The 2020 IDS Release of the Antwerp COR*-Database. Evaluation, Development and Transformation of a Pre-Existing Database

By Sam Jenkinson, Francisco Anguita, Diogo Paiva, Hideko Matsuo and Koen Matthijs

To cite this article: Jenkinson, S., Anguita, F., Paiva, D., Matsuo, H. & Matthijs, K. (2020). The 2020 IDS Release of the Antwerp COR*-Database. Evaluation, Development and Transformation of a Pre-Existing Database. *Historical Life Course Studies*, *9*, 197–217. https://doi.org/10.51964/hlcs9301

HISTORICAL LIFE COURSE STUDIES

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

VOLUME 9, SPECIAL ISSUE 5,
2020

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 http://www.ehps-net.eu/journal.

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 9 (2020), published 08-12-2020

The 2020 IDS Release of the Antwerp COR*-Database

Evaluation, Development and Transformation of a Pre-Existing Database

Sam Jenkinson KU Leuven, Belgium
Francisco Anguita International Institute of Social History, Amsterdam, the Netherlands
Diogo Paiva International Institute of Social History, Amsterdam, the Netherlands & Iscte, University Institute of Lisbon
Hideko Matsuo KU Leuven, Belgium
Koen Matthijs KU Leuven, Belgium

ABSTRACT

The Antwerp COR*-IDS database 2020 is a transformed and harmonized historical demographic database in a cross-nationally comparable format designed to be open and easy to use for international researchers. The database is constructed from the 2010 release of the Antwerp COR*-historical demographic database, which was created using a letter sample of the whole district of Antwerp (Flanders, Belgium). It has a total sample size of +/- 33,000 residents of Antwerp. The sample spans nearly seven decades. The data is collected from historical records: including population registers and vital registration records covering births, marriages, in/external migrations and deaths. The database covers up to three linked generations (in some cases more), and contains micro-data on individual level life courses, and relationships deriving from address-based household composition methods. An important characteristic is the sample's large migrant population, including the timings of their demographic events and living arrangements, whilst resident in the district of Antwerp. In addition, the sample also contains a large array of occupational level information. This paper presents the processes, methodologies and documentation regarding the evaluation and development of a pre-existing historical database. This includes the systematic evaluation of the original samples, methodologies for address based reconstructing of households, and the geocoding of a historical database which took place during the current development of this new version of the database.

Keywords: Historical demography, Historical database management, Intermediate Data Structure

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs9301
The article can be downloaded from here.

1 INTRODUCTION

Historical demography is an increasingly important research discipline for analysing the origins and development of the modern world. This is true both for the analysis of purely historical questions, but also for contemporary research topics that have modern day political, economic and social ramifications. Historical demographic analysis, however, is only as good as the quality of the data resources which are available to us as researchers. This makes it vitally important that our databases are constantly evaluated, updated and developed. It is also crucial that they are readily available and usable for cross national analysis by historical demographic researchers.

The Antwerp COR*-database is a highly unique and sophisticated research infrastructure. The database covers a highly dynamic and significant historical period in Belgian history of nearly seven decades (1806–1920). The historical period of the second half of 19th and the early 20th century is one of rapid societal transformation in Belgium. This is particularly true in the Northern region of Flanders and its port city of Antwerp, which was so central to the process of industrialization across the province and country. The economy of Antwerp had been traditionally based around agriculture and textiles. This changed rapidly in the course of the 19th century, as the city expanded into one of the largest ports in Europe, with the blossoming of multiple economic sectors closely interwoven with the ports activities. The socio-economic development of Antwerp brought mass migration to the city, originating both from neighbouring provinces and also from abroad (Loyen, 2003; Puschmann, 2015; Puschmann, Gröberg, Schumacher, & Matthijs, 2014). Many aspects of the livelihoods of migrants appear to be highly heterogeneous and dependent on the various dynamics of migration flows, involving short and long stays, in and out migration and with much depending on their socio-economic status and place of origin. Examining the expansion of the local opportunity structure and the changing composition of immigrant flows provides insights into the changing and emerging roles of migration into people's life strategies (Winter, 2009). The period bore witness to a multitude of epoch defining social, political, economic and demographic changes, from the industrial revolution to the demographic transition. Not only is the timespan wide, but the size is large. It has a total sample size of +/- 33,000 residents of Antwerp. During the period of coverage, the city quadrupled in size from 56,000 inhabitants in 1800 to more than 273,000 by the end of the century (Matthijs & Moreels, 2010). The database therefore represents a highly unique historical demographic source.

The complexities of the analyses undertaken by historical demographers and the uniqueness of the sources have often prevented researchers from working with multiple databases at the same time, therefore making comparisons across local and national databases difficult. The Intermediate Data Structure (IDS) is a standard data format that has been adopted by several large longitudinal databases on historical populations (Alter & Mandemakers, 2014). It provides a common structure for storing and sharing historical demographic data. This structure facilitates the extraction of information with the purpose of constructing a rectangular file, for example where the episodes of an individual's life course are represented. The rectangular format is the prerequisite for longitudinal statistical analysis.

The purpose of this article is to provide an update and overview of the latest version of the database, in which we seek to satisfy these requirements to continuously develop our database and make it as easy to use and available for researchers as possible. The new version we have produced has undergone a systematic and methodical evaluation in order to provide confidence to researchers of its fundamental strengths and structures. It has also been updated in a number of highly valuable ways. This includes a greater depth and volume of highly important familial and intergenerational relationships. In addition, it has also undergone a geocodification of the dataset to enable important historical geographical analysis. Finally, the dataset has been transformed to the Intermediate Data Structure (IDS) in order to make it as ready and easy to use as possible for comparative historical demographic analysis. This is a work which builds on that initiated by De Mulder and Neyrink (2014).

This article begins with: (i) a discussion of the evaluation of the original sample to identify strengths, weaknesses and areas for further database development; (ii) our methodologies for reconstructing households based on residential information and how we used this to identify familial relationships within the household; (iii) an overview concerning the geocoding of the database to include a coordinate system for the location of the addresses, adding value to the current format of the 2010 release of the Antwerp COR*-database by means of additional variables: a twofold system (street centroids and municipality centroids) in the context data; and (iv) discuss the conversion to IDS format and discuss variables included in the new version of the database, highlighting additional and deviating variables from the standard IDS format, as well as unique challenges concerning the timestamping of observations within the 2010 release of the Antwerp COR*-database (Alter

& Mandemakers, 2014). In addition, we provide illustrative examples of individual and intergenerational life courses which are representative of the strengths of the broader database. The article finally ends with a discussion and conclusion regarding the challenges encountered during the production of 2020 IDS release of the COR*-database and suggesting areas for future work on the database.

2 THE ANTWERP COR*-DATABASE

2.1 SAMPLING METHOD

The Antwerp COR*-database is constructed using a letter sample (see also TRA survey in Bourdieu, Kesztenbaum and Postel-Vinay (2014)). Full information on the process of construction and data collection can be found in both Matthijs and Moreels (2010) and Van Baelen (2007). This involves the sample selection of all individuals with a surname beginning with the letters COR. All sources, including population registers and certificates of vital events such as births, deaths and marriages, which contain at least one individual whose name begins with these letters were selected for the sample. This selection is not restricted to just 'COR*'-family names either, but also all other people who are recorded in these selected certificates and population registers. We note that approximately 30% of the people that are included in the entire sample have some form of COR*-name. This sampling approach has a number of advantages. It makes the data collection more straightforward, by simplifying instructions to data collectors, and thereby ensuring that the risk of potential mistakes is minimised.

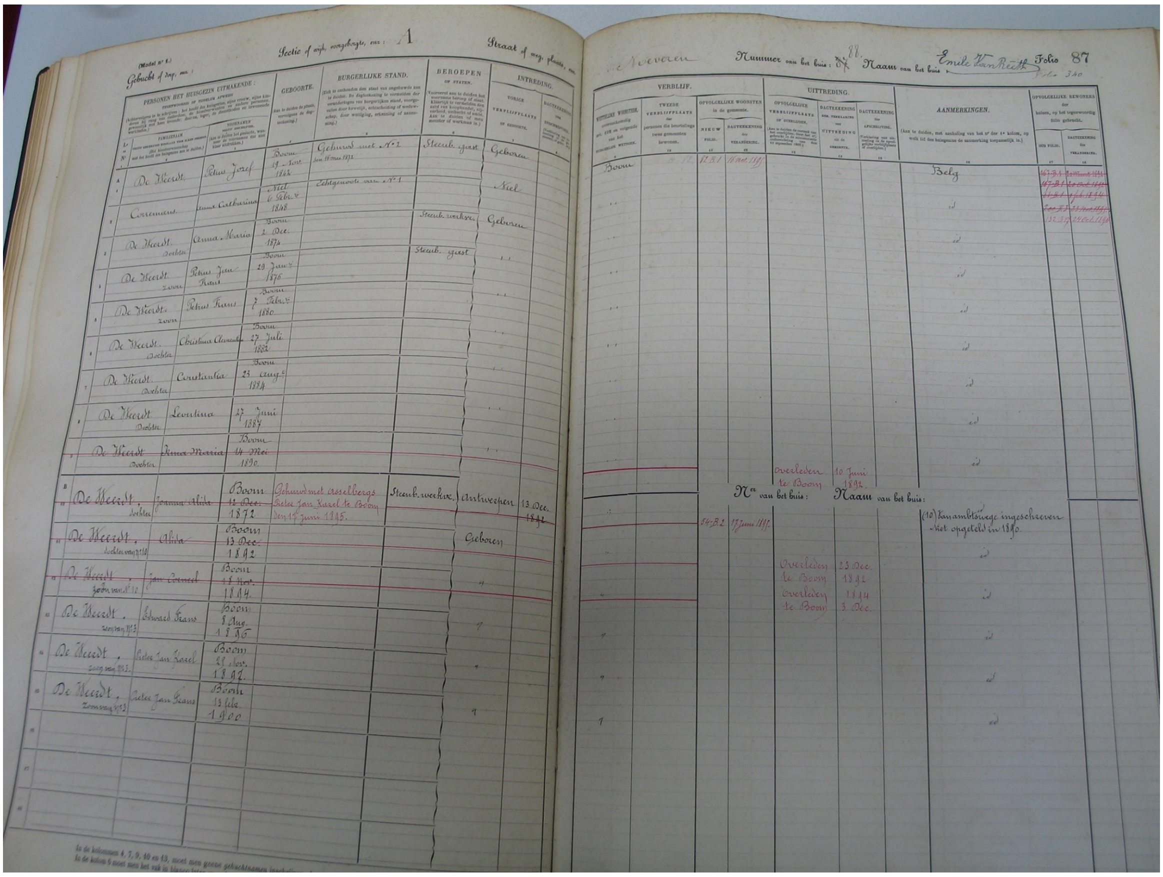
The choice of family names starting with the letters 'COR' had a number of motivations aimed at strengthening the representativeness of the database, and the choice was made in consultation with the Faculty of Linguistics at the University of Leuven. These names were shown to be evenly geographically distributed and highly similar to the distribution of the Flemish historical population as a whole. This letter combination was also sensitive to linguistic and socio-demographic characteristics. Importantly, it was also relatively representative of foreign names in the population, compared to other such names, a particularly important characteristic in multilingual Belgium, and especially Antwerp, with its large volumes of international migration (Matthijs & Moreels, 2010). In addition, it was found to aid with legibility problems in data collection (Van Baelen, 2007). Furthermore, a letter sample made the linking of individuals across different sources much simpler. 19th century Antwerp was a city of rapid growth, almost increasing fivefold from 56,000 inhabitants in 1800 to more than 273,000 by the end of the century. The sample size of COR*-people equates to 0.4% of the Flemish population (Matthijs & Moreels, 2010).

2.2 ORIGINAL SOURCES

A number of different primary sources were used to construct the Antwerp COR*-database covering a number of different periods. 1846 is the year of the first census in Belgium and also the year of the first Belgian population register. The Antwerp COR*-database includes population registers and civil certificates covering the following time periods: population registers 1846–1940, divided over different periods (i.e. they are starting in 1846, 1850, 1856, 1857, 1860, 1870, 1975, 1876, 1880, 1890, 1900, 1910), birth certificates from 1821–1906, marriage certificates from 1806–1913, and death certificates from 1836–1906. This represents a time period of both significant breadth and depth of rich demographic information.

The information contained within these historical records differs depending on the type of source in question. Population registers (see Figure 1) are household registers consisting of all information relating to household members, including demographic and socioeconomic variables: i.e. occupational titles, marital status, relationship to the head of household (available from late 19th century), and a register of recorded vital events including births, deaths, marriages, divorces and internal (within Antwerp) and external (moving out of Antwerp) migration. The certificates for vital registrations of births, deaths and marriages include further information in addition to these registers. This includes details of the event in question, reported at the time of the event, including newborns reported as 'lifeless', multiple births, legitimacy of the child, parental information, witnesses, migration and occupations. Registration of these events began in Belgium in the late 18th century and the records are well preserved and stored in state, town or municipal archives. The distribution of the dates of the events is heavily skewed to the latter part of the 19th century. However this is in line with the aforementioned quadrupling of the city's population over the period that the database covers.

Figure 1 *Picture of a Belgian population register*



3 DATABASE DEVELOPMENTS

In order to prepare the latest version of the Antwerp COR*-IDS database, several steps have been taken which will be discussed at length in this section. These began with an evaluation of the original database to identify strengths, weaknesses and areas for further development. Following the successful implementation of this exercise, work was undertaken in three areas which we identified as important and strategically beneficial. The first of these was a familial reconstitution in order to identify additional households and familial relationships, using address based register information originating from the registration of events within the population registers where internal migrations are recorded. The second was the geocodification of the database using address based information and historical GIS sources and maps. A final exercise was to convert the database into an IDS structure, to make comparative historical demographic analysis using the 2010 release of the Antwerp COR*-database simpler and of greater ease.

3.1 SAMPLE EVALUATION

The first step we took was to methodically evaluate the quality of the original sample in order to ensure that all of our efforts and developments would not be undone by any underlying weaknesses in the database. We began with an assessment of the robustness of the original record linkage. This involved examining several key variables by the already established individual numerical identifier (IDNR) across different sources. These key variables included dates and locations of vital events, such as births and deaths, and also gender. We chose these essential variables as we believe that any significant inconsistencies here would be highly suggestive of greater errors in the initial record linkage.

We began by splitting the database into datasets of the original source records (i.e. source of data). Following this we proceeded by comparing the identifying attributes belonging to an IDNR across sources in order to test for discrepancies. Concerning death dates, the number of discrepancies across records for the same individual was below 1% for day, month and year examined separately. For births this was also the case, with the exception of year, where 3% were inaccurate. Regarding birth and death locations, the discrepancies were much higher, however this is largely due to differing spellings of place names, as well as stray capitalisation and white space, which took place during the data entry process. This is standardised within the database. For gender the number of discrepancies was 0.83%. We believe the low number of discrepancies is supportive of the high quality of the original record linkage.

Following the successful execution of the first exercise a second evaluation was undertaken to test the original individual record linkages through a re-linking of the database. Similarly to above we began by splitting the database into observations from the original historical sources: birth, death, marriage certificates and also population register. The second step was to relink individuals within the database from the original source of data, ignoring the previous identification numbers given during the prior record linkage process. We used the data of birth and death records. The method applied is in line with the initial linkage of 2010 (Van Baelen, 2007), in which we used a stochastic record linkage method provided by Sariyar and Borg (2010) as part of their R package 'record linkage' (Sariyar & Borg, 2016). The record linkage process uses the Fellegi-Sunter Model (Fellegi & Sunter, 1969). It relies on the assumption of conditional probabilities regarding comparison patterns. In the full Fellegi-Sunter model these are used to compute weights which aid in discerning matches and non-matches. The weights within the package are computed using an expectation maximum (EM) algorithm in line with Haber (1984) and Contiero et al. (2005). We then use common variables across individual sources to calculate the likelihood of a record being a match by executing a string comparison tool. The selected variables include given and family names, birth location, birth day, birth month and year separately. These are then used to calculate similarities across different records and to create pairs. To compare strings, the Jaro-Winkler distance was used (Winkler, 1990). This function works by measuring the edit distance between two strings and calculates the minimum numbers of single character transpositions to transform one word into another. Full details of this process, including more detailed description of methods, can be seen in Jenkinson, Matsuo, and Matthijs (2017). Three rounds of matching were carried out on all observations contained within the birth and death records. Of the total matches identified during this process the number of linkages that we found which were not recorded in the original database was 5.2%. Improving the current linkage may be explored in the future, but we believe that the number of potential links is relatively low.

A third and final evaluation exercise was then performed to examine the two main intergenerational variables contained within the database and therefore assess the linkages between individuals; IDmoeder (identification of mothers) and IDvader (identification of fathers). These are constructed variables within the dataset and are largely based on the population registers and also the birth, death and marriage certificates. One important reason for this evaluation concerns the way intergenerational linkages were recorded in and collected from the original sources. These relationship variables in the household registration section of the population register only became standardized in all municipalities after the late 19th century. Before this, no variable is recorded in the source indicating the relationship between co-residents. This means that any information that was documented during the data collection was often actually interpreted by the collector based on the names, genders and positions on the household register of individuals. Full details of this interpretation can be seen in Van Baelen (2007).

Van Baelen (2007) documents the steps implemented to derive familial relationships within the 2010 release of the Antwerp COR*-sample. Two primary approaches were used to obtain kinship relationships: exact family relationships; and the use of family names. The first approach makes use of a schema of relationship codes to derive up to 60 potential types of inter- and intragenerational relationships on the basis of the information contained in the population register (Van Baelen, 2007). The second approach is based on the calculation of an indicator through the information of the family name and the geographical location of the individual.

While this interpretation is considered a reasonable procedure for identifying the familial relationships within the household, it may invite errors for non-standardized living arrangements among complex families that includes non-coresidential parents, out of wedlock births, and migrants. Households with multiple adults are potentially much more difficult to interpret correctly given this format with a higher risk of incorrect adults being identified as parents. This is an important point which must be remembered when using the Antwerp COR*-database. We believe this potential risk of misinterpretation by the original data collectors to be important and as such we investigate it further below.

In order to evaluate this parental linkage contained within the 2010 release of the Antwerp COR*-database, the first step taken was to compare parental information by individuals between different sources. For this exercise a Levenshtein similarity metric is used to compare names. The 2010 release of the Antwerp COR*-database contains 5,883 observations of mother-child relationships, and 5,676 observations of father-child relationships. These observations can be checked against parental names listed within the birth certificate file, which we consider to be the most authoritative source. The names of 89% of mothers and 91% for fathers scored a Levenshtein similarity matching score of higher than 0.8. A large number of those remaining differences appear to be due to non-standardised names and data entry errors. For full details see Jenkinson, Matthijs, and Matsuo (2019). These high number of similarities (89% & 91%) led us to be relatively confident of the constructed variables accuracy and therefore for use in further converting and developing our database.

The outcome of these three exercises was to provide relative reassurance of the quality of the original individual and intergenerational record linkages of the 2010 release of the Antwerp COR*-database. The number of discrepancies by individuals for key variables was low, as too were differences found through the original record linkage and examination of intergenerational record linkages.

This exercise therefore provided reasonable confidence in the core structures of the database around fundamental variables, individual and familial linkages. It also identified a number of areas for potential development. This was particularly true concerning the underutilisation of geographic, address based information, which is an incredibly valuable resource. The following two sections of the paper will focus on how we have harnessed this information to further enrich our database, both with additional and verified familial linkages between individuals and also the geocodification of the 2010 release of the Antwerp COR*-database.

3.2 RECONSTRUCTION OF RELATIONSHIPS WITHIN FAMILIES

The outcome of the above evaluation was a renewed confidence of the core strengths of the 2010 release of the Antwerp COR*-database over a number of highly important areas. In addition, it also gave us a greater understanding of some key aspects which we believed could be developed further for this new release of the database. The first of these was focused around address based geographic information. A large amount of address based information is contained within the database and then attributed to individual records. This is a highly valuable source of information which can be used to reconstitute families and, in turn, further enrich the Antwerp COR*-database with new familial relationships and intergenerational linkages. We explain here the process and the final results.

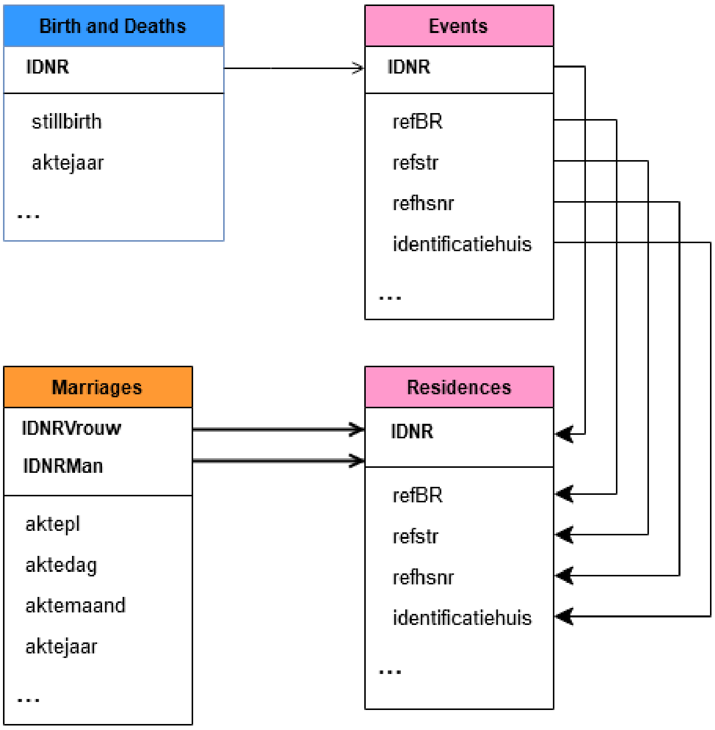
3.2.1 OVERVIEW AND INPUTS

For the goal of the reconstruction of the familial relationships, we made use of four datasets contained in the 2010 release of the Antwerp COR*-database. These tables were the *Events*, the *Residence*, the *Marriage* and the *Births* tables, all originating from different sources. The *Events* and *Residence* tables are compiled from the population registers, whereas the *Marriage* and *Births* tables originate from the certificates. The first two, the *Residence* and the *Events* tables, were merged together in order to combine the individual level information of the residents of the addresses with the events they underwent whilst residing in them. This was then supplemented with additional data contained within the *Marriage* and *Births* tables.

In Figure 2, we can see an elementary outline of the structure of the tables with which we worked. The Antwerp COR*-database consists of a set of tables, all of them linked by means of the personal identifier (or IDNR). In the graph, only links between the personal identifier — in addition to other relevant variables for this study — are depicted.

If we consider that a household consists of the individuals who live in the same dwelling, this presents a number of possibilities to identify family members. In other words, we define households as persons that share the same dwelling, which can include boarders and servants. And we furthermore specify them as 'family units' for our purposes, in order to identify familial relationships that are living in the same time period specific to the register in use. One such example of how we identified families is through the matching of addresses and the timing of migration events. This is possible, as within the database, the dates and destination of internal migration within and between each municipality within 'Antwerp' were registered. For instance, a group of people identified as moving into the same address on the same date are detectable. This type of information is not unique to the Antwerp COR*-database, as it is also recorded in other countries (e.g. in the Netherlands and Sweden). The simultaneous timings and locations of migration events suggests a high degree of likelihood of a family unit. In addition, if these individuals also share a common last name, we can be even more confident that this is a potential family unit.

Figure 2 *2010 release of the Antwerp COR*-database*



This data was combined by means of four key variables from the *Events* and *Residence* tables (see diagram in Figure 2): (i) the unique identifier of individuals (variable 'IDNR'); (ii) the residence address (street name and house number corresponding to the variables 'refstr' and 'refhsnr'); (iii) the pre-defined identification number of the household (the variable 'identificatiehuis'); and (iv) the time period that a population register was in use relating to the moment the address registration took place and that was recorded in the official municipality records (variable 'refBr' in the diagram). The predefined identification number 'identificatiehuis' is a pre-existing variable contained within the 2010 release of the Antwerp COR*-database. It is defined as the particular configuration of members at a unique address within a specific population register at a specified time. This household ID is unique to the specific configuration of the address, the household members and the particular source book of the population register. In addition it is only based on the *Residence* table within the database.

Since the time span of the registration period (variable 'refBr') for each volume of the population register was about ten years, changes in the structure of the familial relationships may remain unnoticed. To overcome this issue we made use of the information contained within the Events table. This table provides the reported demographic events for each individual. Through the common identifier (IDNR), it was possible to connect them to the address that these individuals were reported as officially residing at. In these sources four particular demographic events were of use for identifying potential household members, migration between and within municipalities. In addition, we also made use of a variable contained only within the *Residence file* which defines an individual's relationship to the head of the household. This was used to identify who was the head of the household within the previous version of the database. We also believe that the relationship with the head was a simple relationship for the data collectors to interpret, because the heads are always listed first in the source. This means that it has a low risk of mistakes, and is why we have confidence in its reliability for use in this exercise.

In addition to migration, an event like marriage can also bring information about the composition of family units. When marriage occurs, it is highly likely that the groom and bride will begin to live in the same dwelling. Unfortunately, information about partners is not present in the *Events* table, which is based on the population registers. In order to be able to add marital information to our reconstituted households data we had to integrate it with the *Marriage* certificates table. In doing so, we could confirm the marital relationships of 1,930 persons within the families: out of a total of 13,641 individuals in the marriage registers. This allowed us to identify additional spousal relationships and the age gaps among the individuals within our family units file.

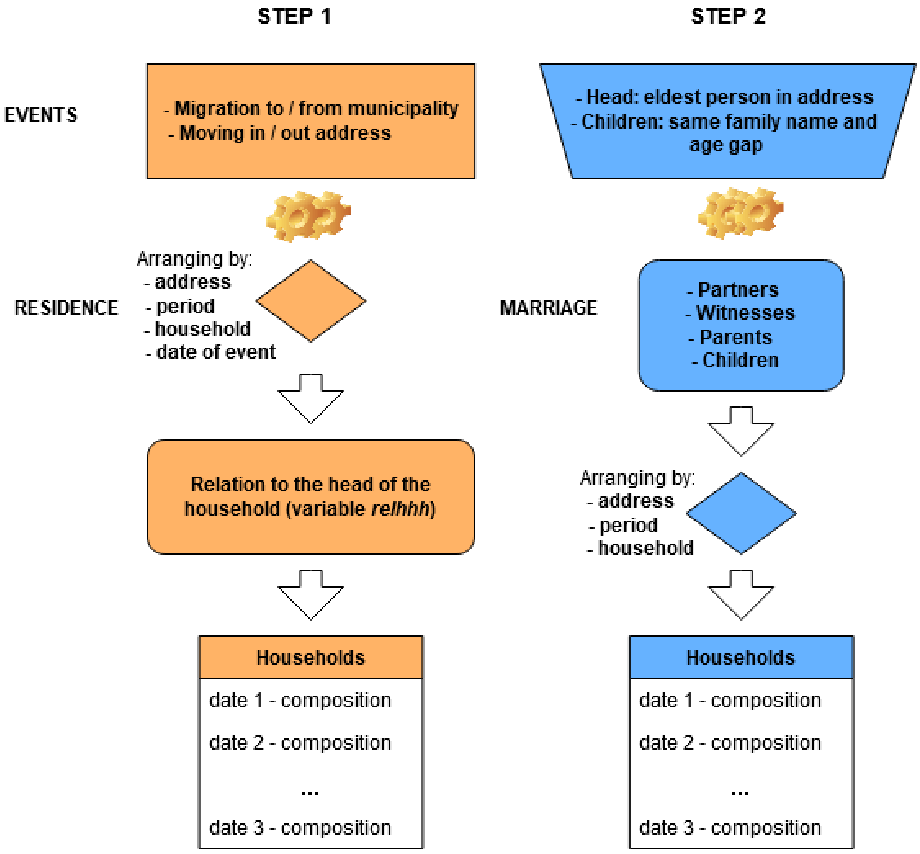
3.2.2 METHODOLOGICAL STEPS FOR THE RECONSTITUTION OF FAMILIAL RELATIONSHIPS

We performed this reconstitution of familial relationships with the development of an algorithm based approach that applied demographic assumptions and cross-referenced information from other sources (population registers, vital registration records, i.e. births and marriages). For the linkage between the tables and the relation among the individuals, we initially relied on three common characteristics: address, heads of household identified from the 2010 release of the Antwerp COR*-database and also the time period of the address registration in the official municipality records.

The use of household information (i.e see above where we define households as 'persons that are sharing the same dwelling') is justified because it was provided directly by the sources with their values clearly stored in the corresponding variable (i.e it was not inferred by the database creator). As a result, where it was possible to confirm the head from the previous version of the database, we did so. Secondly, members sharing the same family name and an age difference of equal to or more than fifteen years with respect to the head, were assumed to be their children. This method may identify more distant kin, such as nieces and nephews, who may meet these same characteristics as daughters or sons. In order to overcome this disadvantage we cross checked observations with relevant information from another dataset, more specifically concerning witnesses from the marital certificates. Unfortunately, rigorous exact name matching between witnesses and members of the reconstructed family units, where the number of units are limited, didn't provide sufficient results to shed light on the distinction between daughters and nieces or sons and nephews.

We also used the data from marital certificates table to ascribe marital and familial relationships to people identified as living together. Figure 3 illustrates the steps taken for reconstituting relationships within families. This depicts a flowchart of the methodology in constructing familial relationships. As a summary, in this process, the most important two data sets are the *Events* and the *Residence tables*, as they provide the core of the information we work with. The first one, mostly provides moves into and out of an address or a municipality, and it is complemented with the dates of birth of the records that it originally lacked.

Figure 3 *Visual representation of reconstitution of family method*



In the first step of Figure 3, we focus on the events of moves, and this information is merged with that contained in the *Residence* tables. Its outcome is arranged by each record's identification (or IDNR), the address, the time period of the address registration in the official municipality records, and the pre-defined identification number of the household. Through examining events involving moves, we were able to identify conjoint actions of different persons.[1] This allowed us to reconstruct some of the relations between individuals sharing family links, by creating groups or clusters of people sharing an address at a particular time. For this goal, the utilization of the information of the role of individuals with respect to the head person was also useful (when this information was available). This is collected by the variable 'relhhh' that we can see in Figure 3. Through the combination of all these actions we managed to reconstruct the first version of the address based clusters (see the final stage of Step 1 in Figure 3, outlining their prospective composition throughout time).

Step 2 of Figure 3 summarizes the practices carried out to finetune the first draft of the clusters. We did so by applying the assumptions outlined at the beginning of this section, the inclusion of other information from the sources (i.e. concerning partners, witnesses, parents or children), and the arrangement of the distribution of the clusters by the address, the time period of the address registration, and the pre-defined identification number of the household.

3.2.3 HOUSEHOLDS AND HEADS

Once the relations are reconstructed, each family unit is assigned a numeric code. We computed 12,396 observations of heads of a family unit derived from our methods, most of them appearing several times due to the manifestation of different events for the members of the same household, such as moves in or out of an address. In other words, when a family moves to a new address, this counts as one observation. When they leave, this is a second observation. Among them, only 728 times involved just one family event. In other words, only one move for the unit and no extra information was given via any other events.

Heads of family units identified by this method can be compared to heads of households in the population registers, but we should not necessarily expect these attributes to match exactly. Many households may have included more than one family unit, and the population registers did not always designate a new head of household when a prior head died or departed. Among our inferred 12,396 observations of family units, we found 9,045 unique heads of families. In contrast, the Residence file identifies heads of households 34,582 times, among whom are 11,798 unique individuals (IDNRs). If we count occurrences, persons newly identified as heads of family units (12,396) matched the heads of households (34,582) 6,374 times or 51.4%. Counting unique individuals, 64.3% (5,814) of newly identified heads of family units (9,045) were matched to heads of households (11,798). This means that we were able to identify 3,231 new potential heads of family units in addition to the numbers known in the 2010 release of the Antwerp COR*-database. It is true that attribution of headship can be questionable when only based on age, but these possible new heads can be seen as the likely cores of newly identified family units.

Concerning the replacement of the head of the household when the head died, for the purpose of the reconstitution of the groups, we decided to focus on the event types that connected several individuals together at the same moment, leaving out those that involved only one individual, like the event of death. In doing so we were seeking for individuals that appear to be part of groups, moving to/from the same address at the same time. For this goal of the reconstruction of groups, or proto family units, single-individual events were not useful.

Furthermore, an evaluative exercise was carried out among the pairs that were obtained by means of our assumptions. We did this by cross checking our final outcome of the linked couples of heads and children (12,396) with that of the *Births* table. By doing so we could confirm 732 pairings of the type head (as a father) and child; and 469 of the type head (as a mother) and child. This accounts for a total of 1,201 pairs of relations between parents and children derived through our model of assumptions that were confirmed with data from our sources.

3.2.4 PARENTS AND CHILDREN

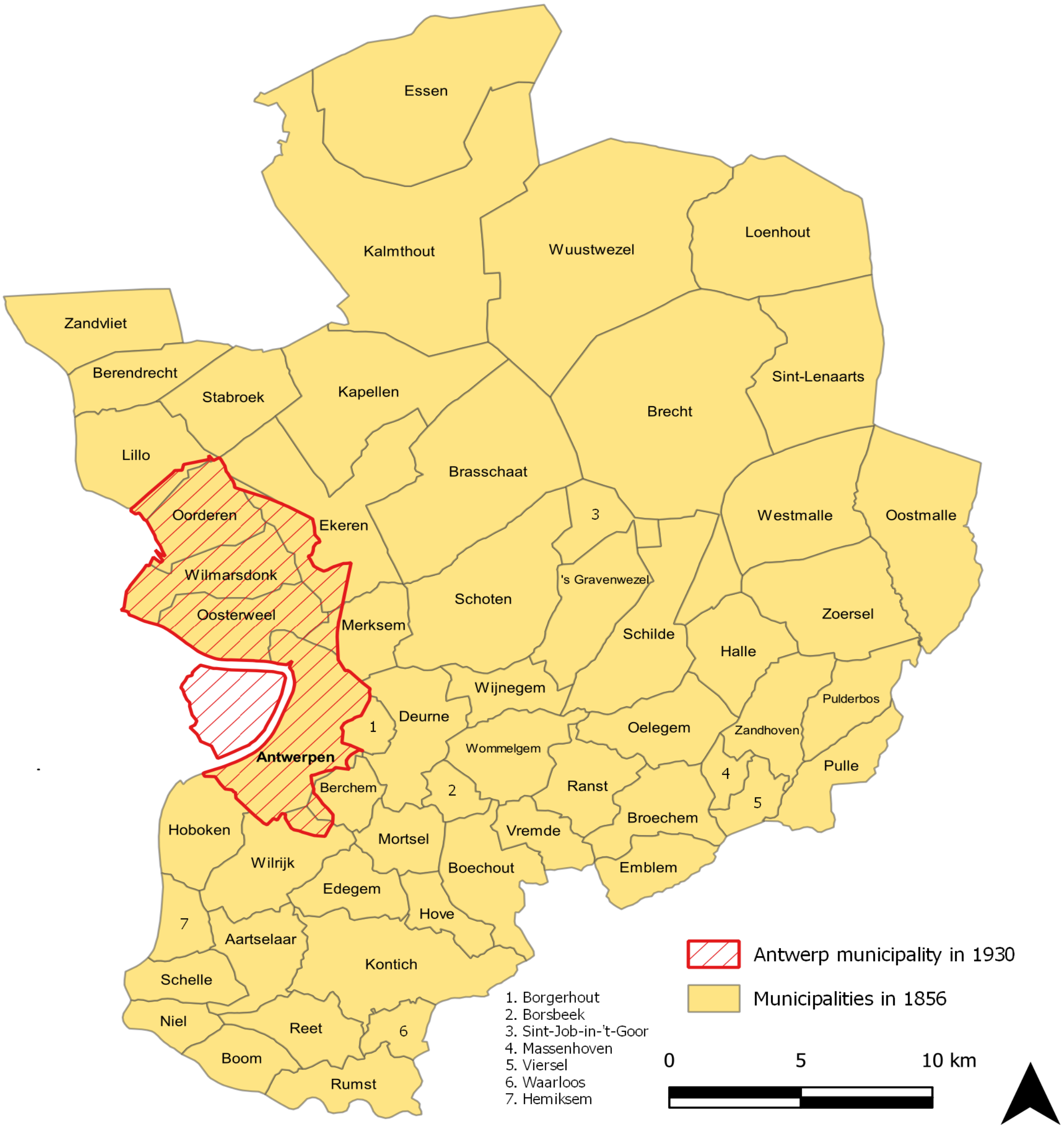
The total sample contained 13,138 observations of sons or daughters. Our methods using migration events from the registration of events file and marriage information from the marriage certificate files obtained 7,654 observations of sons or daughters. Of those 7,654, 4,472 (58.4%) relationships matched the earlier database, and we were able to identify 3,182 new parent-child relationships.

[1] Like several individuals moving into the same address on the same date, and leaving the same address on the same date as well.

3.3 GEOCODING THE ANTWERP COR*-DATABASE

Spatial analysis is continuing to become an ever more important analytical component to an ever increasing number of highly significant demographic research questions. We follow the similar work by Hedefalk, Harrie, and Svensson (2014) (IDS-Geo), namely the inclusion of a coordinate system for the location of address level information. This development adds critical value to the current format of the 2010 release of the Antwerp COR*-database, by specific means of additional variables: a twofold system in the context table (street centroid[2] and municipality centroid[3]) and including individual variables at the very low level (i.e. street level) (i.e. NIS level 7[4]) (Paiva, 2019). Figure 4 represents the municipalities in the Antwerp region, while figure 5 shows the representation of addresses by street centroids.

Figure 4 *Geographical map of the Antwerp COR*-database, borders of the 19th century and 1930*

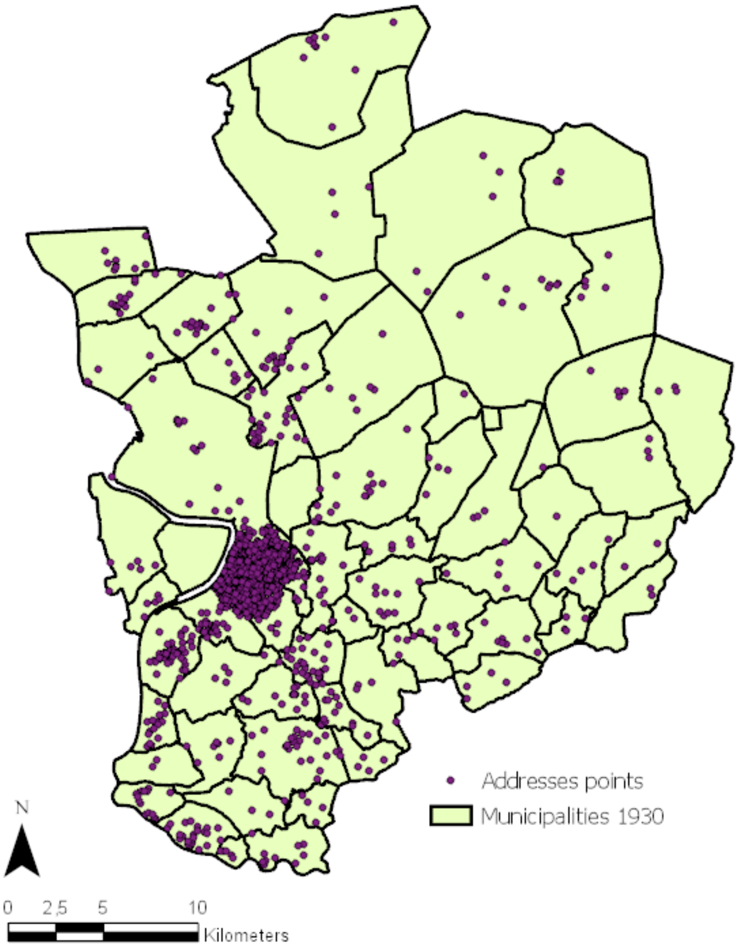


[2] This means for instance, the median point of the line representation of the street.

[3] A point coordinate representing the location of the administrative-political centre of the municipality.

[4] NIS code represents an alphanumeric code for regional areas and consists of 5 digits. The first number refers to the province; the second, the arrondissement within the province and the last three, identifies the community within the arrondissement (Statistics Belgium).

Figure 5 *Geographical map of Antwerp COR*-database: addresses represented by street centroids*



The georeference process uses the data from the 2010 release of the Antwerp COR*-database named 'huisSAMEN2'. This corresponding data contains several fields and variables expressed in brackets:

- ID code for household (identificatiehuis)
- Original and standardized municipality name (gemeente/gem)
- Original and standardized Year of the source where information was retrieved — population register (bevolkingsregister) (bevolkingsregister/BR)
- Original and standardized quarter's name (wijknummer/wk)[5]
- Original and standardized house number in quarter (wijkhuisnummer/wkhsnr)
- Original and standardized street name (straat/str)
- Original and standardized street house number (hsnr/huisnummer).

Additionally, two sets of data were used: the historical borders of Antwerp arrondissement (1856–1930) (Vrielinck, Wiedemann, & Deboosere) and the historical streets for the year 1898, from the GIStorical Antwerp (UAntwerpen/Hercules Foundation) project. Research on historical streets was also possible with Geopunt.be's historical maps: Atlas der Buurtwegen (1841), Vandermaelen kaarten (1846–1854) and Popp kaarten (1842–1879).

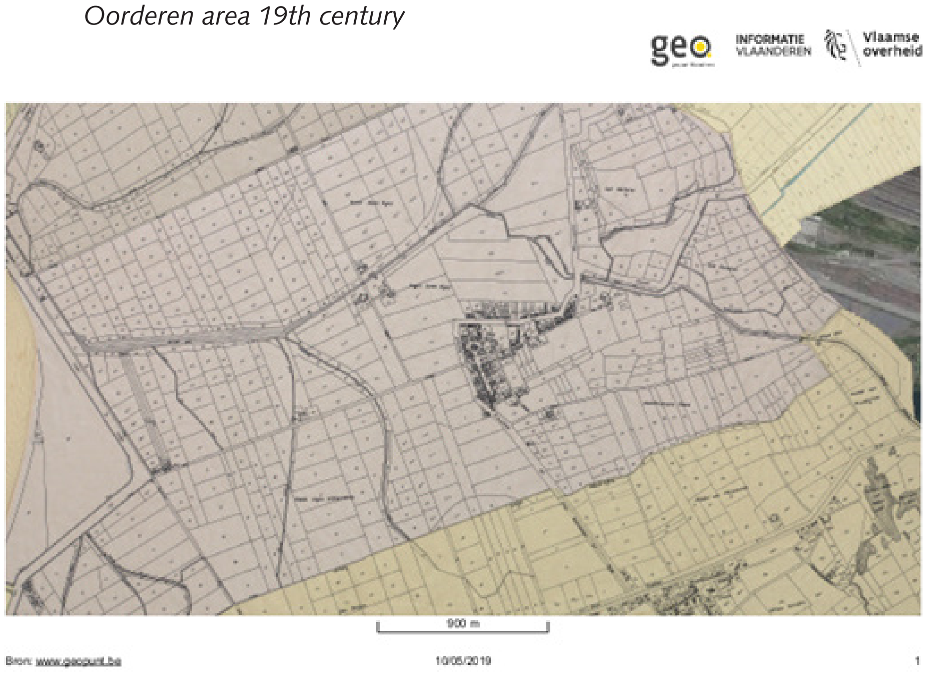
The process of georeferencing the addresses in the 2010 release of the Antwerp COR*-database involves performing several sub-processes. The idea behind this is to link information between addresses (e.g. municipality and street) and geographic data, in the most efficient and least time-consuming form. To implement this, we use the existing aforementioned geo-databases, while a more thorough process is applied to those units which lack information.

[5] The level of 'wijk' (quarter) is not always included, depending on the municipality.

Given that the huisSAMEN2 data (original, 2010 release of the Antwerp COR*-database) already provides the standardized text names for the original municipality (**gem**) and street (**str**) names, derived from the population register a simple record linkage (exact matching) is applied to most addresses. For the remaining addresses with missing values, additional work is implemented to obtain the exact geocodes.

There are three issues that deserve specific attention for the specification of geocodes using missing or incomplete addresses. Firstly, it should be noted that the standardisation process applied to **gem** and **str** text fields was imperfect.[6] This means a further revision of the standardized text in the huisSAMEN2 data (2010 release of the Antwerp COR*-database) was necessary as an intermediary step to implement matching. As an extension of this first step, two conversion datasets are created. These act as dictionaries between the original name in huisSAMEN2 (**straat**) and an alternative standard (in case the street still exists today with the same name — **fix_str**) or the corresponding modern name (if the street name in the sources is outdated — **hist_str**). The latter is based on online research (various media, encompassing blogs, forums, and historical maps including the use of Google (Google Maps) and Esri's (ArcGIS Pro), plus consultation of a list of old street names.[7] A last dataset was created by assigning coordinates manually to specific streets (i.e. a textual description of a broad localization — e.g., between the known streets A and B) or landmarks in historical maps. Secondly, the geographic data available (GIStorical Antwerp 1898 street shapefile) only covered the city of Antwerp and some of its immediate surroundings. In order to obtain the information of the entire Antwerp arrondissement, two other shapefiles (lines) were created: one to link current streets with the huisSAMEN2 table; and another for historical streets (i.e. streets that no longer exist today, but can be found in historical maps) (see Figures 6 and 7 illustrating this phenomenon of disappearing streets, in present day Antwerp north port area, where before the village of Oorderen stood). For each street of these sets, a medium point was obtained resulting in an additional file, including variables of street name, municipality name, longitude and latitude of the medium street point.

Figure 6 *Oorderen area 19th century*

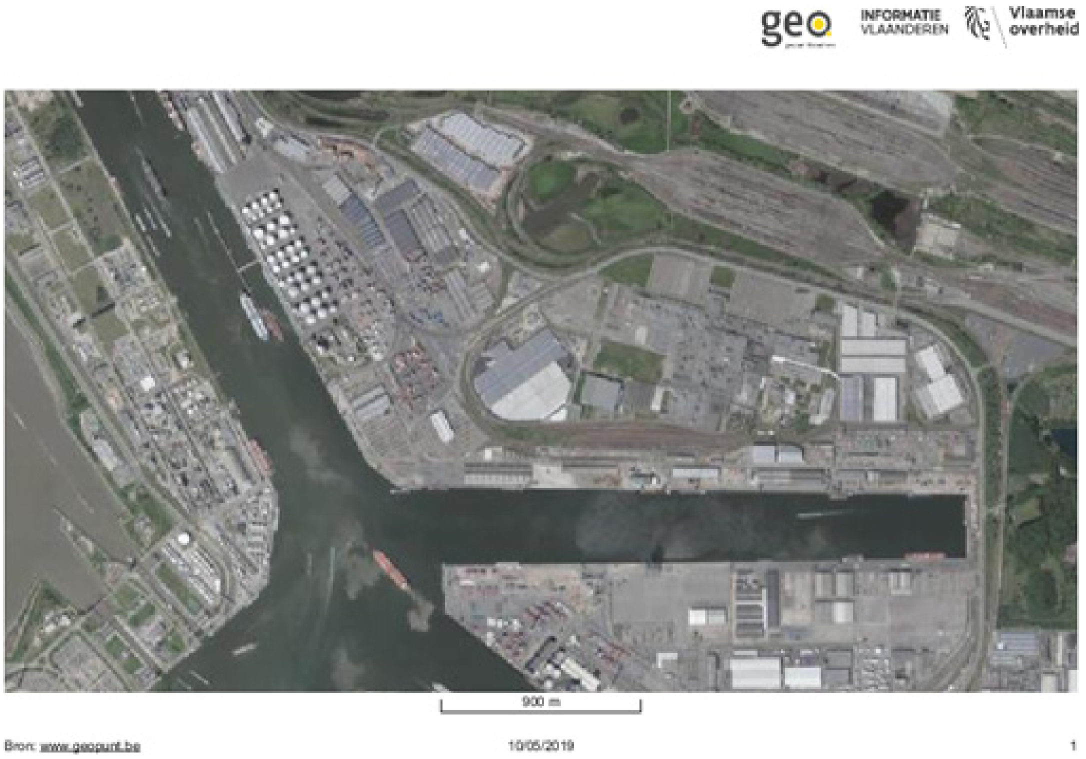


Source: This picture is from P.C. Popp's, 'Atlas cadastral parcellaire de la Belgique', 1842–1879 (www.geopunt.be).

[6] For example: 'Driesch', 'Driessche', 'Drieschstraet' and 'Dries Straat' in Antwerp were all being standardized as 'Driesstraat' while the correct spelling is 'Dries'; more significantly, 'Hagelkruisstraat', 'Hagelkruis' (both as 'Hagelruis'), 'Gr. Hagelkruisstraat' (as 'Klein Hagelkruis') were being divided into two different streets although all are forms of 'Groot Hagelkruis' and the latter was transformed from Groot ('Gr.') to Klein. The process of standardization relied on the document of 'alle plaatsen' recorded for the production of COR*2010.

[7] GIStorical Antwerp project provided an extensive list of old street names (mostly for Antwerp and its surroundings: Berchem, Hoboken, Borgerhout and Oosterweel), Robert vande Weghe's *Geschiedenis van de Antwerpse straatnamen* (1977).

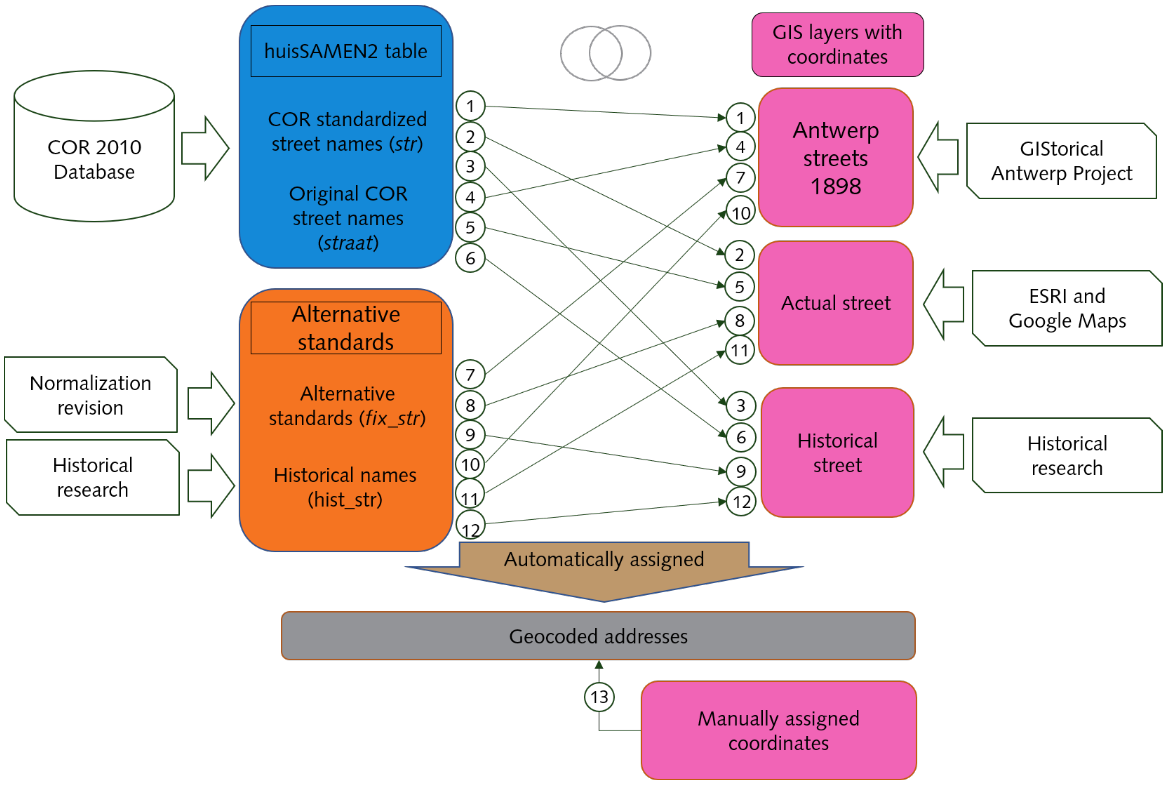
Figure 7 *Oorderen area today*



Source: www.geopunt.be.

Linking geographic data with COR*-street names and their alternatives is an iteration process, ultimately to provide the coordinates to the addresses present in huisSAMEN2. Figure 8 shows the order of the geocoding process sequence. The huisSAMEN2 variables that contain names of streets (**str** and **straat** from the Antwerp COR*-database plus the added alternatives from conversion tables, **fix_str** and **hist_str**) are matched with names contained in the spatial datasets. After implementing 12 steps, links were sought manually (the final 13th step). Finally, for the records successfully linked street coordinates are added (**s_lat** and **s_lon**).

Figure 8 *The Geocoding process*



3.4 2020 IDS RELEASE OF THE ANTWERP COR*-DATABASE

Once the aforementioned developments to the database were completed, our next task was to seek a way to make the database as usable and available as possible to historical demographers undertaking comparative cross national research. For this reason we chose to transform our database into the Intermediate Data Structure (IDS) stipulated in Alter and Mandemaker (2014).

As previously mentioned, this is an internationally standardised format for historical databases. Its purpose is to make it easier for users to perform cross national comparative research, without having to spend months getting to know each individual historical database. This could otherwise be incredibly time consuming, with each database having its own intricacies and unique coding and storage, making cross national research and analysis both time consuming and extremely difficult. Here we discuss some of the key obstacles we faced in transforming our database. Many of these obstacles are important for users of our database, whilst others will perhaps be informative for anyone else seeking to transform their own historical data and who may encounter similar problems.

3.4.1 IMPERFECT DATA AND TIME STAMPS

Time stamp information, the system used to date information within the intermediate data structure, is essential to the construction of individual histories. One issue we have faced in this area relates to imperfect data, as a result of the nature of the historical sources we have. This concerns the dating of information for occupational observations. Several occupation titles are recorded in the source, however the dating of this information is quite problematic.

This is because the ability to date an occupational observation is highly dependent on the source from which it has been collected. Observations of occupations which have been collected from the registration of event based sources, such as death, birth and marriage certificates, have exact dates when an individual was observed with a specific occupation. This means we can reliably date the incidence of an individual having said occupation by the certificate in question. This is not the case for occupational observations collected from the population registers. In these cases we know an individual's occupation only by the source at the start of a population register (roughly ten year periods) and also at the time of any events they experience during the ten years recorded in the population register. This means that persons with no events to register usually will not report a new occupation during the ten-year duration of a population register. Those who reported new events, such as moving to a different household, might have several chances to record changes in their occupations. This situation clearly creates a bias concerning the inclusion of occupational information.

We are also unable, with absolute certainty, to give timings and order to these sequences. We do, however, consider that the order refers to how it is being reported by the individual. Consequently, the only information we are able to provide with absolute certainty is that of the specific source. This reference to the source then enables the researcher to identify which particular population register a particular observation originated from, and therefore the year of it's opening. A researcher can then choose how to specify or estimate the dating of occupational observations themselves.

We know that this is much less than ideal, but with the limited information we have concerning these observations contained within the 2010 release of the Antwerp COR*-database, it is the only_date available to provide and closest to what is actually contained within the historical primary sources. So, the date field in the timestamp in this instance is blank. What we have included is the source the specific observation originated from, which allows the researcher to access the period in which the population register was in use by the administration, which starts with the opening year of the book. This can then be used by researchers to estimate dates for these observations of occupations.

The above refers to the lack of specific dates for occupations, and because of this problem, another related difficulty which we have encountered is in determining occupational spells using this imperfect data resulting from only a very partial view of an individual's occupational trajectory over the life course. We know the source (and year of the source) when individuals were observed with a particular observation, however we do not know for how long they kept this occupation. Furthermore, we cannot observe if there were any occupational changes or periods of unemployment within these unobserved periods, which as has been stated can be up to ten years long.

3.4.2 NEW VARIABLES AND INFORMATION UNIQUE TO THE 2020 IDS RELEASE OF THE ANTWERP COR*-DATABASE

Every historical dataset is constructed from unique sources. This brings a number of opportunities and challenges for a task of transforming data into a standardised structure, such as the IDS. Here we discuss some of these which may be relevant for other historical demographers.

There are a number of variables which exist within the new release of the Antwerp COR*-database resulting from the unique sources used to build it. These include reported locations for events such as birth, death and marriages. This is not the location of the event in question, but the place it was reported. In addition we have the dates for the reporting of events, as a distinct date from the event itself. This means we are able to observe differences in the timing and reporting of events and also the location of the reporting.

Moreover, a number of variables which exist in the IDS have unique meanings within the new release of the Antwerp COR*-database. For certain events the method of reporting is somewhat different to other historical sources. This is true for both legitimation of illegitimate children and also divorce. Most events are reported in their own original record which is registered at the time of the event. However, the legitimation is added to the original birth record and the divorce to the original marriage record.

We transformed and stored all data in line with the IDS guidelines (Alter & Mandemakers, 2014). In total, there are 41 items in the metadata, of which five are new types (variables) and 12 are new values for already existing types. They cover the dates and locations of vital events, such as births, deaths, marriages, migrations and divorces, as well as additional information including occupations, legitimacy and literacy.

3.4.3 2020 IDS RELEASE OF THE ANTWERP COR*-DATABASE: CONTENTS AND CHARACTERISTICS

The final sample consists of 33,583 individuals in 14,537 addresses drawn from population registers and vital registration records covering the period of 1806–1920. Reflecting the increase of the population in Antwerp from mid-19th century onwards, records contained in the Antwerp COR*-database are heavily skewed to the latter period. The total sample consists of birth years of 1734–1937, with slightly more than half of the sample being male.

The sample includes a large migrant population, meaning both Belgians born outside of the city of Antwerp and also international migrants. If one considers migrants as those who are neither born in the city, nor in other suburban areas, but were living in Antwerp arrondissement recorded in the COR*-database, this amounts to almost one third of the total sample (Puschmann et al., 2014). The sample consists of 8,398 individuals for which we observe records of both birth and death (i.e 25% to the total cases). This figure is relatively low for two main reasons, firstly those individuals who were born before and/or died after the sampling period, but also secondly due to the high level of migration within the city of Antwerp during this sample period.

The marriage certificates include information for 2,118 couples, equivalent to 6% of the total sample. This figure is not representative of the total number of married couples within the COR*-sample, and is an underestimate. Firstly, many brides tended to marry in their birthplace (outside greater Antwerp), and many individuals married before the sample period. This means there are likely to be a proportion of individuals who are married, but not observed as being married due to the lack of a marriage certificate.

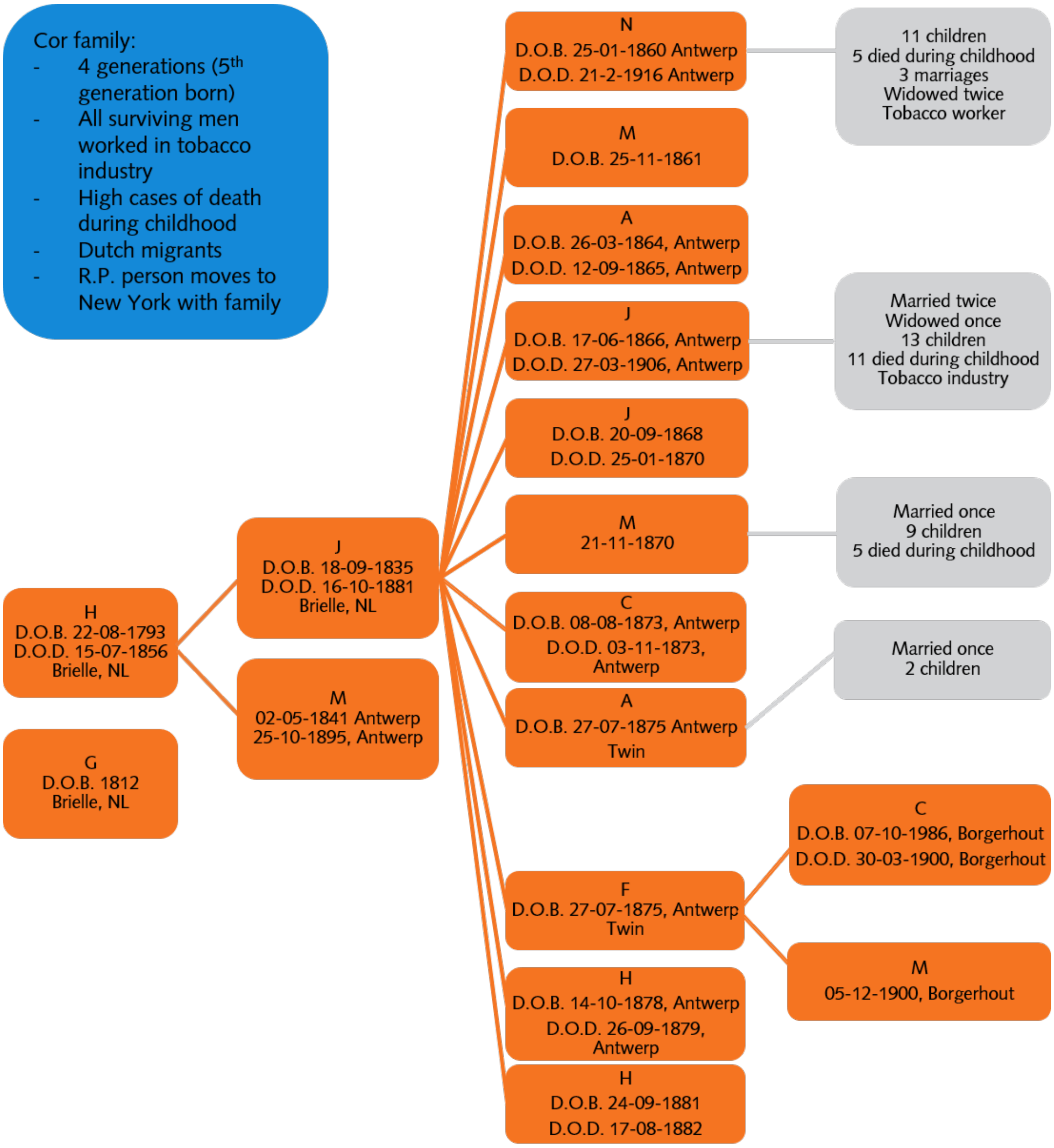
A further unique characteristic of the Antwerp COR*-database concerns occupational level information, recorded from population registers and also vital registration records (i.e. parental information from birth, marriage and death certificates). Depending on the individuals, it is quite often possible to capture at least two jobs for the occupational trajectories during their lifetime. This occupational level information allows for the study of social mobility changes at the micro level, but also trends in social class stratification at the macro level. We believe this could be highly insightful during the rapid economic development of the port city of Antwerp. In addition to this, the sample includes multi-generation families allowing the study of topics such as the intergenerational transmission of demographic behaviour.

3.4.4 MICRO EXAMPLE: LIFE HISTORIES IN THE INTERGENERATIONAL HOUSEHOLD

Another strength of the Antwerp COR*-database is the depth and number of intergenerational links between individuals. The life course of J C (IDNR 11852) is a good, rich and detailed example of this type of quality. As with many families within the COR*-database, several generations are present. In this particular case at least four generations of this family are included. Figure 9 represents individuals in an intergenerational household.

J was born in 1835, in Briele, the Netherlands, and died in 1881. He migrated to Antwerp where he married M. She was born in 1841, Antwerp, and died in 1895 in Antwerp. The information contained within the database encompasses also his parents, including his father H C, born in 1793 and died in 1856, and G G, born in 1812, but no date of death, who were both from Briele, the Netherlands.

Figure 9 *Life histories across generations*



J and M had eleven children during the course of their marriage, five of which died before their fifth birthday. They had eight boys and three girls, including one set of twins. Of their six surviving children we have fertility and occupational histories of five. What is striking about the third generation is also the high rate of child deaths, but also the occupations of the parents, who like J, worked in the tobacco industry. Three of the six individuals of the third generation had at least five of their children die during childhood. N C married three times, is widowed twice and fathers eleven children, five of which died during childhood. J J C marries twice, is widowed once and has thirteen children, eleven of which died during childhood. Both of them worked in tobacco manufacturing. M C married once and mothers nine children, five of which died during childhood. Of the other two, childhood survival is better. The twins A C and F C both have two children who are still alive at the end of the sample. F C and his family, like many in the database, migrate to New York in 1907.

The story of this family highlights two aspects of the COR*-database which are particularly interesting. One includes the richness of the transnational migrations recorded in the database and the other the quality of information concerning infant mortality.

This is an interesting research topic and important feature of the database, which has been investigated in previous empirical research (Donrovich, Puschmann, & Matthijs, 2018). This analysis examined intergenerational transmission of levels of infant mortality risk from grandmothers to mothers and its familial determinants. A further study analysing this topic using this database highlights the potential to conduct cross-national analysis in different context settings (Broström, Edvinsson, & Engberg, 2018; Quaranta, 2018; Sommerseth, 2018; van Dijk & Mandemakers, 2018). The infant mortality research conducted for five cross-national populations is a good example in this direction, for instance (Quaranta et al., 2017).

4 DISCUSSION AND CONCLUSION

This paper discussed the newly constructed the 2020 IDS release of the Antwerp COR*-database, which is a transformed and harmonized historical demographic database in an internationally standardised format that is designed for use in cross-national, comparative, demographic analysis. The database covers up to three generations of families, and contains micro-data on individual level life courses, and address-based household compositions (e.g. names, age, gender, relationships to the heads of households). The sample benefits from the inclusion of a large migrant population, recording the timing of demographic events, living arrangements, and occupational positions and trajectories during their stay in the district of Antwerp. It also includes detailed geographical level information about the city and arrondissement. The data is well-suited to understand the processes of changing demographic behaviour in the context of the first demographic transition.

The paper discusses in detail the processes of preparing the 2020 IDS release of the Antwerp COR*-database. This falls into three broad areas. We began with a thorough evaluation of the pre-existing 2010 release of the Antwerp COR*-database. The benefits of this exercise are to provide evidence and reassurance to researchers as to the strengths of the underlying linkage structures of the database. Our results confirmed the strengths of the 2010 release of the database by examining: i) the consistency of key variables from different sources for each individual in the database; ii) the strengths of the individual linkage; and iii) the consistency of the intergenerational linkage. The second purpose was to identify important areas of development for this latest release of the Antwerp COR*-database which would be beneficial to researchers of historical demography.

The second area of development was the geocodification (i.e. coordinate system (street centroid) for the location of the addresses) of the database. The inclusion of geographical information in the database offers an important new depth to the database and a critical area for potential future analysis. These new research areas which can be explored include the role of neighborhood characteristics in topics such as infant mortality, marriage rates and changing fertility patterns. An example of research in this area is provided by Ekamper and van Poppel (2019). They use a GIS dataset of Amsterdam to illustrate the effects of sociodemographic characteristics, residential environment, as well as health and sanitation conditions on infant mortality and stillbirth outcome. We believe that the addition of this new asset to the 2020 IDS release of the Antwerp COR*-database would allow similar insightful research for the Antwerp arrondissement.

The third area of development was to harvest new relationship information between individuals within the 2010 release of the Antwerp COR*-database. We did this by means of the address based reconstruction of households. This was carried out by developing a theoretically and empirically tested algorithm to reconstruct households in order to obtain relationships within the household. This method has allowed us to identify relationships beyond what was found in the 2010 release of the database. This will be highly important for future research using the Antwerp COR*-database giving an enriched scope and breadth to the intergenerational links and kinship networks between individuals within the database.

The final exercise was to convert the database into the Intermediate Data Structure (IDS) by preparing all necessary input variables and encompassing these new developments to the database, which have occured since the 2010 release of the Antwerp COR*-database. The benefits of this are the increased accessibility and ease of use of the database. In order to promote cross national historical demographic research, the goal of database administrators has to make our datasets available in as simple and coordinated fashion as possible. This indeed is the only way to make the process of cross-national research, using complex and disparate historical data sources more simple.

Whilst we believe we have made important steps forwards with the latest version of our database, there are some issues that are important to consider in the new 2020 IDS release of the Antwerp COR*-database. In the first place, while migrants (i.e. non-Antwerp and foreign born) are included in the sample, this group may not be fully representative when the entire migration movements (e.g. flows and stocks) including the temporal and seasonal movements that are not well understood. More specifically, it is considered that long-distance migrants are under represented. This is because the letter sample and the choice of COR, though more sensitive to foreign languages than other potential choices, may still not be as truly representative of non-native names as native names. This may mean that there are some unobserved sampling biases, affecting which migrants (i.e. types and duration of stay among migrants) are likely to be included. One may note however that the COR*-sample includes a relatively high proportion of non-natives in it.

A second issue concerns a unique difficulty with the 2010 release of Antwerp COR*-database and the nature of Flemish historical sources. As has been discussed, the population register presents a number of difficulties in dating much of the information it contains. The best example of this is occupations. Many observations are only recorded at the beginning of each book, which covers a 10 year window. Others are recorded at the timing of events within the population register. The timing of an event is only known for one point of time which it is recorded. This presents several difficulties for accurately registering occupational trajectories which occurred between the recording of demographic events in the register of the opening of a new book, which could be up to 10 years. There was no requirement to register a change of employment, and as such only intermittent observations of employment are possible. For this reason we do not record periods of observation for occupations as an exact date. This we believe allows researchers working with life course occupational trajectories, who will know considerably more about the relevant assumptions to apply, the freedom to methodologically develop this with assumptions themselves. There is also the unresolvable problem that having an occupational title does not give information about the employment or unemployment status of a person.

A third important consideration concerns the household composition method that has been used to identify familial relationships. The current method (i.e. based on age differences, gender, addresses and names) has been based around simultaneous migration events. Within the database we have a number of other events which provide address based information, which in future may be used to further develop this method and increase the number of familial links even more. Nevertheless, we have not used these events here, because they are events that involve only one person, such as births or deaths. An event like divorce that affects two persons was not used this time. These types of events can be used to configure the composition of a family unit once it is formed, and may help to shape it throughout time. Still, for the process of its reconstruction, a one-individual event does not contribute to identifying multiple family members and their roles, and neither does divorce. This is an important area of future work within this database which we hope can be explored further.

Another important potential area for future development concerns information about marriage witnesses. This information is contained within the marriage certificates about the names of people who witnessed the marriage and their relationship to the bride or groom, as well as some demographic information. This marriage data is a potentially very useful way to establish non-biological links with other members of the household, including family members with different surnames. This is, therefore, a highly important future piece of work which we hope can be undertaken.

Our article has important implications for similar types of research in preparing new versions of historical databases. In undertaking this work we have sought to develop and improve a pre-existing historical demographic database in order to make it more valuable, readily available and easier to use in cross national historical demographic analysis. We believe this exercise provides a useful framework for other historical database administrators. Our example highlights both the challenges, opportunities and methods which may be of use in the task of advancing and developing a pre-existing historical demographic database. We hope that by making and improving this data on the historical population of Antwerp during the 18th to early 20th century, we have provided better resources for both the analysis of historical populations, and also the analysis of modern society concerning its origins and its current demographic complexities. Because of the privacy law in Belgium, the research use on individual data needs to be anonymised. Users also need to register with the research unit. The recent development of Belgium Privacy Law, shortening the individual privacy protection from all vital events to protect from 100 years to different years of protection dependending on the type of vital events: from 100 years (birth acts) to 75 years (marriage acts) and 50 years (death acts). This allows researchers to have much wider access to longitudinal data on vital registration records. The arrival of open digitized big data on demographic records opens up huge opportunities for data access and research on the one hand, and a need to constantly include and upgrade individual records into standardized format on the other. This means it is perfectly possible to extend the scope of the 2020 IDS COR*-database into a much longer historical coverage that enables the examination of the COR*-families across additional further generations**.** For more information about the conditions to use the 2020 IDS COR*-database, see the Appendix.

ACKNOWLEDGEMENTS

This research has been presented at the European Historical Demographic Conference (ESHD) in Pécs, 26–29 June 2019. The research is funded by LONGPOP Methodologies and data mining techniques for the analysis of Big Data based on Longitudinal Population and Epidemiological Registers (EC-Marie Sklodowska-Curie grant agreement No. 676060) and Geconcerteerde Onderzoeksactie (GOA) project New approaches to the Social Dynamics of Long-term fertility change (GRANT GOA/14/001). In producing the 2020 IDS COR*-database, we would like to thank Mr. Iason Jongepier and the GIStorical Antwerp project for providing access to historical data used in the georeference process of the 2010 release of the COR*-database. We thank Dr. Francisco Viciana of the Institute of Statistics and Cartography of Andalusia (IECA) for sharing the R code that served as the basis for the conversion of the database to IDS format. We also thank Dr. Johan Dambruyne, Chief of State Archive in Antwerp-Beveren in providing access to the original source and giving useful insights on the original source materials. And finally, we are thankful for the substantial comments provided by Professor Dr. Kees Mandemakers, Professor Dr. George Alter and Dr. Paul Puschmann in improving this article.

REFERENCES

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

Bourdieu, J., Kesztenbaum, L., Postel-Vinay, G. (2014). *L'enquête TRA, histoire d'un outil, outil pour l'histoire: Tome 1 (1793–1902).* Paris, INED: Classiques de l'économie et de la population

Broström, G., Edvinsson, S., & Engberg, E. (2018). Intergenerational transfers of infant mortality in 19th-century northern Sweden. *Historical Life Course Studies, 7*(2), 106–122. Retrieved from http://hdl.handle.net/10622/23526343-2018-0005?locatt=view:master

Contiero, P., Tittarelli, A., Tabliabue, G., Maghini, A., Fabino, S., Crosignan, P., & Tessandori, R. (2005). The Epilink record linkage software presentation and results of linkage test on cancer registry files. *Methods of Information in Medicine, 44*(1), 66–71.

Donrovich, R., Puschmann, P., & Matthijs, M. (2018). Mortality clustering in the family. Fast life history trajectories and the intergenerational transfer of infant death in late 19th- and early 20th-century Antwerp. *Historical Life Course Studies, 7*, 47–68. Retrieved from http://hdl.handle.net/10622/23526343-2018-0006?locatt=view:master

Ekamper, P., & van Poppel, F. W. A. (2019). Infant mortality in mid-19th century Amsterdam: Religion, social class, and space. *Population, Space and Place, 25*(4). doi: 10.1002/psp.2232

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183–1210. doi: 10.1080/01621459.1969.10501049

Geopunt.be. (n.d.). *Atlas der Buurtwegen (1841). Vandermaelen kaarten (1846–1854) and Popp kaarten (1842–1879)*. https://www.geopunt.be/. Accessed at June 27, 2020.

Haber, M. (1984). Algorithm AS 207: Fitting a general log-linear model. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 33*(3), 358–362. doi: 10.2307/2347724

Hedefalk, F., Harrie, L., & Svensson, P. (2014). Extending the intermediate data structure (IDS) for longitudinal historical databases to include geographic data. *Historical Life course studies, 1*, 27–46. Retrieved from http://hdl.handle.net/10622/23526343-2014-0003?locatt=view:master

Jenkinson, S., Matsuo, H., & Matthijs, K. (2017). *COR technical note for the construction of intermediate data structure (IDS)*. (LONGPOP research output 7.1.). LONGPOP. Retrieved from http://longpop-itn.eu/wp-content/uploads/2018/05/S.Jenkinson_COR_technical_note_construction_IDS.pdf

Jenkinson, S., Matthijs, K., & Matsuo, H. (2019). *COR*-IDS progress report: Inter-generational linkage*. (LONGPOP research output 7.3/4). LONGPOP. Retrieved from https://limo.libis.be/primo-explore/fulldisplay?docid=LIRIAS2818367&context=L&vid=Lirias&search_scope=Lirias&tab=default_tab&lang=en_US&fromSitemap=1

LONGPOP (n.d.). Website: http://longpop-itn.eu/.

Loyen, R. (2003). Throughout in the port of Antwerp (1901–2000): An integrated functional approach. In R. Loyen, E. Buyst & G. Devos (Eds.), *Struggling for leadership: Antwerp-Rotterdam Port Competition between 1870–2000* (pp. 29–61). Heidelberg, Germany: Physica-Verlag.

Matthijs, K., & Moreels, S. (2010). The Antwerp COR*-database: A unique Flemish source for historical-demographic research. *The history of the family, 15*(1), 109–115. doi: 10.1016/j.hisfam.2010.01.002

De Mulder, W., & Neyrink, W. (2014). *Documentation construction IDS database with Antwerp COR*-data.* (WOG report Historical Demography WOG/HD/2014-1). Leuven: CeSO. Retrieved from https://soc.kuleuven.be/ceso/fapos/nasdltfc/files/WOG2014-1OmzettingCORnaarIDS_09012015.pdf

Paiva, D. (2019). *Geocoding COR*-Antwerpen Database*. (LONGPOP research output). Retrieved from http://longpop-itn.eu/wp-content/uploads/2019/07/D.Paiva_Geocoding_COR-2_database.pdf

Puschmann, P., Gröberg, P.-O., Schumacher, R., & Matthijs, K. (2014). Access to marriage and reproduction among migrants in Antwerp and Stockholm. A longitudinal approach to processes of social inclusion and exclusion 1846–1926. *The History of the Family, 19*(1), 29–52. doi: 10.1080/1081602X.2013.796889

Puschmann, P. (2015). *Social inclusion and exclusion of urban in-migrants in northwestern European port cities. Antwerp, Rotterdam & Stockholm ca. 1850–1930.* (PhD thesis, KU Leuven). Retrieved from https://www.researchgate.net/publication/287106072_Social_Inclusion_and_Exclusion_of_Urban_In-Migrants_in_Northwestern_European_Port_Cities_Antwerp_Rotterdam_Stockholm_ca_1850-1930

Quaranta, L. (2018). Program for studying intergenerational transmissions in infant mortality using the Intermediate Data Structure (IDS). *Historical Life Course Studies, 7*, 11–27. Retrieved from http://hdl.handle.net/10622/23526343-2018-0010?locatt=view:master

Quaranta, L., Broström, B., van Dijk, I., Donrovich, R., Edvinsson, S., Engberg, E., Mandemakers, K., Matthijs, K., Puschmann, P., & Sommerseth, H. L. (2017, April 27). Intergenerational transfers of infant mortality in historical contexts: a comparative study of five European populations. Paper presented at the Population Association of America, Chicago. Retrieved from https://paa.confex.com/paa/2017/meetingapp.cgi/Paper/15094

Sariyar, M., & Borg, A. (2010). The record linkage package: Detecting errors in data. *The R Journal, 2*(2), 61–67. Retrieved from https://journal.r-project.org/archive/2010/RJ-2010-017/index.html

Sariyar, M., & Borg, A. (2016). *R package 'record linkage*'. Package retrieved from https://cran.r-project.org/web/packages/RecordLinkage/index.html

Sommerseth, H. L. (2018). The intergenerational transfer of infant mortality in Northern Norway during the 19th and early 20th centuries. *Historical Life Course Studies, 7*, 69–87. Retrieved from http://hdl.handle.net/10622/23526343-2018-0008?locatt=view:master

Statistics Belgium. (2020). *Geographische indeling (geographical information)*. Retrieved from https://statbel.fgov.be/nl/over-statbel/methodologie/classificaties/geografie. Accessed on September 4, 2020.

Van Baelen, H. (2007). *Constructie van een historisch-demographisch longitudinale database: Methodologie van de Demographica Flandria selecta.* Leuven: CeSO.

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

vande Weghe, R. (1977). *Geschiedenis van de Antwerpse straatnamen*. Antwerp: Mercurius.

Viciana, F. (2020). *viciana/RTransposer: Import data from R data.table Github*. Retrieved from https://github.com/viciana/RTransposer. Accessed on November, 5, 2020.

Vrielinck, S., Wiedemann. T., & Deboosere, P. (n.d.). *HISGIS België 1800–2000*.

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354–369. Retrieved from https://files.eric.ed.gov/fulltext/ED325505.pdf

Winter, A. (2009). *Migrants and urban change. Newcomers to Antwerp, 1760–1860.* London: Pickering & Chatto /Routledge.

APPENDIX

Conditions for public data use of the 2020 IDS release of COR*-database following the rules for the 2010 release of COR*-database.

1. Ensure anonymity: all variables for family names must be removed
2. Identification of COR*-names: sample units with COR names must be clearly indicated.
3. Consultation with the research group: to ensure and register clearance on privacy issues to ensure anonymity and research topic.
4. The gatekeeper remains Professor dr. Koen Matthijs, head of the research unit.