A *Kees* Study on Nominal Record Linkage

By Gerrit Bloothooft

To cite this article: Bloothooft, G. (2021). A *Kees* Study on Nominal Record Linkage. *Historical Life Course Studies, 10*, 53-58. https://doi.org/10.51964/hlcs9567

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), 53-58, published 31-03-2021

A *Kees* Study on Nominal Record Linkage

Gerrit Bloothooft
Utrecht University

ABSTRACT

This paper describes a case study on nominal record linkage on data from the Mandemakers family. It is demonstrated how names from birth, marriage and death certificates can be used for fast, probabilistic, ego-based record linkage, with the help of year of birth to arrive at unique identification. The procedure includes name standardization to overcome variation in spelling and the use of probabilities of combinations of given names and surnames, computed from the digitized 19th century Dutch vital register.

Keywords: Nominal record linkage, Ego-based, Probabilitistic, 19th Century, The Netherlands

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs9567
The article can be downloaded from here.

1 INTRODUCTION

Prosopography studies require the identification of the people in the past, preferably as individual life courses. Since a persistent personal identification number did not exist until recently, we have to use other evidence to conclude that historical information refers to the same individual. Name and dates of birth, marriage and death are the most distinguishing information on people, and key in vital registration. It is only with admiration that we can observe the data structure the French developed for this in the Napoleonic era and which they imposed on the Netherlands in 1811. Currently, thanks to numerous volunteers, a significant part of the key data in the historic certificates on birth, marriage and death until about 1940 are online available in the WieWasWie corpus.[1] In population reconstruction the central question then is what evidence is sufficient to conclude that two attestations concern the same individual. This actually is a matter of probability: what is the likelihood that there is only a single person who underlies some set of information. With the current level of digitization of archival materials it is possible to make reasonable estimates on this matter. But still, population reconstruction should always be a dynamic process, where new data or insights may change part of the reconstruction.

2 AN EGO-BASED APPROACH

Here I want to discuss an ego-based method to arrive from the raw archival data to a population register from which life courses, pedigrees and family trees can be derived, for which the original thoughts already have been explored in Bloothooft (1995). This is demonstrated by a case study on a member of the Mandemakers family: Maurits, the *ego* in this story. From the online WieWasWie corpus, two birth certificates, a marriage certificates and a death certificate can be found featuring 'Maurits Mandemakers' (and many more certificates which are ignored for the sake of simplicity) for which we want to know whether these indicate the same person. The first issue is what each of these certificates teaches us about all the actors involved. The first birth certificate introduces us with Maurits, born in Capelle near Waalwijk on May 20, 1820 as the son of Willemijna Nieuwenhuijzen and Arie Mandemakers. This defines three actors, Maurits, Willemijna and Arie (for the moment we ignore the surnames), of which we know that Maurits has Willemijna and Arie as parents, while Willemijna and Arie are partners. From Willemijna as actor we know that she has the partner Arie, and reversely. Besides names and relations, we also know that Maurits is born in 1820 and therefore certainly will die before 1930, while he may marry between 1834 and 1920. These ranges are derived from rules on the absolute time range of life events (such as 14 years for the minimum age at child birth or marriage and 110 years for the maximum age), but may be refined by probability distributions. Also for the parents Willemijna and Arie ranges for birth, marriage and death can be computed. Unfortunately, the ages of the parents, 33 and 35 years respectively at the birth of son Arie, are mentioned in the paper certificate but not digitized in the WieWasWie corpus, and therefore the time ranges are much wider than needed. It is the consequence of the realistic consideration that it is too time consuming to index certificates in full. All results are shown in table 1. The same identification of actors, their relations, and time ranges for a marriage, death and another birth certificate related to Maurits is given in table 1 as well. The simplified data structure of this table is a subset of the implementation in the LINKS project (2018).

Table 1 shows all information we can derive from the four certificates, featuring 16 person mentions. We could hypothesize that these 16 person attestations concern 16 different individuals. This is not likely, but we should very explicitly understand why we may arrive at that conclusion. The base line of the approach is that we try to reduce the number of individuals needed to explain the data to a minimum. This is known from knowledge theory as Occam's razor. If two attestations of a person do not have any differences in the names of the person, his/her parents and partner (if mentioned), dates or any other entities taken into account, we may conclude that these concern the same person. This is not necessarily always true, however, since it depends on how much information is available. If we encounter just Maurits Mandemakers several times without age information nor relatives, this is not necessarily the same person, since three children Maurits Mandemakers were born in the 19th century. The result of Occam's razor is just a start, which can be refined by considering relations between identified individuals and the use of additional information not explored in the analysis (for instance the region where events took place). Using the principle of Occam's razor, all data in the four certificates in table 1 can be reduced to the seven persons in table 2, but we need a general procedure for that.

1 https://www.wiewaswie.nl/en

Table 1 *Person records derived from four certificates featuring 'Maurits Mandemakers'*

Certificate	ID	ID-ego	Full name	Date of birth	Date of decease	Date of marriage	ID	ID-ego	Father	ID	ID-ego	Mother	ID	ID-ego	Partner	^{10}log prob
Birth	1	1	**Maurits Mandemakers**	10-5-1820	<1930	>1834 <1920	2	2	Arie Mandemakers	3	3	Willemijna Nieuwenhuijzen				-14
	2	2	Arie Mandemakers	<1806 >1720	<1906 >1820	<1820 >1784							3	3	Willemijna Nieuwenhuijzen	-11
	3	3	Willemijna Nieuwenhuijzen	<1806 >1770	<1906 >1820	<1820 >1784							2	2	Arie Mandemakers	-11
Marriage	4	1	**Maurits Mandemakers**	10-5-1820	<1930 >1852	24-4-1852	6	2	Arie Mandemakers	7	3	Willemijna Nieuwenhuijzen	5	5	Catharina Hulst	-18
	5	5	Catharina Hulst	14-3-1820	<1930 >1852	24-4-1852	8	8	Willem Hulst	9	9	Teuntje Timmermans	4	1	Maurits Mandemakers	-16
	6	2	Arie Mandemakers	<1806 >1720	<1916 >1852	<1820 >1784							7	3	Willemijna Nieuwenhuijzen	-11
	7	3	Willemijna Nieuwenhuijzen	<1806 >1770	<1916 >1852	<1820 >1784							6	2	Arie Mandemakers	-11
	8	8	Willem Hulst	<1806 >1720	<1916 >1852	<1820 >1784							9	9	Teuntje Timmermans	-9
	9	9	Teuntje Timmermans	<1806 >1770	<1916 >1852	<1820 >1784							8	8	Willem Hulst	-9
Death	10	1	**Maurits Mandemakers**	<1821 >1819	12-1-1894	<1894 >1833	11	2	Arie Mandemakers	12	3	Willemijna Nieuwenhuijzen	13	5	Catharina Hulst	-18
	11	2	Arie Mandemakers	<1807 >1719	<1894 >1819	<1821 >1783							12	3	Willemijna Nieuwenhuijzen	-11
	12	3	Willemijna Nieuwenhuijzen	<1807 >1769	<1894 >1819	<1821 >1783							11	2	Arie Mandemakers	-11
	13	5	Catharina Hulst	<1887 >1753	<1894 >1833	<1894 >1833							10	1	Maurits Mandemakers	-12
Birth	14	14	Arie Mandemakers	10-2-1853	<1963 >10-2-1853	<1953 >1867	15	1	Maurits Mandemakers	16	5	Catharina Hulst				-14
	15	1	**Maurits Mandemakers**	<1839 >1753	<1949 >10-7-1852	<1853 >1803							16	5	Catharina Hulst	-12
	16	5	Catharina Hulst	<1839 >1789	<1949 >1853	<1853 >1803							15	1	Maurits Mandemakers	-12

Note: For each person mentioned in a certificate, all information on father, mother and partner is presented, with estimated date ranges for the own birth, decease and marriage, according to rules (in most cases here simplified to years). Person records each have a unique ID, while after linkage the ID-ego's (orange) of identified individuals are assigned. The ^{10}log of the total probability of all names is shown as well.

Table 2 *Identified individuals derived from table 1, with collapsed date ranges*

ID-ego	Ego	Date of birth	Date of decease	Date of Marriage	ID-ego-father	ID-ego-mother	ID-ego-partner
1	Maurits Mandemakers	10-5-1820	12-1-1894	24-4-1852	2	3	5
2	Arie Mandemakers	<1806 >1720	<1894 >1852	<1820 >1784			3
3	Willemijna Nieuwenhuijzen	<1806 >1770	<1894 >1852	<1820 >1784			2
5	Catharina Hulst	14-3-1820	<1894 >1853	24-4-1852	8	9	1
8	Willem Hulst	<1806 >1720	<1916 >1852	<1820 >1784			9
9	Teuntje Timmermans	<1806 >1770	<1916 >1852	<1820 >1784			8
14	Arie Mandemakers	10-2-1853	<1963 >10-2-1853	<1953 >1867	1	5	

For this, an ego-based approach is followed, that is: choose a person mention (ego) and try to find as many non-conflicting attestations for this ego. The total information derived from an attestation should guide this process. The best person mention to start with is the one with most related information, while preferably this information is rare. A given name 'Unico' would be a good start (as the name itself already indicates). But rare information may also imply erroneous spelling, aliases or nick names and abbreviations, which may seriously hamper ego-reconstruction. This is a difficult problem that can be partly solved by using a (semi-)phonetic form of a name which reduces spelling variation with equal pronunciation. A further reduction of variation is to assign a functional standard to a name. Functional because such a standard is useful for nominal linking but does not pretend to have any etymological or genealogical meaning. The development of functional standards has been pursued in the NAMES project (2019), in which 187,707 given name variants from the WieWasWie corpus (2011 version) are projected on 813 standards only, and 562,676 surnames on 19,016 standards. This implies that Mandemaker, Mandemakers, Mandenmaker, Mandenmakers, Mandemaaker and Mandemaakers all get the same standard MANDEMAKER, while for instance the names Catharina and Trijntje are both standardized to CATHARINA. Standardization may imply that from the genealogical point of view genuine differences are lost: Mandemakers and Mandemaker are said to be two different families (although both refer to the very old profession of basket maker). But by using other entities such as given names and surnames of related persons, high probability links can be made still, while avoiding mismatches because of irrelevant spelling differences or aliases such as Kees and Cornelis.

Our 2011 version of the WieWasWie corpus has 54 million surnames attestations, of which only 1,993 have the standard MANDEMAKER, while 11,458 out of 30 million male given name attestations concern the standard MAURITS. For every person record the total probability of occurrence of the standardized given names and surname involved can be computed on the basis of this kind of frequency information, given in table 2 as ^{10}log(probability). The lower this total probability, the more descriptive the record is. The number of names (the sum of given names and surnames) has the strongest influence on total probability, followed by the probabilities of the constituting names. In table 2 this implies that the record linkage procedure should start with ID=4, Maurits Mandemakers, for which 7 names are known from the marriage certificate (ego surname is by definition father surname), with the lowest total probability. Starting with ID=4 as ego-record, other records with the same ego name standard (MAURITS MANDEMAKER) are ID= 10 (7 names, death), ID=1 (5 names, birth), ID=15 (4 names, birth child). There is no conflict in any entity nor time range between these records and they all take ID=5 from the initial ego-record as ID-ego. Subsequently, the next best ego-record is for Catharina Hulst (CATHARINA HULST), ID=5 (7 names but higher total probability, marriage), with ID=13 (4 names, death husband) and ID=16 as records to be matched (birth child, 4 names each). And so on. Notice that females in the Netherlands keep their maiden name in all registrations.

Multiple marriages may complicate the role of a partner. It is best to distinguish the triple {ego, mother, father} with 3 given names and 2 surnames and the double {ego, partner}, with 2 given names and 2 surnames. The triple can be used to identify the birth, marriage(s) and death of ego, while the double can be used to link to ego's children (from various partners) and to the death of ego's partner. When four actors are known (as in

marriage and death certificates), this makes links stronger of course. With the focus on the 5 (standardized) names of the triple ego and parents, records concerning ego can be quickly found by sorting all records on names, for which the best sorting order is (1) surname ego, (2) given name ego, (3) surname mother, (4) given name mother, (5) given name father. The resulting set of records should be analysed further on time ranges of life events to distinguish between different ego's with the same name. Such an approach, tested in van Boheemen (2016), is substantially faster than any other method that uses the comparison of pairs of records.

3 UNIQUENESS IN WIDER PERSPECTIVE

In line with the above, it may be asked for what percentage of couples the combination of their names is unique. For this, 4.4 million marriage certificates from the WieWasWie 2019[1] corpus were analysed. These provided the names of bride and groom, while in 90% of the certificates their ages (converted to year of birth) were available. For a unique couple the combination of their given names, surnames and years of birth should only occur once. After ignoring 3% duplicate certificates (from different archives), and double registrations for marriage and divorce or a certificate correction for the same couple, nearly all couples proved to be unique, even when their names were standardized. Without using year of birth, 5,248 name combinations of bride and groom occurred twice or more (on top: Cornelis de Boer and Grietje de Boer, 4 couples). When using the initial names only, this number raised to 7,759 (on top: Peter Janssen and Maria Janssen, 4 couples), while converting these names to semi-phonetic forms resulted in 9,446 couples (on top: variants of Peter Janssen and Maria Janssen, 10 couples). With standardized name forms there were 33,766 (0.8%) non-unique couples out of the 4.28 million non-duplicate certificates, with the standardized combination JOHANNES JANSEN and JOHANNA JANSEN on top (32 couples). With the help of time ranges, four standardized names are very distinctive and will identify ego and partner or ego and mother almost always in the Netherlands, without needing geographic information of the places of events. Multiple given names and the given name of the father even make a stronger match. Since standardized names overcome spelling variation and aliases, this enormously facilitates the reconstruction of reliable life courses.

As mentioned before, not all person attestations can be used for matching. When the total number of names involved in linking is less than four there is a risk of non-unique matching, although rare names in combination with time ranges may still lead to a convincing link. Nevertheless, this is a grey area where true links are hard to establish. This also touches upon the issue whether it is possibly to identify a true link at all. Unfortunately, we do not have a gold standard for nominal record linkage, although the Historic Sample of the Netherlands (HSN) may be considered as such. This could be an option when we attempt to reconstruct the full historic population of the Netherlands after 1811 on the basis of the WieWasWie corpus, and to examine whether the subset of the population described in the HSN is replicated.

The matching criteria so far require matching of names (at original, semi-phonetic or standard level) and non-conflicting time ranges. However, if the number of entities for matching is high (when four actors are known for instance), the need for full agreement becomes less. In that case a few mismatches for some entities may be acceptable. This may originate in errors in the original registration or in digitization but also in erroneous standardization. The former may happen in death certificates where reporting neighbours did not know precisely the name of the former wife of a deceased widower. Or in cases where multiple given names are present, and some are missing or mixed up. This property of relaxed matching in the presence of sufficient information has been used by Bloothooft and Schraagen (2015) to gather spelling variants.

When a series of life events of ego has been identified, the life course of ego can be reconstructed. For our ego: *Maurits Mandemakers was born on May 10, 1820 as son of Willemijna Nieuwenhuijzen and Arie Mandemakers. At age 30, he married Catharina Hulst, and they got children among which the first son Arie in 1853. Maurits died in 1894 at age 73.* At this life course level it is still possible to find inconsistencies, such as a partner who dies *after* another marriage of ego. It is the interaction between life courses that is not explored in full in the first phase of an ego-based approach, as the death of someone else is not part of ego's reconstruction. Such conflicts require re-interpretation of data and re-allocation of ID-ego's, and iterative improvements.

4 CONCLUSIONS

The general lesson is that major steps in population reconstruction for the Netherlands can be made on the basis of reasonably reliable data (which is the case for the WieWasWie corpus), nominal standardization, probability computation and a fast, ego-based approach. Yet, life generates numerous rare appearances that ask for dedicated solutions which should not discourage researchers. On the contrary, these challenge for the development of even more flexible reconstruction methods, which could extend to pre-1811 data from parish registers which are much less reliable and complete. This asks for creativity and it may not be by accident that the Mandemakers family orginates from the area where also a World of Wonders resides.

REFERENCES

Bloothooft, G. (1995). Multi-source family reconstruction. *History and Computing, 7*(2), 90–103. doi: 10.3366/hac.1995.7.2.90

Bloothooft, G., & Schraagen, M. (2015). Learning name variants from inexact high-confidence matches. In G. Bloothooft, P. Christen, K. Mandemakers & M. Schraagen (Eds.), *Population Reconstruction* (pp. 61–83). Cham Heidelberg New York Dordrecht London: Springer. doi: 10.1007/978-3-319-19884-2

HSN website (n.d.). https://iisg.amsterdam/nl/hsn

LINKS project (2018). https://iisg.amsterdam/en/hsn/projects/links

NAMES project (2019). Data and manual downloadable from https://ivdnt.org/taalmaterialen/

van Boheemen, J. (2016). *Assembling the pages. A sorting-based approach to historical record linkage* (MA thesis, Utrecht University)