Strength in Numbers? A Short Note on the Past, Present and Future of Large Historical Databases

By Lionel Kesztenbaum

To cite this article: Kesztenbaum, L. (2021). Strength in Numbers. A Short Note on the Past, Present and Future of Large Historical Databases. *Historical Life Course Studies, 10*, 05-08. https://doi.org/10.51964/hlcs9557

HISTORICAL LIFE COURSE STUDIES

Not Like Everybody Else.
Essays in Honor of Kees Mandemakers

VOLUME 10, SPECIAL ISSUE 3
2021

GUEST EDITORS
Hilde Bras
Jan Kok
Richard L. Zijdeman

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 openjournals 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 https://openjournals.nl/index.php/hlcs.

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

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

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

HISTORICAL LIFE COURSE STUDIES
VOLUME 10, SPECIAL ISSUE 3 (2021), 05-08, published 31-03-2021

Strength in Numbers? A Short Note on the Past, Present and Future of Large Historical Databases

Lionel Kesztenbaum
Institut National d'Études Démographiques (INED) & Paris School of Economics

ABSTRACT

Historical demography is inherently associated with constructing large-scale databases from historical records. Although there have been tremendous changes in the way they are constructed, many of the challenges remain. Throughout his career, Kees Mandemakers has been instrumental in facing some of these challenges, particularly those related to the conservation, standardization, and dissemination of databases. This short contribution discusses the evolution of large historical databases in historical demography.

Keywords: Historical demography, Large-scale historical databases, Sources, Cumulative datasets

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs9557
The article can be downloaded from here.

1 DATABASES IN HISTORICAL DEMOGRAPHY

Large historical databases have been at the center of historical demography since the very beginning, when Louis Henry and then the Cambridge group collected parish records on a large scale. In many ways, the very idea of historical demography is linked to the development — recollection, aggregation, combination, etc. — of large-scale databases from historical material.

While the technology, design, and methods of these databases have changed tremendously since Louis Henry, many of the issues arising when constructing such databases and the choices faced by historical demographers remain the same, e.g. representativeness and sampling design; the degree and level of standardization of the information contained in the sources; the degree of selection from a given source; interpreting the original material. One could argue that the precision and subtlety needed to understand past societies — in line with the discussions initiated by, among many others, microhistory, gender history, and global history — have made the task of constituting and analyzing large-scale historical databases more difficult. This short contribution discusses some of these issues.

2 THE THREE AGES OF HISTORICAL DEMOGRAPHY

In the infancy of historical demography, from the late 1950s to the early 1970s, databases were mostly drawn from parish (or civil) registers and the demographic events they record. These sources were a convenient tool for this young field's purposes, namely reconstituting and studying demographic behaviors of past populations, and they were easy enough to access in large numbers in many European countries. This first age of historical demography was Eurocentric and dominated by the issue of population dynamics, in particular the study of fertility, but its results and methods helped strengthen its place in the social sciences.

As the field expanded, new research questions stimulated a thirst for more complex and elaborated datasets and innovative methods for analyzing them. For instance, the life cycle perspective has become key for analyzing demographic practices — today or in the past — and the role of social and economic factors that influence them. Kees Mandemakers was instrumental in this second age, as one of the promoters of the Historical Sample of the Netherlands (HSN), certainly emblematic of the large historical databases designed as multipurpose representative samples, i.e. a complete research infrastructure (Mandemakers, 2000). Other examples abound — the Programme de Recherche en Démographie Historique in Québec (PRDH), the China Multigenerational Panel Database-Liaoning in China (CMGPD-LN), the Scanian Economic Demographic Database in Sweden (SEDD), and many others throughout Europe, Asia, and America — demonstrating that this second age, the last third of the 20th century, was a golden age for historical demography.

Many of these projects are still ongoing but face new challenges as a third wave of databases, based on artificial intelligence, machine learning, and big data, is slowly taking prominence. This third age, roughly covering the past decade, bears witness to technical changes and improvements in the processing of large data — to expand studies on historical populations to previously uncharted, or at least neglected, territories, e.g. migration, social mobility, early life effects, etc. As a result, more and more datasets are based either on crowdsourcing or on machine learning — used for anything from automatic transcription of printed (OCR) or handwritten (HTR) sources to record linkage, through automated coding of variables.

These changes go together with two major shifts. The first is a higher degree of standardization of information, both a cause and a consequence of the increasing automatization of database creation. The source material is being processed according to standard scales impervious to the historical and spatial context. A case in point is that of occupations, which are more and more often transformed into a single number based on the same scale (e.g. HISCO). In addition, too often, what cannot be quantitatively measured or imputed is discarded from the data. The second is a change in the dominant sources, from vital registers to censuses. It results in part from the technological and financial changes that make complete-count censuses easier to process and access, but it also reflects the geopolitical evolution of historical demography (some would say of the social sciences as a whole), as the 'disciplinary' center of historical demography shifted from Europe to the United States. Because the latter does not have reliable vital event registration until the 20th century, censuses are the main source available to study demographic change on a large scale. Consequently, they have become the dominant source of contemporary historical demography.

Both changes have important consequences for observation, analysis, and interpretation of past societies. Impressive as they are for the number of individuals observed, censuses are usually weak on the type of information they gather. Too often, the large number of observations tends to conceal the weakness of the information they contain. Although researchers may find clever ways to exploit every piece of information from historical sources, it is difficult not to think that censuses, even with hundreds of millions of records, provide a rather limited picture of past societies.

3 BUILDING 'MATRICES' FOR CUMULATIVE RESEARCH

Both gains and losses mark the transitions between these ages. For instance, the move from Henry's *fiches de famille* to longitudinal individual databases greatly expanded the research questions and interests in historical demography but tended to promote a return to analyses based on local data rather than nationally representative databases, given how costly it is to accumulate longitudinal data. More damagingly, many of the databases built during the first age of historical demography were forgotten and some were lost. More recently, the development of big datasets based on machine learning has dramatically increased the size of the datasets, but it also has entailed an increasing disconnection between research and data construction. In many ways, data construction has become, or is becoming, an independent part of the work. One consequence is the potential reduction in interdisciplinarity and a stronger monopoly of technicality that would exclude not only historians but also many scholars trained primarily in the social sciences; indeed, for historical demography, interdisciplinarity is as crucial as it is vulnerable (Saito, 1996).

These changes involve potential losses of previous methods, databases, and knowledge, an issue beyond the scope of this brief paper. Here, we will simply give two examples of how we could face this challenge by building cumulative processes.

The first relates to the transmission and maintenance of old databases so that they can be used again, either alone or with new ones. Here, Kees has been instrumental in designing ways to standardize and, perhaps more importantly, help preserve and distribute past historical demography datasets (e.g. Alter, Mandemakers, & Gutmann, 2009). By applying the same structure to very different databases, the proposed Intermediate Data Structure (IDS) is an important tool to ensure easy access to existing databases.

Secondly, we must allow databases to be extended, expanded, reused and repurposed. This means building datasets that are themselves cumulative. This is exactly what the HSN has been doing, and still does. Built around a core of individuals identified and collected in standard historical demographic sources, it can be expanded at will by adding new sources — and, as a result, new information — on these individuals. Indeed, 'all possible sources in which individuals are named and have an address can be linked to the HSN'. (Mandemakers, 2000, p. 159). The same has been done in France, where the TRA survey was also designed to be able to combine as many sources as possible. It is built around a core sample of individuals whose surname starts with the letters TRA, selected in marriage and fiscal records. In this case, it is the last name that enables new sources and data to be added to the existing sample (Bourdieu, Kesztenbaum, & Postel-Vinay, 2014). In doing so, it can take advantage of the large variety of archival sources in France throughout the 19th century: election rolls, pension records, military records, marriage contracts, etc.

Both the HSN and the TRA survey illustrate the advantages and importance of smaller but carefully designed surveys that allows us to obtain rich data. Building a cumulative framework is thus a possible solution: the combination of tens of different sources on a few, carefully selected, individuals instead of having standardized, simple sources, such as censuses, for hundreds of millions of individuals. The key idea here is to link as many sources with as much information as possible to obtain small — in relation to current standards — but very rich samples. To understand past societies, it is worth the additional cost of collecting information from complex historical material. Richer and more detailed sources, such as those on which the HSN and TRA are based, should not be relegated to the background by the explosion of simple and standardized datasets such as those based on censuses. Both ways of building large historical databases could be complementary, the latter being used as a core sample for building more complex datasets related to specific research questions.

4 CONCLUDING REMARKS: LARGE HISTORICAL DATABASES, NOW AND THEN

In what would be an ideal world for historical demographers, automated methods would decipher and transcribe any source, whether printed, handwritten, or in another format (e.g. maps), transform them into gigantic databases linked together, and automatically code every piece of information they contain so it can be processed and analyzed by statistical tools. Whether this will happen soon is open for debate, but the fact that it has become part of the conversation on the evolution of the field must give historians, demographers, and other social scientists pause for thought. This possibility would not produce a perfect observatory of societies of the past, given that no society is simply a collection of individuals (see for instance the discussion in Prost, 2014), but these technical developments have major implications for the way historical research is done and, even more so, for the way it will be done in the near future.

We must consider how large historical datasets have evolved with the development of historical demography. What is the trade-off between data quantity and quality? Census data, such as those disseminated by NAPP or IPUMS, are the new standard; but they are relatively poor, as they give limited information on considerable numbers of individuals. In contrast, the 'matrix' databases, such as the HSN or the TRA, aim at aggregating as many sources as possible into a central and common 'core'. They highlight the advantage of combining sources for restrictive, but representative and cumulative, samples. Preserving this way of doing research in historical demography is certainly an important legacy of Kees Mandemakers.

ACKNOWLEDGEMENT

The author thanks Jan Kok for helpful comments and Christopher Leichtnam for his careful revision of the text. The usual disclaimer applies.

REFERENCES

Alter, G., & Mandemakers, K., & Gutmann, M. (2009). Defining and distributing longitudinal historical data in a general way through an intermediate structure. *Historical Social Research, 34*(1), 78–114. doi: 10.12759/hsr.34.2009.3.78-114

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

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. 149–177). Minneapolis: Minnesota Population Center.

Prost, A. (2014). Des registres aux structures sociales en France. Réflexions sur la méthode. *Le Mouvement Social, 246*(1), 97–117. doi: 10.3917/lms.246.0097

Saito, O. (1996). Historical demography: Achievements and prospects. *Population Studies, 50*(3), 537–553. doi: 10.1080/0032472031000149606