding projects, such as the Balsac survey, have also been enhanced thanks to the integration of AI technologies (Tarride et al., 2023; Vézina & Bournival, 2020). The importance of these developments is reflected in the growing number of conference sessions devoted to the topic in recent years. The POPP project positions itself within this broader movement, contributing to the ongoing transformation of historical demography through the use of artificial intelligence.

One of the major advantages of using these techniques to build historical demography databases — beyond the obvious time savings — is their capacity to handle very large populations. Once the software has been developed for a coherent dataset, it makes little difference whether it processes 300 or 30,000 pages — aside from processing time, which remains an important consideration. However, data curation remains particularly demanding, both in terms of time and the need for skilled research support. Before returning to the challenges of curating and adapting the POPP database, we first discuss the source material itself: the nominative lists of the population census.

2 In the case of certain articles previously presented at conferences, the discussions that followed the presentation are reported after the article in the journal. This makes them a very rich historiographical source.

3 This is also the case for France with recent surveys such as the Charleville survey (Boudjaaba, et al., 2010) or Socface (Boillet et al., 2024).

3 SOURCE MATERIAL

3.1 NOMINATIVE LISTS OF THE POPULAR CENSUS IN FRENCH HISTORICAL DEMOGRAPHY

Population censuses are among the richest sources available for historical demographic research. Yet, until recently, they remained largely underutilised by French historical demographers (Gourdon & Ruggiu, 2015). This limited use can be attributed in large part to the dominance of the family reconstitution method developed by Louis Henry (Fleury & Henry, 1956; Henry, 1953). This helps explain why the earliest major French databases were constructed from parish and civil registers (Séguy, 2001). Louis Henry, however, was already attentive to the potential of nominative lists in the early 1960s. In a letter to the director of the Pas-de-Calais Archives dated 14 June 1961, he wrote: "It now appears to me that nominative lists, even incomplete ones, are destined to play a major role as a complement to the reconstruction of families based on civil records" (Biraben, 1963, p. 313). Shortly thereafter, Jean-Noël Biraben (1963) produced his well-known inventory of nominative lists — a sequence of events that is unlikely to be coincidental. The strong focus on this approach, which relied heavily on parish registers, likely contributed to the lack of response from French scholars to Peter Laslett's influential appeal at the 1969 Cambridge Conference (Laslett, 1965; Laslett, & Wall, 1972). Then, Laslett urged his colleagues to examine household structures in their own countries to test his hypothesis that, as in England, pre-industrial households were more often nuclear than previously assumed.

Critics of census-based analysis have often drawn on Lutz Berkner's argument, contending that censuses offer only static "snapshots" of the population and household structure at a single moment in time (Gourdon, & Ruggiu, 2015). As such, they are seen as quickly outdated and unable to capture the dynamic changes in household composition over time (Berkner, 1972, 1975). Since then, and especially since the 1990s, a growing body of research has demonstrated the value of population censuses for historical demography, particularly because they reveal patterns of co-residence (Mandemakers, 2025). In response to earlier criticisms of the census as a static source, researchers have developed methods to trace individuals across multiple census waves, thereby recovering aspects of their life histories, even over relatively short periods. This approach will be presented and critically assessed in Section 5. Before turning to these methodological considerations, we begin with an examination of the French and Parisian censuses, and in particular those of the interwar period, which serve as the foundation for the POPP database.

3.2 A BRIEF HISTORY OF THE FRENCH AND PARISIAN POPULATION CENSUSES

Two types of historical sources provide individual-level information on the French population at the national scale for the post-revolutionary period: civil registers — birth, marriage, and death records — and the nominative lists of the population censuses. Civil registers have been recorded continuously since 1789 — and even earlier in the form of parish registers — and are generally well preserved, largely thanks to their legal status as official administrative documents admissible in court rulings (Esmonin, 1964). In contrast, census nominative lists lack such legal standing and are therefore not protected by preservation regulations, although their use may in fact predate that of civil records. As a result, many of these lists were not systematically preserved, especially those from the early 19th century (Biraben, 1970). In fact, nominative lists remain an intermediate document with only an immediate utilitarian function, typically to maintain an up-to-date account of a municipality's population. Once the information they contained became obsolete, there were no legal or institutional barriers to their destruction.

Following a decree in 1822, nominative lists were to be drawn up alongside the population census, which was conducted every five years in years ending in 1 and 6, and were to include all inhabitants regardless of age. In practice, however, it was not until 1836 that these lists began to be compiled regularly. This continued until 1946, with a few exceptions: the 1871 census was postponed to 1872 due to the Franco-Prussian War, and the censuses scheduled for 1916 and 1941 were cancelled because of the First and Second World Wars. After 1946, the frequency of censuses declined, as their financial and logistical costs were deemed too high to maintain the five-year interval.

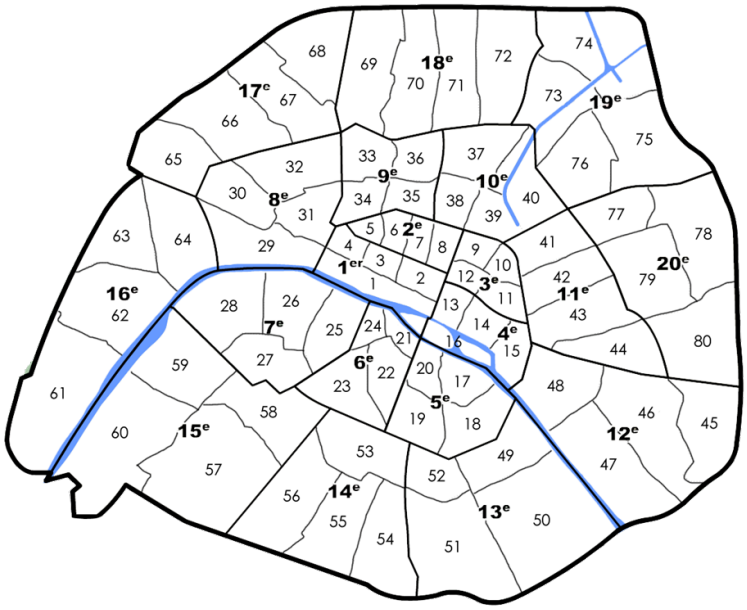
In Paris, the first municipal statistical office was established in 1816 (Biraben, 1963). The following year, in 1817, the city conducted its first population census. Unlike elsewhere in France, this census was not based on a general nominative list but rather on a survey of individual and household records. This data collection method was still praised a century later and eventually served as a model for

nationwide implementation. Although praised for its quality, this system meant that no nominative list of the population of Paris was drawn up until 1926. Indeed, beginning with the 1866 census, the city of Paris obtained an exemption from producing nominative lists due to the city's size and the high cost of such an enterprise (Biraben, 1963). As a result, no nominative list from Paris prior to 1926 has survived, most likely because they were never created. Fortunately, the nominative lists from the 1926, 1931, and 1936 censuses were preserved and are available in the Paris municipal archives, hence the focus of the POPP project on the interwar period.[4] Although statistics have been compiled for the city of Paris since 1817, including particularly detailed district-level data between 1881 and 1896 (see Figure 1), the published statistics for the interwar period are much sparser and limited to the city level.[5] The individual-level information contained in the nominative lists of the interwar censuses thus offers unprecedented opportunities for analysis, both at the individual and at highly granular aggregated levels — down to the neighbourhood, street, or even building level — as we discuss in Section 5.

3.3 THE INTERWAR PARISIAN POPULATION CENSUSES

In France, the census nominative list records all the inhabitants of a commune and specifies the household to which each person belongs.[6] In Paris, however, the population is not classified by commune but by its 80 districts (*quartiers*), themselves grouped into 20 *arrondissements* (Figure 1).

Figure 1 *The 80 districts of Paris, grouped into 20 arrondissements*



Source: Nagaï (2002, p. 325).

4 The nominative lists from the 1946 census are also available. We hope to process them using OCR technology and integrate them into a structured database as well. These lists were not included in the current analysis alongside the three interwar censuses due to the specific historical context of the Second World War. Moreover, the layout and formatting of the 1946 census tables differ significantly from those of the interwar period, which requires the development of an alternative deep learning algorithm tailored to their structure.

5 Statistical publications for the city of Paris have been produced since the early 19th century and contain a wealth of information on the population but also on hospitals, the economy, weather, prisons, schools, and other aspects of urban life. Depending on the period, population data from the census were included therein. See *Recherches statistiques sur la ville de Paris* (six volumes published between 1821 and 1860; a seventh volume, completed and ready for publication, unfortunately burned along with the other archives of the Parisian population during the Commune in May 1871); *Statistique municipale de la ville de Paris* (1865–1879; much less detailed and only on the movement of the population. The results of the 1866, 1872 and 1876 censuses do not appear therein). Finally, the *Annuaires statistiques de la ville de Paris* (1880–1967). Four very detailed publications exist for the 1881, 1886, 1891 and 1896 censuses (see Brée, 2016).

6 On the origins, development, and availability of this archival source, see Biraben (1963), Haug (1979), and Pinchemel (1954).

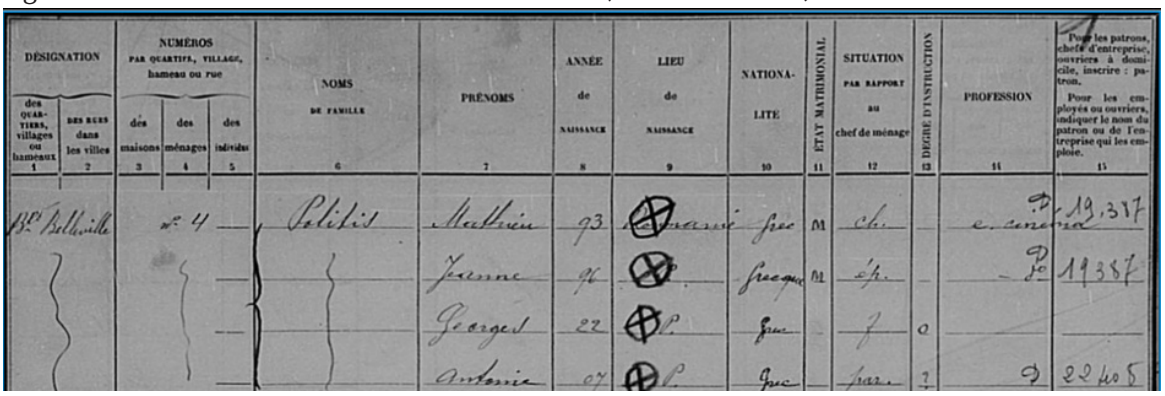
The census distinguishes between three categories of population: the population of usual residence, the population counted separately (collective dwellings), and visiting guests. In practice, individual and household census forms were distributed by enumerators to places of residence a few days before the actual census date and collected on that date. In the following two weeks, enumerators compiled the nominative list. This document consists of tables with 30 rows — each row corresponding to one individual (Figure 2). It contains 13–15 columns, depending on the year. The first five columns record address and identifier information: street name, street number, and the household and individual identifiers. The remaining columns provide individual characteristics: last name, first name(s), year of birth, place of birth, nationality, marital status, educational attainment (omitted in 1936), and occupation, along with additional occupational details and the relationship to the household head, which allows the reconstruction of household composition.

Beyond the information recorded in the columns, the nominative lists include additional marks used for statistical tallying. For instance, dashes — blue for men and red for women — are used to count the number of men and women.[7] Crosses — again, colour-coded by gender — tally foreigners, while the letter "N" marks naturalised citizens. Finally, occupational codes were added to the last column in 1926 and 1931. Several features suggest that these tallies were produced immediately after the lists were completed, while census agents still had access to the individual and household forms, which contained more information than the lists themselves. In particular, they distinguished men from women even though gender is not recorded explicitly in the lists. These counts were then used to construct summary tables at the beginning and end of each register.

The Parisian nominative lists are distinctive in that their tables are more complete than those of the rest of France. Specifically, two additional columns appear in the Parisian tables that are absent from the lists compiled in other *départements*: marital status and level of education (the latter absent in 1936). Unfortunately, however, a gender column is missing in Paris, as elsewhere in the country, even though this information was collected on individual forms.[8]

Like many historical administrative records, the nominative lists of the Parisian population censuses of 1926, 1931, and 1936 are now available online in digital form.[9] Indeed, institutions "have invested a lot of resources in the last decade in digitising large collections of historical documents not only for preserving them in a digital format, but also to give access to scholars and citizens at large through web-based digital repositories" (Fornés et al., 2019, p. 2). However, these online platforms are typically limited to browsing and searching based on image metadata and do not allow for direct queries of the actual content of digitised documents.

Figure 2 *Nominative list of the 1926 census (Belleville district)*



Source: Archives de Paris, D2M8 307.

Note: The corresponding annotation of the first row is "Politis/Mathieu/93/Roumanie/grec/M/ch./¤/e. cinéma/?19.387".

7 These colors do not appear in the online nominative lists, which have been scanned in black and white.

8 In fact, the information recorded on individual bulletins and family forms was far more detailed than the 15 entries included in the nominative lists. While this richer information was used to produce aggregate statistics published in the official census reports, it is no longer accessible at individual level or at any aggregate level than that of the commune.

9 The nominative lists of the Parisian population censuses of 1926, 1931, and 1936 are accessible at https://archives.paris.fr/archives-numerisees/sources-genealogiques-complementaires/recensement-de-population.

During the interwar period, Paris reached its historical population peak, with almost 3 million inhabitants: 2,871,429 in 1926, 2,891,020 in 1931, and 2,829,746 in 1936. Each census therefore represents approximately 50,000 images, typically composed of two double pages, and thus covering up to 60 individuals. In total, processing the three censuses required handling about 150,000 images, that is, 300,000 handwritten pages. To make this exceptionally rich material amenable to empirical research, the POPP project created a structured database of nearly 9 million individual records by applying OCR technology to the 300,000 images from the 1926, 1931, and 1936 Parisian censuses.

4 ADAPTATIONS AND CORRECTIONS OF THE DATABASE

4.1 CREATION OF THE DATABASE BY THE IT TEAM

The three censuses processed by the POPP project share an almost identical structure. This fixed structure significantly facilitated the recognition of handwritten information, as each cell was expected to contain a specific type of information — a name, date, address, etc. — that could be modelled using dictionaries or regular expressions to guide and constrain the recognition process.

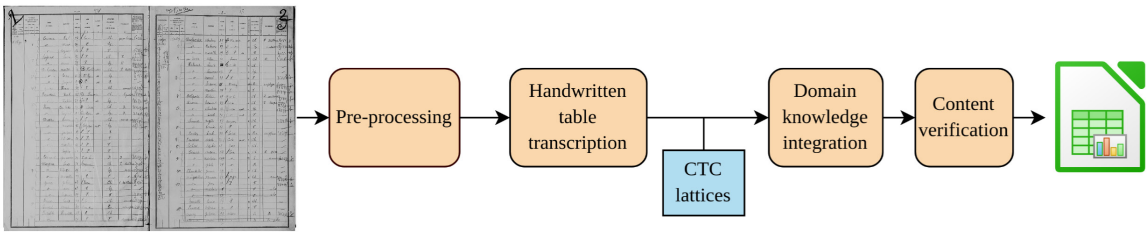
Because the original tables were filled in by hand, the column layout was often only loosely followed. Some words frequently spilled over into adjacent columns or were written between rows, making automated processing challenging. In addition, since the tables were filled out by different enumerators — sometimes multiple within the same district — certain fields were completed differently. For instance, in the "place of birth" column, some entries include both the commune and the *département*, while others provide only one of the two.

Transforming digitised images into a structured CSV database involved several distinct stages (Figure 3). While we provide a brief overview here, a detailed account of the technical workflow is available in an article dedicated to the IT component of the POPP project (Constum et al., 2022). The processing pipeline comprises four main stages: image pre-processing, handwriting recognition, integration of domain-specific knowledge, and content verification.

The pre-processing stage focuses on detecting table structure, cropping images, and detecting the individual lines within each table. As previously mentioned, the row structure of the table was generally well respected. In contrast, the column structure was often compromised due to the limited writing space, leading to frequent overlaps between entries in adjacent columns. As a result, it was not feasible to identify table cells based on visual boundaries. Instead, cell identification had to rely on interpreting their textual content. Although nominative lists do not contain actual paragraphs, we decided to treat them as blocks of text, each row corresponding to a line. This approach required implementing automated line detection as an important step in the recognition process.

To adapt the handwriting recognition model, an initial set of training data was required. Although the structure of the tables is highly consistent across pages, there is great variability in writing style, background colour, ink type, and image resolution. To account for this variability, we therefore annotated one double page for each of the 80 districts in the 1926 census in order to create a training set that was as representative as possible of the corpus as a whole. This initial dataset comprises 160 pages, for a total of 4,800 lines of handwritten text. In addition, to conduct experiments with a single writing style, we annotated another dataset based on 49 pages (1,470 lines) from the Belleville district, all written in a single style. Of these, 39 pages were used for training, 5 for validation, and 5 for testing. Training our model on the generic dataset yielded a character error rate (CER) of 7.08% and a word error rate (WER) of 19.05% on the test set. These initial results are acceptable, though the error rate is higher than the state of the art on other datasets. To improve performance, we then applied a self-training approach, a technique in which a model trained on manually labelled data (the "teacher") is used to generate pseudo-labels that are then used to train another model (the "student"). This technique is particularly effective when large quantities of unlabelled data are available, as is the case in the POPP project. The unlabelled dataset consisted of 2.4 million linear images randomly selected from the 1926 census. Using this self-training approach, we reduced the CER from 7.08% to 4.52%.

Figure 3 *Workflow of the OCR pipeline*



Source: Constum et al. (2022, p. 5).

We now turn to domain knowledge extraction and content verification. To improve the output of the OCR model, we used a combination of column-specific grammar rules and dictionaries — constructed using multiple sources, including the initially annotated pages. In addition, we implemented a rejection mechanism that compared the output of the optical model with that of the grammatical model, enabling us to reject sequences for which the divergence between the best optical grammatical paths exceeded a predefined threshold. This approach helped filter out implausible outputs and improved the overall reliability of the extracted data.

Note that the first five columns — relating to addresses and household identifiers — were not processed using the automated method described above due to several challenges specific to these fields. Enumerators often recorded street names inconsistently: sometimes at the top of the page, sometimes in the middle, and either horizontally or vertically. These variations made automatic recognition particularly difficult. As a result, we relied on a specialised scanning company to manually extract address information. Similarly, household identification posed other difficulties, as enumerators used a range of notations to indicate households — numbers, brackets, or sometimes more ambiguous marks. Where available and clearly recorded, we used the "head of household" information to infer household membership automatically (see below).

4.2 COUNTING THE POPULATION

The POPP database includes 2,845,057 individuals in 1926, 2,828,614 in 1931, and 2,784,276 in 1936 (Table 1).[10] The nominative lists divide the population into three categories: the "usual population" (about 96% of the total population), the "population counted separately" (residents of hospices, prisons, and similar institutions, representing 1.4–2.2%), and "temporary visitors" (1.8–2.3%). The latter group consists of individuals present in Paris at the time of the census but who cannot properly be considered part of the "Parisian population."

Within the present population of Paris — those normally residing in the city — some individuals were absent at the time of the census. In the nominative lists, their entries are crossed out and often marked "ABS" (Figure 4). On average, absentees account for 2.8% of the present population across the three censuses (3.3% in 1926, 2.8% in 1931, and 2.3% in 1936). As in most census studies, the results presented in this article focus mainly on the present, or *de facto*, population, that is, the population usually residing in Paris, excluding those who were absent at the time of the census.

Table 1 *Population counts in Paris in 1926, 1931, and 1936 (POPP database)*

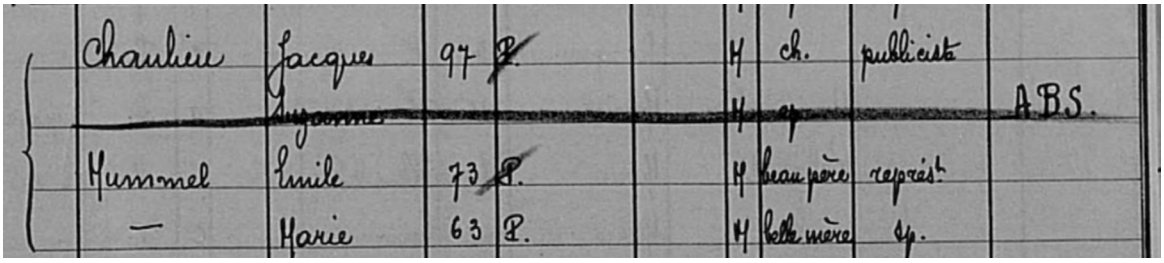
Census	Usual population (A)	Population counted separately (B)	Total resident or legal population (A + B)	Present or *de facto* population (A + B - abs.)	Temporary visitors (D)	All population (A + B + D)	All absent people
1926	2,739,706	41,098	2,780,804	2,688,370	64,253	2,845,057	92,434
1931	2,714,401	63,511	2,777,912	2,700,071	50,702	2,828,614	77,841
1936	2,678,222	51,887	2,730,109	2,668,494	54,167	2,784,276	61,615

Source: POPP database.

Note: For the usual population (A), the population counted separately (B) and the legal population (A + B), absent people are included; but they are excluded from the present, or de facto *population.*

10 The results presented in this article were obtained from computations based on a preliminary version of the POPP database and should therefore be considered provisional, although subsequent changes are likely marginal.

Figure 4 *Absent individual in the 1936 census (Père Lachaise district)*



Source: Archives de Paris, D2M8 702.

Details on population categories are available in the *Annuaire Statistique de Paris* (Paris Statistical Yearbook), but only for 1926. In that year, the officially recorded total resident population exceeds that found in the POPP database by about 3.2% (Table 2): 125 individuals are missing from the separately counted population (0.3% of that group), and 90,500 from the usual population (3.2%). In 1931 and 1936, the discrepancies for the entire legally domiciled population are 3.9% and 3.5%, respectively. In 1931, part of the difference arises from the missing nominative lists for the Enfants Rouges and Sainte-Avoye districts (3rd arrondissement) and the Saint-Georges district (9th arrondissement). In 1936, the Sainte-Marguerite district (11th arrondissement) list is very incomplete, and parts of the Javel (15th arrondissement) and Belleville (20th arrondissement) lists are also missing.

Table 2 *Population counts in Paris in 1926, 1931, and 1936 (Statistic yearbooks)*

Census	Usual population (A)	Population counted separately (B)	Total resident or legal population (A + B)	Present or *de facto* population (A + B - abs)
1926	2,830,206	41,223	2,871,429	2,838,416
1931	No data	No data	2,891,020	No data
1936	No data	No data	2,829,746	No data

Source: Annuaires statistiques de la ville de Paris*, several years.*

*Note: Temporary visitors are not indicated. Total resident population or legal population (*Population domiciliée ou de droit*): Usual population + population counted separately; absent people included; Present population: Usual population + population counted separately; absent people excluded.*

4.3 CORRECTING AND ADAPTING THE DATABASE

Although OCR error rates are very low for the POPP database (Table B1), various types of errors persist. To improve data quality and enable reliable statistical analysis, we made a series of corrections and adjustments to the initial database. First, we created new variables — most notably gender, which was not included in the original nominative lists. Second, we corrected the existing variables. The corrections all follow the same general principle: we retain the variable as read by the OCR, then create an additional variable that corrects spelling errors while remaining faithful to the source. Finally, we harmonised the categories to facilitate statistical analysis. Examples are provided below.

Data normalisation is not merely a technical operation aimed at reducing noise or addressing missing values. Rather, it often requires navigating ambiguity and making explicit choices about what kind of information to prioritise: standardisation for analytical comparability or preservation of original forms for interpretive richness. In our case, while the primary objective is to produce usable structured variables, we were cautious not to erase potentially meaningful variation, especially when it could signal historically or culturally situated forms of identity.

The remaining errors fall into two main categories: incorrect character recognition and misalignment of text into the wrong columns. To prepare the database for statistical analysis, we thus carried out a canonicalising process to standardise expressions — ensuring that all variations referring to the same entity were written consistently. During the data extraction phase, the algorithm inputted the symbol `"$"` to flag entries that were not validated by the system. These flagged strings received specific attention during the post-processing phase and were systematically reviewed and corrected. In addition, some instances of column misalignment — marked with the symbol `"/"` — were identified.

In these cases, the system failed to detect column breaks, resulting in data from one column spilling into another. Other types of errors were specific to individual columns. The corrections made during post-processing are documented in detail in the variable dictionary accompanying the database. However, several are worth highlighting here, including the correction of first names, the creation of the gender variable, the specific corrections of places of birth, the procedures used to clean and harmonise occupations, and those used to reconstruct household structures.

The following paragraphs provide a brief overview of the proposed corrections to the various variables in the database. Details of these corrections and adjustments can be found in Appendix A.

4.3.1 FIRST AND LAST NAMES

Last names were retained as read by the OCR and were not corrected. By contrast, first names required substantial post-processing due to frequent recognition errors, abbreviations, and spelling variants — first names are crucial for reconstructing other variables, most notably gender. After correction and validation procedures, the share of unrecognised first names was reduced from 8.5% to 3.5%.

Instructions to census agents for these two columns relate only to the order in which individuals should be recorded: "First, list the household head, male or female. Then list the head's spouse, followed by their children, if any. Next, list any ascendants, relatives, or in-laws who live with the family. Finally, list any servants, employees, or labourers who live with the family" (Instructions appearing on the first page of the nominal census lists). In households with a married couple, the husband is generally treated as the household head, although this convention is less systematic for cohabiting couples, as we show below. In addition, women are generally recorded under their husband's last name, though in some neighbourhoods the maiden name is also specified.

4.3.2 GENDER

Gender is not recorded in the original nominative lists and was reconstructed using a combination of the gender of first names and — when unambiguous — household position and occupational titles. This approach makes it possible to identify gender for more than 95% of individuals in each census year (see Table 3). The remaining unidentified cases mainly reflect misspellings, missing or ambiguous first names, and disproportionately concern men of foreign origin. These limitations should be borne in mind when analysing gender differences, particularly among foreign-born populations.

Table 3 *Identified and unidentified genders in the POPP database*

Gender	**1926 (%)**	**1931 (%)**	**1936 (%)**	**1926 (%)**	**1931 (%)**	**1936 (%)**
				Among valid:		
Men	43.52	43.46	43.31	45.22	45.41	45.21
Women	52.73	52.25	52.48	54.78	54.59	54.79
Total non missing (%)	*96.25*	*95.71*	*95.79*	*100*	*100*	*100*
Undeclared first name	0.39	0.49	0.7			
First name not found in database	2.67	3.15	2.91			
Ambiguous gender from first name	0.69	0.65	0.6			
Total missing (%)	*3.75*	*4.29*	*4.21*			
Total	*100*	*100*	*100*			
Total (n)	2,688,370	2,700,071	2,668,494			

Source: POPP database.

Note: Population: De facto population = Usual population + population counted separately; absent people excluded.

4.3.3 YEARS OF BIRTH AND AGE

Years of birth were extracted from heterogeneous formats and, when necessary, reconstructed from reported ages. After correction, fewer than 1.2% of individuals have missing or invalid age information. Comparisons with published aggregate statistics indicate a small but persistent underrepresentation of men aged 20–39, largely linked to gender identification issues among foreign-born individuals (see Table 4). This discrepancy remains below 1% for all age groups after correction.

Table 4 *Age groups and undefined age in the POPP database, 1926 census*

Age groups	Men (%)	Women (%)	Missing gender (%)	Total (%)	Men (%)	Women (%)	Missing gender (%)	Total (%)
					Among valid:			
0–9	10.44	8.36	5.63	9.16	10.52	8.42	6.27	9.26
10–19	13.44	11.81	8.91	12.41	13.55	11.89	9.92	12.55
20–39	40.34	40.79	43.43	40.69	40.67	41.08	48.34	41.15
40–59	27.12	27.22	23.49	27.03	27.34	27.41	26.15	27.33
60+	7.86	11.12	8.38	9.6	7.92	11.20	9.33	9.71
Total non missing (%)	*99.2*	*99.3*	*89.84*	*98.89*	*100*	*100*	*100*	*100*
Undeclared year of birth	0.5	0.4	7.87	0.72				
Undefined year of birth	0.27	0.28	2.24	0.35				
Negative value	0.03	0.03	0.05	0.03				
Value higher than 100	0.01	0.01	0.01	0.01				
Total missing (%)	*0.81*	*0.72*	*10.17*	*1.11*				
Total (n)	1,169,909	1,417,486	100,975	2,688,370				

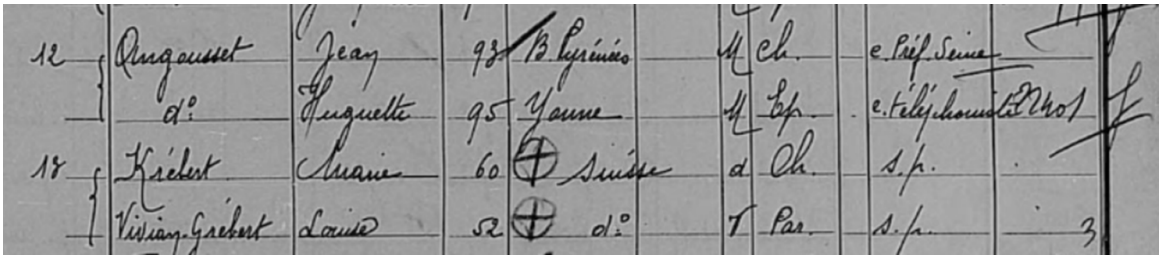
Source: POPP database.

Note: Population: 1926 census, De facto population = Usual population + population counted separately; absent people excluded.

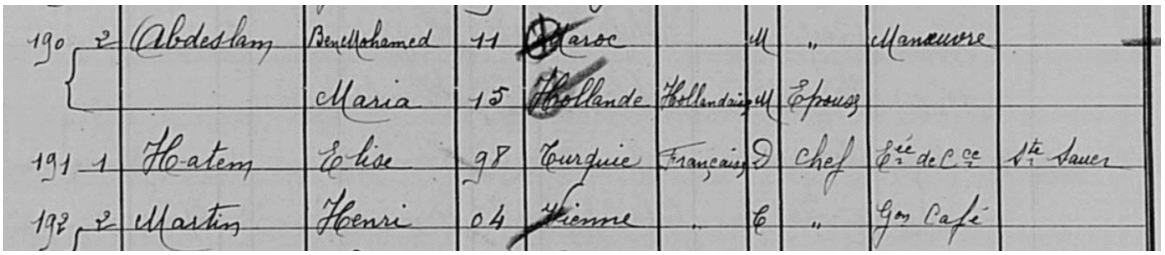
4.3.4 PLACES OF BIRTH AND NATIONALITIES

No instructions were provided for completing these columns in 1926 or 1931. However, in 1936, the column header specified "*Département* or nation," although some enumerators still recorded municipalities, especially for large French or foreign cities. In most cases, the entries nonetheless follow the appropriate format — French *départements* or foreign countries (see Figure 5). We standardised information on places of birth into four fields: commune, *département*, country, and other/undefined. About 3–4% of entries remain missing or ambiguous, with higher rates among individuals whose gender could not be identified (see Table 5).

Figure 5 *Examples of place of birth and nationality*



Source: Archives de Paris, D2M8 230, 1926, Jardin des Plantes (5e arrondissement).



Source: Archives de Paris, D2M8 679, 1936, Grandes Carrières (18e arrondissement).

Table 5 *Place of birth in the POPP database*

Place of birth	1926 (%)	1931 (%)	1936 (%)	1926 (%)	1931 (%)	1936 (%)
				Among valid:		
Seine	35.8	35.51	35.89	37.44	36.81	37.05
France (mainland, excluding Seine)	49.41	49.75	50.32	51.67	51.58	51.94
Other country	10.42	11.2	10.67	10.90	11.61	11.01
Total non missing	*95.63*	*96.46*	*96.88*	*100*	*100*	*100*
Undefined	3.46	2.68	2.36			
Undeclared	0.91	0.86	0.76			
Total missing (%)	*4.37*	*3.54*	*3.12*			
Total (%)	*100*	*100*	*100*			
Total (n)	2,688,370	2,700,071	2,668,494			

Source: POPP database

Note: Population: De facto population = Usual population + population counted separately; absent people excluded.

4.3.5 MARITAL STATUSES AND HOUSEHOLD STRUCTURES

Marital statuses required limited correction and are available for more than 99.7% of individuals aged 15 and over (see Table 6). Household structures were reconstructed from household identifiers combined with information on the relationship to the household head. While this procedure is reliable for most observations, household boundaries could not always be inferred automatically and, in a minority of cases, required manual intervention. We therefore provide a reliability indicator distinguishing inferred households from those explicitly identified.

Table 6 *Marital status in the POPP database, individuals older than 15*

	All			Men			Women		
	1926 (%)	1931 (%)	1936 (%)	1926 (%)	1931 (%)	1936 (%)	1926 (%)	1931 (%)	1936 (%)
Married	53.62	54.49	55.51	60.5	60.94	62.93	48.67	49.67	50.19
Separated	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.02	0.02
Divorced	1.79	2.3	2.69	1.27	1.68	2	2.23	2.85	3.3
Widowed	9.79	10.09	10.42	3.83	3.93	4.05	14.7	15.25	15.73
Never married	34.7	32.88	31.12	34.32	33.3	30.85	34.29	32.01	30.54
Total non missing (%)	*99.91*	*99.77*	*99.76*	*99.93*	*99.86*	*99.85*	*99.9*	*99.8*	*99.78*
Missing status	*0.1*	*0.22*	*0.24*	*0.07*	*0.14*	*0.16*	*0.1*	*0.2*	*0.22*
Total (%)	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*
Total (n)	2,310,268	2,318,954	2,256,139	983,851	984,653	950,589	1,234,030	1,226,738	1,200,962

Source: POPP database.

Note: Population: De facto population = Usual population + population counted separately; absent people excluded; older than 15.

4.3.6 OCCUPATIONS

Occupational information is highly heterogeneous, reflecting both enumerator practices and the complexity of handwritten occupational descriptions. We therefore prioritised normalisation of the most frequent occupational strings — the current harmonised coding covers about 77% of individuals with a reported occupation. In ongoing work, we are further expanding the coverage of this normalisation procedure. Overall, our data cleaning strategy prioritises analytical usability while preserving historically meaningful variation. We now discuss remaining biases and limitations.

4.3.7 REMAINING BIASES

As with any historical source, the nominative lists of the population census are subject to inherent biases stemming from the data collection itself. Despite administrative efforts to ensure comprehensive coverage, it is likely that some individuals were inadvertently omitted from the records.[11] Others may have submitted incomplete or inaccurate information.[12] Public reluctance to the census was already significant at the time — just as it remains today — prompting several poster campaigns aimed at reassuring the population about the confidentiality and intended use of their personal data (see Appendix Figure B1).[13]

Creating a database using artificial intelligence introduces another set of potential biases. Among these are errors produced by machine reading, which are likely comparable — if not slightly lower — to those encountered in manual data entry. However, unlike the human brain, which can intuitively interpret abbreviations or contextual cues and convert them into standardised categories, machines lack this level of semantic flexibility. Moreover, the correction of errors in large-scale datasets cannot match the precision achievable in smaller datasets, simply due to the volume and complexity of the data involved.

5 GOING FURTHER

5.1 THE VALUE OF THE POPP DATABASE FOR HISTORICAL DEMOGRAPHY

The POPP database will enable historical demographers to analyse the Parisian population by sex, age, place of birth, marital status, household position, and occupation by linking these characteristics at the individual level (Brée & the POPP Team, 2025). One of the strengths of the POPP database is its comprehensiveness. Historical demographic databases often focus on villages or small towns, or rely on samples, because of the time required to collect and link individual records — especially when individuals must be traced across multiple sources. Recent advances in artificial intelligence make it possible to identify the entire population of Paris at specific points in time — here, the censuses of 1926, 1931, and 1936. This choice was not driven simply by technical feasibility, as full-population coverage increases the complexity of correction and harmonisation efforts given the large number of enumerators and the scale of the underlying sources. Instead, the POPP project was developed to enable research that would otherwise be infeasible. Indeed, observing the entire population makes it possible to study subpopulations that are typically too small for reliable statistical analysis — for instance, specific occupational or origin groups, or groups defined by marital and cohabitation statuses such as divorcees and cohabiting couples, which was the primary motivation for constructing the database (Brée, 2024). This comprehensiveness also makes it possible to conduct fine-grained analysis across all 80 Parisian neighbourhoods without having to pre-select study areas, as would typically be necessary with sample-based designs. Moreover, the planned integration of a GIS (see below) will enable analyses of population and household structures at street level.

11 It is particularly true for historical but also contemporary censuses for young children and young adults (Dupâquier & Dupâquier, 1985; Héran & Toulemon, 2005; Toulemon, 2017). The non-response rate was 3.9% for the 2019 survey and was higher in cities. The introduction of a new census system, unprecedented anywhere in the world, which replaced the exhaustive census of the population with a five-yearly survey, has not improved the situation (Dumont, 2018).

12 In rare cases, individuals explicitly refused to return their census forms — instances that are occasionally noted directly on the nominative list (only three such refusals have been identified to date in the Parisian census).

13 In France today, 36% of non-responses are explicit refusals (Dumont, 2018).

Finally, linking individuals across the three censuses will make the database genuinely dynamic, allowing researchers to track changes in individual characteristics and household structures over time. Most importantly, linking POPP to the M-POPP database — currently built from all marriage records in Paris and its suburbs between 1870 and 1940 — will substantially extend POPP's longitudinal scope and enable richer reconstructions of Parisian life histories during this period.

5.2 LINKING INDIVIDUALS ACROSS CENSUSES

An additional objective of the POPP project is to link individuals across all three censuses. Tracking individuals from one census to the next raises complex methodological questions about which characteristics can reliably establish the uniqueness of an individual (Antonie et al., 2014; Dillon, 2002; Dillon & Roberts, 2002; Ruggles, 2002; Ruggles et al., 2018). These challenges are compounded when working with noisy data produced through digitisation processes such as the one used to construct the POPP database. To address this, we adopted an automated approach based on identifying individuals who share a set of stable characteristics: the greater the number of shared attributes between two records, the higher the probability that they refer to the same individual. Given the size of the Paris population — 2.9 million inhabitants at the time — we prioritised the development of a search algorithm with limited computational complexity. The identifying variables were selected based on their temporal stability: last name, first name, date of birth, and place of birth. These were chosen specifically to minimise the risk of false matches across census years.

Still, the issue of identifying individuals by their first and last names is complex. In Paris, nominative lists were compiled based on forms filled out by individuals themselves, which minimises the risk of potential misunderstandings between respondents and enumerators (Dillon, 2002; Dillon & Roberts, 2002). However, names of foreign origin can be more challenging to interpret in writing, especially when they were unfamiliar to the enumerator. Furthermore, although people generally reported the first name they commonly use, they may at other times have provided their maiden name, or vice versa. This inconsistency can result in individuals not being matched across censuses, rendering them untraceable over time. Finally, the recognition of foreign names is also hindered by limitations in the dictionaries provided to the machine, as they are primarily based on the names of people who died in France and lived in the 1970s (see Footnote 17). This is also true for foreign last names.

A more serious bias affecting last names concerns women who changed their names upon marriage. Unfortunately, there is no straightforward method to track women who transition from singlehood or cohabitation to marriage. A woman can only be identified by her maiden name if she remains unmarried, or by her married name if she stays married throughout the period. For those who change marital status — a population of particular interest — the probability of not matching these women across censuses is relatively high. These women are thus classified as "not found" not because of an OCR or migration error, but because of their last name change, which significantly compromises the analysis. In fact, even if it is possible to track women whose marital status remains stable, the inability to quantify their share in the population limits their analytical value.

The linking procedure relied on calculating the Levenshtein distance (Wagner & Fischer, 1974) across four characteristics. In order of their importance, these are:

- **Place of birth.** We considered both *départements* and countries of birth. For individuals with only one place of birth recorded, we assigned either a *département* (if the city was in France) or a country (if abroad). The score is binary: 1 if the place of birth matches, 0 otherwise. Since the names of *départements* and countries were standardised, no tolerance for typographical variation was required.
- **Year of birth.** The year had to match within one digit (i.e., an error tolerance of 1 digit out of 4, or 0.25). The score was calculated in proportion to the number of differing digits.
- **Last name.** The last name had to match with a tolerance of 1 character in 5 (0.2). The score was calculated in proportion to the number of differing characters.
- **First name.** The same rule as for last names: a tolerance of 1 character in 5 (0.2). The score was likewise calculated in proportion to the number of differing characters.

Table 7 *Number of individuals linked across censuses*

	Number of individuals matched	% of the population actually present in year t	% of men
1926–1931	1,175,015	41.8%	47%
1931–1936	1,142,517	42.3%	48%
1926–1936	980,588	36.1%	48%
1926–1931–1936	771,882	28.3%	48%

Notes: Usual population + population counted separately; absent people excluded. The proportion of men in the general population is of 46%.

The analysis was conducted for each pair of census years, going forward: between 1926 and 1931, then between 1931 and 1936, then between 1926 and 1936, and finally for all three censuses. Linking 1926 to 1931, 1,175,015 individuals were matched, representing 42% of the 1926 population (47% men and 53% women). Next, linking 1931 to 1936, 1,142,517 individuals were matched, also representing 42% of the 1931 population (48% men and 52% women). Linking 1926 to 1936, 980,588 individuals were matched, representing 36% of the 1926 population (47% men and 53% women). Finally, linking all three censuses, 771,882 individuals were matched, accounting for 28.3% of the 1926 population (Table 7).

These results indicate that at least 42% of the Parisian population remained in Paris five years later, and 28% ten years later. However, they do not support the conclusion that just under 60% of the population left Paris within five years, as it is not possible to distinguish between individuals who are untraceable because of recognition or linkage issues and those who were genuinely absent. It is also important to bear in mind that there are gaps in the nominative lists for 1931 and 1936, meaning that some individuals who were in fact present may not have been identified.

That said, the proportion of individuals identified between successive censuses remains remarkably stable at 42%, as does the overall gender distribution among linked cases. Women are nevertheless less likely to be linked than men: although they account for 54–55% of the total population (see Table 4), they represent only 52% of linked individuals. This discrepancy is largely attributable to changes in last name following marriage.

The method adopted thus far has prioritised minimising false links. This conservative strategy, however, is likely to have resulted in a non-negligible number of missed true links (Bailey et al., 2020). While these initial results are encouraging, ongoing work seeks to improve linkage rates, notably by incorporating household-level characteristics to strengthen matches (Darroch, 2002), and by linking the POPP database to the M-POPP database in order to retrieve both maiden and married names for women (Bailey & Lin, 2025).

5.3 LINKING POPP TO OTHER DATABASES

In the longer term, the POPP database will be linked to the M-POPP and N-POPP databases, which are currently under development as part of the EXO-POPP project led by Sandra Brée.[14] These complementary databases are based on marriage (M-POPP) and birth (N-POPP) records from Paris and its suburbs between 1870 and 1940. Once completed, individuals appearing in the Parisian censuses of 1926, 1931, and 1936 will be matched to their birth and marriage records following standard record-linkage approaches used in other contexts (Bailey et al., 2023; Boudjaaba et al., 2010; Garrett & Reid, 2015; Reid et al., 2002). Such linkages hold considerable promise for research on migration into and out of Paris. They will also make it possible to reconstruct individual and family trajectories, in particular to examine whether families tended to leave Paris after the birth of children.

In addition, the POPP database will be linked to the nominative census records of the Socface database (Boillet et al., 2024), which aims to bring together all nominative lists for French communes between 1836 and 1936, with the exception of Paris. This linkage will be especially valuable given that 70% of the Parisian population was not born in Paris and that half was born outside the Seine *département*. It will therefore allow for more precise tracking of migration flows into the capital, as well as outward migration, since many individuals lived in Paris only for a limited period of their lives.

14 For more information on the EXO-POPP Project, see https://exopopp.hypotheses.org/1#anglais.

5.4 GEOGRAPHIC INFORMATION SYSTEM

Finally, the POPP database will be integrated into a Geographic Information System (GIS), in collaboration with the Paris-Time Machine team.[15] This GIS will enable the precise geolocation of every building in every street in Paris. Individuals will thus be located at their exact addresses, opening the way to highly detailed spatial analyses at a very fine scale.

6 DATA AVAILABILITY

The POPP database will be made openly available on the Progedo platform at the end of 2026 under the DOI 10.13144/lil-1719. Part of the database — specifically names and addresses — has also been transferred to the Paris Archives to enable name-based searches of census images that have not yet been indexed. This new feature, made available in October 2025 through collaboration with the Paris Archives, opens up a wide range of possibilities for both researchers and genealogists.[16] It was launched in conjunction with the exhibition *People of Paris, 1926–1936, through the lens of population censuses*, created around the POPP database and held at the Carnavalet Museum – Histoire de Paris (8 October 2025 to 8 February 2026).

7 CONCLUSION

Historical demographers have long embraced advances in information technology, not only for statistical analysis but also to construct larger databases more efficiently. The development of the POPP database illustrates how recent advances in machine learning are opening up new possibilities for collecting historical data at an unprecedented scale and level of detail. The POPP database alone contains around 9 million records, each with more than 30 variables.

At the same time, the POPP project highlights the limitations of machine learning technologies when applied to large-scale historical data. While AI-generated databases share many of the biases found in more traditional data-collection methods, they also introduce specific challenges. As with manual approaches, machines may misread or misclassify information, leading to errors. Moreover, the difficulties associated with managing and analysing very large datasets are not unique to AI-based projects and are well documented elsewhere (Bailey et al., 2020; Bailey et al., 2023; Ruggles, 2002; Ruggles et al., 2018). Because of scale, corrections and record linkage cannot be carried out manually and must be automated. A small database comprising a few hundred or thousand individuals can be corrected or linked with far greater precision than one containing several million records. One of the key advantages of AI-based database construction is the considerable time saved, making exhaustive manual correction neither feasible nor desirable. Nevertheless, targeted corrections remain both possible and necessary for specific subpopulations when systematic biases are identified — for example, in cases of incomplete gender identification among foreign-born individuals, particularly men aged 20–29 in the POPP database.

Taking all these factors into account, the resulting database is estimated to contain at most 3% errors or missing data. As measurement error remains relatively limited, the dataset offers substantial opportunities to advance our understanding of the Parisian population during the interwar period. More broadly, the POPP project demonstrates the considerable potential of artificial intelligence for the creation of large-scale historical demographic databases.

ACKNOWLEDGEMENTS

This research was funded by the CollEx-Persée, the GED-Campus Condorcet, Progedo and the CNRS.

15 For more details on the Paris-Time Machine, see https://ptm.huma-num.fr/.

16 See https://archives.paris.fr/voila-paris.

REFERENCES

Antonie, L., Inwood, K., Lizotte, D. J., & Ross, J. A. (2014). Tracking people over time in 19th-century Canada for longitudinal analysis. *Machine Learning, 95*(1), 129–146. https://doi.org/10.1007/s10994-013-5421-0

Bailey, M. J., Cole, C., Henderson, M., & Massey, C. (2020). How well do automated linking methods perform? Lessons from US historical data. *Journal of Economic Literature, 58*(4), 997–1044. https://doi.org/10.1257/jel.20191526

Bailey, M. J., & Lin, P. Z. (2025). Marital matching and women's intergenerational mobility in the late 19th- and early 20th-century US. In M. J. Bailey, L. P. Boustan, & W. J. Collins (Eds.), *The economic history of American inequality: New evidence and perspectives* (pp. 165–198). University of Chicago Press.

Bailey, M., Lin, P. Z., Mohammed, A. R. S., Mohnen, P., Murray, J., Zhang, M., & Prettyman, A. (2023). The creation of LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database Project. *Historical Methods, 56*(3), 138–159. https://doi.org/10.1080/01615440.2023.2239699

Berkner, L. K. (1972). The stem family and the developmental cycle of the peasant household: An eighteenth-century Austrian example. *American Historical Review, 77*(2), 398–418. https://doi.org/10.1086/ahr/77.2.398

Berkner, L. K. (1975). The use and misuse of census data for the historical analysis of family structure. *Journal of Interdisciplinary History, 5*(4), 721–738. https://doi.org/10.2307/202867

Biraben, J.-N. (1963). Inventaire des listes nominatives de recensement en France [Inventory of nominative lists of the census in France]. *Population, 18*(2), 305–328. https://doi.org/10.2307/1527137

Biraben, J.-N. (1970). La statistique de population sous le Consulat et l'Empire [Population statistics under the Consulate and the Empire]. *Annales historiques de la Révolution française, 199*(1), 30–45. https://doi.org/10.3406/ahrf.1970.3892

Boillet, M., Tarride, S., Blanco, M., Rigal, V., Schneider, Y., Abadie, B., Kesztenbaum, L., & Kermorvant, C. (2024). The Socface project: Large-scale collection, processing, and analysis of a century of French censuses. *arXiv:2404.18706*. https://doi.org/10.48550/arXiv.2404.18706

Boudjaaba, F., Gourdon, V., & Rathier, C. (2010). Charleville's census reports: An exceptional source for the longitudinal study of urban populations in France. *Popolazione e Storia, 11*(2), 17–42. https://doi.org/10.4424/ps2010-9

Bourdieu, J., Kesztenbaum, L., & Postel-Vinay, G. (2014). The TRA project, a historical matrix. *Population, 69*(2), 191–220. https://doi.org/10.3917/popu.1402.0217

Bourdieu, J., Postel-Vinay, G., & Kesztenbaum, L. (2013). *L'enquête TRA. Histoire d'un outil, outil pour l'histoire* (Tome I, 1793–1902) [The TRA survey: History of a tool, a tool for history (Vol. I, 1793–1902)]. INED.

Brée, S. (2016). La population de la région parisienne au XIXe siècle [The nineteenth-century population of Paris]. In S. Brée (Ed.), *Paris, l'inféconde* (pp. 59–93). Ined. https://doi.org/10.4000/books.ined.1576

Brée, S. (2024). *Mariage, concubinage et célibat dans le Paris de l'entre-deux-guerres* [Marriage, cohabitation and singlehood in interwar Paris]. [Unpublished habilitation dissertation]. Sorbonne Université.

Brée, S., & the POPP Team. (2025). Paris 100 years ago: More people than today — and mostly born elsewhere. *Population & Societies, 636*(9), 1-4. https://shs.cairn.info/journal-population-societies-2025-9-page-1?lang=en

Constum, T., Kempf, N., Paquet, T., Tranouez, P., Chatelain, C., Brée, S., & Merveille, F. (2022). Recognition and information extraction in historical handwritten tables: Toward understanding early 20th-century Paris census. In S. Uchida, E. Barney, & V. Eglin, V. (Eds.), *Document Analysis Systems. DAS 2022. Lecture Notes in Computer Science* (Vol. 1323, pp. 143–157). Springer. https://doi.org/10.1007/978-3-031-06555-2_10

Couturier, M. (1966). Vers une nouvelle méthodologie mécanographique. La préparation des données [Towards a new mechanographic methodology. Data preparation]. *Annales. Histoire, Sciences Sociales, 21*(4), 769–778. https://doi.org/10.3406/ahess.1966.421421

Darroch, G. (2002). Semi-automated record linkage with surname samples: A regional study of 'case law' linkage, Ontario 1861–1871. *History and Computing, 14*(1–2), 153–183. https://doi.org/10.3366/hac.2002.14.1-2.153

Dillon, L. (2002). Challenges and opportunities for census record linkage in the French and English Canadian context. *History and Computing, 14*(1–2), 185–212. https://doi.org/10.3366/hac.2002.14.1-2.185

Dillon, L., & Roberts, E. (2002). Introduction: Longitudinal and cross-sectional historical data: Intersections and opportunities. *History and Computing, 14*(1–2), 1–7. https://doi.org/10.3366/hac.2002.14.1-2.1

Dumont, G.-F. (2018). Une exception française: Son recensement de la population. Quelle méthode? Quelles insuffisances? Comment l'améliorer? [A French exception: Its population census. What method? What shortcomings? How to improve it?]. *Les Analyses de Population & Avenir, 3*(13), 1–26. https://doi.org/10.3917/lap.003.0001

Dupâquier, J. (1984). L'enquête des 3000 familles [The 3,000-family survey]. *Population, 39*(2), 380–383. https://doi.org/10.2307/1532304

Dupâquier, J., & Dupâquier, M. (1985). Histoire des recensements [History of censuses]. *Revue française d'administration publique, 36*, 9–23. www.persee.fr/issue/rfap_0152-7401_1985_num_36_1

Edvinsson, S., Mandemakers, K., & Smith, K. R. (2023a). Introduction: Major databases with historical longitudinal population data: Development, impact and results. *Historical Life Course Studies, 13*, 186–190. https://doi.org/10.51964/hlcs14840

Edvinsson, S., Mandemakers, K., Smith, K. R., & Puschmann, P. (Eds.) (2023b). *Harvesting. The results and impact of research based on historical longitudinal databases*. Radboud University Press. https://doi.org/10.54195/HYLR8777

Esmonin, E. (1964). Statistiques du mouvement de la population en France de 1770 à 1789 [Statistics on the movement of the population in France, 1770–1789]. *Annales de Démographie Historique*, 27–130. https://doi.org/10.3406/adh.1964.882

Fauve-Chamoux, A. (1972). La reconstitution des familles: Espoirs et réalités [Family reconstruction: Hopes and realities]. *Annales. Histoire, Sciences Sociales, 27*(4–5), 1083–1090. https://doi.org/10.3406/ahess.1972.422582

Fleury, M., & Henry, L. (1956). *Des registres paroissiaux a l'histoire de la population: Manuel de dépouillement et d'exploitation de l'état civil ancien* [From parish registers to the history of the population: Manual for counting and exploitation of the ancient civil status]. INED.

Fleury, M., & Henry, L. (1985). *Nouveau manuel de dépouillement et d'exploitation de l'état civil ancien* [New manual for counting and using of the ancient civil status] (3rd ed.). INED.

Fornés, A., Lladós, J., & Pujadas-Mora, J. M. (2019). Browsing the social network of the past: Information extraction from population manuscript images. In A. Fischer, M. Liwicki, & R. Ingold (Eds.), *Handwritten historical document analysis, recognition, and retrieval: State of the art and future trends* (Vol. 89, pp. 195–220). World Scientific. https://doi.org/10.1142/9789811203244_0011

Garrett, E., & Reid, A. (2015). Introducing 'movers' into community reconstructions. In G. Bloothooft, P. Christen, K. Mandemakers, & M. Schraagen (Eds.), *Population reconstruction* (pp. 263–283). Springer. https://link.springer.com/chapter/10.1007/978-3-319-19884-2_13

Gourdon, V., & Ruggiu, F. J. (2015). Richard Wall en France: Retour vers le futur? [Richard Wall in France: Back to the future?]. *Revista de Demografía Histórica, 33*(2), 65–86.

Haug, J. C. (1979). Manuscript census materials in France: The use and availability of the listes nominatives. *French Historical Studies, 11*(2), 258–274. https://doi.org/10.2307/286604

Henry, L. (1953). Une richesse démographique en friche: Les registres paroissiaux [An untapped demographic resource: Parish registers]. *Population, 8*(2), 281–290. https://doi.org/10.2307/1524765

Héran, F., & Toulemon, L. (2005). What happens when the census population figure does not match the estimates? *Population & Societies, 411*, 1–4. https://doi.org/10.3917/popsoc.411.0001

INSEE. (2022). *Fichier des prénoms* [First name file] [Database]. https://www.insee.fr/fr/statistiques/7633685

INSEE. (2023). *Fichier des personnes décédées depuis 1970* [Deceased persons since 1970 file] [Database]. https://www.insee.fr/fr/information/4190491

Kesztenbaum, L. (2021). Strength in numbers: A short note on the past, present and future of large historical databases. *Historical Life Course Studies, 10*, 5–8. https://doi.org/10.51964/hlcs9557

Laslett, P. (1965). *The world we have lost*. Methuen.

Laslett, P., & Wall, R. (1972). *Household and family in past times*. Cambridge University Press.

Mandemakers, K. (2025). Overview and comparison of 85 databases with historical population longitudinal microdata. *Historical Life Course Studies, 15*, 281–321. https://doi.org/10.52024/hlcs21660

Mandemakers, K., Alter, G., Vézina, H., & Puschmann, P. (Eds.) (2023). *Sowing: The construction of historical longitudinal population databases*. Radboud University Press. https://doi.org/10.54195/BJYF5752

Nagaï, N. (2002). Catégories socioprofessionnelles [Socio-professional categories]. In N. Nagaï (Ed.), *Les conseillers municipaux de Paris sous la Troisième République (1871–1914)* (pp. 323–355). Éditions de la Sorbonne. https://doi.org/10.4000/books.psorbonne.1329

Perrenoud, A. (1979). *La population de Genève du XVIe au début du XIXe siècles: Étude démographique* (Vol. I: Structure et mouvements) [The population of Geneva from the 16th to the early 19th century: Demographic studies (Vol. I: Structure and dynamics]. Société d'histoire et d'archéologie de Genève.

Pinchemel, P. (1954). Les listes nominatives des recensements de population [The nominative lists of population censuses]. *Revue du Nord, 36*(142), 419–431. https://doi.org/10.3406/rnord.1954.2150

Pujadas-Mora, J. M. (2019, February 14). *The big data of the past: A journey through historical population documents driven by Computer Vision* [Presentation]. Workshop Automated Registration of Historical Population Registers: New Prospects and Possibilities, Lund, Sweden.

Puschmann, P., Matsuo, H., & Matthijs, K. (2022). Historical life courses and family reconstitutions: The scientific impact of the Antwerp COR*-Database. *Historical Life Course Studies, 12*, 260–278. https://doi.org/10.51964/hlcs12914

Reid, A., Davies, R., & Garrett, E. (2002). Nineteenth-century Scottish demography from linked censuses and civil registers. *History and Computing, 14*(1–2), 61–86. https://doi.org/10.3366/hac.2002.14.1-2.61

Ruggles, S. (2002). Linking historical censuses: A new approach. *History and Computing, 14*(1–2), 213–224. https://doi.org/10.3366/hac.2002.14.1-2.213

Ruggles, S. (2014). Big microdata for population research. *Demography, 51*(1), 287–297. https://doi.org/10.1007/s13524-013-0240-2

Ruggles, S., Fitch, C. A., & Roberts, E. (2018). Historical census record linkage. *Annual Review of Sociology, 44*, 19–37. https://doi.org/10.1146/annurev-soc-073117-041447

Sandholt Jensen, P., & Nørmark Sørensen, E. (2019, February 14). Digitizing and analyzing historical documents at scale: The power of AI [Presentation]. Workshop Automated Registration of Historical Population Registers: New Prospects and Possibilities, Lund, Sweden.

Schofield, R. S. (1972). La reconstitution de la famille par ordinateur [Computer-based family reconstruction]. *Annales. Histoire, Sciences Sociales, 27*(4–5), 1071–1082. https://doi.org/10.3406/ahess.1972.422581

Séguy, I. (2001). *La population de la France de 1670 à 1829: L'enquête Louis Henry et ses données* [The population of France from 1670 to 1829: The Louis Henry survey and its data]. INED.

Tarride, S., Maarand, M., Boillet, M., McGrath, J., Capel, E., Vézina, H., & Kermorvant, C. (2023). Large-scale genealogical information extraction from handwritten Quebec parish records. *International Journal on Document Analysis and Recognition, 26*, 255–272. https://doi.org/10.1007/s10032-023-00427-w

Toulemon, L. (2017). Undercount of young children and young adults in the new French census. *Statistical Journal of the IAOS, 33*(2), 311–316. https://doi.org/10.3233/SJI-171054

Vézina, H., & Bournival, J.-S. (2020). An overview of the BALSAC population database. *Historical Life Course Studies, 9*, 114–129. https://doi.org/10.51964/hlcs9299

Wagner, R. A., & Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the Association for Computing Machinery, 21*(1), 168–173. https://doi.org/10.1145/321796.321811

APPENDIX A VARIABLE CONSTRUCTION AND CORRECTION PROCEDURES

A.1 FIRST AND LAST NAMES

No corrections were applied to the "last name" variable. By contrast, substantial post-processing was required for "first names." Initially, 8.5% of individuals in the dataset had unrecognized first names, defined as names not included in the dictionaries used during the OCR stage. Longer first names (e.g. Marguerite), names ending in "-e," and abbreviated forms such as "M" for *Marie*, "Jh" for *Joseph*, or "Lse" for *Louise* were particularly prone to recognition errors.

To address this issue, we visually inspected approximately 400 of the most frequent problematic strings in order to identify the most plausible correct first names (e.g. interpreting "Gacton" as *Gaston*). We then cross-referenced unrecognized first names against an extended list of 50,000 first names compiled from the *Fichier des prénoms* (INSEE, 2022) and the *Fichier des personnes décédées* (INSEE, 2023).[17] If an unrecognized string appeared in either dataset, it was validated. In addition, when appending a final "-e" yielded a valid first name in one of the reference datasets, this form was also accepted.

This procedure reduced the share of unrecognized first names from 8.5% to 3.5%, representing a decline of nearly 60% in the number of initially invalid entries. Tables of the most frequent last names by census year and first names by gender are provided in Appendix Tables B2 and B3.

A.2 GENDER

The reconstruction of the missing "gender" variable relied primarily on first names. We first constructed a gendered dictionary of first names using the INSEE datasets mentioned above and assigned gender when a first name was associated with a given gender in at least 75% of cases. Individuals whose first names fell into an ambiguity range (between 75% and 90%), such as "Camille", or whose names were absent from the dictionary — particularly foreign first names — were not immediately classified.

For ambiguous cases, we relied on contextual information, notably household position and occupation. Certain household roles — such as "wife," "mother," "sister," "aunt," or "sister-in-law" — unambiguously indicate female gender, while others — such as "husband," "father," or "brother" — indicate male gender. We applied a similar logic to clearly gender-specific occupational titles (e.g. *domestique* or *bonne* versus *valet*), restricting this approach to unambiguous cases and excluding occupations where gender was indicated only by morphological variation (such as a final "-e").

Using this procedure, gender was identified for 96.25% of individuals in 1926, 95.71% in 1931, and 95.79% in 1936. Most unidentified cases stem from misspelled first names absent from the INSEE databases. Other cases, though less frequent, are due to missing or ambiguous first names. For example, in 1926, gender could not be identified for 3.75% of individuals: 2.67% because the first name was not found in the reference database, 0.39% because the first name was missing, and 0.69% because it was ambiguous (see Table 3).

A.3 YEAR OF BIRTH AND AGE

Fewer than 3% of records in this variable did not follow the YYYY format. When the cell contained one or two digits followed by the letter "A," we interpreted this as an age and reconstructed the year of birth by assuming that the reported age corresponded to the age reached in the census year. When dates were recorded in MM or MMYY formats, we extracted the year component. After these corrections, 1.1% of values remained missing.[18] Table 4 presents the age distribution derived from the POPP database for 1926. Less than 1% of individuals with an identified gender have an invalid age, compared with about 10% among those whose gender could not be identified, largely due to missing birth dates. A comparison with published aggregate statistics revealed a notable discrepancy:

17 The *First names file* contains the first names of all babies born since 1945 and individuals born after 1900 who are still alive in 1945. The *Deceased persons files* contains the last and first names of individuals who died in France since 1970.

18 This error rate includes about 90,000 individuals and many unread signs or words (about 80,000).

men aged 20–39 were underrepresented in the POPP-derived age distribution. This bias pointed to limitations in the automated gender attribution process for men of foreign nationality, whose first names are underrepresented in French name databases. Manual gender attribution for these first names substantially reduced the number of unclassified individuals and corrected much of the observed distortion. Although discrepancies remain larger for men than for women, they are below 1% for all age groups (see Appendix Table B5).

A.4 PLACES OF BIRTH

The "place of birth" column contains heterogeneous information, including communes (especially for large cities), *départements*, and countries for individuals born abroad. To standardize this information, we created four separate fields: "commune of birth", "*département* of birth", "country of birth", and "other birth". The latter category captures entries that could not be classified or that appeared to belong to another column. Ambiguous machine outputs — such as "Core," which could refer to *Corée* (Korea) or *Corse* (Corsica) — were not forcibly normalized. Instead, such entries were either retained in the "other birth" category or marked as missing, depending on the likelihood of a correct interpretation.

Manual review allowed us to reassign a substantial share of ambiguous entries, improving overall data quality. Approximately 4% of places of birth remain unidentified (Table 5). Missing birthplace information is more frequent among individuals whose gender could not be identified, particularly when the first name was missing (Appendix Table B6). Comparisons with official statistics indicate that the POPP database slightly underestimates the share of men born outside France, again reflecting limitations in gender attribution for foreign first names.

Data cleaning in this domain required balancing analytical harmonization against preservation of historically meaningful variation. This tension is especially pronounced for "country of birth". While *départements* could be matched against a fixed list with minimal information loss, country names are historically contingent. Enumerators and respondents often used historically or politically specific denominations such as "Dahomey" or "Prussia." We therefore chose not to normalize these entries to contemporary country names. Instead, we harmonized orthography while preserving the original denomination as much as possible and created an additional variable containing a standardized country name.

A.5 NATIONALITY

Nationality posed an additional layer of ambiguity. Beyond historical naming issues, discrepancies frequently arose between nationality and place of birth, especially for individuals born abroad. Enumerators often left the nationality column blank for individuals born in mainland France, implicitly assuming French nationality, while explicitly recording nationality when it deviated from this default. In other cases, the same country name was entered in both the place of birth and nationality columns.

As a result, nationality cannot always be identified unambiguously. It is clear for French individuals born in mainland France (*département* recorded, nationality blank), for foreigners born in mainland France (*département* recorded, nationality specified), and for foreigners whose nationality differs from their country of birth. Ambiguity is greatest for individuals born abroad with no nationality recorded, as this may indicate either French nationality or the nationality of the country of birth. Similar ambiguities arise for individuals born in former French colonies, such as Algeria, where legal status varied.

Given these uncertainties, the "nationality" variable is likely biased. We therefore adopted a cautious approach: nationality was corrected only in unambiguous cases, while ambiguous cases were left unchanged and explicitly flagged. This variable should be interpreted with care, and "place of birth" should be preferred whenever possible.

A.6 MARITAL STATUS

Because of its limited number of categories, the "marital status" variable required corrections in only 0.7% of cases. We identified irregular entries through a flat sort and corrected values to match the four-category grammar (married, single, widowed, divorced), while retaining the abbreviation "sep" for separated individuals who were still legally married but not cohabiting.

All values occurring at least 10 times across the three censuses were manually reviewed and corrected where necessary. Column misalignments were addressed on a case-by-case basis. For example, the value "$FILLE$" (daughter) was reassigned from the marital status column to the relationship-to-head column. These procedures reduced the share of incorrect values to 0.1% across the three censuses (Table 6).

A.7 RELATIONSHIP TO THE HOUSEHOLD HEAD

This variable captures the relationship between each individual and the household head (e.g. head, spouse, child, parent, in-law, domestic servant). Household heads are predominantly male, particularly in married couples, although women appear as heads when living without a partner.

Initially, 5.4% of entries did not match the dictionary. We applied the same cleaning strategy as for marital status, and manually reviewed frequent unmatched values. For terms occurring more than 100 times, we consulted the original images to determine appropriate corrections. Given the time-intensive nature of this process, it was applied only to this variable because of its central role in household reconstruction. After cleaning, the share of incorrect values was reduced to 1.5%.

Household reconstruction relied primarily on household identifiers when present. In districts lacking explicit identifiers — accounting for 36.4% of the usual resident population — we inferred household boundaries from transitions in the relationship-to-head variable. A binary indicator flags whether a household identifier was explicitly recorded or inferred. When neither method was feasible, household groupings were manually reconstructed by the scanning company that processed the address data.

A.8 OCCUPATIONS

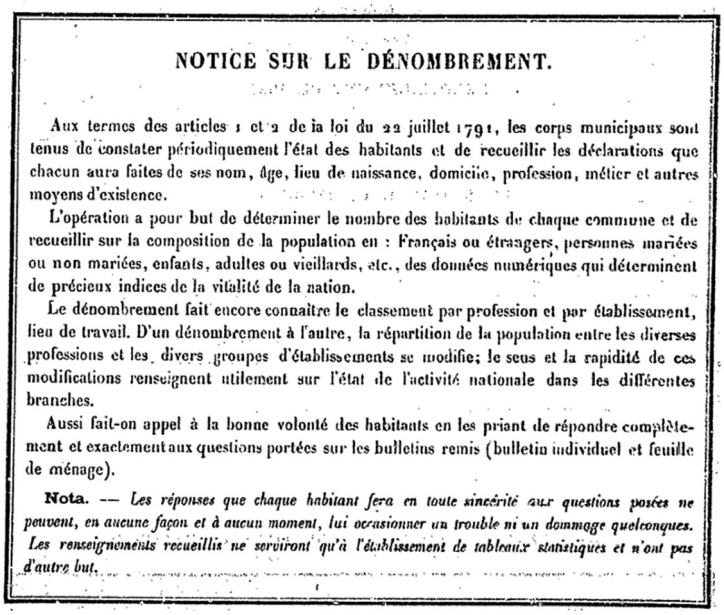
OCR performance was relatively weak for the "occupation" variable (Table B1), reflecting the complexity and heterogeneity of occupational descriptions. The raw dataset contains around 680,000 unique occupational strings, 85% of which occur only once. Nevertheless, concentration among individuals with a recorded occupation is substantial: the 100 most common strings account for 45% of cases, the 1,000 most common for 71%, and the 10,000 most common for 82%.

To address this challenge, we manually reviewed and standardized all occupational strings occurring at least 30 times, corresponding to roughly 1,000 distinct strings. This process allowed us to assign a normalized occupation to about 77% of individuals with an occupational entry. Work is ongoing to extend this coverage further.

Finally, the last column of the nominative lists — indicating self-employment status or the name of the employer — was retained as read by the machine and was not corrected.

APPENDIX B ADDITIONAL FIGURES AND TABLES

Figure B1 *Notice on the 1926 census*



NOTICE SUR LE DÉNOMBREMENT.

Aux termes des articles 1 et 2 de la loi du 22 juillet 1791, les corps municipaux sont tenus de constater périodiquement l'état des habitants et de recueillir les déclarations que chacun aura faites de ses nom, âge, lieu de naissance, domicile, profession, métier et autres moyens d'existence.

L'opération a pour but de déterminer le nombre des habitants de chaque commune et de recueillir sur la composition de la population en : Français ou étrangers, personnes mariées ou non mariées, enfants, adultes ou vieillards, etc., des données numériques qui déterminent de précieux indices de la vitalité de la nation.

Le dénombrement fait encore connaître le classement par profession et par établissement, lieu de travail. D'un dénombrement à l'autre, la répartition de la population entre les diverses professions et les divers groupes d'établissements se modifie; le sens et la rapidité de ces modifications renseignent utilement sur l'état de l'activité nationale dans les différentes branches.

Aussi fait-on appel à la bonne volonté des habitants en les priant de répondre complètement et exactement aux questions portées sur les bulletins remis (bulletin individuel et feuille de ménage).

Nota. — *Les réponses que chaque habitant fera en toute sincérité aux questions posées ne peuvent, en aucune façon et à aucun moment, lui occasionner un trouble ni un dommage quelconques. Les renseignements recueillis ne serviront qu'à l'établissement de tableaux statistiques et n'ont pas d'autre but.*

Translation of the *Nota*: "The answers that each resident provides, in full sincerity, to the questions asked cannot, under any circumstances or at any time, cause them any inconvenience or harm. The information collected will be used exclusively for the preparation of statistical tables and for no other purpose."

Table B1 *Error and recognition rates across columns*

	Caracter Error Rate	Caracter Error Recognition	Word Error Rate	Word Error Recognition
Last names	4.47%	95.53%	14.23%	85.77%
First names	2.25%	97.75%	11.43%	88.57%
Years of birth	2.31%	97.69%	4.53%	95.47%
Places of births	11.07%	88.93%	20.29%	79.71%
Nationalities	2.28%	97.72%	1.42%	98.58%
Marital status	10.67%	89.33%	7.08%	92.92%
Level of education	4.51%	93.11%	2.91%	97.09%
Relationship to the household head	6.89%	95.49%	8.26%	91.74%
Occupations	6.23%	93.77%	16.76%	83.24%

Source: POPP database.

Table B2 *Most frequent last names*

Last names	1926		1931		1936	
	Martin	0.35%	Martin	0.30%	Martin	0.30%
	Petit	0.18%	Petit	0.18%	Petit	0.18%
	Bernard	0.17%	Bernard	0.16%	Bernard	0.16%
	Moreau	0.16%	Moreau	0.16%	Moreau	0.15%
	Dubois	0.16%	Thomas	0.15%	Thomas	0.15%
Missing family names	0.14%		0.33%		0.39%	

Source: POPP database.

Note: Population: Usual population + population counted separately; absent people excluded.

Table B3 *Five most frequent first names by gender*

	1926			1931			1936		
	First name	% in the category	Share of males with this first name	First name	% in the category	Share of males with this first name	First name	% in the category	Share of males with this first name
First names identified as male	Jean	6.78	99.94%	Jean	7.14	99.94%	Jean	7.56	99.94%
	Louis	5.08	99.92%	Louis	4.67	99.92%	Pierre	4.42	99.88%
	Georges	4.18	99.86%	Pierre	4.23	99.88%	André	4.26	99.84%
	Henri	4.16	99.93%	Georges	4.08	99.86%	Louis	4.24	99.92%
	Pierre	4.12	99.88%	Henri	3.94	99.93%	Georges	4.02	99.86%
Total male (n)		1,169,909			1,173,474			1,155,789	
First names identified as female	Marie	13.59	1.42%	Marie	12.49	1.42%	Marie	11.83	1.42%
	Jeanne	5.67	0.06%	Jeanne	5.48	0.06%	Jeanne	5.31	0.06%
	Louise	4.04	0.07%	Louise	3.61	0.07%	Louise	3.25	0.07%
	Marguerite	3.39	0.09%	Marguerite	3.29	0.09%	Marguerite	3.2	0.09%
	Suzanne	2.65	0.07%	Suzanne	2.7	0.07%	Suzanne	2.77	0.07%
Total female (n)		1,417,486			1,410,668			1,400,468	
Undefined gender	Camille	60.7	46.64%	Camille	57.32	46.64%	Camille	57.08	46.64%
	Dominique	6.74	67.26%	Dominique	7.01	67.26%	Dominique	7.74	67.26%
	Alix	3.06	29.85%	Alix	2.81	29.85%	Alix	3.0	29.85%
	Modeste	1.45	60.44%	Modeste	1.24	60.44%	Modeste	1.28	60.44%
	Irénée	1.11	70.60%	Irénée	1.22	70.60%	Irénée	1.16	70.60%
Total undefined gender (n)		18,567			17,649			15,992	
No first name		10,507			13,345			18,575	
First name not found in database		71,901			84,935			77,670	
Total (n)		2,688,370			2,700,071			2,668,494	

Source: POPP database, INSEE Base des prénoms.

Notes: Population: Usual population + population counted separately; absent people excluded. "Share of males with this first name" gives the proportion of male individuals with this first name found in the INSEE Base des prénoms database. A high or a low proportion of male individuals unambiguously identified a first name as a "male" or a "female" first name.

Table B4 *Distribution of place of birth by gender (1926 census)*

	Men (%)	Women (%)	Undeclared first name (%)	First name not found in database (%)	Unclear gender from first name (%)	Total (%)
Department of the Seine (incl. Paris)	37.07	36.02	24.79	14.68	27.43	35.8
France (mainland, excluding Seine)	47.37	52.14	35.8	30.42	50.54	49.41
Other country	11.33	7.7	11.88	47.57	16.05	10.42
Total non missing (%)	*95.77*	*95.86*	*72.47*	*92.67*	*94.02*	*95.63*
Undefined	3.43	3.33	6.35	5.56	4.67	3.46
Undeclared	0.8	0.81	21.19	1.77	1.31	0.91
Total missing (%)	*4.23*	*4.14*	*27.54*	*7.33*	*5.98*	*4.37*
Total (%)	*100*	*100*	*100*	*100*	*100*	*100*
Total (n)	1,169,909	1,417,486	10,507	71,901	18,567	2,688,370

Source: POPP database, 1926 census.

Note: Population: Usual population + population counted separately; absent people excluded.

Table B5 *Age groups — comparison with official statistics (1926 census)*

	Official statistics		POPP database	
	Men (%)	Women (%)	Men (%)	Women (%)
0–9	10.05	8.38	10.52	8.42
10–19	12.94	11.81	13.55	11.89
20–39	41.63	41.38	40.67	41.08
40–59	27.59	27.26	27.34	27.41
60+	7.79	11.16	7.92	11.20
Total	100.00	100.00	100.00	100.00

Source: POPP database and Résultats statistiques du recensement de la population de 1926.

Note: Population: Usual population + population counted separately; absent people excluded.

Table B6 *Places of birth — comparison with official statistics (1926 census)*

	Official statistics		POPP database	
	Men (%)	Women (%)	Men (%)	Women (%)
Department of the Seine (incl. Paris)	36.91	35.40	37.07	36.02
France (mainland, excluding Seine)	46.86	53.12	47.37	52.14
Other country	13.33	8.30	11.33	7.7
Total non missing (%)	*97.10*	*96.83*	*95.77*	*95.86*
Total missing (%)	*2.90*	*3.17*	*4.23*	*4.14*
Total (%)	*100*	*100*	*100*	*100*

Source: POPP database and Résultats statistiques du recensement de la population de 1926.

Note: Population: Usual population + population counted separately; absent people excluded.

Table B7 *Marital statuses – comparison with official statistics*

	Official statistics						POPP database					
	Men			Women			Men			Women		
	1926 (%)	1931 (%)	1936 (%)	1926 (%)	1931 (%)	1936 (%)	1926 (%)	1931 (%)	1936 (%)	1926 (%)	1931 (%)	1936 (%)
Married	61.3	61.2	64.4	48.8	49.6	51.1	60.5	61.0	63.0	48.7	49.8	50.3
Divorced	1.6	1.7	2.0	2.6	2.9	3.4	1.3	1.7	2.0	2.2	2.9	3.3
Widowed	4.2	4.1	4.2	16.6	16.4	16.8	3.8	3.9	4.1	14.7	15.3	15.8
Never married	32.9	32.9	29.3	32.0	31.0	28.7	34.3	33.3	30.9	34.3	32.1	30.6
Total (%)	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*

Source: POPP database and Résultats statistiques du recensement de la population de 1926.

Population: Usual population + population counted separately; absent people excluded.