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

By George Alter, Kees Mandemakers and Hélène Vézina

To cite this article: Alter, G., Mandemakers, K., & Vézina, H. (2023). Introduction: Content, Design and Structure of Major Databases with Historical Longitudinal Population Data. *Historical Life Course Studies*, *13*, 228–234. https://doi.org/10.51964/hlcs15759

HISTORICAL LIFE COURSE STUDIES

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

VOLUME 13, SPECIAL ISSUE 5

GUEST EDITORS
George Alter
Kees Mandemakers
Hélène Vézina

EHPS NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

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

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

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

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

hislives@kuleuven.be

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

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

HISTORICAL LIFE COURSE STUDIES
VOLUME 13 (2023), published 17-07-2023

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

George Alter

University of Michigan

Kees Mandemakers

International Institute of Social History, Amsterdam & Erasmus University Rotterdam

Hélène Vézina

Université du Québec à Chicoutimi

ABSTRACT

In recent years the development of historical databases reconstructing the lives of large populations accelerated. These considerable investments of time and money have greatly expanded possibilities for new research in history, demography, sociology, economics, and other disciplines. This special issue describes the content and design of 23 important historical databases. Authors were given the freedom to discuss a range of practical and technical decisions from evaluating archival sources to crowdsourcing data entry. The most common issue is nominative record linkage, but we find different choices between semi-automatic and fully automatic linkage techniques and various approaches for connecting diverse sources. Some databases describe special problems, like linking Chinese names, handwritten text recognition or the construction of a release in IDS-format. Other databases offer detailed descriptions of sources or discuss prospects for including new datasets.

Keywords: Historical demography, Historical microdata, Life course, Social science history, Record linkage, Standardization historical data, Longitudinal research

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs15759

1 AIMS AND CONTENT

Over the last 60 years several major historical databases with reconstructed life courses of large populations have been launched. 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. At the annual meeting of the Social Science History Association in Montréal in 2017, the session "Development of Major Databases and their Results from the beginning till now" brought together presentations from some of the largest and most well-established databases with life course data, databases that have also been at the forefront of the development in this field. We were well aware in 2017 of numerous additional databases that had been established around the world in recent decades. In his valedictory speech Kees Mandemakers (2023) made an inventory of a total of 54 databases and even this compilation is not exhaustive.

In order to collect, organize, and then publish information on these major databases in a single collection, invitations were first sent to the leaders of about 25 of these databases. We received in most cases positive and enthusiastic reactions and, when the leaders of a database declined cooperation, it was mostly due to time constraints. We had no specific selection criteria, except that databases had to be actively used and maintained and the primary purpose of the database had to be the (re)construction of individual-level historical life courses. Archived databases, like the Louis Henry dataset (Séguy, 2001), were therefore excluded. Following the first round of invitations, others joined the collective endeavour, expanding the geographic coverage of our collection. We are now very pleased to present contributions representing 24 databases in two special issues of *Historical Life Course Studies.* The number and diversity of databases represented here is truly impressive!

Our overall strategy of describing these major databases resulted in creating two separate special issues. One, *Major Databases with Historical Longitudinal Population Data: Development, Impact and Results,* edited by Sören Edvinsson, Kees Mandemakers and Ken Smith (2023), deals with how the databases contributed to discoveries, responded to changing research questions and facilitated the development of novel lines of inquiry in historical demography and related fields. The present issue focuses on the technical and organizational aspects of these databases, such as their origins and evolution, content and database designs, as well as any setbacks and dependence on external funding. Some recently developed databases appear in this issue with information about both their impact and technical aspects, and several of the impact articles included technical information that is not repeated in this issue. The Chinese database consists of five datasets, three of which are described in the impact issue and two in the technical issue.

This special issue presents 23 articles, including seven databases with counterparts in the impact issue. The six databases appearing in the impact one are the Historical Chinese Micro Database, the Demographic Database of Umeå, the Utah Population Database, the Scanian Economic Demographic Database of Lund (SEDD), the Norwegian Historical Micro Database and the Antwerp COR*-database. The Historical Sample of the Netherlands (HSN), covered in the first published article of the impact issue, does not have a technical counterpart, but basic information about the HSN can be found in that article and the article in this special issue on the LINKS database, which is an offshoot of the HSN.

The 24 databases described in these issues represent a larger number of datasets. Whether the database is described as one or many datasets depends primarily on the strategy of the database managers. In general, when data from multiple sources are linked and integrated, the database is considered one big dataset. This is usually the result of a strategy that extended a core database with other data. This could be the result of systematic planning like the Umeå database, a result of the possibilities of crowd sourcing like the Tasmanian database, the result of funding by researchers to extend the dataset with specific data like the HSN, or a combination of these approaches.

The technical issue allows authors freedom to discuss a wide range of issues. The most common issue is nominative record linkage, but we find different choices between semi-automatic and fully automatic linkage techniques (LINKS, BALSAC, SEDD) and different approaches for connecting diverse sources (Utah, Tasmania). Some contributions describe special problems, like linking Chinese names, handwritten text recognition (Barcelona), and the construction of a release in IDS-format. Other contributions offer detailed descriptions of sources (Taiwan, Korea, Finland, Suriname) or discuss prospects for including new datasets.

2 EARLIER INITIATIVES

This is by far the most comprehensive but not the only collection of technical descriptions of databases with micro-data on historical populations. Twenty years ago, an overview of historical databases, the *Handbook of International Historical Microdata for Population Research* (Kelly Hall, McCaa, & Thorvaldsen, 2000), was published by the Minnesota Population Center. It included 16 databases of which five are included in this special issue. These are the databases from Norway, Sweden (Umeå), the Netherlands (HSN), Italy and Canada (PRDH). Most of the databases are not included here, because they concentrate on relatively modern 20th century census data. Four databases with historical census data were not included because they have not linked their data into life courses (like the UK dataset and Stockholm dataset), had no time to participate (Denmark), or are already described in an extensive way, like the IPUMS database initiated by Steven Ruggles (Helgertz et al., 2022; Roberts et al., 2003; Ruggles, 2014; Sobek et al., 2011).

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. The group was formed at the 1998 Social Science History Association conference by a group of researchers who desired discussions focused on the problems and potential of micro-data. The IMAG group concentrated on census data gathered by the IPUMS group, and the first IMAG workshop was hosted by the University of Ottawa in 1999 (Dillon, 2000). In November 2003, a second IMAG workshop in Montréal 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 & Roberts, 2006).

Since these first initiatives, cooperation between historical micro-databases intensified enormously. In May 2001, the International Institute of Social History organized a workshop on the results and practices of large databases, resulting in an overview of best practices for these databases (Mandemakers & Dillon, 2004). This was followed by a full day session on "New sources for historical demographic research" with four panels, organized by the International Commission for Historical Demography at the World History Conference in Sydney in 2005.

In March 2006 a second workshop, "Disseminating and Analyzing Longitudinal Historical Data", took place at IISH Amsterdam. Although participants recognized the complex nature of longitudinal databases, the workshop ended with a consensus on how to make progress. First, it was agreed that standardization in the products of the different databases would help researchers enormously. Second, an intermediate data structure (IDS) was proposed to mediate between the original databases and the data sets required for analysis. On 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 meeting in Miami, October 2008 (Alter & Mandemakers, 2014; Alter, Mandemakers, & Gutmann, 2009). Part of the 2006 workshop was an initiative to publish questionnaires with key information about the databases the participants were representing. This included the Historical Database of the Liège Region (Belgium), Scania Database (Sweden), Registre de la population du Québec ancien (PRDH), Historical Sample of Flanders, Demographic Database Umeå (Sweden), Victorian Scotland database, Connecticut River Valley Project (USA), Texas Longitudinal Data Project (USA), Migration Database (USA, based on genealogies), Danjuro Database Japan, Historical Sample of the Netherlands (HSN), Koori Health Research Database (KHRD) 1855–1930 (Australia), Melbourne Lying-In Hospital Cohort: 1857–1900, Utah Population Database, Geneva Database, IPUMS database (census USA), Norwegian Census Database.

The initial IDS working group was succeeded 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, a Marie Curie Innovative Training Networks project from 2017 to 2020 on "Methodologies and Data mining techniques for the analysis of Big Data based on Longitudinal Population and Epidemiological Registers".

One of the spin-offs of the EHPS-network is the journal *Historical Life Course Studies* which is publishing these two special issues on the technical aspects and impacts of large historical population databases. Both are the result of the already mentioned session on the "Development of Major Databases and Their Results from the Beginning till Now" at the Social Science History Conference in Montréal 2017,

which had presentations from the Utah Population Database (UPDB), the Historical Sample of the Netherlands (HSN), the Demographic Database of Umeå (DDB) and the BALSAC database of the Université du Québec à Chicoutimi.

The database questionnaires collected for the IISH conference in 2006 were taken over by the EHPS-network. The number of participating database increased to 32 (https://ehps-net.eu/databases), and much more detail was added. This special issue offers a new overview by providing an opportunity to update and expand information on many of the databases previously described and by including presentations on recent initiatives. Fourteen of these 29 databases are included in the technical or impact special issue. Four databases were not included because their authors could not produce an article (TRA database France, Denmark, Liège & Verviers region, Isle of Skye which became a subset of Digitizing Scotland). Census data projects that have not been linked into life courses, like MOSAIC and the Canada Family project, are also omitted. Eight databases are missing, because contact with the database managers has been lost or the database was judged too marginal for inclusion in these special issues (see Mandemakers (2023) for estimates of the numbers of persons and households/families included). However, we include a number of databases that did not appear in the questionnaires. Databases developed since 2006 describe Chinese Officials, the military in Finland, slavery in Suriname, family reconstitutions in the Netherlands, Barcelona, Taiwan, Korea, South-Africa, and the Urals. We also have articles on older datasets not described in the questionnaires, like databases on Utah, Tasmania, China, and the Xavier collection in Japan.

3 TYPOLOGY OF INCLUDED DATABASES

Life course databases may be divided into three types: (I) longitudinal data, (II) family reconstitutions, and (III) semi-longitudinal data. The first type, 'pure' longitudinal data, are based on sources like population registers that record continuous observation of vital events (births, marriages, and deaths) as well as migration, such as HSN, SEDD Lund, DDB Umeå, and the Antwerp COR*-database. In contrast, family reconstitutions, such as LINKS, must be analyzed under the restrictive rules developed by Louis Henry, because they only include vital events (Fleury & Henry, 1956, 1985; Henry, 1970; Henry & Blum, 1988). Semi-longitudinal databases (e.g., Utah, Norway and China) combine vital registration with censuses, taxes, and other nominative lists that identify the population under observation (see Alter, 2019).

We can also divide databases by geographic coverage (Mandemakers, 2023). Most historical databases cover only selected communities or specific areas within a country (e.g., Barcelona, Geneva, Québec, Tasmania, Utah). Several databases are national in scope, such as the HSN, LINKS, and the Norwegian database, but national databases may focus on specific cohorts, like government officials in China, babies born in a charity hospital, aboriginals or deported convicts (Australia), and Finnish and Australian military recruits.

The following table presents an overview of all 32 databases/datasets in our two special issues. We see that five of the included databases are typical of family reconstitution, and nine are based on fully longitudinal sources. We classify a majority of the databases as "semi-longitudinal", because they use censuses or other nominative lists, often in combination with vital registration. Within this group of semi-longitudal datasets we distinguish a subgroup of five datasets having only data from linked censuses. Nationwide coverage is available in five databases, 19 databases cover only part of a country, and eight are nationwide but only include a special cohort.

Table 1 *Overview of all included databases*

Nature_basic	Coverage	Name	Country	Technical	Impact
Longitudinal	Nationwide	Historical Sample of the Netherlands	Netherlands		X
Longitudinal	Nationwide	Historical Databae of Suriname	Suriname	X	
Longitudinal	Regional	Antwerp COR*-database	Belgium	X	X
Longitudinal	Regional	Baix Llobregat Demographic Database (BALL)	Spain	X	
Longitudinal	Regional	POPLINK DDB Umeå	Sweden	X	X
Longitudinal	Regional	POPUM DDB Umeå	Sweden	X	X
Longitudinal	Regional	Scanian Economic Demographic Database	Sweden	X	X
Longitudinal	Regional	Taiwan Historical Household Registers Database	Taiwan	X	
Longitudinal	Regional	Xavier Database of Japan	Japan	X	
Family Reconstitution	Nationwide	LINKS	Netherlands	X	X
Family Reconstitution	Regional	Registre de la Population du Québec Ancien (RPQA)	Canada	X	
Family Reconstitution	Regional	Historical Population Database of Transylvania	Rumania	X	
Family Reconstitution	Regional	Ural Population Project	Russia	X	
Family Reconstitution	Regional	BALSAC Population Database	Canada	X	
Semi-longitudinal	Nationwide	South African Families Database	South-Africa	X	
Semi-longitudinal	Regional	Utah Population Database	USA	X	X
Semi-longitudinal	Regional	Barcelona Historical Marriage Database	Spain	X	
Semi-longitudinal	Regional	Integral History Project Groningen	Netherlands	X	
Semi-longitudinal	Regional	Italian Historical Population Database	Italy	X	
Semi-longitudinal	Special cohort	China Government Employee Datasets-Qing (CGED-Q) Jinshenlu (JSL) and Examination Records (ER)	China	X	
Semi-longitudinal	Special cohort	China Multigenerational Panel Database-Imperial Lineage	China		X
Semi-longitudinal	Special cohort	Diggers to Veterans	Australia	X	
Semi-longitudinal	Special cohort	Finnish Army in World War II Database	Finland	X	
Semi-longitudinal	Special cohort	Founders and Survivors (Linked datasets)	Australia	X	
Semi-longitudinal	Special cohort	Founders and Survivors (Ships cohort)	Australia	X	
Semi-longitudinal	Special cohort	Koori Health Database	Australia	X	
Semi-longitudinal	Special cohort	Melbourne Lying-In Hospital Cohort	Australia	X	
Semi-longitudinal Census	Nationwide	Norwegian Historical Population Register, 1800–1964	Norway	X	X
Semi-longitudinal Census	Regional	China Multigenerational Panel Database-Liaoning	China		X
Semi-longitudinal Census	Regional	China Multigenerational Panel Database-Shuangcheng	China		X
Semi-longitudinal Census	Regional	Geneva Demographic Database	Switzerland	X	
Semi-longitudinal Census	Regional	Korean Historical Archives Visualization Network Database	Korea	X	

Each database represented in this issue has its own unique genesis that is well described in the various papers. For example, the launch of the DDB at Umeå University was initially motivated by an interest in the development of literacy. For the Utah Population Database, the impetus was the value of genealogies and family histories for genetics and medicine. At the same time, several common elements and developmental arcs connect these distinct databases. In many respects, the scientific relevance for the development of historical population databases rests on the shoulders of giants who championed quantitative history and the history of the ordinary person. This includes members of the Cambridge Group for the History of Population and Social Structure (Wrigley, Davies, Oeppen, & Schofield, 1997), the Annales School with its advocacy of social history (Séguy, 2016), and the proponents of the life course perspective arguing for the plasticity of human development and the role of history (Kok, 2007). With these intellectual foundations as bedrock, technological advances proved to be a catalyst for accelerating the insights of quantitative history by digitizing archival records and through record linking methodologies that reveal the diversity of human life courses.

The origins of historical longitudinal databases vary, but six in our issue originate from the 1970's when computers and software first facilitated data entry, processing and database management. These are the Xavier database of Japan, the Demographic Database Umeå, the Utah Population Database and the Norwegian Historical Data Centre in Tromsö, as well as the two databases about Québec (the Registre de la population du Québec ancien (Université de Montréal) and the BALSAC database (Université du Québec à Chicoutimi)). These projects created a legacy through numerous publications, large numbers of trainees, and the development of stable and reliable infrastructures. In the following two decades, they were followed by new databases like the SEDD database in Sweden, the Chinese datasets, and the HSN database in the Netherlands. Although databases continue to be added in western Europe (e.g., LINKS, Antwerp COR*-database, and Barcelona), the most impressive expansion has been in other parts of the world, including Asia, Australia, South Africa, and Eastern Europe. The technical special issue describes new challenges encountered in these areas as well as the opportunities offered by new technologies, like machine learning and natural language processing. While the expansion of these infrastructures is impressive and benefits the research community broadly, significant portions of the globe are not yet represented, largely due to lack of resources needed to create and maintain complex databases.

REFERENCES

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

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

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

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

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* (XI–XIV). Minneapolis: Minnesota Population Center. Retrieved from https://international.ipums.org/international/microdata_handbook.shtml

Dillon, L. Y., & Roberts, R. (2006). Introduction: Longitudinal and cross-sectional historical data: Intersections and opportunities. *History and Computing, 14*(1–2), 1–7.

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]. Paris: Editions de de l'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 old civil status] (3rd ed.). Paris: Institut National d'Etudes Démographiques.

Kelly Hall, P., McCaa, R. & Thorvaldsen, G. (Eds.). (2000). *Handbook of International Historical Microdata for Population Research.* Minneapolis: Minnesota Population Center. Retrieved from https://international.ipums.org/international/microdata_handbook.shtml

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. doi: 10.1080/01615440.2021.1985027

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

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

Kok, J. (2007). Principles and prospects of the life course paradigm. *Annales de démographie historique, 113*(1), 203–230. doi: 10.3917/adh.113.0203

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

Mandemakers, K. & Dillon, L. (2004). Best practices with large databases on historical populations. *Historical Methods, 37*(1), 34–38. doi: 10.3200/HMTS.37.1.34-38

Roberts, E., Ruggles, S., Dillon, L. Y., Gardarsdóttir, Ó., Oldervoll, J., Thorvaldsen, G., & Woollard, M. (2003). The North Atlantic Population Project. An overview. *Historical Methods, 36*(2), 80–88. doi: 10.1080/01615440309601217

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

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]. Paris: 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). Bern: Peter Lang.

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, 44*(2), 61–68. doi: 10.1080/01615440.2011.564572

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