Historical Databases Now and in the Future

By Kris Inwood & Hamish Maxwell-Stewart

To cite this article: Inwood, K. & Maxwell-Stewart, H. (2021). Historical Databases Now and in the Future. *Historical Life Course Studies, 10*, 09-12. https://doi.org/10.51964/hlcs9558

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), 09-12, published 31-03-2021

Historical Databases Now and in the Future

Kris Inwood
University of Guelph

Hamish Maxwell-Stewart
University of New England

ABSTRACT

Kees Mandemakers has enriched historical databases in the Netherlands and internationally through the development of the Historical Sample of the Netherlands, the Intermediate Data Structure, a practical implementation of rule-based record linking (LINKS) and personal encouragement of high quality longitudinal data in a number of countries.

Keywords: Digitization, Longitudinal, Record linkage, Intermediate Data Structure, Data accessibility

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs9558
The article can be downloaded from here.

1 INTRODUCTION

As George Orwell once said, 'who controls the present controls the past'. In the world of historical demography this might be a good description for the role that Kees Mandemakers has played. It was Kees who provided much of the inspiration behind the development and wide acceptance of the Intermediate Data Structure for longitudinal data. It is no mean feat to get researchers from different institutions, nationalities and disciplines to follow the same rules in any context. Working with George Alter, Kees managed to persuade academics spread across the globe to adopt a common set of rules for organizing and maintaining the acceptability of diverse longitudinal data — the Intermediate Data Structure — a task that made herding cats look easy (Alter & Mandemakers, 2014).

Kees has championed standards in other ways too. We recall the many times he has railed against hand-linking records. Whereas there are numerous studies that have shown that while individuals can match patterns in ways that are difficult to train a computer to replicate, humans are not very good at applying rules consistently. Kees reminds us that in principle an algorithm can replicate any human decision to recognize (or not) a match, as is illustrated by his wonderful LINKS (Mandemakers & Laan, 2017; van Dijk & Mandemakers, 2018). An automated record linkage system has the additional advantage of documenting all decisions enabling them to be unpicked later if necessary.

Standards are important because datasets take many hundreds of hours to build. The product of our collective labours needs to be durable in order for that investment to pay dividends. For data to last, the ways in which they have been put together need to be readily discoverable.

Organized datasets are important in other ways too, they enable the results of research to be repeated — a gold standard for scholarly endeavors. They also provide researchers with the opportunity to compare outcomes across different societies. Finally, standards provide an opportunity for researchers to think big. As Kees has shown by example, it pays to have the audacity to think on a national and international scale. The pioneering Historical Sample of the Netherlands has an impact far beyond the boundaries of the Netherlands (Mandemakers & Kok, 2020).

In the rest of this short article we will outline some ways in which we think Kees's legacy might help to shape the future of historical data collection and the manner in which those collections are employed.

2 REFLECTIONS

Datasets are not just becoming larger, they are also becoming more complicated. In part this has been driven by the increasing availability of digitized data. Over the course of the last decade or so big data has become a humanities and social science reality. Very soon all or almost published work that is out of copyright will be available in machine-readable form. The huge expansion in family history has created incentives for commercial companies to digitize large volumes of records, including births, deaths and marriages, census data, criminal records, street directories, passenger lists and other record groups that name large numbers of people. The amount of digitized newspaper content also increases each year. Tens of thousands of birth, death and marriage notices are now available, for example, through the National Library of Australia's much acclaimed Trove project which has sought to make all Australian historical newspaper content available on-line. Scale, however, has brought with it a new set of challenges, as well as exciting opportunities.

Whereas it has been commonplace for historical datasets to consist entirely of records that were structured in ways that were highly compatible, this is no longer necessarily the case. While the longitudinal data we worked with in the past were usually constructed by linking birth, death and marriage records or joining decadal censuses together, it is now common for researchers to link information originally collected by different agencies and institutions. Military enlistment papers might, for example, be linked to birth, marriage and death records, as well as census returns, voter registration cards and even criminal records. These more complex data collections allow life courses to be followed in much greater detail although this is often achieved at a price. Every attempt to link to a new record group will inevitably bring with it a new set of selection challenges. Newspaper notices of deaths, for example, might provide researchers with important details but — as family members had to pay by the word — the length of each notice is likely to reflect available resources and hence social standing. The newly available digital source materials often provide a

great deal of extra detail, although that detail is shaped by selection processes that need to be documented, explored and carefully controlled for (Inwood & Maxwell-Stewart, 2020).

A lesson here is that, because large linked datasets can be complicated if not downright messy, it is particularly important to adhere to the standards in data collection, cleaning, linking and curation advocated so passionately be Kees Mandemakers. To do so might even open up the opportunity to provide new insights into old source materials — one of the goals of the Intermediate Data Structure. The ability to triangulate between different record groups also provides the opportunity to explore the composition of each. Differences in the way that occupations, levels of literacy and even age are reported in different contexts can be revealing.

The record sets of the future are likely to be linked directly to the archives and library collections from which they were derived. In part this is because transcriptions are no longer taken directly from the original hardcopy. Record groups are now first imaged and then transcriptions made from those images. This is useful in that it provides future users with a visual facsimile of the original, an important attribute if a user wants to verify a step in the transcription process. Indeed, many datasets now contain persistent links enabling researchers to connect at the click of a button with the archival source from which the data are derived.

While data continue to become more discoverable the teams that we work with are becoming bigger. Although typeset sources have for long been converted to machine readable text using optical character recognition software, it is still commonplace for the output from this process to be cleaned by hand. The same will be true of handwritten text recognition (HTR) outputs which are likely to require considerable levels of manual correction. At least for the foreseeable future HTR approaches will rely on large volumes of training data sourced by using more traditional harvesting techniques. Thus, the challenge of putting together big datasets is likely to continue to rest on the ability to key large volumes of information by hand. Historians of all types are banding together to rise to that challenge. Crowd sourcing platforms such as Zooniverse (https://www.zooniverse.org/) are providing ways of linking the research endeavors of those based inside the academy with family historians and other constituent groups. As these 'volunteer' historians want the fruits of their labour to be available widely and permanently, the increase in citizen history is likely to lead to more research databases becoming linked to archival access platforms.

Such data are likely to be used in other ways too. Tuition fees paid by undergraduate students commonly cross-subsidize and enable funding for research at many universities. In order to derive the necessary income to support future research indicatives, it is likely that we will need to fashion innovative teaching tools out of the datasets that we put together. Apart from anything else, this provides an opportunity to train a new generation of historians to appreciate the need for standards. It might even lead to the expansion in student numbers. Dedicated units aimed at the family history market have proved popular in many places for example. In turn these have helped to increase volunteer recruitment, creating what might be referred to as self-funding virtuous data circle.

In some places historical data has found other uses. As 19th-century prisons are repurposed as heritage sites and boutique accommodation, digitized historical records can be used to drive interpretation strategies or expand public engagement in other ways. A search for a convict or prison record can provide users with a visualization of that individual's life course. This in turn can be used to communicate information to a user about each site the subject of their search was detained in, including advice about how to visit that place. While such user interfaces may not directly contribute to historical research, they are likely to be highly important indirect contributors. The University of Tasmania, for example, recently topped the first Australian impact and engagement exercise in the History and Archaeology field, largely because of the way that researchers had been able to imaginatively repurpose their historical data.

Visualisations based on digitized archival information are likely to make increasing use of spatial data. While this might help to engage a wider public the creation of geo-locatable variables might also facilitate data linkage. Geo-referenced polygons can be used for example to connect micro-data to macro census returns for a particular district. They can also be useful in mapping boundary changes over time, enabling static series to be animated in ways that facilitate understandings of change over time. The linking of birth, death and marriage records to housing and other spatial data opens up a plethora of new research opportunities. This might include explorations of ways in which epidemics have spread or the manner in which characteristics of the household or community of birth might shape life expectancy. In order to get the most out of such complex data, however, historians may have to start to rethink underlying dataset architecture.

SQL databases have been at the core of attempts to use quantitative techniques to understand the past for at least four decades. One of their strong attributes is that they can be structured to represent many common forms of tabulated data. A drawback, however, is that modeling relationships across multiple linked tables can be time consuming and complex. This is particularly the case for queries that involve many to many relationships. Graph databases provide the opportunity to reconfigure data into a series of nodes and relationships that could inject a degree of flexibility into future historical datasets (Zhu, Yan, & Song, 2016). There is a danger, however, that this might also promote the indiscriminate mining of historical information without an adequate understanding of underlying data constraints. Again, a care and attention to processes and the adoption of common standards in marking up datasets so passionately advocated by Kees are likely to hold the discipline in good stead as it digests the implications of these new technologies.

3 CONCLUSION

The contributions of Kees Mandemakers cast a long shadow over all of the research areas identified above, and in the meeting rooms of the European Social Science History Association, the International Institute of Social History, the European Historical Population Samples Network and many other organizations. We will continue to benefit for some time from his quiet leadership, hard work and strategic good sense. Kees is the rare scholar whose personal research contributes in tangible ways to collective understandings and, in addition, his commitment to the building of institutions and standards will enable future research by many others.

REFERENCES

Alter, G., & Mandemakers, K. (2014). The Intermediate Data Structure (IDS) for longitudinal historical microdata, version 4. *Historical Life Course Studies, 1*, 1–26. Retrieved from http://hdl.handle.net/10622/23526343-2014-0001?locatt=view:master

Inwood, K., Maxwell-Stewart, H. (2020). Selection bias and social science history. *Social Science History, 44* (3), 411–416. doi: 10.1017/ssh.2020.18

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

Mandemakers, K., & Laan, F. (2017). *LINKS dataset Genes Germs and Resources. WieWasWie Zeeland. Civil Certificates. version 2017.01. IDS version* [Data file and code book]. Amsterdam: IISH.

van Dijk, I. K., & Mandemakers, K. (2018). Like mother, like daughter. Intergenerational transmission of infant mortality clustering in Zeeland, the Netherlands, 1833–1912. *Historical Life Course Studies, 7*, 28–46. Retrieved from http://hdl.handle.net/10622/23526343-2018-0003?locatt=view:master

Zhu, Y., Yan, E., & Song, I.-Y. (2016). The use of a graph-based system to improve bibliographic information retrieval: System design, implementation, and evaluation. *Journal of the Association for Information Science and Technology, 68*(2), 480–490. doi: 10.1002/asi.23677