Overview and Comparison of 84 Databases with Historical Population Longitudinal Micro Data

By Kees Mandemakers

To cite this article: Mandemakers, K. (2025). Overview and Comparison of 84 Databases with Historical Population Longitudinal Micro Data. *Historical Life Course Studies, 15*, 281–321. https://doi.org/10.52024/hlcs21660

HISTORICAL LIFE COURSE STUDIES

VOLUME 15

2025

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 15 (2025), 281–321, published 27-10-2025

Overview and Comparison of 84 Databases with Historical Population Longitudinal Micro Data

Kees Mandemakers
International Institute of Social History, Amsterdam

ABSTRACT

In the last 65 years several major historical databases with reconstructed life courses of large populations have been launched. Around 1990, we could find two important types of databases with longitudinal micro-data. The first type were event databases aimed on family reconstructions and usually based on baptism, marriage and funeral registers or on civil certificates introduced after 1800. The second type were databases with life courses: persons are observed on a more permanent basis using church examination or population registers. After 1990 a third type, databases with census data, really took off. In first instance in the form of samples, in second instance by entering full count samples which makes it possible to link the several censuses into one system, creating semi-longitudinal databases. Another development was the growth of special purpose samples into semi-longitudinal ones by following sampled persons from one source during their life course through linking with all kind of other sources. The development of these databases is indicative of considerable investments that have greatly expanded the possibilities for new research within the fields of history, demography, sociology, as well as other disciplines. In this paper I will compare 84 of these databases on several key figures like included sources, year of foundation, period of observation, area of observation, sample fraction and number of included observations, families and unique persons. An overview of all databases with all key figures is presented in the Appendix.

Keywords: Historical demography, Longitudinal historical data, Population census data, Event history data, Family reconstitution, Intermediate Data Structure, Intergenerational historical data

e-ISSN: 2352-6343
DOI article: https://doi.org/10.52024/hlcs21660

1 INTRODUCTION

In the last 65 years several major historical databases with reconstructed life courses of large populations have been launched. In this paper[1] I will present an overview of these databases on several key figures and sketch the major developments in this field since 1960.

Around 1990, we could find two important types of databases with longitudinal micro-data. The first type was event databases aimed at family reconstructions and usually based on baptism, marriage and funeral registers or on civil certificates introduced after 1800. The second type was databases with life courses: persons are observed on a more permanent basis using church examination or population registers. After 1990 a third type, databases with census data, really took off. In first instance in the form of samples like IPUMS with 1% samples of the American censuses, in second instance by entering full count samples which makes it possible to link the several censuses into one system, creating semi-longitudinal databases. Another development was the growth of special purpose samples into semi-longitudinal ones by following sampled persons from one source during their life course through linking with all kinds of other sources.

This is not the first overview or databases of this kind, but it is the most complete one so far. Section 2 lists previous limited enumerations and elaborates the cooperation in this field and standards as they have been developed. Section 3 discusses the structure and development of these datasets with micro-data. Section 4 explains the criteria used to include databases with life course data in this overview, the main types in which they can be distinguished and the sources that have shaped the content. In Section 5 these databases will be compared on several key figures like included sources, year of foundation, period of observation, area of observation, sample fraction, number of included observations and unique persons. An overview of all databases with key figures is presented in the Appendix and more detailed in the accompanying spreadsheet.[2]

2 EARLIER OVERVIEWS

This is the most comprehensive overview of databases with historical microdata so far. It is also a summary of earlier published technical descriptions of databases. A global overview of the development of these databases was delivered by Kesztenbaum (2021). More limited overviews were already made as a kind of introduction to a specific database (for example Bourdieu et al., 2014; Dillon et al., 2018). Other overviews were more topical as Dong et al. (2015) who compared five household- and individual-level historical panel data sets for East Asian populations not only with each other, but also with their European and North American counterparts. Pfister and Fertig (2024) presented an overview of five German family databases based on the *Ortsfamilienbücher* (local genealogic collections; before *Ortssippenbücher*). Most of these were also included in the report of Voland (2000) about the contributions of micro-data to the field of evolutionary reproductive ecology. Klancher Merchant and Gutmann (2022) developed an overview over the American censuses from a societal and technical perspective, where Gutmann et al. (2018) discussed several mainly census databases from the perspective of the availability and use of big data in economic history. Alter (2019) mentioned several databases as an introduction and context to his evaluation of the models used in historical demography. Ruggles (2012) presented in his view of the future of historical family demography the most important census and longitudinal databases. Song and Campbell (2017) provided in their discussion of the importance of genealogical data for social science, especially social mobility, an overview with key figures about 15 historical databases mainly based on genealogical data, and about 10 retrospective surveys with data going back to the 1960s. And the national overviews of the history of historical demography in *A Global History of Historical Demography* (Fauve-Chamoux et al., 2016) also included a lot of details of country-specific databases.

1 This paper is a rework and upgrade of the first part of my farewell speech "*You Really got me*". *Ontwikkeling en Toekomst van Historische Databestanden met Microdata* [Development and Future of Historical Databases With Microdata] as professor on the endowed chair Large Historical Databases at the Erasmus School of History, Culture and Communication (Mandemakers, 2023a). A first upgraded version, *60 Databases With (Semi-)Longitudinal Data Compared*, was presented at the conference of the European Society for Historical Demography (ESHD), Nijmegen (Mandemaker, 2023b). The author thanks commentators, reviewers and audience for their fruitful suggestions to improve this paper.

2 The spreadsheet is also published in Dataverse (IISH Data Collection), see Mandemakers, 2025.

The first real collection with detailed information concerning multiple databases was the *Handbook of International Historical Microdata for Population Research* (Kelly Hall et al., 2000). However, over half of the included 16 databases concentrated on relatively modern 20th century census data. The *Handbook* was a product of IMAG, the International Microdata Access Group, that was formed to realize international collaboration between researchers working with historical micro-data. In first instance the IMAG group concentrated on census data gathered by the IPUMS group, but a second IMAG workshop went a step further by concentrating on record linkage, i.e., the ways multiple appearances of the same persons and households were linked in various databases (Dillon, 2000; Dillon & Roberts, 2002).

Since these first initiatives, cooperation between historical micro-databases have intensified enormously (Inwood & Maxwell-Stewart, 2021). A full day session on "New sources for historical demographic research" with four panels, was organized by the International Commission for Historical Demography at the World History Conference in Sydney in 2005. The International Institute of Social History (IISH) organized two workshops in 2001 and 2006. Main result of the first workshop was an overview of best practices for these databases (Mandemakers & Dillon, 2004), the second one was an agreement on the basic principles to construct a scheme for standardization of the data of the different databases, the so-called Intermediate Data Structure (IDS). May 2008, the Inter-university Consortium for Political and Social Research (ICPSR) hosted a planning group to continue working on the IDS. This resulted in a model for data sharing, which was presented to an open meeting of historical databases at the Social Science History Association conference in Miami, October 2008 (Alter et al, 2009; Alter & Mandemakers, 2014).

Part of the 2006 IISH workshop was an initiative to publish questionnaires with key information about the databases the participants were representing (Alter & Mandemakers, 2025). The initial IDS working group was embraced by the ESF funded European Historical Population Sample Network project (EHPS-Net), which ran from 2011 to 2016. This project gathered almost all historical micro-databases with a European background. Networking activities around historical population databases and the IDS continued with the LONGPOP project. The database questionnaires collected for the IISH conference in 2006 were also adapted by the EHPS-Network. The number of participating databases increased to 32 (Moisseenko & Koster, 2025), and much more detail was added.

One of the spin-offs of the EHPS-Network was the journal *Historical Life Course Studies* which published two special issues on the technical aspects and impacts of large historical population databases. One, *Content, Design and Structure of Major Databases With Historical Longitudinal Population Data*, edited by George Alter, Kees Mandemakers and Hélène Vézina (2023), describes 24 databases with historical longitudinal data, while the other *Major Databases with Historical Longitudinal Population Data: Development, Impact and Results,* edited by Sören Edvinsson, Kees Mandemakers and Ken Smith (2023a), deals with how the databases that contributed to research responded to changing research questions and facilitated the development of novel lines of inquiry in historical demography and related fields.[3]

3 STRUCTURE AND DEVELOPMENT OF DATA FILES WITH MICRO-DATA

Around 1990, there were two main types of databases with longitudinal data, which are data that describe multiple points in a person's life: databases with event related data and databases based on life courses in which people are followed on a more or less permanent basis.

Event databases are usually based on the baptism, marriage and burial books, or later after 1800, the civil status records in a particular parish or municipality. Families are reconstructed by adding the births to the marriages and matching the moments of death to the right birth, bride or groom. However, especially with the often not too accurate or incomplete church registers, this is easier said than done. Within families the same first names are often reused, for example a newborn child gets the

3 Both volumes were reissued by Radboud University Press as Kees Mandemakers, George Alter, Hélène Vézina and Paul Puschmann (2023), *Sowing. The Construction of Historical Longitudinal Population Databases* and Sören Edvinsson, Kees Mandemakers, Ken R. Smith and Paul Puschmann (2023b), *Harvesting. The Results and Impact of Research based on Historical Longitudinal Databases*.

first name of a previously deceased child, sources are not always complete, and many people leave a village before their marriage or death and can therefore no longer be found. It was Louis Henry who dealt with these problems in a systematic way. Together with Michel Fleury he developed a form for making family reconstitutions (Fleury & Henry, 1956, 1985) and developed a set of restrictive rules to construct and analyse this kind of data (Alter, 2019; Henry, 1970; Henry & Blum, 1988; Séguy, 2001).

Well known databases existing in 1990 were those of Louis Henry containing the data of 34,812 families from 39 French parishes for the period 1640–1829 (Séguy, 2001) and the database built by Wrigley and colleagues from the Cambridge Group for the History of Population and Social Structure including 26 parishes for England and Wales over the period 1580–1837 (Wrigley et al., 1997). And in Quebec a very special project had been started, a complete family reconstruction for the entire population of European origin living in the valley of the St. Lawrence River. The database starts in 1621, when the first church registers were created (Naul & Desjardins, 1989).

But connected events are not longitudinal data in the sense that everyone is tracked moment by moment. The number of moments depends on whether persons marry and the number of children they have. A person who remains unmarried has only two measuring moments: birth and death. And also in earlier times there was a lot of migration, the number of adult people who died in the native municipality could be less than half. And if a person left at a young age, he only appears in the database with a name and a birth date. Nevertheless, datasets with this semi-longitudinal character allowed good research to be done into the development of fertility, extramarital conceptions, deaths and all this in relation to changing economic conditions. For example, what were the effects of a food crisis after a failed harvest? How large was the excess mortality and which groups were particularly affected?

In 1990, there were already several databases which kept considerably more changes than just event data and where no families needed to be reconstructed, because the people who belonged together were already brought together in the source. These are what you might call the "real" longitudinal databases. In addition to the family compositions, they also provide data on the profession and changes in it and data on origin and departure. In addition, population registers also include persons living in the household as boarders and servants. Sources that provide this kind of data are only found in a few countries. This mainly concerns population registers, existing in the Netherlands (Mandemakers, 2002), Belgium (Jenkinson et al., 2020) and some regions in Italy (Breschi et al., 2020; Derosas, 1989). Comparable are the catechism registers in Sweden (Dribe & Quaranta, 2020; Vikström et al., 2002). In China and Japan, there are similar registries (Campbell & Lee, 2020; Kurosu et al., 2021), although these were not kept permanently but on an annual or triennial basis, so that, for example, prematurely deceased children may be missing from the dataset observation (Alter, 2019). You could call these, like the family reconstitutions, semi-longitudinal.

The first important database in the United States was the Utah Population Database (UPDB). It was set up in the early seventies of the last century as a combination of a genealogical database and databases with medical data, in order to investigate to what extent family factors play a role in various forms of cancer and cardiovascular diseases. The genealogical data was largely provided by the Church of Jesus Christ of Latter-day Saints, or Mormons, who had been copying historical sources with personal data for years for religious reasons and developing them into genealogies. The database was then expanded with numerous other historical sources and linked to contemporary administrations (Smith et al., 2022). BALSAC, Chicoutimi's database in northern Quebec, also started from a medical background. Here the database was built on the basis of marriage certificates that were linked to pedigrees. It was after 1990 that more and more historical and demographic research was carried out here (Vézina & Bournival, 2020).

Traditionally, in the United States and also in Canada, the censuses were of great importance for historical, demographic, sociological and economic research. In the United States, all censuses, except the one from 1890 that was destroyed by fire, have been preserved on a personal level. Since 1960, the US Census Bureau made samples from the census results selecting at least 1% of households. These data files called PUMS (from Public Use Microdata Samples) are available in anonymized form for scientific research. Historians then built similar PUMS for older census years. One of the involved researchers was Steven Ruggles who created a 1% sample for the counts of 1850 and 1880 (Ruggles & Menard, 1995).

A big problem with these PUMS datasets was that they were hardly comparable. All definitions of variables and associated values differed per census. Only the samples from 1960 and 1970 were set up in a comparable way which made research possible on social and economic developments between 1960 and 1970. In 1990 Ruggles realized funding to make all PUMS data comparable over the period 1850–1990. This became IPUMS (*Integrated* Public Use Microdata Samples). A variable such as 'type of household' had 161 different categories in 1910 and only 15 in 1960. To make these comparable without losing the details, a double coding system was set up; the first part with the most general code and the second part with the possibly different and more detailed code per census (Ruggles et al., 1995). In addition, it also includes all kinds of census-related subsamples that ask much more detailed questions and have become available in 5% samples (Ruggles, 2014).

In 1998, Bob McCaa started a major project to save South American censuses from 1960 onwards whose tapes were about to disappear and so preserve them for posterity. This project has grown into a worldwide collection with more than 300 samples of censuses from 100 countries with a total of more than 800 million personal data (McCaa & Ruggles, 2002; Ruggles, 2014). A second major leap forward with another 1.1 billion personal data began in 2000 with the start of the North Atlantic Population Project (NAPP), which used the fully entered censuses of the United States (1880), Canada (1881) and the United Kingdom (1881). These were entered by volunteers organized within the aforementioned Church of Jesus Christ of Latter-day Saints which offered the data in exchange for being harmonized, improved, and enriched by IPUMS, which was able to obtain ample funding for this. In addition, other samples from US censuses and countries (Iceland and Norway, 1860 and 1900 also 100%) were integrated into the IPUMS system (Roberts et al., 2003). Next steps were 100% samples for other countries (Sweden, Denmark) and the USA (1850 and 1920) in which other, more commercial parties also collaborated, such as Findmypast which made the dataset of all censuses from the period 1851–1921 of the United Kingdom available in anonymized form for scientific research (I-CeM project, see https://www.essex.ac.uk/research-projects/integrated-census-microdata). The big blow came when Ancestry and Family Search teamed up and brought together all US censuses from 1790–1930 in one large database and then made them available for research (Sobek et al, 2011; Ruggles, 2014). Diverse projects were initiated to link these censuses into semi-longitudinal datasets (Foxcroft et al., 2022; Goeken et al., 2011; Helgertz et al., 2022; Longley et al., 2022; Robinson et al., 2022).

Since the beginning, IPUMS has grown into a data institute that has increasingly become an example for other large databases (Ruggles, 2018). This huge development in numbers is exemplary but not unique. Various data files that already existed in 1990, such as the UTAH Population DataBase, the Demographic Databases Umeå and the Canadian databases, have also experienced a huge growth. Other examples are the PHRD project in Quebec which grew between 1990 and 2020 from 65,000 unique individuals (Naul & Desjardins, 1989) to 438,193 from 74,000 families spanning four to five and sometimes as many as nine generations in linked form over the period 1621–1799 (Dillon et al., 2018). The digitization of the Roteman archive with 9 million family cards over the period 1878–1926 in Stockholm, grew from 1.2 million cards in 1990 (Fogelvik, 1989) to 2.8 million in 1999 (Geschwind & Fogelvik, 2000) and to 6.3 million in 2015 with the expectation that the process will be completed in 2025 (Moisseenko & Koster, 2025). These are just a few examples of the enormous growth that historical longitudinal databases have experienced in the last 30 years, apart from the newly established data projects after 1990.

4 84 DATABASES WITH HISTORICAL POPULATION LONGITUDINAL MICRO-DATA

4.1 SELECTION

Presently, I have distinguished 84 databases with historical population longitudinal micro-data. Although this is not all there is and we can be sure that such a list will never be complete given all initiatives that are developing all over the world, I am quite satisfied with this result. Especially relatively small databases and databases that are in the first stage of development, may have been below the radar. Main criteria to include a database or not was that a) it needed to be a historical one which was operationalized by the condition that it included data from 1940 or before, b) it needed to have a longitudinal character which meant that they should have minimal two points in time measuring data

about persons, for example a birth and death record and c) it needed to be open access or seriously intends to become open access (e.g., in the case when a database is still under development), and d) it needed to be incorporated in historical demographic research.

Overlap may occur in countries with more databases with micro-data. And in some countries databases are in a process of fusing at least for the end of the data pipeline, making combined data releases. Examples are Sweden and Canada: SWEDPOP combines the five main Swedish databases (Dribe & Quaranta, 2020; Edvinsson & Engberg, 2020) on the basis of the Intermediate Data Structure (IDS) and PRDH and BALSAC and two other Canadian institutions are united into the Infrastructure intégrée des Microdonnées historiques de la Population du Québec (IMPQ), covering the period 1621–1965 (Dillon et al., 2023; Vézina & Bournival, 2020). Because the main effort of this cooperation is not the creation of new datasets but combining, standardizing and enriching existing data, it was decided to include each database separately.

Besides these fusions one can discern families of databases which share the same characteristics or are based on the same kind of sources. Most important is the already discussed IPUMS family that is centered on census databases and is now working on linked datasets between these censuses. Newest developments are the integration of civil certificates, for example the Life-M project (Bailey, Lin, et al., 2023) and the UK censuses (Diduch, 2024; Schürer, 2007). Other datasets derive their data from a common source as the German datasets which usually are based on the Ortsfamilienbücher, which are genealogical sources of high standards (Pfister &Fertig, 2024). Others usually small or medium sized datasets are grouped on the basis of common approach, such as the ICPSR repository of *Historical Demography Longitudinal Data Series* (https://www.icpsr.umich.edu/web/ICPSR/series/326) or their usefulness for anthropological research (Voland, 2000).

The foregoing criteria implied that genealogical datasets were not selected unless they were included in a scientific database. This implies that we may have missed serious genealogical datasets, but also that we prevented including datasets that will not always pass quality standards, since only a small portion of the available material is really suitable for scientific research (Calderón-Bernal et al., 2023; Gellatly, 2009, 2015; Kaplanis et al., 2023). For this reason, we also excluded the Iceland Genealogy Database, because it is only open for medical research (Garðarsdóttir, 2016; Tulinius, 2011) and of which figures about the content could not be found anyway. It also means that we did not include any specific anthropological dataset in our research. Most of them tend to be quite small, moreover, unfortunately the obvious entry point to track datasets (kinsources.net) was not operational at the time of writing which hampered tracking of these kind of datasets.[4] Only the already mentioned overview of Voland (2000) delivered some datasets. Size as such was only used as a criterium, when the number of unique persons was lower than 500. Compared with the earlier versions of this overview (Mandemakers, 2023; see also note 1) four databases were skipped.[5] Basic principle for a dataset to be included is that person observations have been linked to create unique persons or are very near to this stage.

The Appendix lists all included databases with the most important characteristics, such as name, year of foundation, number of personal observations, families, marriages, unique persons, period and area of observation, sample fractions, etc. The smallest one is an Australian database containing the demographic biographies of 899 women with 7,315 person observations over the period 1880–1938. The largest is the aforementioned IPUMS USA with 798 million person records which in case they could be fully linked, would amount to about 200 million unique persons over the period 1790–1950. Could be linked, because linking data from Anglo-Saxon censuses is a rather difficult task because of the many inaccuracies in names and ages, but also because the woman usually started using the

4 Kinsources is an open, peer-reviewed and interactive repository to archive, share, analyze and compare kinship data and designed for comparative and collaborative research, see https://eadh.org/projects/kinsources.

5 One because it was included by mistake and three others because, although as a census database important for the scientific community, they will probably not turn into semi-longitudinal data soon. These are the MOSAIC database counting more than 92 most 19th-century regional or local censuses from all over Europe with 733,107 persons (Szoltysek & Gruber, 2016), the Canadian Century Research Infrastructure with the 3–5% samples of unconnected censuses for the period 1911–1951 (Darroch et al., 2007; Gaffield, 2007; https://ccri.library.ualberta.ca/enoverview/microdata/index.html, retrieved December 13, 2024) with about 18 million observations and the National Sample of the 1901 census of Canada (5% sample; see Sager, 1998).

husband's name at marriage, thus breaking the link with the parental family and thus the first 20 years of the life course (Antonie et al., 2015; Goeken et al., 2011). Presently the focus of IPUMS and other Anglo-Saxon databases is not anymore to link persons on an individual basis but within the context of complete families. Through this strategy the linkage results have improved drastically but nevertheless are still far away from an ideal result (Akgün et al., 2020; Helgertz, 2022).[6] And of course the results inherently contain a bias on families and on families that don't geographically move between censuses. Wisselgren et al. (2014) compared the results from two Swedish linked censuses with the life courses from the Umeå Demographic Database. They found that missing links were more of a problem than faulty ones, but that on the other hand missing ones could easily be established by using constructed names (where surnames were lacking) and relations within households.

Figures about the databases are fixed at some moment in time. This could have been relatively long ago if the development of a database was closed or working on it stopped for a definitive time. Also, some databases lag behind in publishing figures about actual data entry and linked data. For this reason, the year of measurement was also included in the Appendix. Besides fields for the main figures indicating the size of the database, special fields were created to give an impression of the used sources, the possibilities of links with contemporary registrations, the application of the IDS structure and the presence of links between more than two generations. In case numbers of personal observations and/or unique persons were lacking these numbers were estimated mainly on the basis of the average ratio for each type of database. For detailed information about all databases, see the Appendix and the detailed table published in combination with this article.

4.2 MAIN TYPES

There are roughly four types of sources on which the databases have been and are being built. The first is based on the entry of censuses, the second on the entry of church registers (baptism, marriage and burial books) and/or civil certificates, the third on the entry of population or comparable registers in which all or an at random sample of persons are permanently monitored and the fourth on the entry of a specific source from which persons are followed as closely as possible during their lives, for example registers of maternity hospitals or conscription registers. Quite a lot of databases using church records or civil certificates present themselves as "Family Reconstitutions". However, it is to be preferred not to use this concept for typing datasets since it is a concept belonging to the practice of research and not inherently bound to event structured databases, even more because quite a lot of them do not completely follow the basic rules of reconstituting families (Alter, 2019). So, in case the content of a database only consists of family reconstructions, it is included in our overview as one based on vital events.

Not all databases use the sources in an equal way. For example, the Historical Sample of the Netherlands (HSN) begins in 1812 with civil certificates whereas the population register only started in 1850 (Mandemakers, 2000). So, for the categorization of the databases in main types some additional criteria were used. The first one was that the source with the richest input was used to type the database and the second one that in case more different types of sources were included in a database, the development and goal of the database served as the main criterion for definition. This resulted in four types: semi-longitudinal datasets consisting of a) linked vital event datasets, or b) linked census datasets, c) longitudinal datasets based on population registers and d) special selection of groups of persons, for example conscripts or migrants. More and more common are combinations of these types, especially the combination of census or register data with event data from church records or civil certificates. Besides these main sources all kind of other sources could be linked to the basic dataset, such as tax registers, crime data, examination registers, etc. The division in four categories is visualized in Figure 1.

6 See also https://usa.ipums.org/usa/mlp/mlp_data_description.shtml, retrieved December 13, 2024.

Figure 1 *Main types of historical (semi-)longitudinal micro-databases (n = 84)*

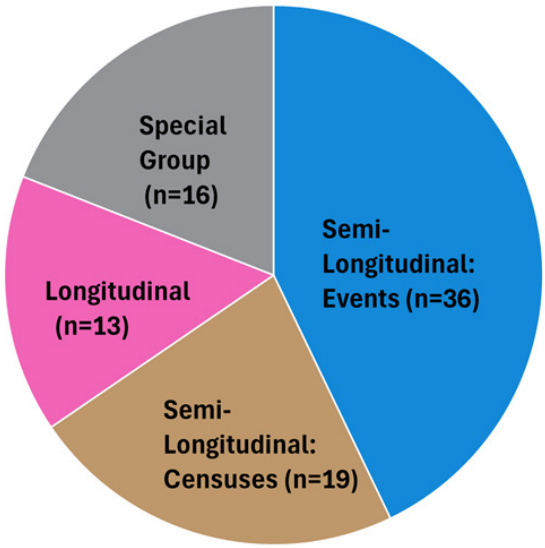


Figure 1 shows that the event databases are by far the largest group. About a quarter of them (n = 10) link only church registers and/or civil certificates, especially the smaller ones. But the others link with other sources as well. Especially links with census data (n = 19) are common. Other linked sources are tax and land registers, and death registers including causes of death (Janssens & Reid, 2025). Half of the census group only has census data, the other half consists of databases with a diverse range of links to other registers, including links to the vital event registers. The next type are the databases with a longitudinal character. Although these types of registers only occur in a few countries, it was still possible to distinguish 13 of these databases. Just three of them contain only population registers, all the others link to other sources: at the top the vital events, in addition mainly causes of death, censuses and tax registers. Finally, (semi-)longitudinal databases are created by starting with a specific cohort or source and subsequently linking records of the same individuals appearing in different sources over time. These databases vary widely in form and scope. We classify them as a "Special Group" and they range from a large birth cohort of a Melbourne hospital, lists of persons deported from England to Tasmania, Finnish military registers to Quaker families in the USA, see further the Appendix. A further condition for inclusion in this group was that the source on which the dataset was founded is linked with other sources and the database is open to research by third parties.

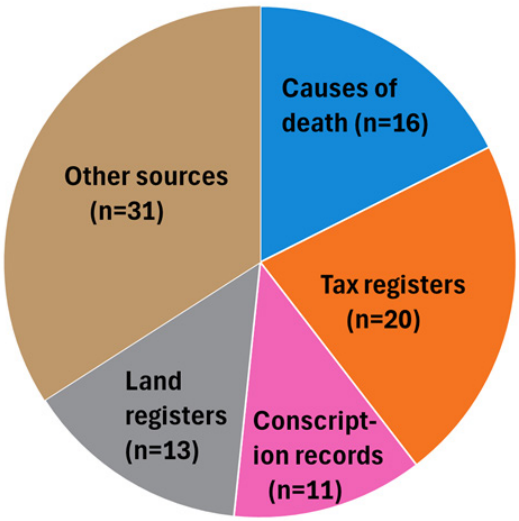
4.3 LINKAGE WITH OTHER SOURCES AND CONNECTIVITY

From the 84 included databases 60 are linked with other sources than the main one with an average of 3.3 per database. Top databases are the Scanian Economic Demographic Database with at least eight different sources (Dribe & Quaranta, 2020) and the Historical Database of the Liège Region with seven linked sources (Alter et al., 2004; van de Walle, 1976). However, it is only because links with external databases are not included in our overview, that the Utah Population Database is not at the top of this ranking (Smith et al., 2022).

Figure 2 presents an overview of the count of the most important sources linked to historical longitudinal micro-databases, excluding the basic registers on which these databases were founded (censuses, vital events and population registers). The most popular links are made to tax registers (20 databases) and registrations with causes of death (16 databases).

The connectivity of the databases is distinguished fourfold: a) linking with contemporary registers with personal data which as a rule are not available for the public, b) internal linking to create links over more than two generations, c) creating links between datasets from different countries and d) standardizing the output structure.

Figure 2 *Different sources linked to by historical longitudinal micro-databases*



Linking with contemporary registers, which are usually rich in data, greatly expands the scope of the database (for European contemporary systems, see Poulain & Herm, 2013). Eleven databases were found to employ this type of linkage. Those are the Scandinavian databases of Norway, Lund and Umeå, the Uppsala Birth Cohort Study, the deCODE database of Iceland, the UTAH population database, USA CenSoc project, the HSN, the Scottish Historic Population Platform (SHiPP), Diggers to Veterans (Tasmania), and the Valserine Valley database in France. The nature of the linking differs per database, e.g., the Scandinavian ones link directly to the population register, the CenSoc project links the Social Security records with the US census of 1940 and the HSN connects with the system of Social Statistical Datasets (SSD) of Dutch Statistics. However, because of the risk of revealing identities it is not always possible to use the full content of the historical data.

Since almost all databases include two-generational links (mother and child, etc.), a database should link at least three generations to get the predicate intergenerational. More than half of the databases (n = 46) fulfill this condition, especially the longitudinal and events databases.

Linking and thus tracking migrants from one country to another is another form of connectivity that can be realized if one of the two countries has a dataset that covers the entire country. The US censuses are an example of this, and there are now successful attempts to find both Dutch (from the HSN sample) and Norwegian migrants in these censuses (Paiva et al., 2020; Roberts, 2018). A complete link between the Canadian and American censuses is also currently being worked on (Fitch et al., 2024).

Connectivity between databases can also be reached by using the same definitions in the specifications of the variables and values. Most evident is the IPUMS case when US census descriptions are the example to follow for the Canadian and European censuses. This internationalization process started with the North Atlantic Population Project which standardized and released the full count of the 1880 censuses including the censuses of the US, UK, Norway, Iceland and Canada (Roberts et al., 2003; Ruggles et al., 2011). The use of the Intermediate Data Structure (IDS; Alter & Mandemakers, 2014) as a way of standardizing the output of (semi-)longitudinal micro-data has now been adopted by 25 databases, including 13 vital events databases and 6 longitudinal ones.

5 COMPARISON OF 84 DATABASES

Figure 3 shows the period of creation of a database in relation to the main type of the database. The first ones were created in the sixties and were initially executed on paper and only in second instance digitized. Examples are the files of Louis Henry already discussed and the Xavier database with Japanese household registers for four villages from the period 1708 to 1870 (Kurosu et al., 2021). From 1970 the data was entered by means of computers, first by means of punched cards, but after 1980 input

took place directly on the mainframes. In addition, the Personal Computer made its appearance and decentralized input became possible. Nowadays, the input usually takes place via web-based data entry programs, where one logs in to the database server.

In the period until 1990 the majority of the databases are based on event data records or population registers. After 1990 the two other types, census based and special groups or research cohorts, became more important. Especially in case of the census databases this is clearly a consequence of the large-scale data entry that became possible with the PC. Volunteers were activated on a large scale through crowd source projects (Prats López et al., 2024) or whole censuses were entered in cheap labour countries as India.

Figure 3 *Number of databases according to main type and period of foundation (n = 84)*

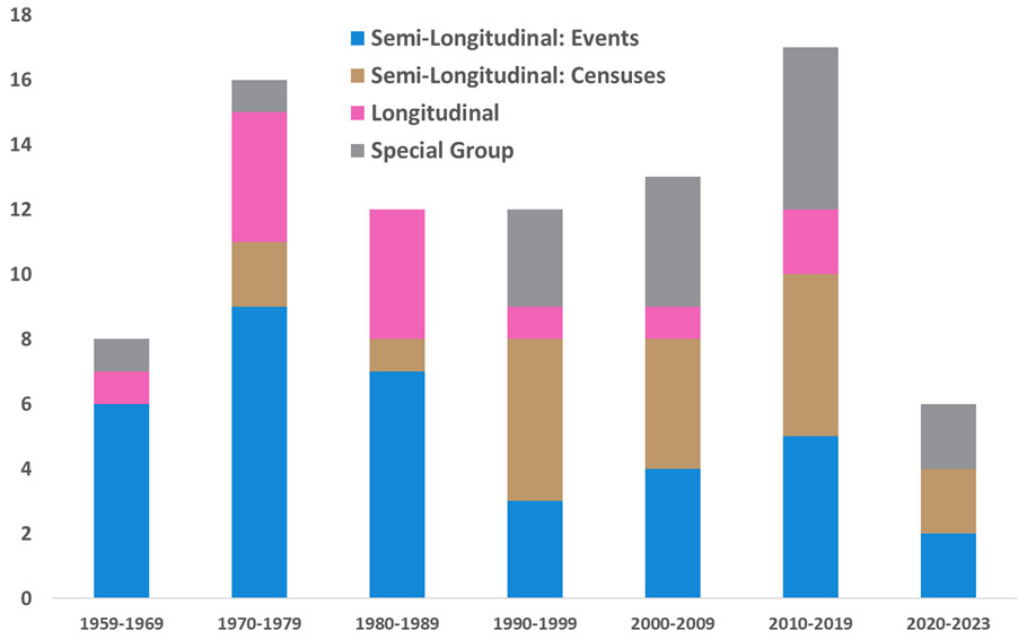


Figure 4 *Number of databases according to sample area and period of foundation (n = 84)*

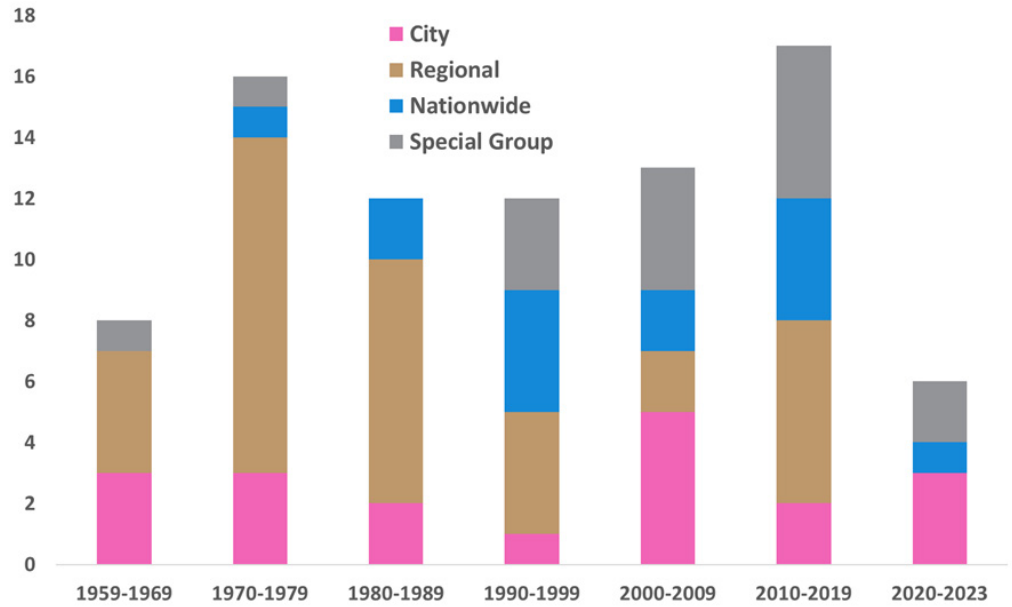


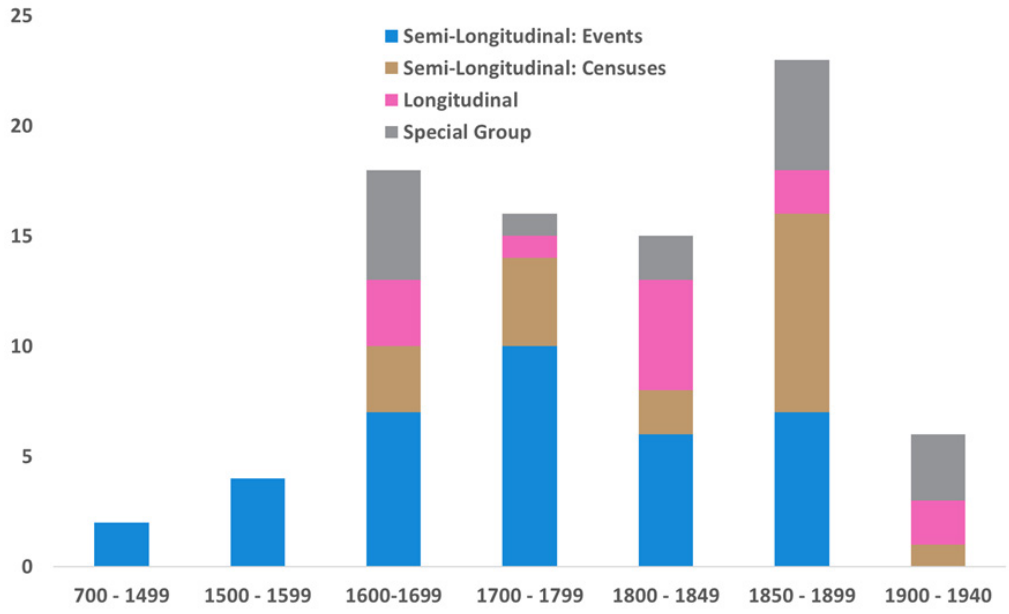
Figure 4 shows the area covered by a database and the period of creation. One may wonder whether a particular area should be considered a region or a country. In this paper the contemporary bordering is followed which means that databases covering large regions as Quebec or certain US states are included as regional databases. We see that in the first 30 years of establishment, almost all databases were limited to a specific area. The criteria on which a region or a city is chosen varies per database and often depends on the presence of sources. Quite often the choice was made for a clustered sample in which extensive consideration is given to which parishes should be selected to become as representative as possible. Within the chosen areas, there usually is no more selection, which makes it a 100% sampling.

The first truly nationally rolled out database is TRA. This project started in 1980 and concerns a letter sample in the marriage certificates of France: brides or grooms whose surnames begin with the letter combination TRA. Originally, the sample was known as Enquête des 3,000 Familles and amounts to about 40,000 marriage certificates (Dupâquier & Kessler, 1992). Subsequently, the database followed different paths, with data being added by various researchers, such as the other civil documents (birth, death), asset data from the succession files and military registers (Bourdieu et al., 2014).

After 1990, the number of databases covering an entire country got off to a good start. An example is the Historical Sample of the Netherlands (HSN), in which the first birth certificates were entered in 1991. The HSN consists of an at-random sample of on average 0.6% of all birth certificates drawn up throughout the Netherlands between 1812 and 1922 (Mandemakers, 2000; 2002), which sample can be enlarged in all kind of ways (Mandemakers & Kok, 2020). Subsequently, the marriage and death certificates are searched and from 1850 onwards also data from the population register, which makes the HSN a true longitudinal database. The big advantage of national coverage is that migration can also be monitored. Other nationally set up databases are the Icelandic deCODE project and the various already mentioned census-based databases.

After 2010 new developments were the creation of two databases with civil certificates, but not for a small region but for the whole country. This concerns the Scottish Historic Population Platform (SHiPP) and in the Netherlands LINKS. In Scotland, funding became available to fully digitize and transcribe all civil status documents for the period 1855–1973, in the Netherlands volunteers created indexes of the certificates which were collected and disseminated by the Dutch Family Centre. These indexes contain the names and roles of all persons mentioned in the certificates, for example in case of the marriage certificates this concerns the bride, the groom and their parents. As soon as the indexation is finished (the speed differs per province) family reconstitutions can be made for the whole of the Netherlands over the period 1812–1924 (birth certificates are only public after 100 years). Ultimately, this will involve about 120 million person observations and 23 million unique persons (Mandemakers, Bloothooft, et al., 2023).

Figure 5 *Number of databases according to main type and start observation period (n = 84)*



The "Special group" databases in Figure 4 are mostly datasets that were not set up regionally but from the point of view of a specific research group. Examples are the Koori in Australia or all men who served in the Finnish army in the first half of the 20th century. These databases were mainly developed after 1990. Sixteen are included in this overview, but quite a lot will be missing, including ones from before 1990. Almost all files have a semi-longitudinal character. Some of these databases grow into very large collections that are moving towards nationwide coverage. We can think of the original Founders and Survivors project with the population of Tasmania largely descended from petty criminals, vagrants and other needy persons transferred by the English crown (Bradley et al., 2010; Cowley et al., 2021).

Figure 5 illustrates the main types of databases according to their observation periods. Over a quarter of the databases begin observations before 1700. Two predate 1500: the Iceland deCODE project, with sources dating back to 740, and the Barcelona Historical Marriage Database, which starts in 1451. Most census databases begin after 1850, whereas most event and longitudinal databases commence before 1850.

Table 1 presents key figures for each database per main type: sample area, sample fraction, length of observation, number of person observations and number of unique persons. When we look at the sample areas covered by the databases, we see that events and longitudinal databases are more than average of a regional character, whereas census databases are more likely to concentrate on a whole country. These are the big census databases from the USA, UK, Canada, France and the Scandinavian countries. The only nationwide *longitudinal* ones are the Historical Sample of the Netherlands and the Historical Database of Suriname and the Caribbean.

The second block shows the average sample frequency. Of all databases, 75% consists of a sample of 100%, the event databases stand out with an average of 89%. This means that if a sample design is used at all, it is a clustered one: sampling parishes in a — as good as possible — representative way and then including all persons. But in most cases, it is only a city or a small region with relatively good sources which makes it worthwhile to start a data entry project. Especially the event structured databases concentrate on a specific area. Databases that are using a sample design on the individual level, quite often opt for a design of a letter sample as the Antwerp COR database and the TRA database in France.

The third group of key figures in Table 1 shows the length of the observation period covered by the databases. The relative size of the three distinguished categories shows that only 24% of the databases covers a range of more than 200 years, 36% belong to the middle category (100–200 years) and 41% cover 100 or fewer years. It is the census main type that has on average a relatively short observation period of less than 100 years, the event-oriented databases cover the longest periods. The two with the greatest coverage are the Islandic database, ranging from 740(!) until the present day (Gudbjartsson et al., 2015) and the Barcelona Historical Marriage Database starting in 1451 and ending in 1905 (Pujadas-Mora et al., 2022). The three with the smallest coverage are an American and Canadian database connecting two censuses (Darroch & Ornstein, 1984; Ferrie, 1996) and the Würzburg database with 16 years concentrating around the census of 1701 (de Vries, forthcoming).

The fourth and fifth group in Table 1 present the estimated numbers of person observations and unique persons. If the number of person observations is not directly known from a statement of the database itself or from the literature, it is mostly estimated by extrapolation of the number of unique persons per database type. In a comparable way the number of unique persons was extrapolated on the number of person observations. See the Appendix for a detailed description how these estimations were made. The Appendix also offers a comprehensive overview per database of all directly found numbers of person observations, marriages, families and unique individuals. The size of the databases varies enormously, but some pattern can still be discovered. More than half of them (56%) can be considered as a fairly small database with a maximum of 1,000,000 person observations, of which most are below the 250,000. Slightly over a quarter are the middle category ranging between 1 and 10 million observations. All except one each, longitudinal and "Special Group" databases contain less than 10 million observations. The big databases with more than 10 million observations are the census databases and the events one with country wide coverage. The same pattern can be seen in the number of unique persons. About two-third of the databases do not exceed the 250,000 unique persons. The longitudinal databases have a relatively big share in the category of between 0.25 and 10 million unique persons and the census databases in the category with over 10 million unique persons.

Table 1 *Key figures (in percentages) by main type of database (n = 84)*

	Events	Censuses	Longitudinal	Special Group	Total
Sample Area					
Nationwide	13.9	36.8	15.4		16.7
City	30.6	26.3	23.1		22.6
Regional	55.6	36.8	61.5		41.7
Special Group				100	19.0
Sample fraction					
1%	2.8	10.5	7.7	12.5	7.1
2–10%		5.3	7.7	12.5	4.8
11–80%	8.3	15.8	15.4	18.8	13.1
100%	88.9	68.4	69.2	56.3	75.0
Length of observation in years					
< 101 years	25.0	57.9	53.8	43.8	40.5
101–200 year	44.4	31.6	15.4	37.5	35.7
> 200 years	30.6	10.5	30.8	18.8	23.8
Estimated number of person observations					
< 0.25 million	30.6	36.8	15.4	56.3	34.5
0.25–1 million	30.6	5.3	30.8	12.5	21.4
1–10 million	22.2	21.1	46.2	25.0	26.2
10–100 million	11.1	21.1	7.7	6.3	11.9
> 100 million	5.6	15.8			6.0
Estimated number of unique persons					
< 0.05 million	27.8	36.8	38.5	56.3	36.9
0.05–0.25 million	38.9	10.5	15.4	25.0	26.2
0.25–1 million	13.9	15.8	23.1	12.5	15.5
1–10 million	5.6	10.5	23.1	6.3	9.5
> 10 million	13.9	26.3			11.9
Total N (= 100%)	36	19	13	16	84

Because in most large databases a distinction is made between the input, processing and output of data, the number of records is in practice a multiple of the "pure" numbers included here (Mandemakers & Dillon, 2004). However, size does not always matter. The large national databases mentioned above often contain relatively few variables in many cases. The censuses up to 1900 give little more than personal data such as name, age, occupation, relationship in the household and place of birth and residence (Roberts et al., 2003; Ruggles, 2014). The longitudinal files based on population registers at least provide migration data and all events occurring between censuses such as deaths, marriages, etc. For scientific research, the databases do not need to be that extremely large, as the Eurasia study has shown, in which five regions in Europe and Asia were analysed in a comparative way (Bengtsson et al., 2004; Lundh & Kurosu, 2014; Tsuya et al., 2010).

Figure 6 *Number of databases according to observation period and number of person observations (n = 84)*

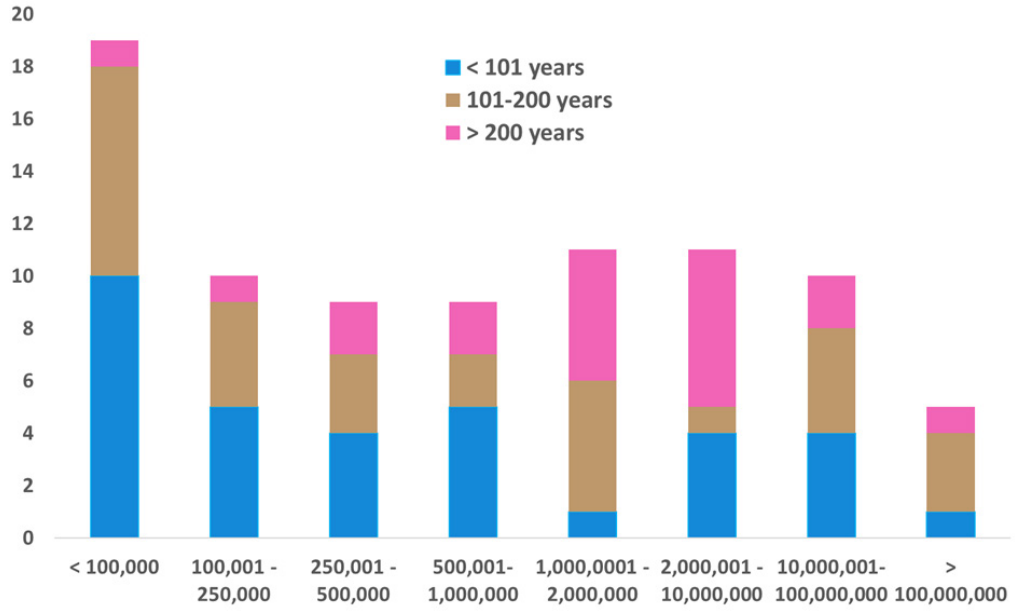


Figure 6 shows the number of included person observations in relation to the observation period covered by a database. Although not perfect, we see a clear relationship between the observation time covered by the data and the size of the database. The smaller the dataset, the smaller the period from which the observations are collected. A clear exception to this rule is the already mentioned UK census database with 230 million person records and an observation period of only 71 years (1851–1921). The other very large ones are the census databases of France and the USA, and the events databases of Utah and the Netherlands (LINKS).

Table 2 presents the totals of estimated person observations and unique persons per main type database. All databases counted we arrive at a total of rounded 2.3 billion person observations. If we subtract the census databases, there are still 727 million observations left, of which 624 million for the five big databases based on events. The longitudinal ones count almost 44 million person observations based on 6 million unique individuals. This is still a huge achievement when we consider that most of the sources for these databases have been collected and edited by research institutes themselves, while the data from the large projects have either been entered by the crowd or by companies such as Findmypast or Ancestry.

Table 2 *Total estimations of the number of person observations and unique persons by main type of database, in millions (n = 84)*

Main type	Number databases	Estimated person observations	Estimated unique persons
Events	36	650.7	100.7
Censuses	19	1,604.9	587.2
Longitudinal	13	43.6	6.0
Special Group	16	32.0	9.6
Total	84	2,331.2	703.5

6 CONCLUSION

An impression has been given of the enormous flight databases with historical population longitudinal data have taken since the first ones were introduced in the sixties of the previous century. A distinction was made in semi-longitudinal types of databases: event databases, census databases and special cohort databases and the real longitudinal ones based on population or comparable registers. All in all, 84 databases were inventoried and compared on key figures like year of foundation, sample fraction and sample area, total number of observations and unique persons and the period of observation. The inventory in the appendix presents these databases with many more details.

It was also inventoried whether and how databases are connected with other datasets. Ten databases are already linked to contemporary registers, providing access to a large number of variables that can be analyzed within a historical micro-data context. Over half of the databases include intergenerational links spanning three or more generations. Structurally, almost all North American and European census databases have implemented the IPUMS standardization and integration efforts. Additionally, 25 databases have adopted the integration of database outputs through the Intermediate Data Structure. And recently links have been realized between North European datasets and the US census data.

All in all, 2.3 billion person observations have been entered over the years and since the speed of data entry is accelerating with the introduction of software for handwritten text and/or character recognition this will probably amount to more than 3 billion before the end of 2025. The development of these databases is indicative of considerable investments that have greatly expanded the possibilities for new research within the fields of history, demography, genetics, sociology, as well as other disciplines.

REFERENCES

Akgün, Ö., Dearle, A., Kirby, G., Garrett, E., Dalton, T., Christen, P., Dibben, C., & Williamson, L. (2020). Linking Scottish vital event records using family groups. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 53*(2), 130–146. https://doi.org/10.1080/01615440.2019.1571466

Alexandre, C., Dupuy, J., & Gourdon, V. (2022). Introduction. Nouveaux regards sur Charleville, Charleville-Mézières [Introduction. New perspectives on Charleville, Charleville-Mézières]. *Cahier d'études Ardennaises, 26,* 5–10. https://hal.science/hal-03905374

Alter, G. (1988). *Family and the female life course: The women of Verviers, Belgium, 1849–1880.* University of Wisconsin Press.

Alter, G. (2019). The evolution of models in historical demography. *Journal of Interdisciplinary History, 50*(3), 325–362. https://doi.org/10.1162/jinh_a_01445

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

Alter, G., & Mandemakers, K. (2025). *Survey of historical databases with longitudinal micro-data* (Version 1) [Data set]. IISH Data Collection. https://hdl.handle.net/10622/VUJHAG

Alter, G., Mandemakers, K., & Gutmann, M. (2009). Defining and distributing longitudinal historical data in a general way through an intermediate structure. *Historical Social Research, 34*(3), 78–114. https://doi.org/10.12759/hsr.34.2009.3.78-114

Alter, G., Mandemakers, K., & Vézina, H. (Eds.). (2023). Content, design and structure of major databases with historical longitudinal population data [Special issue]. *Historical Life Course Studies, Special Issue 5*. https://hlcs.nl//specialissue5

Alter, G., Neven, M., & Oris, M. (2004). Mortality and modernization in Sart and surroundings, 1812–1900. In T. Bengtsson, C. Campbell & J. Z. Lee (Eds.), *Life under pressure. Mortality and living standards in Europe and Asia, 1700–1900* (pp. 173–208). MIT Press. https://doi.org/10.7551/mitpress/4227.003.0013

Alter, G., Newton, G., Oeppen, J., Wrigley, E., Davies, R., & Schofield, R. (2020). *CAMPOP: 26 English family reconstitutions in Intermediate Data Structure format with fertility analysis files 1538–1851* [Data set]. UK Data Service. SN: 854465. https://orcid.org/0000-0001-9338-2164

Amorim, M. N., & de Matos, P. T. (2016). Historical demography in Portugal (1950–2012). In A. Fauve-Chamoux, I. Bolovan & S. Sogner (Eds.), *A global history of historical demography. Half a century of interdisciplinarity* (pp. 533–548). Peter Lang. https://doi.org/10.3726/978-3-0352-0331-8

Antonie, L., Inwood, K., & Ross, J. A. (2015). Dancing with dirty data: Problems in the extraction of life-course evidence from historical censuses. In G. Bloothooft, P. Christen, K. Mandemakers & M. Schraagen (Eds.), *Population reconstruction* (pp. 217–241). Springer. https://link.springer.com/chapter/10.1007/978-3-319-19884-2_11

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: A Journal of Quantitative and Interdisciplinary History, 56*(3), 138–159. https://doi.org/10.1080/01615440.2023.2239699

Bailey, M. J., Leonard, S. H., Price, J., Roberts, E., Spector, L., & Zhang, M. (2023). Breathing new life into death certificates: Extracting handwritten cause of death in the LIFE-M project. *Explorations in Economic History, 87*, 101474. https://doi.org/10.1016/j.eeh.2022.101474

Baskerville, P., & Inwood, K. (2020). The return of quantitative approaches to Canadian history. *The Canadian Historical Review, 101*(4), 585–601. https://doi.org/10.3138/chr-2020-0022

Bengtsson, T., Campbell, C., & Lee, J. Z. (Eds.). (2004). *Life under pressure. Mortality and living standards in Europe and Asia, 1700–1900.* MIT Press. https://doi.org/10.7551/mitpress/4227.001.0001

Bengtsson, T., & Dribe, M. (2021). The long road to health and prosperity, southern Sweden, 1765–2015. Research contributions from the Scanian Economic-Demographic Database (SEDD). *Historical Life Course Studies, 11,* 74–96. https://doi.org/10.51964/hlcs10941

Bideau, A., & Brunet, G. (1998). Les bases de données démographiques sur la vallée de la Valserine et Lehaut-Bugey du XVIIe siècle à nos jours [Demographic databases on the Valserine Valley and Lehaut-Bugey from the seventeenth century to the present day]. *Annales de Démographie Historique, 2*, 175–185. https://www.jstor.org/stable/44385417

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.* Cornell University. https://arxiv.org/pdf/2404.18706v2

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., & Tovey, J. (2014). The TRA Project, a historical matrix. *Population (English Edition)*, *69*(2), 191–220. https://www.jstor.org/stable/43187925

Bradley, J., Kippen, R., Maxwell-Stewart, H., McCalman, J., & Silcot, S. (2010). Research note: The Founders and Survivors project. *The History of the Family, 15*(4), 467–477. https://doi.org/10.1016/j.hisfam.2010.08.002

Brée, S., Gay, V., Leturcq, M., Doignon, Y., & Coulmont, B. (forthcoming). POPP. An OCR-generated database of the population censuses of Paris (1926–1936). *Historical Life Course Studies*.

Breen, C. F., Osborne, M., & Goldstein, J. R. (2023). CenSoc: Public linked administrative mortality records for individual-level research. *Science Data, 10*, 802. https://doi.org/10.1038/s41597-023-02713-y

Breschi, M., Fornasin, A., & Manfredini, M. (2020). The richness of Italian historical demography. *Historical Life Course Studies*, *9*, 228–240. https://doi.org/10.51964/hlcs9304

Brunet, G., Lallich, S., & Bideau, A. (2006). Analyse généalogique et structure de la population. L'ascendance des natifs de la vallée de la Valserine (Jura français), XVIIe-XXe siècles [Genealogical analysis and population structure. The ancestry of the natives of the Valserine Valley (french Jura), 17th-20th centuries]. *Bulletins et Mémoires de la Société d'Anthropologie de Paris, 18*(1–2), 87–102. https://doi.org/10.4000/bmsap.1334

Calderón-Bernal, L. P., Alburez-Gutierrez, D., & Zagheni, E. (2023). *Analysing biases in genealogies using demographic microsimulation* (MPIDR Working Paper WP 2023–034). Max Planck Institute for Demographic Research. https://doi.org/10.4054/MPIDR-WP-2023-034

Campbell, C., & Chen, B. (2022). Nominative linkage of records of officials in the China Government Employee Dataset-Qing (CGED-Q). *Historical Life Course Studies, 12,* 233–259. https://doi.org/10.51964/hlcs11902

Campbell, C., & Lee, J. (2020). Historical Chinese microdata. 40 years of dataset construction by the Lee-Campbell research group. *Historical Life Course Studies, 9*, 130–157. https://doi.org/10.51964/hlcs9303

Cilliers, J. (2021). The South African families database. *Historical Life Course Studies, 11,* 97–111. https://doi.org/10.51964/hlcs11095

Clausen, N. F. (2015). The Danish Demographic Database — Principles and methods for cleaning and standardisation of data. In G. Bloothooft, P. Christen, K. Mandemakers & M. Schraagen (Eds.), *Population reconstruction* (pp. 3–22). Springer. https://link.springer.com/chapter/10.1007/978-3-319-19884-2_1

Constum, T., Kempf, N., Paquet, T., Traounez, 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 (Eds.), *Document Analysis Systems. DAS 2022. Lecture notes in computer science, 13237* (pp. 143–157). Springer. https://doi.org/10.1007/978-3-031-06555-2_10

Cowley, T., Frost, L., Inwood, K., Kippen, R., Maxwell-Stewart, H., Schwarz, M., Shepherd, J., Tuffin, R., Williams, M., Wilson, J., & Wilson, P. (2021). Reconstructing a longitudinal dataset for Tasmania. *Historical Life Course Studies*, *11*, 20–47. https://doi.org/10.51964/hlcs10912

Darrett, B., & Rutman, A. H. (1984a). *A place in time: Middlesex County Virginia, 1650–1750.* Norton.

Darrett, B., & Rutman, A. H. (1984b). *A place in time: Explicatus.* Norton.

Darrett, B., & Rutman, A. H. (2016). *A place in time: Colonial Middlesex County, VA, 1650–1750* [Data set]. Inter-university Consortium for Political and Social Research, 2016-06-15. https://doi.org/10.3886/ICPSR35057.v1

Darroch, A. G., & Ornstein, M. (1984). Family and household in nineteenth-century Canada: Regional patterns and regional economies. *Journal of Family History, 9*(2), 158–177. https://doi.org/10.1177/036319908400900204

Darroch, G., Smith, R. D. B., & Gaudreault, M. (2007). CCRI Sample Designs and Sample Point Identification, Data Entry, and Reporting (SPIDER) software. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 40*(2), 65–75. https://doi.org/10.3200/HMTS.40.2.65-75

de Vries, J. (forthcoming). *Life, death and migration in the early modern city. The urban historical demography of Würzburg.* Cambridge University Press.

del Bosque González, I., García Ferrero, S., Gómez Nieto, I., Martín-Forero, L., Ramiro Fariñas, D. (2010). Cartografía y demografía histórica en una IDE. WMS del plano de Madrid de "Facundo Cañada" [Cartography and historical demography in an SDI. WMS of the Madrid map of "Facundo Cañada"]. *Revista Catalana de Geografia, 15*(40). http://hdl.handle.net/10261/26007

Derosas, R. (1989). A database for the study of the Italian population registers. *Historical Social Research, 14*(4), 59–65. https://doi.org/10.12759/hsr.14.1989.4.59-65

Derosas, R. (1999). Residential mobility in Venice, 1850–1869. *Annales de Démographie Historique, 1*, 35–61.

Diduch, E. (2024). The record linking glass ceiling. Applying automated methods to the census and women's marriage records, 1881–1911. *Historical Life Course Studies, 14*, 126–143. https://doi.org/10.51964/hlcs19189

Dillon, L. Y. (2000). Preface. The origins of IMAG: The International Microdata Access Group. In P. Kelly Hall, R. McCaa & G. Thorvaldsen (Eds.), *Handbook of international historical microdata for population research* (pp. XI-XIV). Minnesota Population Center. https://international.ipums.org/international/resources/microdata_handbook/0_06_introduction_ch01.pdf

Dillon, L., Amorevieta-Gentil, M., Caron, M., Lewis, C., Guay-Giroux, A., Desjardins, B., & Gagnon, A. (2018). The programme de recherche en démographie historique: Past, present and future developments in family reconstitution. *The History of the Family, 23*(1), 20–53. https://doi.org/10.1080/1081602X.2016.1222501

Dillon, L., Amorevieta-Gentil, M., Gagnon, A., & Desjardins, B. (2023). PRDH and IMPQ 1800–1849 Quebec historical family reconstitution. Content, design and biographical completeness. *Historical Life Course Studies, 13,* 103–126. https://doi.org/10.51964/hlcs13984

Dillon, L. Y., & Roberts, R. (2002). Introduction: Longitudinal and cross-sectional historical data: Intersections and opportunities. *History and Computing, 14*(1–2, published 2006), 1–7. https://www.euppublishing.com/doi/abs/10.3366/hac.2002.14.1-2.1

Dong, H., Campbell, C., Kurosu, S., Yang, W., & Lee, J. Z. (2015). New sources for comparative social science: Historical population panel data from East Asia. *Demography, 52*(3), 1061–1088. https://doi.org/10.1007/s13524-015-0397-y

Dribe, M., & Quaranta, L. (2020). The Scanian Economic-Demographic Database (SEDD). *Historical Life Course Studies*, *9*, 158–172. https://doi.org/10.51964/hlcs9302

Dumănescu, L., Hărăguș, M., Lumezeanu, A., Holom, E. C., Hegedűs, N., Mârza, D., Covaci, D., & Bolovan, I. (2022). Historical Population Database of Transylvania. Sources, particularities, challenges, and early findings. *Historical Life Course Studies, 12,* 133–150. https://doi.org/10.51964/hlcs12038

Dupâquier, J., & Kessler, D. (Eds.). (1992). *La société française au XIXe siècle. Tradition, transition, transformations* [French Society in the nineteenth century. Tradition, transition, transformations]. Fayard.

Edvinsson, S., & Engberg, E. (2020). A database for the future. Major contributions from 47 years of database development and research at the Demographic Data Base. *Historical Life Course Studies, 9*, 173–196. https://doi.org/10.51964/hlcs9305

Edvinsson, S., Mandemakers, K., & Smith, K. R. (Eds.). (2023a). Major databases with historical longitudinal population data: Development, impact and results [Special issue]. *Historical Life Course Studies, Special Issue 4*. https://hlcs.nl//specialissue4

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://www.jstor.org/stable/jj.6445823

Fauve-Chamoux, A., Bolovan, I., & Sogner, S. (Eds.). (2016). *A global history of historical demography. Half a century of interdisciplinarity*. Peter Lang. https://doi.org/10.3726/978-3-0352-0331-8

Ferrie, J. P. (1996). A new sample of males linked from the public use microdata sample of the 1850 U.S. federal census of population to the 1860 U.S. federal census manuscript schedules. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 29*(4), 141–156. https://doi.org/10.1080/01615440.1996.10112735

Fitch, C. A., Udalova, V., & Antonie, L. (2024). Leveraging full count census data through record linkage. *International Journal of Population Data Science, 9*(5). https://doi.org/10.23889/ijpds.v9i5.2932

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]. Institut National d'Etudes Démographiques.

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.). Institut National d'Etudes Démographiques.

Fogelvik, S. (1989). The Stockholm Database at work. In P. Denley, S. Fogelvik & Ch. Harvey (Eds.), *History and computing II* (pp. 256–265). Manchester University Press.

Fourie, J., & Green, E. (2018). Building the Cape of Good Hope panel. *The History of the Family, 23*(3), 493–502. https://doi.org/10.1080/1081602X.2018.1509367

Fourie, J., Green, E., Rijpma, A., & von Fintel, D. (2025). Income mobility before industrialization: Evidence from South Africa's Cape Colony. *Social Science History, 49*(1), 22–51. https://doi.org/10.1017/ssh.2024.24

Foxcroft, J., Inwood, K., & Antonie, L. (2022). Linking eight decades of Canadian census collections. *International Journal of Population Data Science, 7*(3), 2076. https://doi.org/10.23889/ijpds.v7i3.2076

Gaffield, C. (2007). Conceptualizing and constructing the Canadian Century Research Infrastructure. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 40*(2), 54–64. https://doi.org/10.3200/HMTS.40.2.54-64

Galli, S., Theodoridis, D., & Rönnbäck, K. (2024). Reconstructing a slave society: Building the *DWI* panel, 1760–1914. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 57*(3), 163–184. https://doi.org/10.1080/01615440.2024.2400188

Garðarsdóttir, O. (2016). Historical demography in Iceland, 1970–2011. In A. Fauve-Chamoux, I. Bolovan & S. Sogner (Eds.), *A global history of historical demography. Half a century of interdisciplinarity* (pp. 315–332). Peter Lang. https://doi.org/10.3726/978-3-0352-0331-8

Garrett, E. M., Blaikie, A., Reid, A., & Davies, R. (2007). *Scottish census enumerators' books: Skye, Kilmarnock, Rothiemay and Torthorwald, 1861–1901* [Data set]. UK Data Service. SN: 5596, https://doi.org/10.5255/UKDA-SN-5596-1

Garrett, E. M., Razzell, P., & Davies, R. (2007). *Sociological study of fertility and mortality in Ipswich, 1872–1910* [Data set]. UK Data Service. SN: 5413. http://doi.org/10.5255/UKDA-SN-5413-1

Garrett, E., & Reid, A. (2015). Introducing "movers" into community reconstructions: Linking civil registers of vital events to local and national census data: A Scottish experiment. In G. Bloothooft, P. Christen, K. Mandemakers & M. Schraagen (Eds.), *Population reconstruction* (pp. 285–298). Springer. https://link.springer.com/chapter/10.1007/978-3-319-19884-2_13

Garrett, E., & Reid, A. (2022). What was killing babies in Ipswich between 1872 and 1909? *Historical Life Course Studies*, *12*, 173–204. https://doi.org/10.51964/hlcs11592

Gehrmann, R. (1984). *Leezen 1720–1870. Ein historisch-demographischer Beitrag zur Sozialgeschichte des ländlichen Schleswig-Holstein* [A historical-demographic contribution to the social history of rural Schleswig-Holstein]. Wachholtz.

Gellatly, C. (2009). Trends in population sex ratios may be explained by changes in the frequencies of polymorphic alleles of a sex ratio gene. *Evolutionary Biology, 36*, 190–200. https://doi.org/10.1007/s11692-008-9046-3

Gellatly, C. (2015). Reconstructing historical populations from genealogical data files. In G. Bloothooft, P. Christen, K. Mandemakers & M. Schraagen (Eds.), *Population reconstruction* (pp. 111–128). Springer. https://link.springer.com/chapter/10.1007/978-3-319-19884-2_6

Geschwind, A., & Fogelvik, S. (2000). The Stockholm Historical Database. In P. Kelly Hall, R. McCaa & G. Thorvaldsen (Eds.), *Handbook of International Historical Microdata for Population Research* (pp. 207–230). Minnesota Population Center. https://international.ipums.org/international/resources/microdata_handbook/1_12_sweden_stockholm_ch13.pdf

Glavatskaya, E., Borovik, J., & Thorvaldsen, G. (2022). The Ural population project. Demography and culture from microdata in a European-Asian border region. *Historical Life Course Studies, 12*, 151–172. https://doi.org/10.51964/hlcs12320

Goeken, R., Huynh, L., Lynch, T. A., & Vick, R. (2011). New methods of census record linking. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 44*(1), 7–14. https://doi.org/10.1080/01615440.2010.517152

Greven, P. J. (1972). *Four generations: Population, land and family in colonial Andover, Massachusetts.* Cornell University Press. https://www.jstor.org/stable/10.7591/j.ctv3mt9gt

Greven, P. J. (2023). *Four generations: Population, land, and family in colonial Andover, Massachusetts, 1630–1750* [Data set]. Inter-university Consortium for Political and Social Research, 2023-01-19. https://doi.org/10.3886/ICPSR35070.v2

Gudbjartsson, D. F., Sulem, P., Helgason, H., Gylfason, A., Gudjonsson, S. A., Zink, F., Oddson, A., Magnusson, G., Halldorsson, B. V., Hjartarson, E., Sigurdsson, G., Kong, A., Helgason, A., Masson, G., Magnusson, O. Th., Thorsteinsdottir U., & Stefansson, K. (2015). Sequence variants from whole genome sequencing a large group of Icelanders. *Scientific Data, 2,* 150011. https://doi.org/10.1038/sdata.2015.11

Gutmann, M. P., (2016). *European-origin and Mexican-origin populations in Texas, 1850, 1860, 1870, 1880, 1900, 1910* [Data set]. Inter-university Consortium for Political and Social Research, 2016-06-20. https://doi.org/10.3886/ICPSR35032.v1

Gutmann, M. P., & Fliess, K. H. (1993). The determinants of early fertility decline in Texas. *Demography, 30,* 443–457. https://www.jstor.org/stable/2061650

Gutmann, M. P., Klancher Merchant, E., & Roberts, E. (2018). "Big data" in economic history. *The Journal of Economic History, 78*(1), 268–299. https://doi.org/10.1017/S0022050718000177

Hautaniemi, S. L., Anderton, D. L., & Swedlund, A. (2000). Methods and validity of a panel study using record linkage: Matching death records to a geographic census sample in two Massachusetts towns, 1850–1912. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 33*(1), 16–29. https://doi.org/10.1080/01615440009598943

Hayami, A. (2016). Historical demography in Japan: Achievements and problems. In A. Fauve-Chamoux, I. Bolovan & S. Sogner (Eds.), *A global history of historical demography. Half a century of interdisciplinarity* (pp. 383–410). Peter Lang. https://doi.org/10.3726/978-3-0352-0331-8

Helgertz, J., Price, J., Wellington, J., Thompson, K. J., Ruggles, S., & Fitch, C. A. (2022). A new strategy for linking U.S. historical censuses: A case study for the IPUMS Multigenerational Longitudinal Panel. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 55*(1), 12–29. https://doi.org/10.1080/01615440.2021.1985027

Henry, L. (1970). *Manuel de démographie historique* [Handbook of historical demography]. Librairie Droz.

Henry, L., & Blum, A. (1988). *Techniques d'analyse en démographie historique* [Analytical techniques in historical demography] (2nd ed.). Institut National d'Etudes Démographiques.

Higgs, E., Jones, C., Schürer, K., & Wilkinson, A. (2013). *Integrated Census Microdata (I_CEM) guide*. University of Essex, Department of History. https://core.ac.uk/download/pdf/18528247.pdf

Imhof, A. E. (1998). *Life-expectancy in Germany, 1700 to 1890* [Data set]. GESIS Data Archive ZA8066. https://doi.org/10.4232/1.8066

Inwood, K., & Maxwell-Stewart, H. (2021). Historical databases now and in the future. *Historical Life Course Studies*, *10*, 9–12. https://doi.org/10.51964/hlcs9558

Janssens, A., & Reid, A. (Eds.). (2025). What was killing babies? European comparative research on infant mortality using individual level causes of death [Special issue]. *Historical Life Course Studies, Special Issue 6*. https://hlcs.nl//specialissue6

Jenkinson, S., Anguita, F., Paiva, D., Matsuo, H., & Matthijs, K. (2020). The 2020 IDS release of the Antwerp COR*-Database. Evaluation, development and transformation of a pre-existing database. *Historical Life Course Studies, 9*, 197–217. https://doi.org/10.51964/hlcs9301

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek , M., Bhatia, G., MacArthur, D. G., Price, A. K., & Erlich, Y. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science, 360*(6385), 171–175. https://doi.org/10.1126/science.aam9309

Karlsson, T., & Lundh, C. (2015). *The Gothenburg Population Panel 1915–1943 GOpp. Version 6.0* (Göteborg Papers in Economic History no. 18). University of Göteborg, Department of Economy and Society. http://hdl.handle.net/2077/38488

Kelly Hall, P., McCaa, R. G, & Thorvaldsen, G. (Eds.). (2000). *Handbook of international historical microdata for population research.* Minnesota Population Center. https://international.ipums.org/international/microdata_handbook.shtml

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

Kesztenbaum, L. (2025). *Finding everyone? Automated data processing of millions of records to reconstitute the French population in the Socface project* [Conference presentation]. European Social Science History Conference (ESSHC) 2025, Leiden, the Netherlands.

Klancher Merchant, E., & Gutmann, M. P. (2022). The IT of demography. *EEE Annals of the History of Computing, 44,* 6–15. https://doi.org/10.1109/MAHC.2022.3214634

Knodel, J. E. (1988). *Demographic behavior in the past: A study of fourteen German village populations in the eighteenth and nineteenth centuries*. Cambridge University Press. https://doi.org/10.1017/CBO9780511523403

Koupil, I. (2007). The Uppsala studies on developmental origins of health and disease. *Journal of Internal Medicine, 261*(5), 426–436. https://doi.org/10.1111/j.1365-2796.2007.01799.x

Kross, J. (1974). *A town study in colonial New York: Newtown, Queens County (1642–1790)* [Doctoral dissertation, University of Michigan]. https://dx.doi.org/10.7302/10927

Kross, J. (2016). *Families of Newtown, New York, 1642–1790* [Data set]. Inter-university Consortium for Political and Social Research, 2016-06-15. https://doi.org/10.3886/ICPSR35005.v2

Kurosu, S., Takahashi, M., & Dong, H. (2021). Thank you, Akira Hayami! The Xavier database of historical Japan. *Historical Life Course Studies, 11*, 112–131. https://doi.org/10.51964/hlcs11113

Liczbińska, G., Pankowski, P. S., Antosik, S., & Kowalska, K. (forthcoming). Searching for a needle in a haystack: The intricacies of sources on causes of death, their interpretation, and the construction of a database for the city of Poznań, 1830–1900. *Historical Life Course Studies*.

Lin, C., Chen, S., Chuang, Y., Yang, W., Wilkerson, J., Hsieh, Y., Yap, K., & Huang, Y. (2020). A longitudinal historical population database in Asia. The Taiwanese Historical Household Registers Database (1906–1945). *Historical Life Course Studies, 9,* 218–227. https://doi.org/10.51964/hlcs9300

Longley, P., van Dijk, J., & Lan, T., (2022). Linkage of historical GB census data to present day population registers. *International Journal of Population Data Science, 7*(3), 227. https://doi.org/10.23889/ijpds.v7i3.2002

Lundh, C, & Kurosu, S. (Eds.). (2014). *Similarity in difference: Marriage in Europe and Asia, 1700–1900.* MIT Press. https://doi.org/10.7551/mitpress/9780262027946.001.0001

Mandemakers, K. (2000). The Netherlands. Historical Sample of the Netherlands. In P. Kelly Hall, R. McCaa & G. Thorvaldsen (Eds.), *Handbook of International historical microdata for population research* (pp. 149–178). Minnesota Population Center. https://international.ipums.org/international/resources/microdata_handbook/1_10_netherlands_ch11.pdf

Mandemakers, K. (2002). Building life course datasets from population registers by the Historical Sample of the Netherlands (HSN). *History and Computing, 14*(1–2, published 2006), 87–108. https://www.euppublishing.com/doi/abs/10.3366/hac.2002.14.1-2.87

Mandemakers, K. (2023a). "*You really got me*". *Ontwikkeling en toekomst van historische databestanden met microdata* ["You really got me". Development and future of historical databases with microdata] [Valedictory speech]. Erasmus University Rotterdam, the Netherlands. https://doi.org/10.25397/eur.23256467

Mandemakers, K. (2023b). *60 Databases with (semi-) longitudinal data compared* [Conference presentation]. European Society for Historical Demography (ESHD) 2023, Nijmegen, the Netherlands.

Mandemakers, K. (2025). *Overview databases with historical longitudinal data* (Version 1) [Data set]. IISH Data Collection. https://hdl.handle.net/10622/VEODGX

Mandemakers, K., Alter, G., Vézina, H., & Puschmann, P. (Eds.). (2023). *Sowing. The construction of historical longitudinal population databases.* Radboud University Press. https://www.jstor.org/stable/jj.6445824.28

Mandemakers, K., Bloothooft, G., Laan, F., Mourits, R., Raad, J., & Zijdeman, R. (2023). LINKS. A system for historical family reconstruction in the Netherlands. *Historical Life Course Studies, 13*, 148–185. https://doi.org/10.51964/hlcs14685

Mandemakers, K., & Dillon, L. (2004). Best practices with large databases on historical populations. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 37*(1), 34–38. https://doi.org/10.3200/HMTS.37.1.34-38

Mandemakers, K., & Kok, J. (2020). Dutch lives. The Historical Sample of the Netherlands (1987–): Development and research. *Historical Life Course Studies, 9*, 69–113. https://doi.org/10.51964/hlcs9298

McCaa, R., & Ruggles, S. (2002). The census in global perspective and the coming microdata revolution. In: J. Carling (Ed.), *Nordic demography: Trends and differentials* (pp. 7–30). Scandinavian Population Studies, 13. Unipub/Nordic Demographic Society. https://users.pop.umn.edu/~rmccaa/census_microdata_revolution.pdf

McCalman, J., Morley, R., Smith, L., & Anderson, I. (2011). Colonial health transitions: Aboriginal and 'poor white' infant mortality compared, Victoria 1850–1910. *The History of the Family, 16*(1), 62–77. https://doi.org/10.1016/j.hisfam.2010.09.005

McCalman, J. (2021). Building longitudinal datasets from diverse historical data in Australia. *Historical Life Course Studies, 11,* 1–19. https://doi.org/10.51964/hlcs10939

Mehldau, J. K. (2011). *Wittgensteiner Familiendatei. Eine Datenbank zur Familiengeschichtsforschung* [Wittgensteiner family file. A database for family history research] (Version 1.0.0) [Data set]. GESIS Data Archive ZA8509. https://doi.org/10.4232/1.10721

Meindl, R. D. (1979). *Environmental and demographic correlates of mortality in 19th century Franklin County, Massachusetts* [Unpublished doctoral dissertation]. University of Massachusetts Amherst.

Minvielle, S. (2013). Les ménages de Charleville aux XVIIIe–XIXe siècles [Households in Charleville in the 18th and 19th centuries]. *Histoire & Mesure, 28*(2), 17–52. https://doi.org/10.4000/histoiremesure.4788

Moisseenko, T., & Koster, M. (2025). *Survey of historical databases with longitudinal micro-data: The second questionnaire* (Version 1) [Data set]. IISH Data Collection. https://hdl.handle.net/10622/SEY90H

Naul, F., & Desjardins, B. (1989). Computers and historical demography: The reconstitution of the early Québec population. In P. Denley, S. Fogelvik & C. Harvey (Eds.), *History and computing II* (pp. 143–148). Manchester University Press.

Oris, M., Mazzoni, S., & Ramiro-Fariñas, D. (2023). Immigration, poverty, and infant and child mortality in the city of Madrid, 1916–1926. *Social Science History, 47*(3), 453–489. https://doi.org/10.1017/ssh.2023.9

Oris, M., Perroux, O., Ryczkowska, G., Schumacher, R., Remund, A., & Ritschard, G. (2023). Geneva. An urban sociodemographic database. *Historical Life Course Studies, 13,* 212–227. https://doi.org/10.51964/hlcs15621

Paek, S., Park, J. H., & Lee, S. (2022). Building an archival database for visualizing historical networks. A case for pre-modern Korea. *Historical Life Course Studies*, *12*, 42–57. https://doi.org/10.51964/hlcs11718

Paiva, D., Anguita, F., & Mandemakers, K. (2020). Linking the Historical Sample of the Netherlands with the USA censuses, 1850–1940. *Historical Life Course Studies, 9*, 1–23. https://doi.org/10.51964/hlcs9312

Pakot, L. (2014). Family composition, birth order and timing of first marriages in rural Transylvania. A case study of Szentegyházasfalu (Vlăhița) and Kápolnásfalu (Căpâlnița), 1838–1940. *Hungarian Historical Review, 3*(1), 141–167. https://www.jstor.org/stable/43265192

Paping, R., & Sevdalakis, D. (2022). The Groningen Integral History Cohort Database. Development, design and output. *Historical Life Course Studies, 12,* 78–98. https://doi.org/10.51964/hlcs12033

Park, H., & Lee, S. (2008). A survey of data sources for studies of family and population in Korean history. *The History of the Family, 13*(3), 258–267. https://doi.org/10.1016/j.hisfam.2008.05.005

Pfister, U., & Fertig, G. (2024). Demographic data for the pre-statistical age (late sixteenth century to 1870). *German Economic Review, 25*(4), 255–273. https://doi.org/10.1515/ger-2024-0064

Poulain, M., & Herm, A. (2013). Central population registers as a source of demographic statistics in Europe. *Population-E, 68*(2), 183–212. https://doi.org/10.3917/pope.1302.0183

Prats López, M., van Oort, T., Ganzevoort, W., van Galen, C., & Mourits, R. J. (2024). Understanding patterns of engagement in the citizen humanities: The civil records of Suriname. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 58*(1), 1–16. https://doi.org/10.1080/01615440.2024.2414925

Pujadas-Mora, J. M., Fornés, A., Ramos Terrades, O., Lladós, J., Chen, J., Valls-Fígols, M., & Cabré, A. (2022). The Barcelona Historical Marriage Database and the Baix Llobregat Demographic Database. From algorithms for handwriting recognition to individual-level demographic and socioeconomic data. *Historical Life Course Studies*, *12*, 99–132. https://doi.org/10.51964/hlcs11971

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

Raftakis, M. (2025). The Bologna Parish Population Database (BPPD), 17–19th centuries. *Italian Parish Records, 6*(1), 5–16. https://drive.google.com/drive/u/1/folders/1QCnEyPA_zqFploJ7Cct3SpxtXoWRNNm1

Raftakis, M., Barban, N., & Rettaroli, R. (2025). *Challenges in constructing large-scale parish microdata: The Bologna and its suburbs parish register database* [Conference presentation]. European Social Science History Conference (ESSHC) 2025, Leiden, the Netherlands.

Rappo, L. (2022). *Parenté, proximite spatiale et liens sociaux de l'Ancien Régime à la Suisse modern. Le cas de Corsier-sur-Vevey de 1700 à 1840* [Kinship, spatial proximity and social ties from the Ancien Régime to Modern Switzerland. The case of Corsier-sur-Vevey from 1700 to 1840]. Peter Lang. https://www.peterlang.com/document/1297594

Reher, D. S., & Sanz-Gimeno, A. (2007). Rethinking historical reproductive change: Insights from longitudinal data for a Spanish town. *Population and Development Review, 33*(4), 703–727. https://doi.org/10.1111/j.1728-4457.2007.00194.x

Reid, A., Davies, R., & Garrett, E. (2002). Nineteenth-century Scottish demography from linked censuses and civil registers: A 'Sets of Related Individuals' approach. *History and Computing, 14*(1–2, published 2006), 61–86. https://doi.org/10.3366/hac.2002.14.1-2.61

Rijpma, A., Cilliers, J., & Fourie, J. (2020). Record linkage in the Cape of Good Hope Panel. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 53*(2), 112–129. https://doi.org/10.1080/01615440.2018.1517030

Roberts, E. (2018). Across the Atlantic and back: Tracing the lives of Norwegian-American migrants, 1850–1930. *Journal of Migration History, 4*(2), 289–313. https://doi.org/10.1163/23519924-00402004

Roberts, E., Ruggles, S., Dillon, L. Y., Garðarsdóttir, Ó., Oldervoll, J., Thorvaldsen, G., & Woollard, M. (2003). The North Atlantic Population Project. An overview. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 36*(2), 80–88. https://doi.org/10.1080/01615440309601217

Robinson, O., Mathiesen, N., Thomsen, A., & Revuelta-Eugercios, B. (2022). *Link-Lives Guide* (Version 1). Danish National Archives/University of Copenhagen. https://www.rigsarkivet.dk/wp-content/uploads/2022/08/Link-lives-Guide-v.1.pdf

Rönnbäck, K., Galli, S., & Theodoridis, D. (2024). *The Danish West Indies Panel* (Version 2) [Data set]. University of Gothenburg. https://doi.org/10.5878/05g8-5n03

Ruggles, S. (2012). The future of historical family demography. *Annual Review of Sociology, 38*(18), 1–19. https://doi.org/10.1146/annurev-soc-071811-145533

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

Ruggles, S. (2018). Metadata and preservation. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research* (pp. 635–643). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-54395-6_71

Ruggles, S., Hacker, J. D., & Sobek, M. (1995). General design of the integrated public use microdata series. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 28*(1), 33–39. https://doi.org/10.1080/01615440.1995.9955311

Ruggles, S., & Menard, R. R. (1995). The Minnesota historical census projects. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 28*(1), 6–10. https://doi.org/10.1080/01615440.1995.9955308

Ruggles, S., Roberts, E., Sarkar, S., & Sobek, M. (2011). The North Atlantic Population Project: Progress and prospects. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 44*(1), 1–6. https://doi.org/10.1080/01615440.1995.9955308

Saarti, J., Ropponen, J., & Soivanen, S. (2017). *The Karjala database — Challenges and solutions for digitizing heterogeneous, old genealogical documents for internet use* [Conference paper]. DH. Opportunities and Risks. Connecting Libraries and Research 2017, Berlin, Germany. https://inria.hal.science/hal-01660143/document

Sabean, D. W. (1990). *Property, production and family in Neckerhausen, 1700–1870.* Cambridge University Press.

Sabean, D., & Ball, R. (2024). *Neckerhausen Research Database* [Data set]. UCLA. https://dataverse.ucla.edu/dataverse/NRDB

Sager, E. W. (1998). *The national sample of the 1901 census of Canada: A new source for the study of the working class* [Conference paper]. European Social Science History Conference (ESSHC) 1998, Amsterdam, the Netherlands.

Sapiro, P. (2024a). Development of the Liverpool Jewry Historical Database. *Genealogy, 8*(4), 128. https://doi.org/10.3390/genealogy8040128

Sapiro, P. (2024b). *Liverpool Jewry Historical Database, 1740–1881* [Data set]. UK Data Service. SN: 9304. https://doi.org/10.5255/UKDA-SN-9304-1

Schürer, K. (2007). Focus: Creating a nationally representative individual and household sample for Great Britain, 1851 to 1901 — The Victorian Panel Study (VPS). *Historical Social Research, 32*(2), 211–331. http://www.jstor.org/stable/20762213

Schürer, K., & Higgs, E. (2020). *Integrated Census Microdata (I-CeM), 1851–1911* [Data set]. UK Data Service. SN: 7481. http://doi.org/10.5255/UKDA-SN-7481-3

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.

Séguy, I. (2016). The French School of historical demography (1950–2000). In A. Fauve-Chamoux, I. Bolovan & S. Sogner (Eds.), *A global history of historical demography. Half a century of interdisciplinarity* (pp. 257–276). Peter Lang. https://doi.org/10.3726/978-3-0352-0331-8

Smith, K. R., & Mineau, G. P. (2021). The Utah Population Database. The legacy of four decades of demographic research. *Historical Life Course Studies, 11*, 48–73. https://doi.org/10.51964/hlcs10916

Smith, K. R., Fraser, A., Reed, D. L., Barlow, J., Hanson, H. A., West, J., Knight, S., Forsythe, N., & Mineau, G. P. (2022). The Utah Population Database. A model for linking medical and genealogical records for population health research. *Historical Life Course Studies, 12*, 58–77. https://doi.org/10.51964/hlcs11681

Sobek, M., Cleveland, L., Flood, S., Kelly Hall, P., King, M. L., Ruggles, S., & Schroeder, M. (2011). Big data: Large-scale historical infrastructure from the Minnesota Population Center. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 44*(2), 61–68. https://doi.org/10.1080/01615440.2011.564572

Sommerseth, H. L., & Thorvaldsen, G. (2022). The impact of microdata in Norwegian historiography 1970 to 2020. *Historical Life Course Studies, 12,* 18–41. https://doi.org/10.51964/hlcs11675

Song, X., & Campbell, C. D. (2017). Genealogical microdata and their significance for social science. *Annual Review of Sociology, 43*, 75–99. https://doi.org/10.1146/annurev-soc-073014-112157

Swedlund, A. C. (2008). *The genetic structure of an historical population: A study of marriage and fertility in Old Deerfield* (Research Report 07). University of Massachusetts Amherst. https://hdl.handle.net/20.500.14394/2535

Szołtysek, M., & Gruber, S. (2016). Mosaic: Recovering surviving census records and reconstructing the familial history of Europe. *The History of the Family, 21*(1), 38–60. https://doi.org/10.1080/1081602X.2015.1006655

Taskinen, I. (2023). Construction of the Finnish Army in World War II Database. *Historical Life Course Studies, 13,* 44–60. https://doi.org/10.51964/hlcs13565

Teibenbacher, P. (2010). Fertility decline, illegitimacy, and denomination. The demographic story of the Wald parish between 1880 and 1938. In M-P. Arrizabalaga, I. Bolovan, M. Eppel, J. Kok & M. L. Nagata (Eds.), *Many paths to happiness? Studies in population and family history. A Festschrift for Antoinette Fauve-Chamoux* (pp. 238–260). Aksant.

Thorvaldsen, G., & Holden, L. (2023). The development of microhistorical databases in Norway. A historiography. *Historical Life Course Studies, 13*, 127–147. https://doi.org/10.51964/hlcs14315

Tsuya, O., Feng, W., Alter, G., & Lee, J. Z. (Eds.). (2010). *Prudence and pressure: Reproduction and human agency in Europe and Asia, 1700–1900.* MIT Press. https://doi.org/10.7551/mitpress/8162.003.0023

Tulinius, H. (2011). Multigenerational information: The example of the Icelandic Genealogy Database. In J. Dillner (Ed.), *Methods in Biobanking. Methods in Molecular Biology, 675* (pp. 221–230). Springer. https://doi.org/10.1007/978-1-59745-423-0_11

van de Walle, E. (1976). Household dynamics in a Belgian village, 1847–1866. *Journal of Family History, 1*(1), 80–94. https://doi.org/10.1177/036319907600100106

van Galen, C. W., Mourits, R. J., Rosenbaum-Feldbrügge, M., A.B., M., Janssen, J., Quanjer, B., van Oort, T., & Kok, J. (2023). Slavery in Suriname. A reconstruction of life courses, 1830–1863. *Historical Life Course Studies, 13,* 191–211. https://doi.org/10.51964/hlcs15619

Vézina, H., & Bournival, J.-S. (2020). An overview of the BALSAC Population Database. Past developments, current state and future prospects. *Historical Life Course Studies, 9*, 114–129. https://doi.org/10.51964/hlcs9299

Vikström, P., Edvinsson, P., & Brändström, A. (2002). Longitudinal databases — Sources for analyzing the life-course: Characteristics, difficulties and possibilities. *History and Computing, 14*(1–2, published 2006), 109–128. https://www.euppublishing.com/doi/abs/10.3366/hac.2002.14.1-2.109

Vikström, P., Larsson, M., Engberg, E., & Edvinsson, S. (2023). The Demographic Database — History of technical and methodological achievements. *Historical Life Course Studies, 13,* 89–102. https://doi.org/10.51964/hlcs12163

Voland, E. (1988). Differential infant and child mortality in evolutionary perspective: Data from late 17th to 19th century Ostfriesland (Germany). In L. Betzig, M. Borgerhoff Mulder & P. Turke (Eds.), *Human reproductive behaviour — A Darwinian perspective* (pp. 253–261). Cambridge University Press.

Voland, E. (2000). Contributions of family reconstitution studies to evolutionary reproductive ecology. *Evolutionary Anthropology: Issues, News, and Reviews, 9*(3), 134–146. https://doi-org.ru.idm.oclc.org/10.1002/1520-6505(2000)9:3<134::AID-EVAN3>3.0.CO;2-M

Watkins, S. C., & McCarthy, J. (1980). The female life cycle in a Belgian commune: La Hulpe, 1847–1866. *Journal of Family History, 5*(2), 167–179. https://doi.org/10.1177/036319908000500203

Wells, R. V. (1969). *A demographic analysis of some middle colony Quaker families of the eighteenth century* [Unpublished doctoral dissertation]. Princeton University.

Wells, R. V. (1971). Family size and fertility control in eighteenth-century America: A study of Quaker families. *Population Studies, 25*(1), 73–82. https://doi.org/10.1080/00324728.1971.10405784

Willführ, K. P., & Störmer, C. (2015). Social strata differentials in reproductive behavior among agricultural families in the Krummhörn region (East Frisia, 1720–1874). *Historical Life Course Studies, 2*, 58–85. https://doi.org/10.51964/hlcs9359

Wimmer, L. T. (2003). Reflections on the Early Indicators project: A partial history. In D. L. Costa (Ed.), *Health and labor force participation over the life cycle: Evidence from the past* (pp. 1–11). University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/H/bo3614165.html

Wisselgren, M. J., Edvinsson, S., Berggren, M., & Larsson, M. (2014). Testing methods of record linkage on Swedish censuses. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 47*(3), 138–151. https://doi.org/10.1080/01615440.2014.913967

Wrigley, E. A., Davies, R. S., Oeppen, J. E., & Schofield, R. S. (1997). *English population history from family reconstitution 1580–1837.* Cambridge University Press. https://doi.org/10.1017/CBO9780511660344

APPENDIX KEY FIGURES OF 84 DATABASES WITH HISTORICAL LONGITUDINAL DATA

A.1 INTRODUCTION

Basis for this article is a table with metadata about 84 databases with historical longitudinal data. The full table in the form of a spreadsheet is attached to this article.[7] The following briefly describes the sources for this overview, including the way missing data was estimated. Next comes a description of all the fields included in the table. Fields with names printed in italics are shown in the sequential section with a brief overview of all included databases. The last section explains the way in which data concerning numbers of observed persons and unique persons have been estimated for a number of databases.

In general databases stand for one big dataset, but a database may contain more than one dataset. Examples are IPUMS with several distinct census databases, or HSNDB with the HSN and the LINKS data or the Demographic Database in Umea which makes a distinction between the public datasets (POPUM) and the closed one linking with the modern population register (POPLINK). But, in general, when data from multiple sources are linked and integrated, the database is considered one big dataset.

A.1.1 SOURCES

Information about databases was collected within the context of the EHPS-Net (European Historical Population Samples Network, https://ehps-net.eu/). During the period 2013–2014, for 32 databases a huge range of metadata was systematically collected in the form of questionnaires (Moisseeko & Koster, 2025). Information from these questionnaires is referenced as "EHPS_Q". The column "References" in the table below (and in the spreadsheet) delivers all references to sources that were used to collect these metadata.

A second major source were the articles that were published in two special issues of *Historical Life Course Studies* (HLCS), providing descriptions of 32 databases in 30 articles. Special Issue 4 was edited by Sören Edvinsson, Kees Mandemakers and Ken Smith (2023a) and dealt with the impact of seven of the larger databases. Special issue 5 was edited by George Alter, Kees Mandemakers and Hélène Vézina (2023) and dealt with the technical and methodological aspects of databases.

Besides these major sources, information was gathered from articles or books describing this kind of databases or other earlier published more limited overviews (see Section 2) and personal information from database administrators. Further information was found on websites and in three cases the dataset itself was used. Another possible source, kinsources.net, with anthropological data, could not be reached during the finalizing of this overview.

The classification is based on the current formal situation, which means that large regions/states that form fairly unique entities, such as Quebec in Canada, are nevertheless classified as a region. The regions therefore differ enormously in geographical areas. If a project was only a sample of a few parishes of a country, it was considered as a regional project, because the degree of representativeness for the country as a whole was unclear. These are, for example, the Louis Henry dataset, the German Ortsfamilienbücher and the UK dataset of Wrigley et al. (1997).

The measure points of the data for the overview differ. In the case of the articles in *HLCS* data and included sources are measured around 2020, in case of the questionnaires around 2014. When there were two or more measuring points, the most recent one was chosen. Some — especially smaller —databases were closed over the years, in which case the dates of measurement were taken at an earlier stage.

A.1.2 NUMBERS OF PERSON OBSERVATIONS AND UNIQUE PERSONS

At the time of writing, it proved to be practically impossible to obtain completely accurate numbers concerning observations and unique persons and families, from almost any database. In the case of databases that are still being worked on, this makes sense; the known numbers are always lagging behind. But even with closed databases, it is often unclear what they actually contain without directly inspecting

7 The spreadsheet is also published in Dataverse (IISH Data Collection), see Mandemakers, 2025. See also Footnote 2.

the data (if they are easy to find at all and stored in a clearly manageable format). Only rarely both numbers of family observations and the linked result in terms of unique families are explicitly reported.

Ideally the magnitude and size of a database is indicated by the number of included unique persons, the number of observations of these persons, the number of unique families and the number of linked generations. There were only a few databases that delivered exact data on all four indicators. Although both the questionnaires and the articles in the above-mentioned two issues of *Historical Life Course Studies* tried to get the best possible figures, the result appeared to be disappointing in practice. Even the results from the questionnaires are incomplete and not always clear, although the questions were quite explicit. Documentation made by the organization itself about the database is not always clear on these points either. This implies that for this overview they need to be estimated, which, luckily, is sufficient to get a good impression. In this article we limit ourselves to estimations for the main categories: person observations and unique persons.

A.1.3 ESTIMATION OF NUMBERS OF UNIQUE PERSONS AND PERSONS OBSERVATIONS

The goal was to get an indication of the size and importance of a database through the total number of person observations and the total number of unique persons. Estimating the number of families seemed even more complicated, not in the least because the number of families and/or households were the least published and families and households were mixed up. For this reason, we only estimated data about persons.

In quite a lot of instances no numbers of persons are given at all, but only the number of births, marriages and deaths. If the number of these sources was known, it was possible to calculate the number of person observations, this was done through the multipliers presented in Table 3.

Table 3 *Numbers of person observations per type of source*

Source type	Person observations	
Birth certificate/Baptism records	3	Mother, father child
Marriage certificate, church records	6	Bride, groom with parents
Death certificate/Buriel records	3.5	Mother, father, child, spouse (half of the records)

For the Transylvanian database, such a calculation resulted in 536,718 person observations, while the database itself published a number of 570,036 observations. The difference could be explained by the fact that officials and in some cases also witnesses were included (Dumănescu et al., 2022). So, in the case of databases with only church records our basic calculation was done as explained above. Sometimes adjusted, for example when it was known that the parents were lacking in the church marriage records or when the database included intergenerational links which implied that births would appear as parents in the next generation.

Table 4 provides an overview of what metadata about persons we could get from database organizations or the relevant literature. We see that 38 databases delivered both figures on person observations and on unique persons and the other ones only deliver one of these figures. This reduces the problem to estimating data about persons of 46 databases. For 21 databases the number of unique persons has to be estimated through the number of person observations and for 25 databases the other way round for lacking numbers of person observations. These estimates will be made per main type of databases.

In principle the established average ratio for each type of database was used to estimate the unknown values of either the number of person observations or the number of unique persons. Exceptions to the rule are the big census databases. Only for IPUMS the number of unique persons could be directly estimated (on the basis of the matching results). This resulted in an average of 2.1 person observations per unique person, so an average unique person was only seen in two censuses. This seems very low but in order to be matched in two censuses a person had to reach an average age of 15. So, almost all children that died early are not included. Other main leaks are the last census (1950) which is not linked forward by definition and the bad linking results of persons leaving their birth environment, and — especially in Anglo-Saxon environments — women who started using the last name of the husband with marriage. For the Canadian and the UK census we used the same average as for IPUMS and for the Scandinavian censuses which link better, an average of 3 was taken. For the French census, which was 5-yearly counted, the average was put on 6. For the remaining census databases, we used the average of 3.3.

Table 4 *Number of databases according to main type and presence of person data (n = 84)*

Main type data	Total	Both person observations and unique persons	Only person observations	Only unique persons	Ratio person observations and unique persons*	Range
Events	36	18	8	10	4.4	1.7 to 25.0
Longitudinal	13	4	3	6	6.7	3.7 to 12.3
Censuses	19	9	9	1	3.3	1.4 to 12.5
Special group	16	7	1	8	6.3	1.8 to 13.4
Total	84	38	21	25		

** Ratio is calculated as the average of the ratios of each database excluding the very deviant maximum value of the main types of events and censuses.*

In case of longitudinal data, quite often the number of unique persons was known but not the number of registrations (c.q. observations). Estimation is quite difficult since the number of registrations depends on the life cycle and migration path of a person and the way the population registers or comparable sources were handled. A complication is also that data can be changed (e.g., an occupational title or a changed civil status) without creating an extra line in the population register, although these are new person observations as well. All things considered, for the calculation of observations it seems reasonable not to accept the average based on only four databases, but to put this rate higher, at 8.5.

A.2 METADATA TABLE: DESCRIPTION OF FIELDS

The next table presents all included fields in the overview of included databases (attached spreadsheet) with an explanation where necessary (italicized field names are also included in the table following).

Name field	Explanation	Values and/or more general remarks
Id	Primary key table.	
Name	Name of the database or dataset.	In case no official name was known, a name was constructed.
Year_founding	Year in which the database was established and/or started with data entry.	It was not always exactly clear in which year a database was established, years with an * are assessed.
Region	Main geographical area the data originates from.	
Country	Country the region is situated in.	Present borders are used.
Continent	Continent the country is situated in.	
Sample_information	Information about the structure of the sample.	Typical samples are clustered samples, letter-based samples, at random samples etc.
Sample_area	Typology of the geographical area covered.	C City (could include neighboring areas) R Regional N Nationwide SG Special group study
Sample_frequency	Frequency of the sample in percentages.	It was not always exactly clear which sampling rate was used. Figures with an * are averaged or estimated. Percentages are rounded to whole numbers. Clustered samples are judged as 100%.
Sample_frequency_cat	Sample frequency in categories.	Values: 1, 2–10, 11–50, 80, 100%
Start_observation	Year in which the observation period started.	Period of observation have been taken as broad as possible, which implies that not all subperiods covered by the database have the same data intensity.
End_observation	Year in which the observation period ended.	

Name field	Explanation	Values and/or more general remarks	
Observation_period	Observation period.	Combination of start and end year observations.	
Observation_range	Range observation period in years.	End year minus start year.	
Observation_range_cat	Range observation period in categories.	Values in number of years: < 101, 101–200, > 200	
Main_source	Starting point of a database (also main type of the database).	CC	Civil certificates and/or church records
		CNS	Census records
		LNG	Longitudinal data (population registers or comparable sources)
		SG	Special Group study (semi-longitudinal data)
Main_source_comb	Sophistication of field *Main_source,* in case a second source can also be seen as basic.	CC CC-CNS CNS CNS-CC LNG LNG-CC SG SG-CC SG-CNS	Combined values of first and second basic source of the database, see foregoing field *Basic_source* for the meaning of the different values.
IDS_implemented	IDS stands for the output format of the Intermediate Database Structure (Alter & Mandemakers, 2014).	y	Databases are characterized as IDS when data are partly or fully converted into the IDS format or in the process of being converted.
		n	No IDS format.
Contemporary_linked	Linked with actual national registrations (or a dataset reaching until at least the year 2000).	y	Linked with modern datasets or actual registrations.
		n	Not linked with contemporary data.
Intergenerationally_linked	Database is linked over three or more generations.	y	Linked over generations.
		n	Not linked over generations.

The following fields describe the sources of the database: a value "x" in a field means that the database contains information from the mentioned sources, except the field *Specific_register* which entails short descriptions of the specific register.

Civil_records	Database includes civil certificates or church records.	This may imply that the database is based on these sources.
Pop_registers	Database includes population registers or comparable sources.	This may imply that the database is based on these sources.
Specific_register	Database includes data from a specific register.	Name of specific registers, which may be registers of enslaved persons, conscripts, hospital registers etc., this could imply that the database is based on these sources.
Census	Database includes census records.	This may imply that the database is based on these sources.
Death_causes	Databases includes cause of death registrations.	
Tax_register	Database includes data from tax registers.	
Conscription_records	Database includes conscription registers.	

Name field	Explanation	Values and/or more general remarks
Land_registers	Database includes information about real estate.	This may be registers of type, quality, taxable wealth, size, etc. of real estate (e.g., kadaster registration).
More_sources	Database includes data from all kind of other sources.	

The following fields describe the content of the database, in case of empty fields no figure was published or could be estimated.

Year_data	Year measurement data.	In case of inconsistencies values with the youngest date were given priority.
Persons_sampled	Number of persons sampled.	This is relevant in case a specific sample laid the foundation of the data gathering.
Person_observations	Number of person observations.	The total of all observations of a person in the included sources.
Marriages	Number of marriages.	Total number of reported marriages (difference with families is not always clear).
Family_observations	Number of family observations.	The total of all observations of a family in the included sources (difference with households is not always clear).
Household_observations	Number of household observations.	The total of all observationsof a household in the included sources (difference with families is not always clear).
Unique_persons	Number of unique persons.	As reported by the database.
Unique_families	Number of unique families.	As reported by the database.
Unique_households	Number of unique households.	As reported by the database.
Est_person_observations	Estimated or reported number of all person observations.	Estimated, in case the number is not reported by the database (the field *Person_observations is* empty).
Est_unique_persons	Estimated or reported number of unique persons.	Estimated, in case the number is not reported by the database (the field *Unique_persons* is empty).
Remarks	More specific information about the database, releases etc.	
References	References to sources of the data in this record and relevant literature.	

A.3 LIMITED OVERVIEW

Id	Name	Year_ founding	Region	Country	Observation_ period	Main_ source	Est_person_ observations	Est_ unique_ persons	References
1	Antwerp COR-database	2005	Antwerp district	Belgium	1846–1920	LNG	285,456	33,583	EHPS_Q; Jenkinson et al., 2020; Puschmann et al., 2022
2	Aranjuez Database: Individual and family trajectories	1995*	Aranjuez	Spain	1871–1970	CC	301,233	68,462	EHPS_Q; Reher & Sanz-Gimeno, 2007
3	BALSAC Population Database	1971	Québec	Canada	1621–1992	CC	11,672,172	2,652,766	EHPS_Q; Vézina & Bournival, 2020
4	Base Tra Patrimone (originally: Enquête 3000 Familles)	1980	France	France	1800–1960	CC	730,000	165,909	EHPS_Q; Bourdieu et al., 2014; Dupaqier & Kessler, 1992
5	China Multigenerational Panel Database-Liaoning	1982	Liaoning	China	1749–1909	CNS	1,513,312	266,091	EHPS_Q; Dong et al., 2015; Campbell & Lee, 2020
6	China Multigenerational Panel Database-Shuangcheng	2004	Shuangcheng	China	1866–1913	CNS	1,346,826	107,551	EHPS_Q; Dong et al., 2015; Campbell & Lee, 2020
7	Female Demographic Biographies: Wald Parish	2000*	Wald parish	Austria	1880–1938	SG	7,315	1,161	EHPS_Q; Teibenbacher, 2010
8	Founders and Survivors (Linked datasets)	2008	Tasmania	Australia	1803–1930	SG	1,250,000	200,000	EHPS_Q; Cowley et al., 2021
9	Geneva Urban Sociodemographic Database	2003	Geneva	Switzerland	1800–1880	CNS	55,936	40,000	Oris, Perroux, et al., 2023
10	Historical Database of the Liège Region	1970*	Liège region	Belgium	1806–1900	LNG	170,000	20,000	EHPS_Q; van de Walle, 1976; Alter, 1988; Alter et al., 2004
11	Historical Population Database of Transylvania	2014	Transylvania	Rumania	1850–1914	CC	570,036	129,554	EHPS_Q; Dumănescu et al., 2022
12	Historical Sample of the Netherlands	1991	Netherlands	Netherlands	1812–2022	LNG	7,900,000	1,200,000	EHPS_Q; Mandemakers & Kok, 2020
13	Hungarian Historical Demographic Database	2000*	Szentegyházasfalva and Kápolnásfalva (Eastern Transylvania)	Hungary	1800–1945	CC	189,200	43,000	EHPS_Q; Pakot, 2014
14	Integral History Project Groningen	1987	Province of Groningen	Netherlands	1770–1914	CC	110,000	25,000	EHPS_Q; Paping & Sevdalakis, 2022
15	Italian Historical Population Database	1971	Casalguidi (Tuscany) and Madregolo (Emilia)	Italy	1791–1883	CC	165,000	17,000	EHPS_Q; Breschi et al., 2020

Id	Name	Year_ founding	Region	Country	Observation_ period	Main_ source	Est_person_ observations	Est_ unique_ persons	References
16	Karelian Database	1988	Karelia	Finland	1681–1950	LNG	11,000,000	1,294,118	EHPS_Q; Saarti et al., 2017; https://katiha.kansallisarkisto.fi/tietoa.php, retrieved July 10, 2024
17	Koori Health Database	2010	Victoria	Australia	1870–1930	SG	49,770	7,900	EHPS_Q; McCalman, 2021
18	Melbourne Lying-In Hospital Cohort	1998	Melbourne	Australia	1857–1985	SG	67,000	38,000	EHPS_Q; McCalman, 2021; McCalman et al., 2011
19	Norwegian Historical Population Register	1976	Norway	Norway	1801–1964	CC	87,000,000	19,772,727	EHPS_Q; Sommerseth & Thorvaldsen, 2022; Thorvaldsen & Holden, 2023
20	Odense database: Persons and buildings	1980*	Odense	Denmark	1741–1921	CC	500,000	100,000	EHPS_Q; see also https://www.odensedatabasen.dk/om, retrieved July 16, 2024
21	POPLINK Demographic Database Umeå (DDB)	1973	Skellefteå region	Sweden	1900–1950	LNG	4,408,049	518,594	EHPS_Q; Edvinsson & Engberg 2020; Vikström et al., 2023
22	POPUM Demographic Database Umeå (DDB)	1973	Six different regions	Sweden	1620–1900	LNG	9,823,680	1,155,727	EHPS_Q; Edvinsson & Engberg 2020; Vikström et al., 2023
23	Portuguese Genealogical Repository	1970*	Several regions Portugal, main region Minho	Portugal	1550–1910	CC	1,000,000	227,273	EHPS_Q; Amorim & de Matos, 2016
24	Registre de la population du Québec Ancien (RPQA)	1966	Québec	Canada	1621–1799	CC	2,600,000	483,193	Dillon et al., 2018, 2023
25	Scanian Economic Demographic Database	1983	Scania	Sweden	1646–2015	LNG	1,488,766	175,149	EHPS_Q; Bengtsson & Dribe, 2021; Dribe & Quaranta, 2020
26	Texas Counties Database	1992	Texas	USA	1850–1910	CNS	150,095	38,000	EHPS_Q; Gutmann, 2016; Gutmann & Fliess, 1993
27	Scottish Census Enumerators' Books: Skye, Kilmarnock, Rothiemay and Torthorwald, 1861–1901	1998*	Isle of Skye (7 parishes), Kilmarnock and two small mainland parishes	Scotland	1861–1901	CC	631,000	341,000	EHPS_Q; Reid et al., 2002; Garrett et al., 2007; Garret & Reid, 2015
28	The Roteman Database	1978	Stockholm	Sweden	1878–1926	LNG	4,000,000	1,000,000	EHPS_Q; Fogelvik, 1989; Geschwind & Fogelvik, 2000
29	Utah Population Database	1973	Utah	USA	1790–2022	CC	275,000,000	11,000,000	Smith & Mineau, 2021; Smith et al., 2022

Id	Name	Year_ founding	Region	Country	Observation_ period	Main_ source	Est_person_ observations	Est_ unique_ persons	References
30	Taiwan Historical Household Registers Database	1985	Taiwan 16 regions	Taiwan	1906–1945	LNG	3,139,670	369,373	Lin et al., 2020
31	Barcelona Historical Marriage Database	2011	Barcelona region, 250 parishes	Spain	1451–1905	CC	528,000	120,000	Pujadas-Mora et al., 2022
32	South African Families Database	2010	South-Africa	South-Africa	1650–1949	SG	2,769,530	439,608	Cilliers, 2021
33	Xavier Database of Japan	1960	Three villages and one town current Fukushima prefecture	Japan	1708–1870	LNG	346,700	28,105	Hayami, 2016; Kurosu et al., 2021
34	Korean Historical Archives Visualization Network Database	2010	Danseong and Daegu-bu regions	Korea	1606–1876	CNS	1,756,965	532,414	Paek et al., 2022; Park & Lee, 2008
35	Ural Population Project	1997	Ural and Polar region	Russia	1857–1919	CC	325,000	73,864	Glavatskaya et al., 2022
36	LINKing System for historical family reconstruction (LINKS)	2006	Netherlands	Netherlands	1812–1973	CC	120,000,000	27,272,727	Mandemakers & Kok, 2020; Mandemakers, Bloothooft, et al., 2023
37	Finnish Army in World War II Database	2021	Finland	Finland	1897–1980	SG	26,794	4,253	Taskinen, 2023
38	Venice Family Database	1988*	Venice	Italy	1850–1869	LNG	30,267	3,561	Derosas, 1989, 1999
39	Scottish Historic Population Platform (SHiPP)	2012	Scotland	Scotland	1855–1973	CC	90,000,000	18,000,000	Akgün et al., 2019; https://digitisingscotland.ac.uk/ retrieved July 22, 2024 (a little bit out of date), new website: https://www.scadr.ac.uk/our-research/shipp (but not much information)
40	Baix Llobregat Demographic Database (BALL)	2017	Barcelona region	Spain	1820–1965	LNG	263,786	31,034	Pujadas-Mora et al., 2022
41	China Government Employee Datasets-Qing (CGED-Q) Jinshenlu (JSL) and Examination Records (ER)	2010	China	China	1644–1911	SG	4,433,600	329,851	Campbell & Chen, 2022
42	Founders and Survivors (Ships cohort)	2008	Tasmania	Australia	1818–1853	SG	157,500	25,000	McCalman, 2021
43	Diggers to veterans	2015	Victoria	Australia	1900–2020	SG	75,474	11,980	McCalman, 2021
44	China Multigenerational Panel Database-Imperial Lineage	1990	China	China	1644–1933	SG	1,575,000	250,000	Campbell & Lee, 2020

Id	Name	Year_ founding	Region	Country	Observation_ period	Main_ source	Est_person_ observations	Est_ unique_ persons	References
45	Enquete Louis Henry	1959	France (39 parishes)	France	1640–1829	CC	1,060,000	190,000	Séguy, 2001, 2016
46	English population history from family reconstitution 1580–1837	1964	England (26 parishes)	United Kingdom	1580–1837	CC	2,200,000	500,000	Alter et al., 2020; Wrigley et al., 1997
47	Íslendingabók (Genealogy Iceland)/ deCODE genetics	1988	Iceland	Iceland	740–2024	CC	4,360,400	991,000	Dillon et al., 2018; Garðarsdóttir, 2016; Gudbjartsson et al., 2015; https://www.islendingabok.is/english, retrieved August 4, 2024
48	IPUMS USA Multigenerational Longitudinal Panel (MLP)	1991	USA	USA	1850–1950	CNS	710,000,000	341,000,000	Ruggles, 2014; Helgertz et al., 2022; https://usa.ipums.org/usa/mlp/mlp_data_description.shtml
49	I-CeM/IPUMS Britain	2011	England and Wales	United Kingdom	1851–1921	CNS	230,000,000	109,523,810	Higgs et al., 2013; Longley et al., 2022; Schürer, 2007; Schürer & Higgs, 2020
50	Historical Database Suriname and the Caribbean	2018	Suriname and Caribbean	Suriname and Caribbean	1828–1921	LNG	775,000	210,000	van Galen et al., 2023; Information by mail database manager July 24, 2024; https://www.ru.nl/en/research/research-projects/historical-database-suriname-and-the-caribbean#:~:text=The%20Historical%20Database%20Suriname%20%26%20the,the%20Caribbean%20and%20the%20Netherlands
51	L'enquête Charleville	2007	Charleville	France	1739–1876	CNS	161,479	48,933	Alexandre et al., 2022; Boudjaaba et al., 2010; Minvielle, 2013
52	Gothenburg's Population Panel (GOPP)	2010	Gothenburg	Sweden	1915–1943	CNS	18,605	6,153	Karlsson & Lundh, 2015
53	Canadian Historical Mobility Project	1975*	Central Ontario	Canada	1869–1879	CNS	74,000	16,000	Darroch & Ornstein, 1984; https://international.ipums.org/international/resources/enum_materials_pdf/sample_design_ca1871a.pdf

Id	Name	Year_ founding	Region	Country	Observation_ period	Main_ source	Est_person_ observations	Est_ unique_ persons	References
54	SweCens database	1995*	Sweden	Sweden	1860–1940	CNS	32,600,000	10,866,667	https://swedpop.se/databases/, retrieved July 14, 2024; Wisselgren et al., 2014
55	Valserine Valley Genealogical Database	1985*	Valserine (Jura)	France	1700–2000	CC	184,000	47,037	Bideau & Brunet, 1998; Brunet et al., 2006
56	Danish Demographic Database	2001	Denmark	Denmark	1787–1940	CNS	20,000,000	6,666,667	Clausen, 2015; Robinson et al., 2022; https://link-lives.dk/en/about-link-lives/; https://www.ddd.dda.dk/kiplink_en.htm, retrieved July 16, 2024
57	The Longitudinal Intergenerational Family Electronic Micro-Database (LIFE-M)	2015*	Ohio and North-Carolina	USA	1841–1968	CC	42,000,000	15,000,000	Bailey, Lin, et al., 2023; Bailey, Leonard, et al., 2023; https://life-m.org/data/
58	Early Indicators (EI) project	1992	Northern States	USA	1850–1910	SG	250,000	50,000	Wimmer, 2003; https://www.nber.org/programs-projects/projects-and-centers/Early%20Indicators%20of%20Later%20Work%20Levels%2C%20Disease%20and%20Death?page=1&perPage=50
59	CenSoc project (linking census 1940 with Social Security records)	2015*	USA	USA	1940–2005	SG	19,700,000	8,000,000	Breen et al., 2023
60	Canadian Family Project	2018	Canada	Canada	1852–1921	CNS	40,000,000	19,047,619	Baskerville & Inwood, 2020; Foxcroft et al., 2022; https://thecanadianpeoples.com/
61	Ferrie database (Males 1850–1860 census data)	1992*	USA	USA	1850–1860	CNS	20,000	9,524	Ferrie, 1996
62	Fourteen German village populations (Knodel data)	1974	Germany (14 parishes)	Germany	1700–1925	CC	338,800	77,000	Knodel, 1988
63	Four generations Andover	1963*	Andover, Massachusets	USA	1630–1799	CC	15,400	3,500	Greven, 1972, 2023
64	Families of Newtown	1970*	Newtown, New York	USA	1642–1790	CC	6,600	1,500	Kross, 1974, 2016; Political and Social Research [distributor], June 22, 2016; https://doi.org/10.3886/ICPSR35005.v2

Id	Name	Year_ founding	Region	Country	Observation_ period	Main_ source	Est_person_ observations	Est_ unique_ persons	References
65	A Place in Time, Middlesex County Virginia	1970*	Middlesex and Lancaster Counties, Virginia	USA	1650–1750	CC	12,445	6,586	Darrett & Rutman, 1984a, 1984b, 2016
66	Quaker Families in 18th Century America	1968	New York, New Jersey and Pennsylvania	USA	1650–1830	SG	13,129	2,084	Wells, 1969, 1971
67	Connecticut Valley Historical Demography Project	1965*	Deerfield, Massachusetts	USA	1776–1850	CC	21,696	4,931	Meindl, 1979; Swedlund, 2008; Hautaniemi et al., 2000
68	Krummhörn Region Dataset	1985*	Krummhörn region (East Frisia)	Germany	1720–1874	CC	250,000	100,000	Voland, 2000; Willführ & Störmer, 2015
69	Neckerhausen Research Database (NRDB)	1967	Neckerhausen (Württemberg)	Germany	1560–1870	CC	80,000	11,000	Sabean, 1990; Sabean & Ball, 2024; figures communicated by Roii Ball, August 19, 2024
70	German Family Database 1700–1890	1975*	Berlin area, OstFriesland, Hartum, Ortenau, Saarland, Herrenberg	Germany	1700–1890	CC	1,260,000	149,000	Parts of the dataset are used by a. o. Imhof, 1998; Gehrmann, 1984; Voland, 1988
71	Corsier-sur-Vevey Database	2015*	Region Corsier-sur-Vevey includes Vevey and four villages: Corsier, Corseaux, Chardonne et Jongny	Switzerland	1700–1840	CC	60,000	26,000	Rappo, 2022
72	Korea Multi-Generational Panel Dataset–Tansung	1995*	Tansung County	Korea	1678–1888	CNS	36,5000	136,690	Dong et al., 2015; Park & Lee, 2008
73	Socface (Census France 1836–1936)	2020	France	France	1836–1936	CNS	540,000,000	90,000,000	Boillet et al., 2024; see also https://socface.site.ined.fr/en, retrieved November 10, 2024; Kesztenbaum, 2025
74	Family History of the region Wittgenstein. A genealogical database.	1982	Wittgenstein, Nordrhein-Westfalen	Germany	1525–1875	CC	1,050,000	150,000	Mehldau, 2011; see also https://de.wikipedia.org/wiki/Wittgensteiner_Familiendatei, retrieved October 26, 2024

Id	Name	Year_ founding	Region	Country	Observation_ period	Main_ source	Est_person_ observations	Est_ unique_ persons	References
75	Würzburg Database	1978	Würzburg, Bayern	Germany	1696–1711	CNS	43,143	13,074	de Vries, forthcoming
76	Population Censuses of Paris	2020	Paris	France	1870–1946	CNS	25,400,000	8,500,000	Brée et al., forthcoming
77	Liverpool Jewry Historical Database	2020	Liverpool	England	1745–1881	SG	30,000	7,140	Sapiro, 2024a, 2024b
78	Danish West Indies Panel (DWI panel)	2018	St Croix	Danish West Indies	1760–1914	CNS	1,415,019	428,794	Galli et al., 2024; Rönnbäck et al., 2024
79	Poznań Historical Population Database	2021	Poznań	Poland	1830–1900	CC	88,5000	201,136	Liczbińska et al., forthcoming; see also https://poznandatabase.pl, retrieved February 23, 2025
80	Cape of Good Hope Panel	1975*	South-Africa	South-Africa	1663–1844	SG	760,000	70,000	https://www.capepanel.org/wp-content/uploads/2024/04/0-Standardised-Variables-Descriptions.pdf; Fourie & Green, 2018; Fourie et al., 2025; Rijpma et al., 2020
81	Ipswich Fertility and Mortality Database	2001	Ipswich	England	1871–1901	CC	420,000	220,000	Garrett et al., 2007; Garrett & Reid, 2022
82	Uppsala Birth Cohort Multigenerational Study (UBCoS)	2004	Upsala	Sweden	1915–2009	SG	882,000	140,000	Koupil, 2007; https://www.su.se/english/research/research-projects/uppsala-birth-cohort-multigenerational-study-ubcos-multigen?open-collapse-boxes=research-project-description,research-project-more-about, retrieved July 22, 2025
83	The Madrid Historical Longitudinal Database	2005	Madrid	Spain	1888–1929	CC	3,500,000	2,022,000	Figures communicated by Diego Ramiro-Fariñas, July 30, 2025; del Bosque González et al., 2010; Oris, Mazzoni, et al, 2023; https://idehistoricamadrid.csic.es/
84	Bologna Parish Population Database (BPPD)	2023	Bologna and region	Italy	1650–1899	CC	1,700,000	600,000	Figures communicated by Michail Raftakis, August 4, 2025; Raftakis, 2025; Raftakis et al., 2025

A.4 EXPLANATION

The following offers more explanation and gives examples about the kind of decisions that had to be made in constructing the overview. For some exemplary databases they also explain in more detail how data were estimated.

4 The number of person observations in the TRA database (Enquête 3000 Familles) is estimated through the fixed multipliers as explained in Table 3 and the number of sources: 47,100 marriages * 6 = 282,600 persons + 13,000 extra known families * 2 = 26,000 parents + 125,000 deaths * 3,5 = 420,000 + 46,000 wealth records makes a total of about 730,000 person observations.

10 In the case of the Historical Database of the Liège Region it was only possible to estimate the numbers of person observations on the number of inhabitants, number of deaths and the sample frequencies (ranging for each parish from 20 to 100%) and the assumed mutual proportions.

12 The Historical Sample of the Netherlands (HSN) is an example of a complicated longitudinal database of which the core sources are civil certificates covering the period 1812–1970 and the population register from 1850 until the present, divided in four main sources: a) population registers until about 1910/1920, replaced by b) family cards until 1939, c) person cards until 1994 and d) person lists until present. The basic sample consists of 85,500 birth certificates, each with five persons (child, mother, father and two witnesses), supplemented with about 40,000 death certificates and 35,000 marriage certificates. Excluding the witnesses there are about 0.8 million person observations of which 0.55 million unique persons. The population registers and family cards were entered for about 44,000 sampled persons, which adds up to 216,000 registers including 1.5 million person observations with about 2 million person mutations (occupation, civil status, migration etc.) and 350,000 changes of household address, all to be counted as person observations. The total estimate is 5 million person observations. This stands for 0.6 million unique persons of which about 250,000 are also included in the certificates. Adding the data of the person cards and lists we have an extra 50,000 unique persons and 250,000 person observations. All in all 0.95 million unique persons with 5.25 million observations. Projects adding extra data of brothers, fathers, children, migrants, etc. have probably multiplied these figures with respectively 25% and 50%, which gives total estimates of respectively 1.2 and 7.9 million.

15 Regarding the Italian Historical Population Database it was not always clear how far databases had developed. The Questionnaire provides information on two parishes: Casalguidi (Tuscany) and Madregolo (Emilia), but the technical article only describes Casalguidi for a shorter period. For both parishes 17,000 individuals are reported who had lived at least one year in the villages and 4,500 marriages. For Casalguidi information on 8,015 unique individuals was retrieved from the *Status Animarum* (an annual nominative population and household register) over the period 1819–1859 totaling 96,581 person observations. Linking with tax registers delivered another 18,557 observations and including the births, marriages and death records another 35,000 person observations (using the standard calculation). For Madregolo we have probably only marriages adding roughly 12,000 person observations. All in all 165,000 person observations were estimated.

16 The Karelian database has not been linked yet, so there are only observations. Since they provide parts of life courses, it is considered as a longitudinal database of which the number of unique persons were estimated on the basis of person observations.

17 Of the Koori Health Database we only know that 7,900 unique persons are included in the database. Although it is clear that there are many observations from many sources, no figures are presented, so they had to be estimated in a general way which will probably be an understatement.

18 The Melbourne health cohort includes 16,290 cases of which 8,602 cases could be linked with a death certificate. For about 40% of the linked cases no marriage record of the mother was available. The number of observations for the linked cases may be estimated at 8,602 * 2 (hospital) + 5,602 * 3 (marriage record only parents and infant counted) + 3,000 marriage

records of their own (6 person observations) and 8,602 death records (3.3 person observations per unique person, a little lower than the norm of Table 3), the unlinked cases count 32,580 person observations. Total sum counts about 67,000 observations with 38,000 unique persons. For the sample size we took a number of 8,602 cases, in line with the figures of the database itself.

19 In the near future the Norwegian database will have 87 million records (censuses plus church records). Although presently only 10.7 million records have been entered (1865–1920), we include the potential figures in the table, since we may expect that the complete data entry will be finished within the next two or three years.

20 The Utah Family Database has many links to other datasets. These external ones have not been counted in our overview. Counting this database is not an easy matter. Its own total counts 52,024,217 records plus more than 200,000,000 "All payer claims data". It seems that the first number only counts sources. This implies that for some sources the number of person observations should be higher. We only calculated the extra observations for the 7 million birth, marriage and death records with an average of 4, resulting in 21 million extra person observations, bringing the total to 275 million.

27 The number of unique persons from the certificates is estimated using the fixed multipliers described in Table 3, resulting in 141,750 persons. The number of observations is calculated as follows: 62,500 birth certificates * 3 = 187,500; 12,300 marriages * 6 = 73,800; and 40,500 deaths * 3.5 = 141,750, giving a total of about 403,050 person observations. The database was linked with five censuses containing about 228,000 persons, making an overall total of 631,000. Eighty percent of the individuals in the censuses were linked to the civil records, leaving 20% as unlinked unique persons. This amounts to about 340,000 unique persons.

32 In case of the Barcelona marriage database we estimated the numbers of unique persons under the condition that only the bride and groom are unique records, except for the first generation when also the parents are counted (for each marriage line).

35 The Ural and Polar databases entails 23 parishes of which 15 in Ekaterinenburg; the data entry of the Polar censuses is very limited, so it is not included. The numbers of observations and unique persons was calculated with the norms as described in Table 3, totaling 321,585 with some tax records makes a total of 325,000.

48 IPUMS USA (IPUMS MLP project) presents a table with the number of links between the different censuses. The total number of 10-year links between two censuses sums up to 251 million which stands for 502 million person observations. The 251 million linked persons are not the same as unique persons because they will link over longer periods. There are 109 million persons linked over three census years, this group consist of two 20-year groups, so half of them must be distracted from the number of 251 million, which makes 196 million linked persons. And there are 57 million linked over four census years, which makes three 20-year groups, which implies that 2/3 of 57 million must be distracted which leaves 158 million unique persons. And we may suppose that substantial parts of the group that linked over 40 years will also link over 50, 60 years etc. A conservative estimation of 20 million for 40 years and 12 million for 50 years, brings this number of unique persons further down to 158 – 15 – 10 = 133 million persons. In total there are about 710 million person observations available for the period 1850–1950 of which 502 million are linked, leaving 208 million unlinked. Linked and unlinked persons together add up to 341 million unique persons. Although IPUMS has other data as well, they are not counted here, because they are not linked or only on a very small scale. Anyway, for the magnitude of this database it will not make a difference.

50 Database Surinam and the Caribbean — the registers with enslaved persons from both areas count in total 185,721 person observations from about 104,212 unique persons. From the birth and death certificates we have calculated (see multiplier in Table 3) respectively 189,720 and 264,397 person observations. The emancipation registers entail 35,631 person observations. Population registers of Paramaribo include another 100,260 observations. Total person observations add up to about 775,000. Calculating unique persons is more complicated but considering that a lot of persons have been observed more than once in all these sources the number of unique persons was estimated at 210,000.

56 The Denmark Linked Lives project aims to link all historical persons over the period 1787–1968, in two parts: a public part for 1787–1901 and a non-public one for 1901–1968. All of the public data are more or less entered, totaling about 20 million person observations.

57 The Life-M database linking the civil certificates of Ohio and North Carolina with the census of 1940, counts about 15 million unique persons (15.3 million minus a small portion belong to multiple generations). The number of observations has been estimated by counting 4 persons for each of the 10 million certificates (the database counts 5 million links between generations which makes 10 million certificates), which adds up to 40 million person observations. About 1 million persons are linked to the census 1940 and there also 185,000 linked death certificates. The total estimate is set at 42 million person observations.

58 What comes to the surface very clearly in the Early Indicators (EI) project is that in case a person from a sample, here the soldiers from the Union Army, is linked with census records quite a lot of unique for the relevant studies not so relevant persons are added to the database. For this reason they are not included in the table.

59 The CenSoc project links the America census 1940 with the Social Security records (4.7 million links) and the Numident records (7 million links). Linking these two indices of deaths with the same census will create overlap. For this reason the total number of unique persons was estimated at 8 million. Then the link with the census will provide 8 million person observations, together with the 11.7 linked death records makes a total of 19.7 person observations. Given the specific character the type of the database was characterized as a "Special Group".

68 The Krummhörn Region Dataset comprises 34,708 marriages, 80,840 births and tax registers. Given the source (Ortsfamilienbücher) the number of unique persons is already brought back to 151,000. Together with other linked sources total observations is estimated at 250,000. The number of unique persons must be lower than 151,000 because of the intergenerational links in the dataset (which will repeat children as parents); the estimation is put at a maximum of 90,000 unique persons.

69 The Neckerhausen comprises 2,437 marriages, 8,746 births and 6,533 burials. Based on parents from birth certificates and marriages there are 3,400 reconstituted families. With the standard criterium for church records this adds up to 64,000 person observations. Adding other sources brings this estimate to 80,000. The number of unique persons is more complicated, because the families are linked over multiple generations across the entire period 1560–1870. If we count all births and half of the marriages we arrive at 11,000 unique persons.

70 For the estimation of the figures of the German family database the ratios derived from the dataset of Giessen made by Rolf Gehrmann was used. For Giessen it was estimated that 9,158 persons stand for 7,102 unique persons. For the total database this implied that the 192,401 included person observations stand for 149,000 unique persons. For Giessen, using the multipliers of Table 3, it could be estimated that the number of person observations amounted to 3 * 9,000 = 27,000 from birth certificates, 9,000 from families (both parents) and about 3,5 * 9,000 = 31,500 from death certificates. All in all conservatively added up to 60,000 person observations. Using the known ratio of 1:21 for unique persons, the total for the German family database results in 1.26 million person observations.

72 The Korean Tansung dataset is based on linked observations from 3-yearly censuses. The number of persons with only one observation is 52.4%, especially because of gaps in the sequence of the registers. Although it is a 3-yearly census, because of this large proportion of not linked cases, the number of person observations was conservatively estimate with a multiplier which was only a little higher than the one used for IPUMS (see before); 136,690 * 2.67 makes 365,000 person observations.

73 Socface is the big census project of France and based on data entry through handwriting recognition and automatic record linkage. The data are from 20 5-yearly censuses in the period 1836–1936. Based on French population counts for the census years the potential number of person observations was estimated at a total of 720 million (the population increased from 35 million in 1836 until 40 million during 1900–1936). But about 25% of the census forms are lacking, including those of Paris (see POPP database, number 76); reason to diminish

the number of person observation with 25%, resulting in 540 million observations [Lack of 25% communicated by Lionel Kesztenbaum (2025)].

74 The Wittgenstein database counts 150,000 unique persons from the period 1525–1875. With the known multipliers (Table 3) and given the number of 140,000 birth records, 35,000 marriage records and 120,000 death records we arrive at an estimation of 1,050,000 person observations.

76 The Population censuses of Paris (POPP project) include the censuses of 1926, 1931, and 1936 for Paris; the 1946 census is planned to be included as well. Before 1926, no nominal census forms survived. The database is not part of the SOCFACE project (see explanation at number 73). In the near future, a database containing all marriage certificates (M-POPP) from Greater Paris for the period 1870–1940 will be added. The censuses from 1926 to 1946 cover about 11 million persons, of which — based on the linkage rates between 1926 and 1936 — around 8 million are unique individuals (though, so far, only adults have been linked). Including the 2.4 million marriage certificates from M-POPP will add another 2.4 * 6 = 14.4 million person observations. The number of unique persons will be much lower, given interlinkage and overlap between the certificates themselves as well as with the census data. We estimate about 500,000 additional unique persons, bringing the total to approximately 8.5 million.

77 The Liverpool Jewry Historical Database is originally based on burial registers and supplemented with birth and marriage records. Other sources were Jewish registers and the census 1841–1881. The database includes another 3,000 not systematically followed persons from 1881 and later, Data format is GEDCOM. The number of observations has been estimated at 30,000 of which 15,000 from the five censuses.

78 From the DWI documentation It is not clear how many unique persons were reconstructed through linking. The sample is too complicated to estimate this amount in a comparable way with other databases. The censuses 1846–1911 counted on average 23,000 persons per year. Supposing a complete turnover every 25 years, the unique number of persons will amount to 75,000. Doubling this number for the period before and the number of unique persons will be about 150,000.

79 The number of person observations in the Poznań database is estimated through the fixed multipliers as explained in Table 3 and the number of sources: 157,671 certificates of birth * 3 = 473,013 persons + 17,832 marriages * 6 = 106,992 persons + 87,261 deaths * 3.5 = 305,413 persons makes a total of about 885,000 person observations. As far as known, the database is not linked yet, so the number of unique individuals was estimated.

80 The database Cape of Good Hope Panel includes yearly tax rolls from the period 1663–1844. It contains the names of householders (landowners) and their wives. Other recorded data include the number of sons, daughters, enslaved persons, and servants, differentiated into 27 categories. By the end of 2024, about 257,837 records had been entered, including 246,791 male observations and 130,980 female observations. We estimate that this represents about 60% of all rolls (Figure 1 in Rijpma et al., 2020). This implies a total of approximately 400,000 male observations and 260,000 female observations. Given an average observation period of eight years (based on Figure 1 in Fourie et al., 2025), we estimate around 70,000 unique persons and about 40,000 unique households.

81 The Ipswich Fertility and Mortality Database is archived at UK Data Service. The numbers in the description cover the period 1871–1910; the archived data itself only 1871–1901. Because the last part covers the linked data the figures in the table were calculated from the archived dataset.

83 The Madrid Historical Longitudinal Database is based on birth and death records. The project began with death and burial records, which are almost complete for the period 1888–1927, totaling 615,846 records so far. In a second stage, all birth records for the period 1900–1926 were included and linked with the death records. Currently, it contains 337,882 birth records, which will soon total about 450,000. The death records contain only the deaths, so the number of person observations is equal to the number of certificates. With a multiplier of 3 we add 1.35 million person observations to the deaths, rounding to a total of 2 million. With an

estimated linkage rate of 40% for births (only relatively young deaths are linked at this stage), there will be 2 million – 180,000 = 1.8 million unique persons. From the 1905 census, all individuals from part of the districts of Madrid were added (about 196,000), and from the 1920 census, about 173,000 so far. We roughly estimate that 60% of these will not be linked with the civil records. This brings the estimated total person observations to about 2.37 million and the number of unique persons to 2.021 million. All persons are geocoded and linked with spatial data for 1900 and 1929 from the Historical Spatial Data Infrastructure of Madrid (HISDI-MAD), which also counts as a person observation. This raises the estimated total number of observations to about 3.5 million.

84 The Bologna Parish Population Database (BPPD) is still under development. When complete, it will consist of the records of 31 selected parishes from Bologna and the surrounding region for the period 1700–1899 (with some parishes starting as early as 1650), as well as the *Status Animarum* (an annual nominative population and household register). The database administrators estimate that it will contain 200,000 births, 74,000 marriages, and 180,000 deaths. Applying the fixed multipliers described in Table 3, this amounts to around 1.7 million person observations.The number of unique persons depends heavily on migration patterns, which are not yet known, but is conservatively estimated at around 600,000. This figure is based on 200,000 births, 100,000 parents, and 25% of the death and marriage certificates involving unique persons, adding roughly 300,000 more unique individuals. It is still too early in the development of the database to make even a rough estimate of the number of person observations that will come from the Status Animarum.