The Life Span of Large Historical Databases

By Jan Kok

To cite this article: Kok, J. (2021). The Life Span of Large Historical Databases. *Historical Life Course Studies, 10*, 19-23. https://doi.org/10.51964/hlcs9561

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), 19-23, published 31-03-2021

The Life Span of Large Historical Databases

Jan Kok
Radboud University Nijmegen

ABSTRACT

Large historical databases, although intended to last for a long time, can become obsolete for a variety of reasons. In this essay these reasons are explored and used for a 'health check' of the Historical Sample of the Netherlands (HSN). The HSN leaders are examined for their visionary qualities and their sense of ownership, and the database for its complementarity, versatility and consistency. The essay concludes that, despite challenges ahead, HSN is sound of mind and body.

Keywords: Large historical database, Social science history, Research dreams

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs9561
The article can be downloaded from here.

1 INTRODUCTION

In contrast to databases geared to short-term research projects, large scientific databases are meant to transcend specific research questions and even to last beyond the professional life of their designers. What is the key to a long and healthy life of a large database? In this essay, I search for the components that keep a database alive in the sense that it is continuously expanded and improved. Then, I will do a health check on the Historical Sample of the Netherlands (HSN) and make a prognosis of its life expectancy.

First of all, what defines a 'large historical database'? Obviously, it contains digitized information from historical sources. But what makes it 'large'? In his inaugural lecture, Kees Mandemakers (2009) offered some helpful criteria. In his view, just as important as the sheer size of the dataset are features such as a) several persons are working on it; b) it is embedded in an institution; c) it has a long-term strategy; d) it is used by multiple researchers and for different research purposes and e) it generally builds on a variety of sources. Acquiring funds and setting up such a project may be relatively easy, but how to make sure that the database does not perish in its infancy? An excellent guide for young parents of a database is provided by Mandemakers and Dillon (2004). They suggest a checklist of best practices regarding sampling, standardization, documentation and release versioning. Indeed, the history of historical databases is strewn with examples of short-lived projects that, for instance, had started data-entry immediately with typing labels or categories instead of preserving the original information as completely as possible (Kok & Wouters, 2012). The guidelines of Mandemakers and Dillon form a good medicine against childhood diseases. But do they also guarantee a long life?

2 WHAT DETERMINES THE LIFE SPAN?

A large historical database is not a physical organism and its lifespan is not determined by factors such as genetic make-up, hygiene or nutrition. Instead, we have to look at the interaction between the team responsible for the database (I will call them 'designers') and the research environment. In this interaction, I discern five main elements.

Vision. All databases begin with a 'research dream', a compelling vision and promise that the database will answer timely and specific research questions but can also deal with topics relevant for still unknown parties or handle issues that cannot even be articulated yet. To give an example: when Chad Gaffield proposed the Canadian Century Research Infrastructure, a database with factual and contextual information on historical censuses, he envisioned that 'the CCRI will raise questions about the making of twentieth-century Canada that we have not yet even imagined' (Gaffield, 2007, p. 63). Science and technology studies have shown that such promises not only function to attract funding and to stimulate agenda-setting, but also to build a 'protected space' in which the development and first results of the database can be discussed, e.g. on dedicated websites or in specialist journals (Kok & Wouters, 2012). Such grand visions are indispensable, but they come with two risks. The first one is that the developmental stage is too protracted and that sponsors and potential users loose interest. The second is that designers and first users remain in the 'protected space' too long. In other words, they fail to convey the database's results to the broader community of historians or to the public at large. The result is that the database is perceived as a 'hobby horse' only yielding answers to trivial questions.

Complementarity. A standard element in visions promoting a large historical database is that it fills a major gap in our knowledge, e.g. because of its unique covering of an area or period, or because it contains unique variables. However, there is always the possibility that the database is being overtaken by a new one in the same country or elsewhere. In the latter case, a researcher looking for 'universal' social processes may opt for the 'competition' because of a larger size (statistical significance) or higher user-friendliness. To secure its lasting relevance, the designers must put the database in a scientific ecosystem of related but different databases. Instead of fighting the competition, they must cooperate with other initiatives, by promoting comparative research, e.g. through international standards of the coding of occupations or causes of death. Another course of action is promoting interoperability, that is looking for data formats (such as Linked Open Data) that allow researchers to combine data from different large databases.

Ownership. Proper institutional embedding is vital for a large historical database. Clearly, an organization had to pledge itself to provide regular updates and thereby continuity of a database's accessibility, software, website, privacy regulations, citation rules et cetera. But 'ownership' goes beyond that. The team behind the database ideally consists of technicians, developers, archivists, researchers and visionaries. The team can respond to problems inherent to each database (e.g. errors, inconsistencies), work on new projects, add new variables, and — most important — keep the database relevant even when conditions and requirements for research are rapidly changing. To elaborate on the latter: the team should keep abreast of developments in digitization (such as handwriting recognition), in science itself (e.g. in genetic research) and in the research topics put on the agenda by the large funding agencies.

Consistency. Although already prominent among the best practices of Mandemakers and Dillon (2004), I would like to emphasize here the enormous importance of keeping all components of the database consistent with each other. Thus, consistency in sampling strategy (as far as possible), in harmonization and standardization of locations, occupations and family relations, in the definition of 'observation' of life course spells, in error detecting procedures and so on. The more oversampling projects, auxiliary data and data sources a large database contains, the larger the risk of inconsistencies. And these will lead to frustrated users and/or unwieldy documentation to account for variations and exceptions. This implies that the 'owners' should not only look forward, but also backward in keeping old releases up to date.

Versatility. In my opinion, the releases and accompanying documentation should be highly flexible, that is catering to the needs of researchers from different scientific backgrounds and with varying degrees of experience and skill. This calls for documentation and codebooks that make it easy to start working and experimenting, but that also provide proper warnings (e.g. in textboxes or hyperlinks) when variables need more explanations, for instance because they have different meanings in different periods or because they require more detailed knowledge on historical population administration. To attract new users, as well as users without previous experience, simplified versions of releases or demonstrators should be made available — again with adapted documentation.

In my view, these factors ensure ongoing interest in the database and its new installments. Interest from a new generation of researchers and/or new disciplines will generate a steady flow of grant proposals and appealing output, which in turn generates new interest as well as the funding to keep the database up to standard. How is HSN faring after thirty years of existence? Can we expect it to live much longer? Or should it be nudged gently in the direction of the data archive, to be downloaded when needed but no longer actively involved in new research plans and visions?

3 HSN 2.0?

Clearly, with thirty years of age HSN is in its prime. When in doubt, one needs only to glance at the *Annual Report* of 2019 which lists a (cumulated) total of more than 800 presentations and more than 400 publications based on HSN. Many of these publications and their main findings have been discussed in a recent overview (Mandemakers & Kok, 2020). However, we can hardly overlook the fact that the HSN is the life work of an extremely diligent and dedicated 'owner'/'designer', Kees Mandemakers. To remain in family terms, Kees Mandemakers was at the same time founding father, midwife and doting mother of HSN. Of course, he got help in no small amount from (sometimes numerous) staff, volunteers, the International Institute of Social History, and a large network of senior and junior scholars both in the Netherlands and abroad. Yet, he managed single-handedly to design software to enter population registers, do the bookkeeping of all projects, start negotiations between archivists and information scientists (on LINKS), acquire new funding, do his own research and present it on many occasions, and create networks and interfaces (e.g. the Intermediate Data Structure) across large historical databases. I could easily make this list of accomplishments spread over several pages. Clearly, Kees' retirement will necessitate all kinds of adjustments in how the HSN operates. Will it be possible to keep the HSN at the core of Dutch social science history, perhaps even to usher in a new era of HSN-based research (HSN 2.0)?

The HSN team is not lacking in *vision*. The grant proposals for major investments of the last decade such as 'Spanning Past and Present' (2007), 'Inequality in the Life Course' (2015) and 'The Rise and Fall of Equal Opportunities' (2017) follow the narrative structure of effective research dreams by pointing at gaps in knowledge (e.g. of 20th century cohorts too old to be analyzed through modern surveys and panels),

by hinting at widening research horizons through the addition of new variables (e.g. on income, height and education of HSN sampled persons) and by promising a vital contribution to solving pressing societal problems (e.g. increasing inequality). These, beautiful grant applications were not successful and introspection is needed on whether these visons are (still) really shared by the research community and funding agencies. However, the recently funded ODISSEI project does allow to add a second generation to the HSN sample, which will be an important building block towards, e.g., a multigenerational history of inequality.

The team supporting HSN could also do some extra work on stressing its *complementarity.* It appears to me that, somehow, the notion has arisen that LINKS — or the linking of all civil records in the Netherlands resulting in family trees, kinship networks as well as vital events during the life course — makes HSN obsolete. However, HSN remains unique and indispensable because of its day-to-day reconstruction of events, choices and outcomes of individuals within the setting of households. On the international level, it is to be hoped that the ecosystem that found a form in the European Historical Population Samples Network can be revived and even expanded on a truly global level. Such a network can lead researchers quickly to the (combination of) datasets suited for answering their questions.

For the time being, the *ownership* of HSN seems secured by the International Institute of Social History which has trusted several persons with the tasks associated with maintaining a large historical database. This will involve some growth pains for 'HSN 2.0' as personal differences regarding skills, priorities and ideals will have to be matched. The greatest challenge will to be sustain or even expand the interaction with the research community, especially with younger generations of researchers.

The new HSN team will face a challenge regarding *consistency:* a new (expanded and cleaned) release is urgently required, implying also that regional differences in the completeness of life courses are eliminated. Releases of subprojects may have to be checked for consistency as well. Fortunately, HSN has created important new tools enabling users to harmonize data on occupations and localities themselves. For instance, the current dataset on occupations consists of the standardized and coded values of 281,355 different occupational titles.

Similarly, *versatility* can be an issue. An integrated version of HSN (and the many subproject releases) in the Intermediate Data Structure is certainly needed. In such a structure, life course information from population registers, personal cards (after 1940), civil records, militia records and so on can be brought together in one unified system, which could also be put in a Resource Description Framework. However, simpler formats of HSN and demonstrators for students and beginners will also be required to cater for the needs of as many researchers as possible.

To conclude, the retirement of the HSN's *pater familias/alma mater* does not bode its imminent demise. Perhaps we may speak of a (mild) mid-life crisis. New (technical, institutional, organizational) ways need to be found to solve outstanding issues, to expand interest among (young) researchers, and to attract new funding to realize long-term strategies. Such strategies could involve placing the sample persons' residential histories in a much finer grid of geo-locations using the street addresses instead of just municipalities, adding entire second or third generations to the current database and linking with social statistical individual-level data (thus bridging the 20th century 'gap'), to expand into the former colonies, or to add the 18th century to the HSN. When ways are found to realize a smooth collaboration between researchers, managers and technicians involved in HSN, we can surely foresee a HSN 2.0.

REFERENCES

Gaffield, C. (2007). Conceptualizing and constructing the Canadian Century Research Infrastructure. *Historical Methods, 40*(2), 54–64. doi: 10.3200/HMTS.40.2.54-64

Kok, J., & Wouters, P. (2012). Virtual knowledge in family history: Visionary technologies, research dreams, and research agendas. In P. Wouters, A. Beaulieu, A. Scharnhorst & S. Wyatt (Eds.), *Virtual Knowledge. Experimenting in the humanities and social sciences* (pp. 219–250). Cambridge, MA/London: MIT Press.

Mandemakers, K. (2009, June 11). *Waarom Jan en Cor met elkaar trouwden. Over grote historische databestanden, koudwatervrees en interdisciplinaire samenwerking*. Inaugural lecture at the Faculty of History and Arts of the Erasmus University Rotterdam.

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

Mandemakers, K., & Kok, J. (2020). Dutch Lives. The Historical Sample of the Netherlands (1987–): Development and Research. *Historical Life Course Studies, Online first*. Retrieved from http://hdl.handle.net/10622/23526343-2020-0001?locatt=view:master