Gerrit Bloothooft · Peter Christen
Kees Mandemakers · Marijn Schraagen
Editors

Population Reconstruction

Springer

Population Reconstruction

Gerrit Bloothooft · Peter Christen
Kees Mandemakers · Marijn Schraagen
Editors

Population Reconstruction

Springer

Editors
Gerrit Bloothooft
Utrecht University
Utrecht
The Netherlands

Peter Christen
The Australian National University
Canberra, ACT
Australia

Kees Mandemakers
International Institute of Social History
Amsterdam
The Netherlands

Marijn Schraagen
Leiden University
Leiden
The Netherlands

ISBN 978-3-319-19883-5 ISBN 978-3-319-19884-2 (eBook)
DOI 10.1007/978-3-319-19884-2

Library of Congress Control Number: 2015942214

Springer Cham Heidelberg New York Dordrecht London

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

Preface

People shape societies. They are linked to each other by family ties and networks with social, economic and religious dimensions. People live together in households and form communities. Some own a house, land and other properties, often related to their profession. And all this is in continuous change. People are born, marry, have children and die, and they change houses and addresses, and build careers. For the study of a society in all aspects, people are at the heart of the problem and should be known in the context of their complex relationships. Even today, it is not easy to get this information in an all-enfolding way, but for populations in the past, it is a real challenge. And that is what this book is about. The book addresses the problems that are encountered, and solutions that have been proposed, when we aim to identify people and to reconstruct populations under conditions where information is scarce, ambiguous, fuzzy and sometimes erroneous.

It is not a single discipline that is involved in such an endeavour. Historians, social scientists, and linguists represent the humanities through their knowledge of the complexity of the past, the limitations of sources and the possible interpretations of information. The availability of big data from digitised archives and the need of complex analyses to identify individuals require the involvement of computer scientists. With contributions from all these fields, often in direct cooperation, this book is at the heart of digital humanities and hopefully a source of inspiration for future investigations.

The process from handwritten registers to a reconstructed digitised population has three major phases which shape the three sections of this book. The first phase is that of data transcription and digitisation while structuring the information in a meaningful and efficient way. Little of this phase can be automated. With archives that comprise easily tens of millions of records, the help of volunteers for transcription and digitisation is indispensable, but requires a rigorous management. Experiences from Denmark demonstrate the complexity of this task in Chap. 1. Spelling variation, aliases, abbreviations, errors and typos all generate difficulties in further processing and require data cleaning. Similarity measures can be helpful to identify variants on the fly in further processing, but standardisation of variants of geographical locations, occupations and names—addressed in Chaps. 2, 3 and 4—can make data processing much more efficient, while identifying variants that are not similar at all. Automatic procedures can be helpful for standardisation but generally require expert review.

In the second phase, records that refer to the same person or persons are identified by a process of linkage. Advanced methods for record linkage are reviewed in Chap. 5, with reference to privacy issues that arise when recent data sources are involved. Given the complex reasoning that can underlie genealogical reconstruction, the availability of reconstructions by genealogists in standardised *Gedcom* format can support wider population analyses. The validation and usage of this type of information are discussed in Chap. 6. Whereas family relationships can be deduced from birth, marriage and death certificates from the vital registration or parish registers, the reconstruction of wider social networks may need the analysis of other sources such as notary acts. Multi-source record linkage in this context is addressed in Chap. 7. A comparable complexity was encountered in the challenging project to reconstruct the historical population of Norway, in which data from a wide variety of sources are used. The structure of this process is presented in Chap. 8. Population reconstruction from medieval charters is only possible for the very limited group of people with property worth mentioning in the charters. Probabilistic record linkage on the basis of context information is attempted to arrive at reconstruction in Chap. 9.

In the third and final phase, the information on an individual is combined into the reconstruction of a life course. Whereas record linkage usually focuses on matches between two records or two events, here the full life cycle is taken into account. Catalonia has a unique collection of marriage licences from over 450 years (1451–1905). In Chap. 10, this data collection is analysed to investigate how to utilise this information to reconstruct lifespans in the sixteenth and seventeenth centuries. For many countries, censuses contain key information for population reconstruction, but tracing individuals across censuses over the years is a complex problem. Chapters 11 and 12 report on results using machine learning algorithms for comparisons between nineteenth-century Canadian census records and especially discuss the limitations of the reconstructions and the possible biases, but also the opportunities to study intergenerational social mobility. One way to support the linkage between censuses is the combination with information from the vital registration. An example of such an attempt is described in Chap. 13 for people from the seven parishes on the Isle of Skye and their residence after migration to Scotland. A special population are the 73,000 men, women and children, transported between 1803 and 1853 to the island prison of Van Diemen's Land, now Tasmania, in Australia. The description of the lives of these convicts is discussed in Chap. 14 and encompasses the full process of data collection—including crowd sourcing—linking and life course reconstruction.

The studies and examples in this book originate from a range of countries, each with its own cultural and administrative characteristics, and from medieval charters through historical censuses and vital registration to the modern issue of privacy preservation. Despite all this diversity in place and time, they share the study of the fundamental issues when it comes to model reasoning for population reconstruction and the possibilities and limitations of information technology to support this process.

Gerrit Bloothooft
Peter Christen
Kees Mandemakers
Marijn Schraagen

Contents

Part I Data Quality: Cleaning and Standardization

Part II Record Linkage and Validation

Part III Life Course Reconstruction

Part I
Data Quality: Cleaning and Standardization

Chapter 1
The Danish Demographic Database—Principles and Methods for Cleaning and Standardisation of Data

Nanna Floor Clausen

Abstract Since 2001 seven Danish censuses dating from 1787 till 1880 have been completely transcribed by volunteers. Due to this effort the research community now has access to a large number of demographic data. The census data were digitised according to the principle of literal data transcription in order to leave all interpretations to the users. The disadvantage of this solution is that it induces problems when creating aggregated statistics as the spelling of, e.g. position in household and occupations was not standardised which leads to great variation in the description of the same entities. In order to overcome this obstacle the data were cleaned and standardised. Standardisation consists of adding numeric codes for the gender, civil status and position in household. For occupations, HISCO has been applied to secure that the data can be used in comparative research.

1.1 Introduction

Every country has its own sources containing microdata, also known as nominative data, stemming from the national statistical institutions or public administration. Overall, these sources hold the same kind of information about citizens, such as name, address, age, occupation and civil status, which make it obvious to use these data in comparative analyses. The way the sources were created and later digitised has resulted in the need for standardisation of the metadata using numeric codes for overcoming the great variance in expressing the same entities and the barriers created by the diversity of language. The process of standardisation must be done locally owing to different languages, but the principles behind the standardisation and the standards used can and have been discussed internationally in workshops and at conferences. In the following the standardisation process will be illustrated by the Danish example.

N.F. Clausen (✉)
Danish National Archives, Odense, Denmark
e-mail: nc@sa.dk

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_1

The short description of the Danish Demographic Database (DDD)[1] is that it is the Internet dissemination portal for the data collected and transcribed by the Source Entry Project. The Source Entry Project can be dated back to 1992 when a group of historians, genealogists and H.J. Marker[2] from the Danish Data Archive (DDA) (from 2014 completely integrated in the Danish National Archives) discussed the ongoing work by genealogists in the archives concerning transcription of sources. In the reading rooms in the archives, a large group of individual genealogists transcribed sources for their own purpose each doing it in his/her own manner. Furthermore, the transcribed sources were kept as private documents with the result that the same sources might be transcribed several times. In 1992, computers were beginning to become more widespread and especially genealogists were rather early users of IT-technology. The observations stated above founded the basis for the structure, managing and coordination of the Source Entry Project, in Danish known as KIP. The primary initiator was the Danish genealogist society which had begun the creation of an inventory of transcribed sources and together with the DDA[3] laid the foundations of the work that is still being carried out.

The group started with three initiatives:

1. Compile an inventory of existing source transcriptions. This was done in collaboration with local archives. The result was published in print and on discs.
2. Establish a coordinating group for future source transcriptions in order to avoid transcribing the same source twice. DDA was part of the group and secured the preservation of the transcribed data.
3. Produce guidelines and models based on a thorough knowledge of the sources. A dedicated group was assigned for this task. The need for this was great as transcriptions on a computer involve a lot of problems where the solution prior to these instructions depended on the individual person and his/her purpose of the transcription.

The design of the models has had great impact on the later work of cleaning and standardisation. The original instructions distinguished between a basic model designed for rigid and accurate source transcription and an expanded model for the more advanced users. The expanded model had the same fields as the basic model but with four additional fields: three for names (given name, patronymic and family name or surname) and the fourth for supplying a personal identifier. In the additional name fields the volunteers were allowed to make interpretations and abbreviations as long as they were explained in the documentation. It was furthermore allowed to make normalisation or standardisation using their own rules provided they were explained in the documentation. It soon became clear that only the basic model was used and this is still the only one used. The different source entry

[1] http://ddd.dda.dk

[2] H.J. Marker, senior researcher at DDA 1984–2009. Presently director of Swedish National Data Service.

[3] Danish Data Archive became a member of the Danish National Archives in 1993.

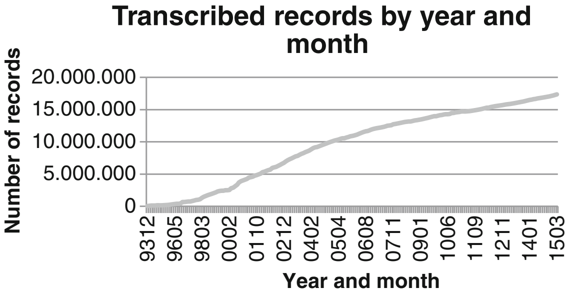


Fig. 1.1 Development in number of transcribed records

programs that have been in use since 1993 have all been developed only for the basic model.

The guidelines for the transcription have as the general rule that everything is transcribed as it is written in the sources. The exceptions for not meticulously transcribing every letter from the sources are obvious errors in the sources, abbreviations should be replaced with the full spelling and the information on marital status where it was decided that the status was more important than the spelling. In the earliest transcriptions we can still find many variations of, e.g. the word for 'Married' ('*givt*', '*giift*') as an example of how difficult it is to make the volunteers adhere to the instructions. If the volunteer had problems reading the source the illegible characters or words are replaced by two question marks: '??'. This makes it possible for other users to identify difficult places and hopefully help reading those characters or words. Two '!!' indicates that the volunteer has been in doubt about the content in the sources and the volunteer informs about the problem in the field for commentaries.

The guidelines were originally published in an issue of the DDA Newsletter in 1993 and since 1996 they have been published on the website for the DDD.[4] The guidelines are updated according to the need of the users or from experience with the transcribed data received at DDA as, e.g. when we discover that too many are making the same mistakes or making wrong interpretations. Updates are done at large intervals—on average every 5 years.

In 1996 the amount of data delivered to the Source Entry Project, primarily from censuses, had reached a level that made it interesting to make the data searchable on the Internet. The website was named DDD, ddd.dda.dk, as it was launched with data from the censuses and emigration records from Copenhagen Police along with links to scanned images of the sources. Since 1996 many more source types have been transcribed and published on the website, which is still open for new types of transcribed sources.

In 2000, it became clear that a dedicated effort in the 1801 census would make it possible to complete the transcription in 2001. This goal was achieved and since then six more censuses have been completely transcribed. The work on cleaning

[4] http://ddd.dda.dk/Vejledning%20i%20kildeindtastning.htm

and standardising those censuses is the focus for this chapter. The censuses are from the years: 1787, 1801, 1803 (Duchy of Schleswig), 1834, 1840, 1845, 1850, 1880 and 1885 that was only taken in Copenhagen. Figure 1.1 illustrates the number of data transcribed since the end of 1993, showing no decrease in the interest in transcribing historical sources.

1.2 The Data—The Historical Censuses

The first census with nominal data was taken in 1787 and the schema and procedures from this census was conducted in such a way that it laid the foundation for the following census takings. One major change, though, was to take the census on 1 February instead of 1 July.

The questions asked in the censuses varied somewhat with the one in 1834 having the fewest questions (or fields) to the most encompassing in 1901 and 1916. Table 1.1 gives an overview of the common fields and in what years they were on the census schemes.

The table only covers the censuses from the nineteenth century which is the period from which we have completely transcribed data. The general fields about name of parish, district and county were given in the census ledgers in the beginning of each parish and in the transcription are automatically added to each record. In Denmark, nearly all the original censuses have been preserved with the following largest exceptions: in 1834 the data for most of Copenhagen have been lost and in 1840 data for 11 towns are lost. The original sources for the censuses and parish registers have been scanned and are publicly available on the Internet: http://www.sa.dk/content/dk/ao-forside. This service is very popular and is a great help for the Source Entry Project.

Table 1.1 Information in the censuses by year and field

	1787	1801	1834	1840	1845	1850	1855	1860	1870	1880	1885
Address	X	X	X	X	X	X	X	X	X	X	X
Name	X	X	X	X	X	X	X	X	X	X	X
Gender										X	X
Age	X	X	X	X	X	X	X	X	X	X	X
Marital status	X	X	X	X	X	X	X	X	X	X	X
Religious community							X	X	X	X	X
Place of birth					X	X	X	X	X	X	X
Position in household	X	X	X	X	X	X	X	X	X	X	X
Occupation/poor relief	X	X	X	X	X	X	X	X	X	X	X
Handicaps					X	X	X	X	X	X	X

Table 1.2 Overview of censuses by year divided on units, population and coverage

	Official population	Number of materials/units	Transcribed records	Coverage in %
1787	841,806	1780	840,989	99.90
1801	925,080	1789	932,055	100.8
1834	1,223,797	1771	1,134,401	92.70
1840	1,289,075	1778	1,256,369	97.46
1845	1,350,327	1790	1,354,150	100.3
1850	1,414,648	1784	1,403,440	99.21
1855	1,489,850	794	594,972	39.94
1860	1,608,362	1766	1,455,985	90.53
1870	1,784,741	614	508,537	28.49
1880	1,969,039	2204	1,979,640	100.5
1885	280,054	972	326,731	116.7
1890	2,172,380	991	1,028,521	47.35

One way of engaging volunteers is a weekly updated list of the progress divided by year. In September 2014 the coverage per year is as shown in Table 1.2. A closer look at the coverage—which is a simple calculation of the population from the official statistics divided by the number of records we have received—illustrates when original data has been lost (like in 1834) or when more people have been included in the censuses than it was supposed (in 1885, parts outside Copenhagen had been counted). For the other censuses there are some minor differences. We have used the official statistics intensely when a census was completely transcribed, where we made a control parish by parish against the statistics. When more than e.g. 25 persons were either missing or were in surplus in the transcription, this parish was controlled against the original source. The table is taken from the DDD website and follows the progress in the project back to 1994. The number of materials is not the same as the number of parishes as the biggest cities are often transcribed streetwise in order to make the work manageable. The percentage of coverage may be more than 100 %, as it is for 1845, e.g. because the official statistics had to count and calculate the data manually and therefore minor differences occur. In 1885 the discrepancy is the biggest, because the enumerators included parishes that should not have been included.

The reason why there is something transcribed for all the censuses with free public access[5] is that the volunteers themselves decide which parish, street and in which year they want to transcribe a source. The steering committee for the source project has since the completion of the 1880 census in 2012 focused on the censuses for 1860 and 1901. 1860 is the last census where the Duchies of Schleswig-Holstein were still a part of Denmark, and 1901 because it is not written in the gothic handwriting and it is from a time when industrialization had really taken off.

[5] A census must be at least 75 years old before every one can get access to it.

The content of the censuses was generally of high quality, although the general educational level is reflected in the censuses in the sense that, e.g. the perception of stating the age clearly improves during the nineteenth century. In the rural areas it was the same person in the parish, normally the vicar, who filled out the census schema and thus making the schemas consistent within a parish. In the urban areas it could be either a central enumerator or from 1880 it was often the owner of the house or the caretaker that filled out the schemas.

Since 1993, there has been a series of source entry programs all developed by volunteers as DDA has not had the resources for developing any applications for the project. The quality in the programs regarding how strictly they reflected the census schemas has been greatly improved with the latest application. The earliest transcriptions do not have the same quality as the later ones as the used applications did not differentiate between the schemas for rural and urban areas, fields were missing in the application (like the information on number of windows—asked in 1880 Copenhagen) but when the data are used in aggregations this is of minor importance (although complicating the work of standardisation). Presently, online transcription is emerging which is very closely linked to the scanned sources diminishing the above-mentioned deviations from the sources.

Researchers get access to the standardised data by contacting the Danish National Archives using the contact information on the DDD website. The SQL-scripts and programs used for the standardisation processes are likewise available in the same way. The programs are made specifically to the database structure used in this project but they can be useful as inspiration and can be modified for other databases.

1.3 Data Cleaning and Standardisation

The ambition of the source entry project has been to create as source a loyal transcription as possible based on literal transcription. The advantage is that the interpretation is not done by the person doing the transcription but is open for everybody using the data. The disadvantage is that it is not possible to use the data as they are for statistical or analytical purpose. When a census is completely transcribed a copy is made which will then be the one used for the necessary cleaning and coding, hereafter referred to as the 'research database'. The procedure of standardising the 1801 census has been described by Marker (2001, 2006). The standardisation process might be called a 'top-down' process as we begin with the highest level, the parish, as this information is given once for every parish and is the same for all the inhabitants. Following this standardisation comes the cleaning and standardisation of the questions in the census schemas. So far we have had neither the intention nor the time for coding, e.g. handicaps or religious community.

The purpose of cleaning is to eliminate variations in spelling. Standardisation consists in replacing the text with a number, a code, which refers to a common description or use of a common description which refers to a number. Before the tasks of cleaning and standardisation are begun it is worthwhile to make a strategic plan. The jobs should be developed in a way that makes them reusable and extendable. The use of referencing tables that hold the common descriptions and codes for ideally each field is recommended. For documentation and later secondary analyses these references are necessary resources.

The data, about one million records per census for the earliest, and almost two million records for 1880, are preserved and managed in a Microsoft SQL database server using all possible features: reference tables, relations between tables, views, stored procedures and functions. One table per census holds the data with the fields from the census schema. To these tables are added fields required for the standardisation as explained in detail in the following. A large variety of specific SQL-programs have been developed in order to automate the coding tasks as much as possible but the process still involves many manual steps, especially when coding position in household and occupation. Not surprisingly, the SQL-jobs require more time to run the larger the table is so that, e.g. when it takes about 1 h to run the coding job for occupations in 1787 the same job runs for almost two whole days for the 1880 census—with more occupations and records to standardise.

The standardisation work began with the 1801 census, which was the first to be completed and the first to be completely coded in-depth for all variables. The next was the one for 1787. The jobs developed for this census have since been used for the following censuses and the coding jobs for each variable have been improved with the new values appearing in those censuses making the jobs larger and larger. The benefit is that one can always use these jobs and only has to add new values; one does not have to start from scratch.

The result of data cleaning is the great reduction in variation:

- 2200 different ages in 1880 were reduced to 174 (ages are given as decimals, e.g. 1 year and 3 months is converted to 1.250 and 77 1/4 as 77.250)
- 39,800 household positions in 1801 to 273
- 43 different civil statuses in 1880 to 6
- 63,000 occupations in 1801 reduced to 2735 occupations that refer to 1135 unique numeric codes.

1.3.1 Management of Place Names

The coordination of the source entry project has included a standardisation of the place names from the beginning. In the censuses a place or parish may be spelled or called something that is different from today. We have applied the administrative division used from 1920 till 1970. Each municipality was given a number from the ministry of internal affairs and the spelling is based on the latest handbook on place names in Denmark (1963). Over time it has been necessary to expand the list of place names and related codes as data entry has been made streetwise in Copenhagen and in a few other large towns. The table with place names and codes is also distributed to persons developing data entry programs so everybody is referring to the same standard.

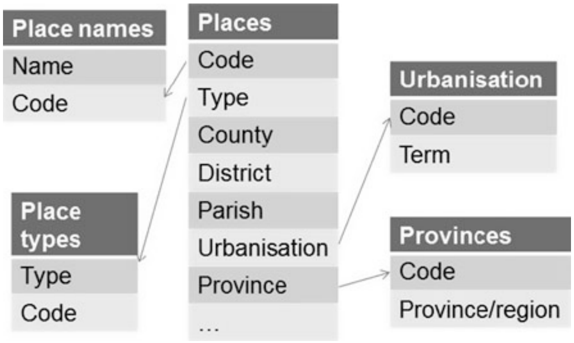


Fig. 1.2 Relational tables managing references to and between place*s*

The tables in Fig. 1.2 show the relations between the different administrative levels. SQL restrictions are applied in order to avoid creating a new place name in the wrong way or forgetting to put it into the proper hierarchy. The table 'place types' refers to the hierarchy: Denmark, part of kingdom of Denmark, county, parish, or street. 'Urbanisation' refers to the three values: Copenhagen, town, or rural district. Each transcription unit is given a unique number (one letter + 4 digits) and at the same time the unique code for parish or street is preserved in the management system for the project. For analysis it is consequently always possible to immediately determine the relevant geography and administrative information about the place where a person lived.

1.3.2 Personal Identifier

In the censuses the enumerators were asked to give each person within the household a record number but this was not done consistently in all censuses and if it exists it is only useful within a household. Each record is uniquely identified using the source entry number and the record number added by the program is used for each entity. In the copy made of the complete census a new record number, an ID-number, is added to each record. This ID is unique through the whole census and not just within an entity. At the same time a field called 'number of persons' is added and the value is given as 1. This value is changed to 0 when an empty house is found. For each census the ID-number begins with 1 so the number must not be mistaken for a unique personal identifier for a person. Linking of persons requires this ID-number and the year of the census in order to make it unique.

1.3.3 Gender

With the 1880 census gender became a field in the census schemas. Earlier, the enumerators were asked to state the gender if a name was too special to immediately identify the gender. The volunteers are instructed to always give information on the gender in the field added for this purpose. Too often this is not the case and for practically all the completed censuses this information has been given in only approximately 50 % of the records. The logical possible values for gender are male, female, or unknown. But the fact is that before cleaning and standardisation there are 18 possibilities where blank and NULL is not included. An example from 1880 could look like this: -, M, (K), (M), -, ?,??, K, k., Kn, M, M!!, mm, U, X. In the Danish censuses we also have transcriptions from the Duchies of Schleswig-Holstein that are made in German. The simple thing to clean is values entered in wrong fields, removal of parentheses (which some volunteers use extensively) and the like. The bigger task is in applying gender when it is missing. The variation in women's name is smaller than for men so the standardisation is begun with women. The first part is done by looking at the names and then by looking at position in household. A name with the suffix 'datter' (daughter) is presumed to be a female. Names beginning with typical female names like 'Anna', are likewise presumed to be females. Most names are typical for one of the genders making the task simpler. The remaining records with missing codes were standardised using household position, like 'hans kone' (his wife), 'hans datter' (his daughter). If the household position contains the suffix 'mand' (man) it is normally coded with an 'M'. Information about occupation was mostly done for men and this information was also used when coding gender. The small number of records not standardised when the preceding procedures had been carried out had to be standardised more or less one-by-one. In the 1880 census only 41 records of nearly two million records do not have a code for gender.

1.3.4 Age

The census schemas before 1880 stated that age should be given as the age you were going to have at your next birthday. From 1880 it was given as achieved age and from 1901 it was the date of your birth. This information is useful when analysing the data and of less importance for the cleaning process except that the enumerators did not follow the instructions. If they had done so we would not have had the ages of, e.g. half a year, 2 weeks, etc., that should have been given as 1. The example above is the reason why the data entry program has to have age entered into a text field. Another reason is the need for using question and exclamation signs when something is illegible or must be wrong like widows at the age of 13. In the research database we add a new field for holding the age as a decimal number in order to be able to specify 2 weeks as 0.04. The age field had in 1845 more than 1000 and in 1880 more than 2000 values. Some of these were due to data entered in the wrong fields, additional characters like apostrophes or parentheses that had to be removed, '!!' or '??' that had to be controlled in the sources. All ages above 100 were also looked up in the sources, which in total means that it is not as simple to clean age as could be expected. When the cleaning has been done it is simple to convert the age into a decimal number. An empirical testing of the accuracy of the age information can be had by making an age distribution. In 1801 the age heaping around the ages ending with a zero is evident. In 1845 this has declined and is almost not visible in 1880. The graphs furthermore show even larger age heaping for women and in general the higher the age the greater the age heaping in the censuses prior to 1880.

1.3.5 Marital Status

The guidelines for data transcription state that the information for marital status should be normalised into six categories. This is likewise specified in the census schema. In 1845 the text in the census was: "Married, not married, widower or widow." In 1880 it was: marital status: Not married (U), married (G), widower or widow (E), divorced (F)." Nevertheless, marital status was entered in a large variety which was, however, fairly simple to standardise. The present version of the data entry program has a drop-down list for the status. We have transformed the information to one of six codes: (Table 1.3).

For children with missing information, living at home and at the age below 15 they are coded with not married. This is normally also the case for servants. When a servant is married it is clearly stated in the sources and the transcriptions.

1.3.6 Position in Household

In the censuses 1787–1801 and again from 1890 the position in household has its own field and heading in the schema, whereas in the censuses in between it is stated in the same field as occupation. In 1801 the enumeration instructions stated the following about this field: "*For each person is specified what it is in the household, like husband, wife, child, kin and how close, servant and journeyman and boy, lodger, quartered and so forth*". In 1880 for example, the instructions were like this: "Position in family (*head of household, housewife, children, relatives, servants, lodgers,* etc.) and title, post, business, trade or by which occupation they

Table 1.3 Possible values for marital status

Code	Marital status
0	Unexplained
1	Not married
2	Married
3	Widow(-er)
4	Separated
5	Divorced

live…" (Johansen 2004). From both examples of instructions the order of positions in the households are made clear: each beginning with the husband who is the same as the head of household, then his wife, children and others in the household. This interpretation of a household has been followed in the development of the reference (Table 1.4).

The combination of position in household and occupation in one field has turned out to have implications on both the detailed information on household compositions, and adds more work and complications to the task of standardisation as we have to split this information before it can be coded. The basis for the coding is a list of possible positions in a household where every person in a household is defined as relative to the head of household. Currently, there are 367 different codes for position in household.

The example in Table 1.4 from the reference table for position in households holds different pieces of information.

Head of household is given the code '1'. Steps refer to the steps a person is away from head of household, in this case 0 steps. The field 'family' is used for stating whether the member was a family member or employed. Generation states if a person belongs to the same generation, which is 0 if it is head of household or his wife or sibling, a child is +1 generation and parents are −1 generation from the head of household. The field 'relation' is an auxiliary variable used for detecting the number of links to the head of household. Number of links is the length of the value minus 1. 1 is the person self, 2 is the spouse, 3 is child, 4 is parent. The value 143 means sibling, that is, two links from head of household.

The database tables in Fig. 1.3 hold the information and relation for position in household. The database table 'Relation' holds the following codes and terms: (Table 1.5).

The work on standardising the position in household is initially done by coding the records that are entered correctly which is the majority. The remaining records are processed via a large SQL-job, which has been developed making use of our experience with the data. The job takes care of correcting spelling to, e.g. one term

Table 1.4 Samples from reference table for standardised positions in household

Code	Term	Steps	Family	Generation	Relation	English
1	Husstandsoverhoved	0	1	0	1	Head of household
5	Husstandsoverhoveds fællesbarn	1	1	1	13	Common child of head of household
17	Husstandsoverhoveds tjenestefolk	1			17	Servant of head of household
9	Husstandsoverhoveds forældre	1	1	−1	14	Parent of head of household

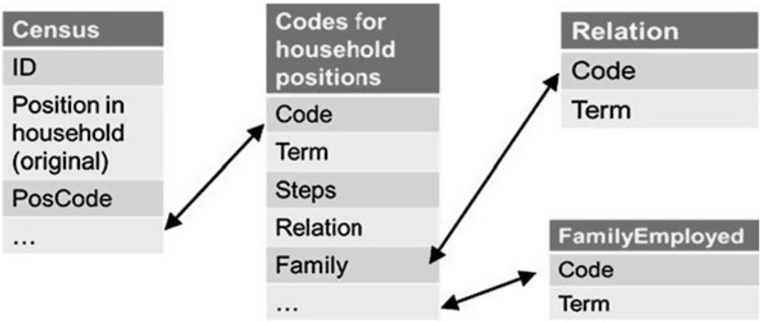


Fig. 1.3 Tables and their relations for position in household

Table 1.5 Possible relations in households

1	Head of household
2	Spouse or similar
3	Child
4	Parents
5	Other family
6	Neither family nor employed
7	Employed or co-owner

for servants, siblings, etc., and correcting typos. There is an SQL-job for each census-year, building on the original SQL job for 1801 and augmented with additional code lines for the variations appearing in the following census. The next step adds the codes mentioned above for those positions having wrong spellings. The programs that correct and standardise the data are given the original data along with the values they should be standardised into.

This job is run iteratively building a temporary dataset normalising the text and using the information from the person before as criteria for adding the code of relationship. Each iteration results in still more records that are coded correctly leaving the remainder of the records to be coded manually, a task that represents many hours of work including iterations of the mentioned SQL-job.

Linking of spouses is done by adding the constructed ID-number from one spouse to the other spouse. A prerequisite is that the households are divided correctly and that position in household is coded. When the position is 'his wife' and the person before is head of household and male then they are one another's spouse. If this is not the case the values will have to be updated by looking at the household divisions and position in household.

1.3.7 Households

In the enumeration instructions it is stated that a person belongs to the household in which he or she sleeps. This would normally not give a person or the head of household any problem of defining the members of his household. In the censuses the enumerators generally have identified each household by giving it a number or drawing a line between each family. In some of the censuses they were asked to give the number of families in each house. The entry program repeats the number of households so that the volunteer just has to change the number when a new household begins. Unfortunately this has not been done in a lot of cases. Using the information on entry number, place name, address, and if possible the household number the data have been given new household numbers—we add a field to hold this number in order not to delete the volunteers' data. The number is unique only within each census. The process of coding for position in household—especially identifying head of household—was used in the process of dividing the data into the correct households. A household could not have two 'heads' or none. If the first person is 'his wife' something is wrong. Some households are very large and are supposed to be so as is the case for hospitals and similar institutions. The selection in the data of large households helped identify households that had not been divided correctly. When we analyse household structure and family types, the large households that are of the type institutions have to be excluded. A problem with the large households as, e.g. hospitals and prisons is that apart from the persons temporarily present there are also families present like the wards, inspectors, etc., who live there with their own household but with an address at the hospital.

We have distinguished five household types and made a reference table with the following household types as shown in Table 1.6.

In a separate table we store the calculated household IDs combining the reference to household ID and household type. By default all households are of the type Family. The relationship between these database tables is shown in Fig. 1.4. These tables form the basis for analyses on households.

Table 1.6 Household types

Code	Description
0	Empty
1	Family
2	Institution
3	Soldiers
4	Without head

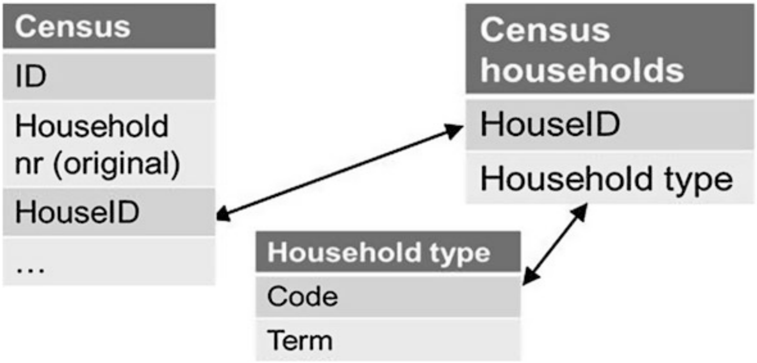


Fig. 1.4 Relational tables controlling households, types and references

1.3.8 Occupation

Standardising occupation is the most complicated task and made more complicated due to the changes in the census schemas. In 1787 and 1801 and again from 1890 position in household and occupation had their own separate fields, but in the censuses in between the question was phrased like this in the census schemas: *Position in family* (head of household, housewife, children, relatives, servants, lodgers, etc.) *and title, post, business, trade or by which occupation they live* as principal or helper (manager, journeyman or boy, etc.) or *if they are supported by the system of poor relief.*

A basic principle for cleaning and standardising the occupation was the division of all the information in the field resulting in each piece of information being preserved and this at the same time in a way that would allow analyses on both the aggregate and the individual level as the information can always be brought together again. Another principle was the construction of a reference table that could be reused for all the censuses. The referencing table holds the standard description, an internal numeric code and the referencing codes for both Historical ISCO[6] (HISCO) and North Atlantic Population project (NAPP[7]). This is explained in detail later.

The first step to do is to copy this information into a table for occupations by census-year. The tables hold fields for the ID-number, standardised text, the order in which the occupation is listed in the original source and a code for standardised occupation. The method for cleaning and standardisation of occupations was to develop a program where we list all the possible variations of spelling and occupations and add SQL codes for correcting these variations to the standard chosen.

[6] History of Work Information system http://historyofwork.iisg.nl/index.php. Codes for historical occupations based on ISCO. Used for comparative research.

[7] www.nappdata.org

The basic program is the one done for 1801. It is used initially for the coding of occupation for all the other completed censuses. The result of each teaching test is the coding of more occupations whenever a match is found and a list of records where no match in values was found. These values are hereafter added with the necessary steps (correction of the spelling, deletion of values, etc.) to the basis job and it is run again. During this process of teaching the coding job we have to decide when we have reached a sufficient level for the quality and quantity: is the standardisation of 75 % sufficient or should it be 95 % or even 100 %? The last 3–5 % not coded are very time-consuming as each remaining value normally only covers one record. On the other hand there might be interesting information left in these values. When odd occupation values are found it is recommended to control the transcribed data against the source. Unusual or unknown values may be a result of misreading of the source.

The original text is gradually converted to the standardised text via an SQL-job where we begin by correcting and harmonising the spelling, delete records that hold information on position in household and splitting occupations when more than one is listed. When the census taker has stated, e.g. *husfader, husmand, hugger og hjulmand* (head of household, cottager, woodcutter and wheelwright) the text for head of household is deleted in the first step and the rest of the text is then split into three records as shown in Table 1.7.

The principle of splitting the information on occupation to its smallest substantial parts and keeping a number for the sequence always allows us to put the text back to its original order but as a standardised expression. The benefit is that we get very detailed information on the occupational structure and status and it allows research for all kinds of purposes.

A modification and extension of the SQL program is required for every specific phrasing and spelling of occupations, which will apply to all records where this text is found in the data. For the rural areas the variation in occupations is limited, so correcting spelling for cottagers and standardising cottagers with or without land cover a large percentage. In the towns and Copenhagen the variation is greater and therefore each line of code covers ever fewer records making the final version of the job for coding occupations expanding for each census-year. The table on occupations by census-year is truncated before each iterative run of the SQL-scripts and the updating and deleting is done on temporary tables only copying the final result into the year and occupation. From this table we can select the number of each occupation, we can see which occupations are still not standardised, and according to this information add more correcting lines to the SQL-scripts and run them again until everything has been completely standardised and coded.

Table 1.7 Example of one person listed with three occupations

ID	Occupation	Number	DDACode
34567	Cottager	1	800
34567	Woodcutter	2	6380
34567	Wheelwright	3	9930

Table 1.8 Synonyms for woodcutter

Synonym	Default
Brændehugger	0
Brændeskærer	0
Huggemand	0
Hugger	1
Pælehugger	0
Skovhugger	0
Skovkløver	0
Tømmerhugger	0
Vedhugger	0

Table 1.9 The codes for woodcutter in HISCO and NAPP

DDA code	Hisco	Danish title	Title	Group	NAPP
6380	63110	Skovarbejder, almindelig	Logger (general)	6	63110

The codes for occupation are managed by reference to a table holding all the standardised text and reference codes to HISCO and the NAPP. An occupation may have synonyms but one is considered the standard that holds the numeric code and the reference to HISCO. An example of this can be seen in Table 1.8. For matter of convenience we use the descriptions for linking between the year-census table and the table of occupational codes, which makes it possible to add more occupations in the correct places and expand the Danish number codes without interfering with the relations to HISCO.

Number 1 designates that this description is the standard used in analyses, and it must be used explicitly in order to avoid that the synonyms muddle the number of records.

The NAPP-field in Table 1.9 refers to the definitions[8] made in the NAPP project and listed on the project’s website. The definition for the shown code is: 63110, Woodsmen and workers in the woods, not further specified (nfs).

The group code in Table 1.9 refers to the code in the table ‘group reference’ shown in Table 1.10. In this way it is simple to aggregate data to a higher level in the hierarchy of occupations. The number of standardised occupation descriptions is (at the end of 2014) 3154, of which 1204 represent the standard for an occupation and the others are variations and very specific descriptions.

Data from the DDD is on the way to be included into NAPP, which again will be integrated with the IPUMS project (Integrated Public Use Microdata Series).[9] The NAPP project has developed a set of occupational codes close to the HISCO codes. The codes to be used in that project have been added as an extra field into the

[8] https://www.nappdata.org/napp-action/variables/OCCHISCO#codes_section

[9] https://www.ipums.org/

Table 1.10 Example of the hierarchy grouping of occupations based on HISCO

Code	Danish term	Occupational group
6	Skovbruger	Forrester

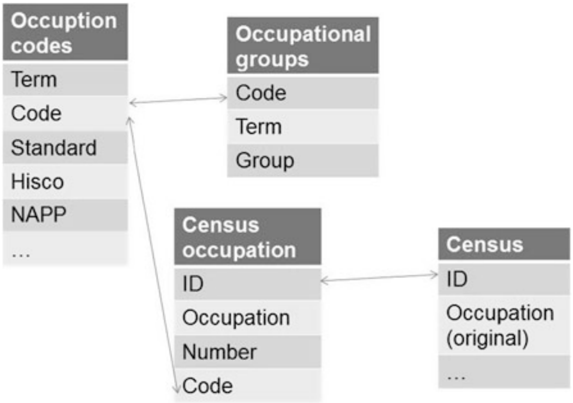


Fig. 1.5 Diagram for the tables used for referencing occupations

occupational codes-table. The concordance between the DDA codes and NAPP-codes was done by using the HISCO codes and the DDA codes in combination. In Fig. 1.5 it is shown how all three kinds of codes are in the same table making it simple to make crossovers from one code system to the other. The system is likewise well-prepared for future updates.

1.4 Discussion on Standardisation

Standardisation and cleaning of the transcribed historical census data is necessary for enabling research on the data and making the data comparable. In this chapter is described how standardisation is carried out on the Danish census data that have been completely transcribed. The only exception from not correcting from the start is the geographical reference and the adding of the record number. It could be argued that for public use (genealogical online search) standardisation of some crucial fields like age and the division of names into first and last names could be done simultaneously with the transcription, but this has been turned down for fear of too many errors and misinterpretations being introduced.

Norway had the census from 1801 transcribed many years before Denmark, which made it still more interesting to have also the Danish 1801 census completed. Solli (2003) has described how he has used the standardisation and made his own contribution to the process which is in line with the procedure used by Thorvaldsen (2006, 2012). The standardisation of the Norwegian census and the Danish census has in general been done in the same way. In both cases the encoding has been done using scripts and programs that encode each variable separately resulting in numeric codes for the variables. These scripts are enlarged with new values for each census and in Norway for each municipality in each census. In Norway it has turned out that the enumerators more often than not did not state the 'normal' occupations, like being a fisher as everybody was fishing in West Norway. The task of standardising has been done both automatically and manually as described above in the Danish example. Although the Danish and Norwegian languages are very close they are not close enough to allow using the same coding programs. The major difference between the Danish and Norwegian coding of occupations is that in Denmark every occupation listed in the field was encoded as a separate record and in Norway either the first listed occupation was coded or the string was coded. E.g.: 'fisherman, cottager' has one code and 'cottager, fisherman' another code. The Norwegian Historical Data Centre[10] began with coding the first occupation listed in a person's occupation in the census schemas as the enumerator should have underlined the most important occupation which was supposed to be the first one listed. They have since started with coding the next mentioned occupation. The data from Norway which can be accessed via the NAPP project has the same code for 'fisherman, cottager' and 'cottager, fisherman'.

In the 1990s, work on historical census microdata was taking a big step forward as more and more sources were digitised. The awareness of the growing number of smaller or larger databases with digitized microdata led in 1998 to the creation of 'The International Microdata Access group', IMAG.[11] IMAG was formed to foster the international collaboration of researchers who work with individual-level electronic data in order to facilitate transnational comparative research. The mission of IMAG is to coordinate and facilitate international standards for data providers and users, to preserve original population microdata and their supporting documentation. IMAG seeks to combine multiple data files with a common set of comparably coded variables. The IMAG group has met and is still meeting at conferences where the present situation for cooperation is discussed. Members of the group are researchers representing national microdata databases with an interest in cooperation. In 2000, a reference handbook (Hall et al. 2000) was published presenting census databases with their background, structure and size. The handbook further included an inventory and specification of the databases known in 2000. Since then the work on expanding the databases has continued as has the work of creating compatible variables.

The largest project with historical census data is NAPP and Integrated Public Use Microdata Series (IPUMS).[12] NAPP gets the standardised and coded data from their partners and afterwards harmonises the data. The NAPP project exploits the experience and coding programs developed at IPUMS when working with US census data. The big challenge for IPUMS is that the data from each country differ from the US census data their routines were based on, so they had to expand their

[10] http://www.rhd.uit.no/

[11] http://www.prdh.umontreal.ca/IMAG/

[12] https://www.ipums.org/

harmonisation jobs. With the huge amount of data they are dealing with each task that might look simple becomes complicated and time-consuming (Ruggles 2005; Roberts 2006). In order to overcome the many languages in the microdata numeric coding of the essential fields is necessary. The documentation of each variable is very extensive, which helps making international and comparative analyses reliable and doable. The documentation furthermore is done in a way that allows owners of datasets not participating in NAPP or IPUMS to create the same constructed variables and/or to implement the same coding schemes on their own data. The inclusion into NAPP and IPUMS-International of as many census data as possible from around the world tends to have a great impact on the standardisation and coding schemes in use. It has almost become a de facto requirement that the nationally encoded variables can be made compatible with the IPUMS-variables. The variables in IPUMS and NAPP are structured on several levels: household versus person, integrated variables versus unharmonized variables and constructed variables versus census questions. The NAPP project adds, e.g. family pointers that indicate the person-number within the household of each person's co-resident mother, father, or spouse.

Mandemakers and Dillon (2004) made a list of the best practices for handling large databases on historical populations. The guidelines in their article encompass the three steps distinguished when planning and creating a large database: (1) definition of the object and content of the database, (2) data entry, standardisation and storage and (3) dissemination of the data. Their recommendations on standardisation are in line with the description of the work done with the Danish censuses. Especially the separation of the literally transcribed data and the cleaned and standardised data is seen as important as we do and have described above.

In recent years the Scottish project (Kirby et al. 2015 this book, chapter 3) is working on developing a method for automatic transcriptions and especially standardisation of the data.

Throughout the years there have been several debates on the issue of the definition of a household and a family. In the Danish case we have used the households as laid out in the censuses. In the sources this is done by drawing a line in the schema between two households or an empty line, and in some years also the number of families in a house is listed indicating F1 for the first family and F2 for the second family, for example. In each household the head of household is listed as the first person—normally the father in the family. Servants, lodgers and journeymen are counted as members of the household. These instructions are valid up till 1930, which is the latest census publicly available. This understanding of a household is in accordance with the one used for the IPUMS data. The United Nations has made the following definition of a household: 'a person or group of people who live together and make common provision for food or other essentials for living' and for a family: 'a group of people residing in the same household who recognise a kin relationship, ordinarily through descent, marriage or adoption' (Ruggles 2012). This definition differs from the one given for the Danish censuses, as the censuses focus on the sleeping place and thus includes lodgers in the earliest censuses.

1.5 Conclusion

Standardisation done in the way described in this chapter does not limit secondary analyses of the data as all the coding is documented by use of the reference tables and the ever present reference to the personal identifier. The understanding and interpretation of the data is left for the individual researcher. The purposes of standardisation and the use of numeric coding are to reduce the number of answers to every question making statistical analyses possible and making the data comparable, reusable and documented. We recommend maintaining a link with the original transcribed version and if possible also with the (scanned) sources.

References

Hall, P. K., McCaa, R., & Thorvaldsen, G. (Eds.). (2000). *Handbook of international historical microdata for population research: A project of IMAG*. Minneapolis: University of Minnesota, Minnesota Population Center.

Johansen, H. C. (2004). Early Danish census taking. *History of the Family, 9*, 23–31.

Kirby, K., Carson, J., Dunlop, F., Dearle, A., Dibben, C., Williamson, L., et al. (2015). Automatic methods for coding historical occupation descriptions to standard classifications (This book, Chap. 3).

Mandemakers, K., & Dillon, L. (2004). Best practices with large databases on historical populations. *Historical Methods*, *37*(1), 34–38.

Marker, H. J. (2001). Folketællingen 1801 – på vej mod en forskningsressource. *Metode and Data, 84.*

Marker, H. J. (2006). Klargøringen af folketællingen 1801. *Metode and Data, 92.*

Roberts, E. (2006). Reflections on coding 90 million historical occupations. In *Paper at 31st Social Science History Conference, Minneapolis.*

Ruggles, S. (2005). The Minnesota population center data integration projects: Challenges of harmonizing census microdata across time and place. In *Proceedings of the American Statistical Association, Government Statistics Section* (pp. 1405–1415). Alexandria, VA: American Statistical Association.

Ruggles, S. (2012). The future of historical family demography. *Annual Review of. Sociology, 38*, 18.1–18.19.

Solli, A. (2003). Livsløb - familie - samfunn. Endring av familiestrukturar i Norge på 1800-tallet. *Tillegg* A,B,C.

Thorvaldsen, G., & Erikstad, M. (2006). Statistikk basert på individdata fra folketellingene – nye muligheter. *Heimen, 43*, 41–53.

Thorvaldsen, G., & Solli, A. (2012). Norway: From colonial to computerized censuses. *Revista de Demografía Histórica*, *XXX*(1), 107–136.

Chapter 2
Dutch Historical Toponyms in the Semantic Web

Ivo Zandhuis, Menno den Engelse and Edward Mac Gillavry

Abstract The standardisation of historical, geographical names or toponyms facilitates both the comparison and combination of datasets that have a spatial dimension. In a digital infrastructure, online web services can be provided that not only standardise the historical, geographic names in these datasets to a canonical name, but also enrich the dataset with geographical coordinates through a process of geocoding. Finally, these online web services can be used to create cartographic visualisations of the datasets. Since there were no convenient web services available for the standardisation of historical, geographic names in the Netherlands, the websites gemeentegeschiedenis.nl and histopo.nl were launched in 2013. Gemeentegeschiedenis.nl presents a uniquely identifiable web page (using a so-called "Uniform Resource Identifier" or URI) for every municipality in the Netherlands since 1812. The web pages provide relations between former and current municipalities and present maps of all the boundary changes over time. The website histopo.nl was launched to provide a disambiguation service for toponyms of settlements and their historical spelling variants. Both gemeentegeschiedenis.nl en histopo.nl present the information as web sites for regular visitors and as web services for computers to be incorporated in an online information system. The aims of the introduction of gemeentegeschiedenis.nl and histopo.nl are to demonstrate the advantages of a standardisation service for historical, geographic names on the one hand and building up our own expertise and experience on the other hand. In this chapter we present a description and evaluation of the information and centralised services presented on these websites and a description of potential services to create a more useful tool.

I. Zandhuis (✉) · M. den Engelse · E. Mac Gillavry
Rijksstraatweg 79, 2024 DB Haarlem, The Netherlands
e-mail: ivo@zandhuis.nl

M. den Engelse
e-mail: menno@islandsofmeaning.nl

E. Mac Gillavry
e-mail: edward@webmapper.net

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_2

2.1 Introduction

2.1.1 Problem

Almost every historical research question has a geographical dimension. When datasets are created, this geographical dimension is often reflected by the introduction of a field containing a geographical name.

When this geographical name or toponym is standardised, historians can compare their dataset with other datasets. This leads to additional information that can be used in the analysis. Standardised toponyms support the creation of an aggregated view on the data, for instance by means of geographically clustering records into larger geographical units. This enables the comparison with knowledge obtained from other resources. Visualisations of the dataset facilitate the historian in interpreting data.

In order to support the historian in the comparison or combination of datasets, and analyse it by aggregation or visualisation, at some point in time the historian wants to create a standardised form of the geographical name that is found in his source. This could be done at the moment of creation of the information, but maybe the standardisation is realised in a later stadium.

The construction, combination, comparison or visualisation of datasets can be realised with the historian's favourite data manipulation tool. The use of a centralised source of the standard in this tool prevents that an older version of inferior quality is used.

2.1.2 Presented Solution

An online web service combining and presenting all the geographical standards in a coherent and consistent way could prove a big help in the standardisation of a dataset and during the interpretation.

For the Dutch context we have introduced two websites: gemeentegeschiedenis.nl and histopo.nl. Gemeentegeschiedenis.nl presents information about municipalities in the Netherlands, hence the name "gemeentegeschiedenis", meaning "municipal history". We aimed to clearly distinguish between a platform for the administrative units and other types of *his*torical *topo*nyms, and therefore we introduced a second website called histopo.nl that primarily focuses on settlements.

2.1.3 Gemeentegeschiedenis.nl

Gemeentegeschiedenis.nl has two main characteristics.

First, the primary geographical entity is the second order administrative area, i.e. "municipality" in the Netherlands. Gemeentegeschiedenis.nl holds all consecutive stages of the Dutch municipalities, from 1812 until now. The scope of geographical entities within gemeentegeschiedenis.nl is therefore strictly defined. Municipalities were instituted by law and geographic changes to their boundaries, mergers and separations all require changes in the law. Thus, the contents are more or less complete already and have been compiled based on datasets that have been released as open data or that have been provided by various parties to this project. Hence, the geographic representation of these administrative areas used in our service platform is a two-dimensional polygon instead of a one-dimensional point location.

The second aspect of this initiative is its ability to interrelate diverse, de facto standards for geographical names. A standardised name in one standard is related to the same name in another standard.

2.1.4 Histopo.nl

In contrast to municipalities, the history of settlements is not governed by law and their rise and fall often goes unrecorded. Because of their specific characteristics, we introduced the separate website histopo.nl.

Histopo.nl consists of a database with names for all kinds of historical settlements including administrative units, with all available spelling variants through time. The spelling variants resolve into a standardised, canonical name.

2.1.5 Purposes of the Websites

By inter-relating these standards we introduce the historical dimension into the international standardisation of contemporary geographical names, combining the best of both worlds. On the one hand, the websites we developed aim to present information about the historical development of municipalities and settlements for a broader public. On the other hand, the websites could facilitate researchers in the humanities by means of more elaborate search and standardisation functionality. Therefore, the websites provide geographical services for cultural heritage institutes and research in the humanities dealing with historical toponyms.

Thus, the websites are able to assist historians who aim to standardise the historical toponyms in their data set. Besides that, the website can play a role in more elaborate data entry projects and is able to standardise historical toponyms semi-automatically. After creation, the data can be used for instance by converting one coding standard for historical toponyms into another. Or data related to municipalities or settlements can be mapped geographically using thematic mapping techniques in order to reveal spatio-temporal patterns in datasets.

In this chapter, we describe the online geographical services we provide. We evaluate the usage and the quality and quantity of data and services. After that we elaborate on the potential of the platform we anticipate.

2.2 Prior Work and Development Elsewhere

The importance of the standardisation of historical toponyms is broadly recognised. There are various (online) resources for historical place names. In this chapter we focus on a relevant selection. For a broader discussion of resources for historical toponyms, see Southall et al. (2011).

2.2.1 *Standards for Toponyms in Present Time*

The Getty Thesaurus of Geographic Names,[1] generally referred to as "TGN", is a structured vocabulary currently containing over two million names and other information about places. These places include not only administrative, political entities (e.g. cities, nations), but also physical features (e.g. mountains, rivers). Both current and historical places are included. The temporal coverage of the TGN ranges from prehistory to the present and the scope is global. While many records in TGN include geographic coordinates, these coordinates are approximate and are intended for reference only. These geographic coordinates in TGN typically represent a single point, corresponding to a point in or near the centre of the inhabited place, political entity, or physical feature.

A widely adopted open dataset for geographical entities is the Geonames Geographical Database.[2] The data set has been created through crowd-sourcing and contains all kinds of geographical entities, each with a unique identifier. One category of entities is the "administrative division". The second order administrative division is equivalent to the municipality in the Amsterdam Code. However, there are very few historical municipalities available in the data set.

Closely related to Geonames is Wikipedia,[3] in the sense that collecting and maintaining information about toponyms in Wikipedia is also a community effort. There are numerous web pages for geographical entities, amongst others municipalities. Due to its nature, Wikipedia does not provide any indication about the completeness and accuracy of the contents. Most data in Wikipedia is also available as Linked Data via the Semantic Web portal Dbpedia.[4]

[1] http://www.getty.edu/research/tools/vocabularies/tgn/

[2] http://www.geonames.org

[3] http://www.wikipedia.org

[4] http://www.dbpedia.org

2.2.2 Websites with Historical Toponyms

Various initiatives are developed to augment the contemporary view on toponyms with a historical dimension.

The community-built gazetteer of ancient places Pleiades[5] for instance, gives scholars, students and enthusiasts worldwide the ability to use, create, share, and map historical geographic information primarily about the ancient world.

In the United Kingdom the main entry into historical toponyms is The Historical Gazetteer of England's Place-Names.[6] All British toponyms surveyed by the English Place-name Society are uniquely identified with a URI and described in terms of their relation with an upper geographic division. Future objective is to provide the data not only in a web page but in a machine-readable format as well.

For Danish toponyms the Digitalt atlas over Danmarks historisk-administrative geografi (DigDag)[7] is constructed with presentation of the corresponding geographic polygons. Searching a place and its upper geographic divisions is enabled through a click on the map. Web services are available to obtain data about toponyms, including the geographical polygons.

2.2.3 Dutch Toponyms

Before the launch of gemeentegeschiedenis.nl and histopo.nl there have been various lists, "gazetteers", available for standardising Dutch historical toponyms, but these were not available online and were scattered across various cultural heritage institutes. Thus, an important aspect of this initiative is its ability to interrelate these diverse, de facto standards:

On gemeentegeschiedenis.nl for municipalities:

- CBS code: assigned by Statistics Netherlands, the Dutch national statistical office
- Amsterdam code: assigned by Van der Meer and Boonstra as part of the Historical Geographic Information System project (Van der Meer and Boonstra 2011).
- GeoNames code: an online, crowd-sourced database of over 10 million toponyms

On histopo.nl for settlements:

- GeoNames code, again

[5] http://pleiades.stoa.org

[6] http://www.placenames.org.uk

[7] http://www.digdag.dk

- Kloeke code: assigned by the Meertens Institute for research and documentation of Dutch language and culture (Kloeke 1926).

All these standards for the Dutch context are existing systems that all have their quality and have been used in datasets in the past.

2.3 Functionality

Gemeentegeschiedenis.nl and histopo.nl provide various functions in order to help historians and cultural heritage experts to create and use better geographical data.

2.3.1 Geographical Standardisation with a Historical Dimension

The comparison of historical data related to municipalities over time and space is a complex matter due to the heterogeneity of historical toponyms. Historical documents that describe people, economic activity or political votes contain references to the contemporary administrative areas. In many instances, the place of birth (actually, the municipality where the birth was registered) recorded in historical documents presents a peculiar anachronism as it reflects the administrative subdivision at the time of birth, not at the time at which the place of birth was recorded.

Digitising these historical sources, historians not only have to copy the geographical description as precisely as possible as it was recorded in the original source, but also have to interpret the original and provide a standardised spelling of this geographical description. At a later stage this facilitates statistical analysis or spatial analysis using a Geographic Information System (GIS). Within the context of the dataset, this standardised form must be unambiguous and unique. Drawing this standardised spelling from a widely adopted thesaurus, historians are then able to combine and contrast their own datasets with other datasets that use the same thesaurus.

2.3.2 Combining Existing Sources on gemeentegeschiedenis.nl

The main source for gemeentegeschiedenis.nl is the thesaurus of Dutch municipalities, generally referred to as the Amsterdam Code (Van der Meer and Boonstra 2011). Although previously published in a book and available under an open license as a downloadable PDF file, there was no online resource in a machine-readable format. In the Amsterdam Code, identifiers are reused over time:

a municipality that only changed name therefore kept the same identifier. When two municipalities were combined into one, the Amsterdam Code of the municipality with the largest number of inhabitants was inherited by the newly formed municipality. This implies that the municipalities are comparable through time, but the codes are not unique. They are only unique in combination with a time-stamp indicating the period of the existence of a municipality.

Each of the municipalities on gemeentegeschiedenis.nl is further identified by their CBS code. This code is assigned and maintained by Statistics Netherlands, the Dutch national statistical office. The historical dimension is limited, but the code is used in most of the currently available datasets, which enables the temporal comparison between the current situation and the past. While the CBS codes do not cover municipalities before 1830, the Amsterdam Code encompasses a broader, temporal range as it identifies municipalities from as early as 1812.

GeoNames contains and classifies various types of toponyms. The geographical coordinates, stored with all the names, enabled us to combine the municipality with the place names that are used for an inhabited place which is not an administrative unit. As a result of this spatial combination, the web pages provide lists of all known settlements in a municipality.

2.3.3 Combining Existing Sources on histopo.nl

GeoNames mainly contains names for settlements with an administrative function. These are the basis for combining information on settlements on histopo.nl.

The Kloeke Code for settlements, assigned by the Meertens Institute,[8] is used in the research and documentation of geographical patterns in Dutch language and culture (Kloeke 1926). Since these settlements are no official administrative areas, disambiguation is ensured by combining the Kloeke code identifier with the coordinates of the settlements. Furthermore, Kloeke codes are not available for all settlements in the Netherlands.

On histopo.nl two additional lists of names of settlements are incorporated. The first list is constructed by Simon Hart and published by the City Archives in Amsterdam.[9] Hart made a list of all the toponyms he found in sources in the Amsterdam City Archives and resolved them to contemporary names of settlements. A similar list was constructed with the names found in the conscription registries from the 1830s until around 1930s.[10] A selection of these registries throughout the Netherlands was digitised and made available through data-entry. Toponyms were organised and resolved to a standardised name.

[8] http://www.meertens.knaw.nl

[9] http://stadsarchief.amsterdam.nl/archieven/hulp_bij_onderzoek/herkomstonderzoek/ (in Dutch).

[10] http://militieregisters.nl/en/

2.3.4 Access to Information

On gemeentegeschiedenis.nl, the Dutch municipalities have their own web page. Each of these web pages contains a set of maps that do not only reflect the changes in the delineation of the municipalities, but also the mergers of municipalities over time (Boonstra 1992). Furthermore, all settlements that are located within the boundaries of a municipality are listed together with links to other online resources that describe that particular municipality or settlement. Visitors to the website can learn about the history of the Dutch municipalities and researchers can obtain information to assist them in standardising the administrative areas in their datasets.

Each of these web pages is available at a short and readable URL that also acts as a Unique Resource Identifier (URI) (Berners-Lee et al. 1998). For example, the former municipality of Schoten is available at and is identified with the URI http://www.gemeentegeschiedenis.nl/gemeentenaam/Schoten.

For every available standard a separate web page is constructed to enable users to resolve a code defined according a specific standard. The Amsterdam Code 10382 identifies the former municipality of Schoten, which is learned by following the URI http://www.gemeentegeschiedenis.nl/amco/10382. The same goes for the CBS code 1173: http://www.gemeentegeschiedenis.nl/cbscode/1173 (Fig. 2.1)

There are also URIs that provide a list of municipalities in one province, for instance: http://www.gemeentegeschiedenis.nl/provincie/Noord-Holland

2.3.5 Providing the Data

Using a process referred to as "content negotiation" a client application obtains the information available in the appropriate data format. In the case of a web browser this is HTML. By manipulating the *accept-header* of the client application requesting the URI RDF-XML or JSON is provided. Presenting these formats in a

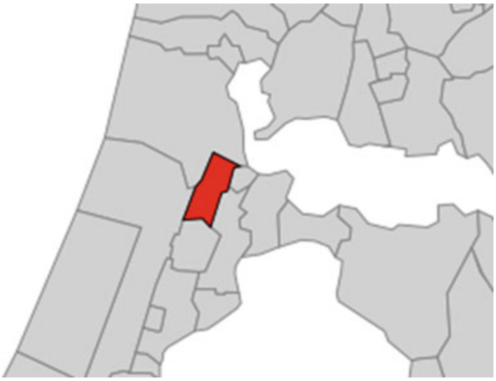


Fig. 2.1 Map of the municipality of Schoten in 1885 as presented on http://www.gemeentegeschiedenis.nl/gemeentenaam/Schoten

web browser can be forced by using a different URI. This is good practice and performed by Dbpedia as well. For RDF-XML this is http://www.gemeentegeschiedenis.nl/gemeentenaam/rdfxml/Schoten and for JSON this is http://www.gemeentegeschiedenis.nl/gemeentenaam/json/Schoten. Notice that in this last format extra information is available in GeoJSON (Butler et al. 2008) about the geographic polygon(s) of the consecutive stages through time.

Corresponding URIs for information concerning one province are: http://www.gemeentegeschiedenis.nl/provincie/rdfxml/Noord-Holland and http://www.gemeentegeschiedenis.nl/provincie/json/Noord-Holland. As an extra service the overview of municipalities in one province is distributed in CSV-format: http://www.gemeentegeschiedenis.nl/provincie/csv/Noord-Holland.

2.3.6 Providing the GIS Data

Maps presented on the HTML-pages are based on existing files containing the geometries of municipalities through the years (Boonstra 1992). These shape-files are imported into a PostgreSQL database,[11] with a PostGIS extension.[12] In this environment geometries can be combined with data. A MapServer[13] is used to enable serving this information on the World Wide Web. This server application provides the standardised interfaces Web Map Service (WMS) (OGC 2009) and Web Feature Service (WFS) (OGC 2014).

The WMS interface provides, based on a query, a map in a specified image format like PNG. This technique is used to present the images on the HTML-pages of the website.

The WFS interface provides, again based on a query, geometries and requested features encoded in a standardised format like GeoJSON (Butler et al. 2008).

2.3.7 Search User Interface

At the moment, the website provides a simple search on the names of municipalities and toponyms. Users can enter a search term and various names of municipalities and settlements are returned. If any of the codes (Amsterdamse Code, CBS-code) is provided instead, the matching municipality or settlement is returned. For every municipality, the search function returns the web page with information about the right information.

[11] http://www.postgresql.org

[12] http://www.postgis.org

[13] http://www.mapserver.org

The user interface provides widgets that enable historians to narrow down the search result set by selecting a year, a region, and/or a standard. Thus, the responsibility remains with the historians for the correct interpretation of the historical source and for the selection of the right standard and standardised form of the geographical description.

2.3.8 Search API

The search query entered by a user on gemeentegeschiedenis.nl, is handled by an Application Programming Interface (API) serviced on histopo.nl.[14]

A user searching for information about a place name does not know whether he is searching for a municipality beforehand, let alone that he used the right spelling. By using the API of histopo.nl as a service on gemeentegeschiedenis.nl, users searching for a place name on gemeentegeschiedenis can be offered alternatives if no municipality is retrieved.

2.3.9 Architectural Overview

In the paragraphs above, we described the various techniques we used to realise the service platform. This could be visualised in the following architectural overview (Fig. 2.2).

2.4 Evaluation

After two years of development and preliminary use of gemeentegeschiedenis.nl and histopo.nl we experienced solutions and problems in four areas: the quality and quantity of the data, organisational issues and the quantity of the services. But first we elaborate on the usage of the platform in various situations at this moment.

2.4.1 Usage

After introduction of the platform, various projects were interested in using the data-services we provide.

[14] The API is documented on http://api.histopo.nl/docs/

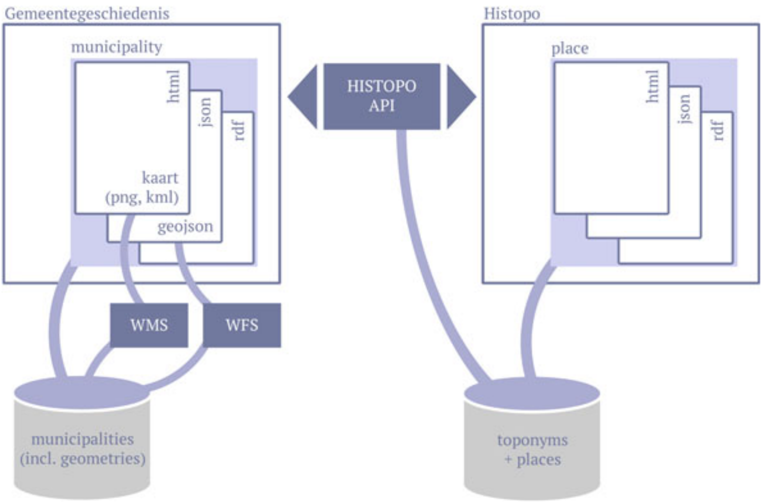


Fig. 2.2 Architectural overview of the websites gemeentegeschiedenis.nl and histopo.nl

A website with information about archival institutions in the Netherlands[15] uses gemeentegeschiedenis.nl to refer to archival holdings. An archival institution provides access to archival collections concerning a particular (former) municipality. By combining the archival institution and the names of former municipalities users can discover which institution is holding the archive of their interest.

In the CEDAR-project[16] (Merono 2013) historical Dutch census-data, which was originally available as scans and in spreadsheets, is converted into Linked Open Data. References to municipalities are common and made with the Amsterdam Code. They experimented with the usage of URIs of gemeentegeschiedenis.nl. By relating the Census data to the municipalities in gemeentegeschiedenis.nl, data in both datasets can be combined. Furthermore future datasets with the same standardised toponym can be combined as well.

As part of the LINKS-project[17] a dataset was constructed with historical Dutch toponyms. The relation between settlements and municipalities available on gemeentegeschiedenis.nl was included (Huijsmans 2013). Unfortunately, new knowledge that is incorporated in our databases must lead to a new release of Huijsmans dataset.

[15] http://www.archiefwiki.org

[16] http://www.cedar-project.nl

[17] http://www.iisg.nl/hsn/news/links-project-nl.php

2.4.2 Data Quality

Combining the various datasets resulted in some surprising errors and inconsistencies in the data. In the presentation of the data these errors are extra visible. For instance, the municipalities of Oosterwolde in the province of Friesland and Oosterwolde in the province of Gelderland somehow ended up in one map presented on the web page of Oosterwolde[Ge].[18]

The link between a toponym, provided by GeoNames, and a municipality is calculated from the maps: if the coordinates of an inhabited place are within the boundaries of the shape of a municipality a link is constructed. The precision of the maps, however, turned out to be insufficient to construct a trustworthy link. In the calculations an inaccuracy of at most 1 km should be taken into account. The maps should be replaced to improve this function.

It is strange to use a specific dataset oriented to municipalities, when toponyms are handled that are older than 1812 or in situations where standardisation to municipalities is not relevant. Standardising historical toponyms in general should be handled by a separate dataset and service which we introduce on the website histopo.nl.

We have already learned that people are very eager to help improve the information on gemeentegeschiedenis.nl when it is incorrect, particularly the maps. Additional functionality for visitors to provide feedback on the information, or even to help us correct it, could prove to be very helpful. New information can be added as well. New names, with different spellings could be provided by historians doing research in a particular source, with unknown names.

2.4.3 Data Quantity

Another area of future growth of our service platform lies in the addition of new names of settlements and their spellings variants. Of particular interest are Dutch exonyms for foreign settlements—e.g. the Dutch toponym "Jarmuiden" refers to the British port of Yarmouth—and foreign exonyms for Dutch settlements, e.g. French toponym "Nimegue" refers to the Dutch city "Nijmegen". This enables historians to use historical sources from all around the world to combine information about the Netherlands. We made a start with this in developing histopo.nl.

The service platform could also be extended to smaller geographic entities, like neighbourhoods, streets or even individual houses. That way, we gradually create a historical gazetteer or geocoder. We hope and aim that these services will enable

[18] http://www.gemeentegeschiedenis.nl/gemeentenaam/Oosterwolde_Ge and http://www.gemeentegeschiedenis.nl/gemeentenaam/Oosterwolde_Fr

historians to interpret, enrich and visualise their historical data in new and innovative ways. Moreover, the standardisation services provide them with a means to actually publish their historical data as linked open data.

2.4.4 Organisational Issues

At this moment the project is not financially supported. This endangers future development and continuity of the system. If the usefulness is recognised by a broader audience in humanities and cultural heritage and long term continuity is desirable, financial and organisational support is needed. This support can be found in a national organisation taking responsibility for this task. Another scenario might be that we develop a more commercially oriented business model, in order to provide the necessary financial means.

Not all the datasets we used have a clear copyright statement. In all cases we got informal, personal approval for reusing the available dataset. This means we are unable to provide a clear copyright statement on our dataset as well.

2.4.5 Service Quantity

We envision that the historical databases containing toponyms and the geographic representation of Dutch municipalities with various types and links between the entries should be available to the general public in a variety of ways.

Historians can use the search interface to standardise their dataset, but have to search for every entry one by one. Therefore, extra functionality is needed to improve the usability and applicability of the websites.

2.5 Potential

We envision that we (or others) can develop all kinds of services based upon the data-services provided by gemeentegeschiedenis.nl and histopo.nl. We distinguish two types of future functionality: services that assist historians to standardise the geographical entities in their datasets and services that assist them to process their standardised datasets. The remainder of this chapter is used to explain and describe these potential services.

In order to obtain a dataset with standardised forms of the original geographical entity in the historical document, historians should add at least one additional field in their records. (A field with extra information, like the standardised toponym, could be added as well). The original records contain the contemporary and authentic description of the geographical name that needs standardising. In the additional field, the service will store the interpretation of this original description by means of the right URI referencing the toponym. The main advantage of using a service platform approach is that new knowledge about toponyms and relations is continuously incorporated into the databases. Therefore, the geocoding success rates of the service platforms will improve over time without additional efforts on the part of the historians.

Datasets that contain standardised geographic names can be processed more consistently. For example, one could create a service that converts the standardised geographic names in a dataset into another standard. Furthermore, historians might want to visualise their data set on a map in order to identify geographic patterns.

2.5.1 Indexing or Large-Scale Data Entry Projects

Adding a standardised geographical code or name can be performed at data entry. In that case it would be convenient to incorporate the search into the user interface of the data entry software. In the Netherlands an interesting example where this service could come in handy is http://www.velehanden.nl. On this crowd-sourcing platform, archival services can start a data entry project on a specific archival source.

For developing this service the API can be used. A more general service would be a SPARQL-endpoint. A SPARQL-endpoint can be used to formulate a specific query that is not provided by the API. That way the current crude Semantic Web functionality of RDF-XML export is extended.

2.5.2 Automatic Standardisation

Additionally, the API can be used to provide automatic standardisation. Therefore, the result set should be organised according to a certain relevance function: the standardised form with the highest probability will be presented as first entry in the result set. The automatic detection can be improved by two extra variables that provide information about the context of the source:

1. place: the location where the source was created
2. target area: the area where the result must be located
3. time: the period in which the source was created and therefore the period in which the toponym or spelling variant is used

Results are selected that are located in the requested geographical target area and existing in the requested time period, prior to the end date of the period in which the source was created.

The order in which the result set is presented could be determined as follows. Of course, the name with the smallest amount of differences with the search term is highest on the result list. To start with, this difference can be expressed in for instance the Levenshtein distance (Levenshtein 1966). When more results are delivered, the geographical unit that has the smallest geographic distance to the place of origin is higher on the result list.

An extra service could enable historians to upload a CSV file. As they indicate which column contains the geographical name to be standardised and the standard in which it is to be standardised, the service platform could return a CSV file that now contains an extra column that holds the automatically generated standardisation. In this case, only the result with the highest probability is returned.

2.5.3 Converting from One Standard into Another

Combining various datasets with differently standardised geographic names can be very complex. A service enabling historians to convert from one standard into another could be very useful. Historians would upload a CSV file and indicate the column that contains the standardised name and select the standard and period in which it is to be converted. A new CSV file is automatically generated that contains an extra column holding the additional standardised value.

Historians have to be aware that the different standards are not equivalent: a settlement that is converted into a municipality results in the loss of precision of the geographical location. However, the service platform enables historians to convert whatever they like (Fig. 2.3).

A scholar could, for instance, want to connect her dataset containing GeoNames ids with Kloeke-referenced linguistic data. In order to do so, she can use the API to find the Kloeke code for every GeoName-id in her dataset. All the records in her dataset are augmented with an extra field containing the connected Kloeke code. If she doesn't want to use the API, she should be able to download a list combining the two standards. This list can be used as table in a database.

2.5.4 Enrichment with Geographic Coordinates

For visualisation purposes the standard codes can be converted into geographic coordinates: in a GIS these coordinates can be used to draw a settlement or municipality on a map, together with the information related to the settlement or municipality. For every geographic entity the service platform is able to provide the geographic coordinates. Uploading a file containing multiple geographic entities, the standard forms of the toponyms are then subsequently enriched with the coordinates of their geographic representation in batch.

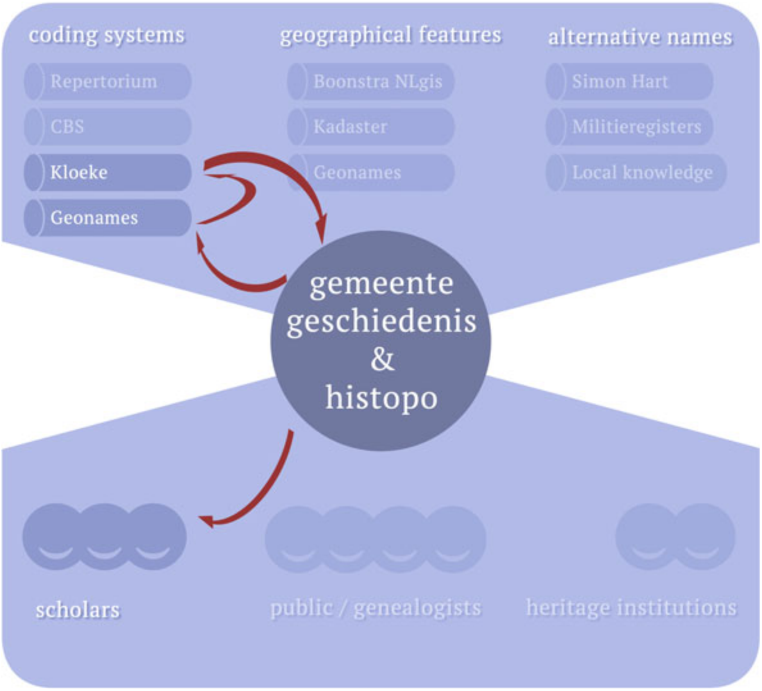


Fig. 2.3 Visual representation of the example of a scholar combining two coding systems for geographical names

2.5.5 Visualising Information

One of the many functions a GIS can perform is the visualisation of geographic datasets. However, the operation of a GIS system can be fairly cumbersome for the uninitiated. A future functionality of a service platform to automatically generate a map based on historical datasets that have been uploaded could be very useful, either to draw the main conclusions in a historical research project, or to assist in the decision whether to involve professional mapping expertise (Fig. 2.4).

If, for instance, a scholar wants to visualise historic data on municipalities in the Dutch province of Limburg, he should be able to download the right selection of geometries of municipalities from our platforms. These geometries can be used to construct a map with the information contained in his dataset. In the dataset a type of municipality encoding, like the Amsterdam Code or the CBS Code, must be included. If this is not the case, the platform could be used to standardise the geographic fields in the dataset first.

Historians would upload their datasets containing both the unique identifiers for the geographic entities, i.e. the Amsterdam Code, CBS code or Kloeke code, and the corresponding attribute values to be represented on the map. Once the specific

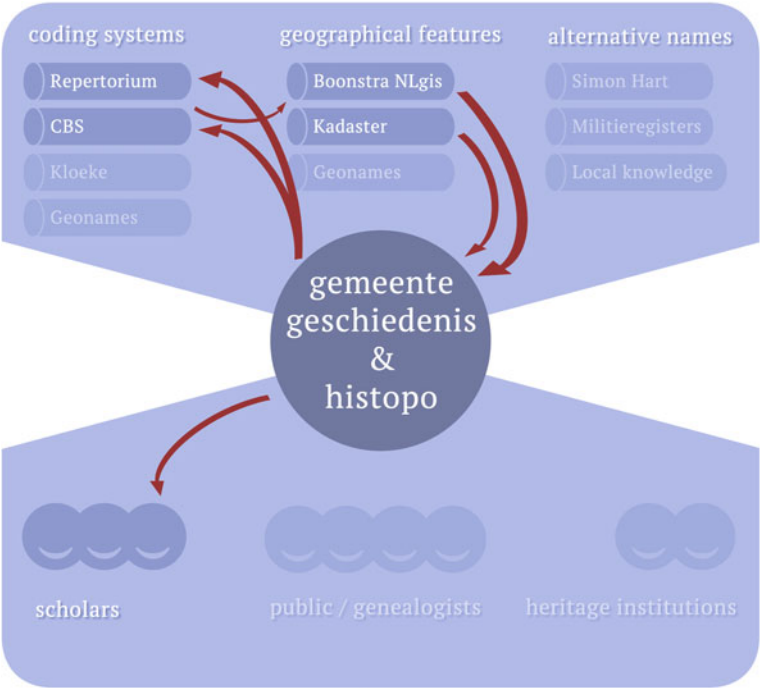


Fig. 2.4 Visual representation of the example of a scholar using the geometries of municipalities

year for the dataset has been selected, the values will be matched to the contemporary geographic delineations of the municipalities. Depending on the nature of the values the appropriate map type is created to visualise the data set geographically.

In case the dataset contains nominal values, a chorochromatic map is created. Municipalities that have the same nominal value are filled with the same colour. An example is a map coloured-coded according to the largest political party per municipality. If the data set contains ordinal or relative values per municipality, a choropleth map is created. Municipalities with increasing values are filled with colours of increasing saturation or lightness. An example is a map that shows the relative growth of the population per municipality. Finally, if the data set contains absolute values, a proportional symbol map is created. An example is a map that shows increasing number of conscripts per municipality using increasing sizes of bars, squares or circles.

The resulting map could be easily downloaded in Scalable Vector Graphics (SVG) or PDF formats for inclusion in scientific publications or presentations. Interactive versions of the maps could be embedded in third-party websites using a JavaScript-based mapping API that would request map images from the service platform.

2.6 Conclusion

This chapter describes the current and future functionality of two websites we envision to be useful for historians with interest in a geographical component in their research. We evaluated the current functionality and identified aspects that can be elaborated. A large amount of these aspects cover the augmentation and improvement of the data itself. Considering the amount of unique visitors and the length of their stay, the website satisfies a need. Also, the amount of feedback and exposure we experienced, as well as projects that used some of our functionality suggest that an historical gazetteer for the Netherlands is much appreciated.

Hopefully, future services will be implemented for our service platform in the near future. Historical research and access to cultural heritage would improve. A financially viable organisation of the initiative is essential.

Acknowledgment We would like to thank Onno Boonstra (Radboud University Nijmegen) who applied for financial support in a “Klein Data Project” (Small Data Project) at DANS-KNAW for further development of some of the web services described.

References

URLs checked March 20, 2015.

Berners-Lee, T., et al. (1998). *Uniform resource identifiers (URI): Generic syntax*, http://www.ietf.org/rfc/rfc2396.txt.

Boonstra, O. (1992). NLKAART. A dynamic map of the Netherlands, 1830–1980. In H. J. Smets (ed.), *Histoire et Informatique V. Montpellier* (pp. 315–324). The shape files of this research are deposited at DANS: https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:43063.

Butler, H., et al. (2008). *The GeoJSON format specification*, http://geojson.org/geojson-spec.html.

Huijsmans, D. P. (2013). *IISG-LINKS Data-set Historische Nederlandse Toponiemen Spatio-Temporeel 1812-2012*, Release 2013.2 http://www.iisg.nl/hsn/data/place-names.html.

Kloeke, G. G. (1926). De totstandkoming van de “kaart van het Nederlandsche taalgebied ten behoeve van het dialectgeografisch onderzoek” In Grootaers, L., & G. G. Kloeke, *Handleiding bij het Noord- en Zuid-Nederlandsch Dialectonderzoek*, (pp. 57-65) Den Haag : Martinus Nijhoff http://www.meertens.knaw.nl/projecten/mand/LITkloeketotstandkoming.html.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*, 707–710.

Meer, A. v.d., & Boonstra, O. (2011). *Repertorium van Nederlandse gemeenten vanaf 1812.* Den Haag : DANS (Data Archiving and Networked Services) http://www.dans.knaw.nl/nl/over/organisatie-beleid/publicaties/DANSrepertoriumnederlandsegemeenten2011.pdf.

Meroño-Peñuela, A., Guéret, C., Ashkpour, A., & Scharnhorst, A. (2013). *Publishing, harmonizing and consuming census data: The CEDAR project*. Open data on the web workshop, world wide web consortium, Open Data Institute, Open Knowledge Foundation, 23–24 April 2013, Google Campus, Shoreditch.

OGC (Open Geospatial Consortium). (2009). *Web map service implementation specification—Version 1.3.0* http://portal.opengeospatial.org/files/?artifact_id=14416.

OGC (Open Geospatial Consortium). (2014). *Web feature service 2.0 interface standard—Version 2.0.2* http://docs.opengeospatial.org/is/09-025r2/09-025r2.html.

Southall, H., Mostern, R., & Berman, M. L. (2011). On historical gazetteers. *International Journal of Humanities and Arts Computing, 5*(2), 127–145.

Chapter 3
Automatic Methods for Coding Historical Occupation Descriptions to Standard Classifications

Graham Kirby, Jamie Carson, Fraser Dunlop, Chris Dibben, Alan Dearle, Lee Williamson, Eilidh Garrett and Alice Reid

Abstract The increasing availability of digitised registration records presents a significant opportunity for research in many fields including those of human geography, genealogy and medicine. Re-examining original records allows researchers to study relationships between factors such as occupation, cause of death, illness and geographic region. This can be facilitated by coding these factors to standard classifications. This chapter describes work to develop a method for automatically coding the occupations from 29 million Scottish birth, death and marriage records, containing around 50 million occupation descriptions, to standard classifications. A range of approaches using text processing and supervised machine learning is evaluated, achieving classification performance of 75 % micro-precision/recall, 61 % macro-precision and 66 % macro-recall on a smaller test set. Further development that may be needed for classification of the full data set is discussed.

G. Kirby (✉) · J. Carson · F. Dunlop · A. Dearle
University of St Andrews, St Andrews, UK
e-mail: graham.kirby@st-andrews.ac.uk

J. Carson
e-mail: jamiekcarson@gmail.com

F. Dunlop
e-mail: fraser.dunlop@gmail.com

A. Dearle
e-mail: alan.dearle@st-andrews.ac.uk

C. Dibben · L. Williamson
University of Edinburgh, Edinburgh, UK
e-mail: chris.dibben@ed.ac.uk

L. Williamson
e-mail: lee.williamson@ed.ac.uk

E. Garrett · A. Reid
University of Cambridge, Cambridge, UK
e-mail: eilidh.garrett@btinternet.com

A. Reid
e-mail: alice.reid@geog.cam.ac.uk

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_3

3.1 Introduction

We describe work being carried out to develop a method for automatically processing the 50 million occupations recorded on Scottish birth, death and marriage records from 1855 to the present day.

There are two key problems: first, how to consistently code occupations over the entire 150-year period so that researchers can explore changing patterns and trends; and second, how to automate this process so that the majority of records do not need to be manually coded. In this chapter, we focus on the second of these problems: developing methods to automatically classify narrative occupation descriptions into a fixed set of standard classifications.

This work builds on previous efforts to automatically classify causes of death contained in the Scottish records mentioned above (Carson et al. 2013).

3.2 Data Sets

The target data set is the entirety of births, deaths and marriages recorded in Scotland from 1855 to present day, comprising 29 million events and around 50 million occupations. These records will ultimately be coded to the Historical International Standard Classification of Occupations (HISCO) (van Leeuwen et al. 2002; HISCO 2013).

In order to develop our approach and to trial various methodologies, we have conducted experiments on a smaller data set, originating from the Cambridge Family History Study (Bottero and Prandy 2001), cited by Prandy and Lambert (2012). This set contains 243,000 records dating from the early eighteenth century to the present, with 30,200 unique occupation descriptions. The occupational titles were anonymised, and any additional contextual information ignored, when processed in our classification experiments.

All of the records were previously classified to SOCN, an extension of the SOC coding system (US Bureau of Labor Statistics 2010) defined in the Cambridge Family History Study to include historical terms. SOCN includes 1000 distinct codes, of which 453 occurred in the data set. A subset of the data, comprising 64,000 of the records with 9400 unique occupational titles, was also classified to HISCO, via a fixed SOCN-HISCO mapping. The HISCO classification includes 1675 codes, of which 337 occurred in the data subset. This subset of 64,000 records was used in the classification experiments.

Tables 3.1 and 3.2 show the ten most common occupational titles, and words, respectively.

Table 3.1 Most common occupational titles

Occupational title	Number of occurrences
Labourer	4683
Farmer	3948
Ag lab	1533
Lab	1501
Carpenter	1227
Servant	907
Gardener	907
Tailor	830
Shoemaker	766
Coal miner	756

Table 3.2 Most common words

Word appearing in occupational title	Number of occurrences
Labourer	5559
Farmer	4297
Lab	3349
Servant	2063
Maker	1818
Ag	1564
Miner	1536
Clerk	1519
Carpenter	1358
Gardener	1092

3.3 Approaches to Classification

Based on previous experience with automatic classification of causes of death (Carson et al. 2013), a number of classification techniques were evaluated, and their results compared with the assumed ‘gold standard’ of the domain expert coding. The following aspects were investigated:

- the effects of various types of data cleaning
- the accuracy of three different automatic classifiers
- the accuracy of three different ensemble classifiers
- the effects of preparatory feature selection

3.3.1 Preparation

3.3.1.1 Cleaning

The previous work on classification of causes of death involved textual descriptions written by doctors and other officials. These were often verbose, containing various filler words and parenthetical comments. In that work, it proved useful to employ a number of simple cleaning rules using a regular expression processor, to remove some of these non-relevant words, and to expand commonly used abbreviations, as illustrated in Table 3.3.

In the current work on classifying occupations, after studying example phrases, it was concluded that this type of cleaning would not be necessary, since most occupation descriptions only contain a few words. For example:

- wharfingers clerk
- wheat bag stacker
- wheeler
- wheelright
- wheelsmith (railway works)
- wheelwright

3.3.1.2 Human Coding Consistency

The consistency of human coders was also considered. The HISCO-coded data contained 9400 unique occupational titles, of which 273 occurred at least 25 times in the total set of 64,000 occupational titles. From these 9400 unique titles, 119 have been multiple-coded to more than one HISCO code.

Of the multiple-coded titles, there were only 10 for which an alternative code was used in more than 5 % of cases. The most common multiple-coded descriptions are shown below, with the most frequent codes for each:

- "fireman"
  - railway steam-engine fireman (37 %)
  - fire-fighters (26 %)
  - boiler fireman (13 %)

Table 3.3 Example cleaning of cause of death strings

Original text	Cleaned text
(Cardiac paralysis) diphtheria	Diphtheria
1 paralysis 2 smallpox	Paralysis
Both flu; acute pneumonia	Influenza
Injury caused by being run over by a railway truck	Injury

- "machinist"
  - sewers and embroiderers (53 %)
  - machine-tool operators (46 %)
  - printing pressmen (1 %)
- "iron founder"
  - metal casters (73 %)
  - metal smelting, converting and refining furnacemen (27 %)

Two approaches to dealing with this variation by cleaning the data set were investigated:

- discarding all titles for which multiple HISCO codes were assigned in a significant number of cases (511 records, i.e. 0.8 % of the data set)
- altering the code for those titles with multiple codes to the most frequently assigned code

The rationale for the former is that such records should be prioritised for further human review for coding errors, while the rationale for the latter is that it might be assumed that the most frequent coding is correct and the others are errors.

Furthermore, some of the variations in coding may be legitimate; the examples in the Table all appear to be possible valid interpretations. It is not known whether there was additional context available to the coders, or whether they made an arbitrary choice in such cases.

3.3.2 Edit Distance Classifier

The first automatic classifier tested was a relatively simple string similarity Edit String algorithm, the intuition being to select the HISCO class whose definition most closely matched the input string. The similarity measure used was *Levenshtein distance*: the number of single-character insertions, deletions or replacements needed to transform one string into another.

The simplest approach is to compare the title to be classified with all the HISCO definition strings, and to select the one with the smallest edit distance from the title. However, this would not work well for cases where the title occurs as a substring embedded in the definition string, or vice versa. For example, the title *machinery fitter* should probably be coded to the HISCO description *Machinery Fitters and Machine Assemblers*, which has a large edit distance from the former.

To address this issue, each of the words in the title is compared pair-wise with each of the words in each candidate definition. The definition with the greatest number of close-matching words, as defined by edit distance, is selected.

3.3.3 Individual Machine Learning Classifiers

A machine learning classifier, in this context, is an automatic system that is first trained on a set of examples. Each example contains a particular text string together with its correct classification. The classifier builds a predictive model, which is then used to classify unseen strings.

To apply this approach to classifying occupational titles, we used the Mahout machine learning framework (Apache Software Foundation 2011) and evaluated two separate classifiers supplied with the framework:

- Logistic Regression using Stochastic Gradient Descent (SGD) (Komarek 2004; Zhang 2004)
- Naive Bayes (NB) (Langley et al. 1992)

In general, NB is expected to yield better results than SGD when the number of available training examples is small. However, SGD is thought to perform better on larger training sets, and in cases where features in the data set are correlated with each other (Ng and Jordan 2001).

3.3.4 Ensemble Approaches

The ensemble approach (Dietterich 2000) is based on the premise that better accuracy may result from combining classifications obtained independently from multiple algorithms. The following ensemble methods were evaluated, each combining results from some or all of the individual classifiers described in the preceding two sections:

- majority voting
- confidence-weighted (1)
- confidence-weighted (2)

The voting method used the three individual classifiers: edit distance, SGD and NB. The need for more than two participants to obtain a majority was one of the motivations for employing the string similarity classifier in addition to the machine learning classifiers.

The confidence-weighted ensembles combined the individual machine learning classifiers (SGD and NB). The general idea was to allow the overall decision of the ensemble to be influenced by a degree of confidence attached to each of the individual classifier decisions. However, while the SGD classifier generates an indication of confidence for each classification decision, the NB classifier does not, and so it was not possible to have an ensemble that simply selected the classifier decision with the highest confidence. Different approaches to addressing this problem were taken by the two confidence-weighted ensembles.

The first ensemble made a choice between the two individual machine learning classifiers based on the confidence expressed by SGD. If this was sufficiently high, the ensemble selected the SGD classification and the NB otherwise. Thus the ensemble decision was driven by the confidence of the SGD classifier.

To do this, the relationship between the accuracy of the SGD classifier's decisions and its corresponding confidence levels was examined offline, by analysis of previous experimental results. A threshold value for the confidence measure was identified, above which the classifier's decisions were correct in 50 % of cases.[1]

The approach taken by the second confidence-weighted ensemble was to synthesise an approximation to a confidence measure for the NB classifier, so that the ensemble could then select the decision of the individual classifier with the higher confidence.

To do this, the trained NB classifier was initially run offline over its own training set, to determine its relative accuracy on each HISCO class. For each HISCO class *X*, the proportion of records that it classified as *X* correctly was recorded in a table, which was then used during the actual classification process.

During that process, the ensemble classifier took the decision of the NB classifier and looked up the corresponding correctness ratio. Interpreting this as a crude approximation to the probability that the NB classification decision was correct, the value was used as a confidence measure, allowing the ensemble to decide between the NB and SGD classifications.

3.3.5 Feature Selection

Due to the high dimensionality of the feature space in a text classification system, it is desirable to reduce the feature space without sacrificing classification accuracy. Terms that have no relationship to any given classification provide little information to the system and can often reduce the accuracy of the model as they introduce noise. Removal of low-value terms can be achieved using automatic feature selection, a process implemented here using the X^2 statistic (CHI) as described by (Yang and Pedersen 1997) to calculate the 'goodness' of a term. Terms with a goodness value below a fixed threshold were removed from the feature set during the training process. Examples of such terms include "and", "at", "in", and "of".

3.3.6 Summary

To summarise, the approaches evaluated were:

- individual classifiers: string similarity, SGD, Naive Bayes;
- ensemble classifiers: majority voting and two confidence-weighted versions.

[1] The 50 % threshold was arbitrarily chosen; the effect of varying this may be investigated in future work.

In addition, the effects were evaluated of applying feature selection to the machine learning classifiers, and correcting or discarding multiple-coded occupational titles in the human-coded data.

3.4 Evaluation

A first group of experiments measured the accuracy of the individual and ensemble classifiers, without feature selection or any cleaning of the data set. A second group assessed the effects of using feature selection and cleaning, with the same six classifiers.

There are several well-known ways to validate machine learning results, including k-fold cross-validation, random sub-sampling and the leave-one-out method (Kohavi 1995; Witten and Eibe 2005).

The reported results were calculated using the repeated random sub-sampling method. The data set was repeatedly partitioned in a random 80/20 training/testing split, resulting in a sub-sample of the data set being used for training and the complement of this random sub-sample being used in testing.

Unlike crossfold validation methods, the repeated random sub-sampling method has been shown to be asymptotically consistent, which can result in more pessimistic predictions of the test data compared with cross validation (Shao 1993).

There are various approaches to measuring the performance of automatic classifiers (Sokolova and Lapalme 2009). Here we employ *precision* and *recall*:

- precision is a measure of the probability that an occupational title classified as some HISCO class X, really does belong to class X;
- recall is a measure of the probability that an occupational title that really belongs to class X, is classified as X.

Precision and recall can be calculated in two different ways, using micro- and macro- measures. Micro-precision and micro-recall are calculated over the test data set as a whole. The micro-precision is $TP/(TP + FP)$, where TP is the total number of true positives (correct classification decisions) and FP is the total number of false positives (incorrect classification decisions). For this type of classification problem, micro-recall is equal to micro-precision.

Alternatively, the macro- versions of precision and recall are obtained by calculating precision and recall separately for each HISCO class, and taking the corresponding means of these values.

The effect of using the macro- measures in evaluating classification performance is that the individual performances for all HISCO classes are treated equally. This can lead to the performance for classes containing small numbers of occupational titles having undue influence on the overall measure. For this reason, micro-precision/recall is considered to give a more representative summary of classification performance in these experiments.

Table 3.4 Classification performance of individual classifiers (all records)

Classifier	Micro-precision/recall (%)	Macro-precision (%)	Macro-recall (%)
String similarity	29.0 ± 5.0	18.9 ± 0.6	14.1 ± 0.9
SGD	29.7 ± 22.1	20.7 ± 15.6	18.2 ± 13.7
Naive Bayes	90.1 ± 0.3	66.6 ± 0.6	69.4 ± 0.9

Table 3.4 shows the classification performance obtained using the three individual classifiers over five independent runs, with 95 % confidence interval. The confidence interval gives an indication of the degree of variation between runs: the smaller the confidence interval, the lower the variation. The interval is calculated such that there is a 95 % probability that the mean of a very large number of runs would lie within it.

In each run, 80 % of the records (51,200) in the HISCO-coded data set were randomly selected for training,[2] and the remainder used for testing (12,800). No feature selection or data cleaning was used.[3]

As expected, string similarity matching performed significantly better than chance, but still not particularly well. SGD performed very inconsistently, with micro-precision ranging between 4 and 50 % in individual runs. Naive Bayes performed consistently well, and indeed the results here were not significantly improved upon in subsequent more complex approaches.

It should be noted, however, that this method of evaluating classification performance is not entirely appropriate in the context of this project, because the data set contains duplicate occupational titles. There is no need to re-classify an occupational title that has been previously classified; it should always be assigned the same class. Thus calculating performance using all titles in the data set may yield over-optimistic results if there are many duplicates of titles that have been correctly classified.

Table 3.5 shows the results of applying automatic classification to the more representative problem of classifying each of the unique occupational titles (of which there were 9400 in total).

As might be expected, the classification performance is significantly poorer when measured using only unique occupational titles. Again, the performance of the Naive Bayes classifier is much better than the other individual classifiers.

Table 3.6 shows the classification performance obtained using the three ensemble classifiers, using the same procedure as in the previous set of experiments.

[2] For the string similarity classifier, which does not involve any training, these records were ignored.

[3] Where a record had been originally multiple-coded, one of the codes was arbitrarily selected as ‘correct’.

Table 3.5 Classification performance of individual classifiers (unique records)

Classifier	Micro-precision/recall (%)	Macro-precision (%)	Macro-recall (%)
String similarity	4.1 ± 0.4	6.5 ± 1.1	4.3 ± 0.7
SGD	19.7 ± 15.1	18.0 ± 13.5	15.1 ± 11.2
Naive Bayes	74.5 ± 0.9	60.5 ± 0.9	64.5 ± 1.1

Table 3.6 Classification performance of ensemble classifiers (unique records)

Classifier	Micro-precision/recall (%)	Macro-precision (%)	Macro-recall (%)
Majority voting	74.4 ± 0.9	60.4 ± 0.9	64.4 ± 1.1
Confidence-weighted (1)	64.7 ± 27.6	52.7 ± 21.3	54.7 ± 25.9
Confidence-weighted (2)	14.1 ± 0.8	19.6 ± 1.5	27.0 ± 3.1

None of the ensembles were able to improve on the performance of Naive Bayes. Rather than being able to combine the strengths of individual classifiers, the ensembles were dominated by Naive Bayes, which performed so much better than the others.

The majority voting ensemble produced the same results as Naive Bayes, due to an accident of implementation whereby that classifier is picked whenever there is no agreement. The first confidence-weighted ensemble also yielded the same results; clearly SGD was never sufficiently confident to outweigh Naive Bayes. The second confidence-weighted ensemble performed much worse, precisely because the SGD decision was sometimes selected.

Table 3.7 shows the effect of enabling feature selection for the relevant individual and ensemble classifiers.

This did not yield any significant improvements. Table 3.8 shows the effect of performing initial data cleaning by correction for the individual and ensemble classifiers, without feature selection. The classifications of 364 records in the data set were altered from an alternative classification to the most popular one for that occupation description. This represents 0.6 % of the data set.

Table 3.7 Classification performance with feature selection (unique records)

Classifier	Micro-precision/recall (%)	Macro-precision (%)	Macro-recall (%)
SGD	22.1 ± 7.5	19.2 ± 6.0	16.7 ± 5.0
Naive Bayes	74.1 ± 0.6	60.1 ± 0.6	65.1 ± 1.9
Majority voting	74.0 ± 0.6	59.9 ± 0.9	64.9 ± 2.3
Confidence-weighted (1)	73.9 ± 0.4	59.7 ± 0.8	64.4 ± 2.2
Confidence-weighted (2)	15.1 ± 1.4	20.5 ± 1.6	29.8 ± 4.7

Table 3.8 Classification performance with data cleaning by correction (unique records)

Classifier	Micro-precision/recall (%)	Macro-precision (%)	Macro-recall (%)
String similarity	4.3 ± 0.2	7.3 ± 0.4	4.7 ± 0.2
SGD	18.7 ± 7.0	17.1 ± 7.1	14.8 ± 5.4
Naive Bayes	75.0 ± 0.8	60.7 ± 1.4	65.4 ± 1.4
Majority voting	75.1 ± 0.4	61.2 ± 1.4	66.1 ± 0.9
Confidence-weighted (1)	75.3 ± 0.5	61.1 ± 1.2	66.0 ± 0.8
Confidence-weighted (2)	14.8 ± 0.4	20.4 ± 0.5	26.2 ± 0.9

Correction of multiple-coded descriptions reduced the performance of SGD, while for Naive Bayes it gave a very marginal improvement.

In conclusion, our measurements indicated no convincing reason to use any more complex approach than the straightforward Naive Bayes classifier. The wider question is whether the current levels of classification performance are good enough to be useful in this domain.

3.5 Future Directions

3.5.1 Potential of New Occupational Titles

Recalling that the context for this work is the classification of a large-scale data set, with the data becoming available in batches over a period of years, we are interested in automating management of the process of obtaining human-coded gold standard classifications, for model training and evaluation of accuracy.

At a given stage during the process, a certain proportion of the data has become available, and a certain proportion of that data has been human coded. The automatic classification process can be run on the human-coded data currently available, and its accuracy evaluated. For forward planning of the overall process, it would be useful to be able to estimate the following:

- the number of new occupational titles that will be encountered in the next batch of data (since duplicates of previously seen records can be classified trivially);
- the length of additional model training time that will be required when the next batch of data becomes available;
- the number of new occupational titles that will need to be human coded, and the cost in time and/or expenditure.

The eventual aim would be to produce an automatic classification system that would be able to estimate, at any point during the overall process, the current level of accuracy achieved, and the possible improvement that could be achieved with additional time and/or money.

To investigate the potential for estimating the number of unseen new occupational titles, we analysed the numbers of unique records in a range of data sets, and examined how these unique records were distributed. We hypothesised that as the records in a given data set were processed, the rate at which new unique records were encountered would gradually decline.

Figure 3.1 shows the progression of unique records as the Cambridge data set is processed. The x-axis represents the number of records that have been processed, and the y-axis the number of unique records that have been encountered thus far. The analysis was repeated ten times, with the data set processed in a different random order each time, to eliminate any peculiarities to do with the original record order. Error bars show 95 % confidence intervals, though the intervals are too narrow to be shown clearly.

Initially, the slope of the curve appears to decrease as the processing progresses. Although from visual inspection it may appear that the curve eventually settles into approximately linear growth, we show later in Fig. 3.4 that the proportion of unique records tends to be lower for larger data sets. This indicates that the rate of growth tends to continue falling.

We performed the same analysis on a number of other data sets. Figure 3.2 analyses a set of 40 million occupational titles from US censuses 1850, 1860, 1870, 1889, 1900 and 1910 (Minnesota Population Center 2008; Ruggles et al. 2010).

We also analysed the following data sets:

- 20 million occupational titles from Great Britain censuses 1851 and 1881 (Minnesota Population Center 2008; Schürer and Woollard 2003, 2008)
- 1.8 million occupational titles from Canada censuses 1852, 1881, 1891 and 1901 (Minnesota Population Center 2008; Darroch and Ornstein 1979; Dillon 2008; Canadian Families Project 2002)

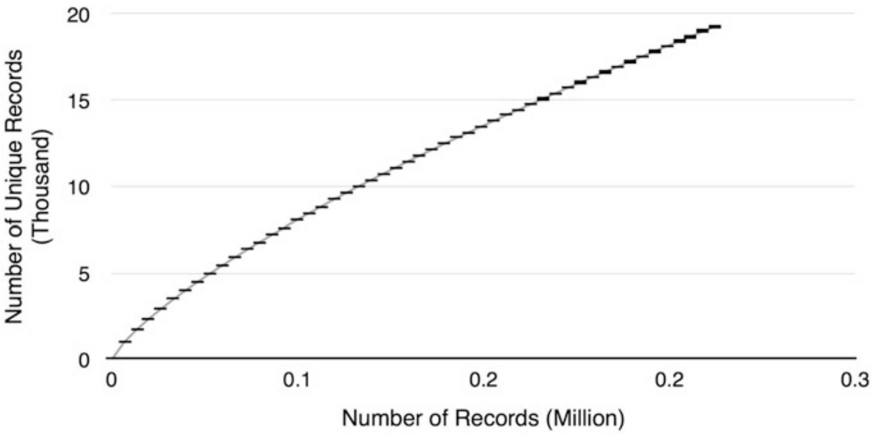


Fig. 3.1 Progression of unique records in Cambridge data set

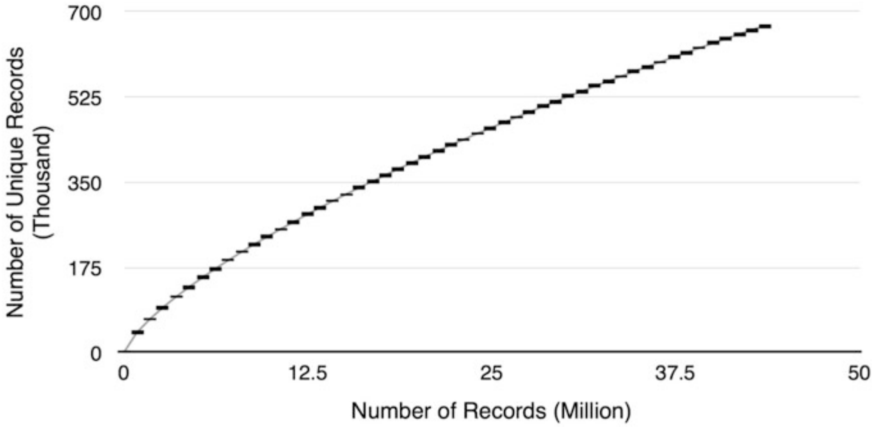


Fig. 3.2 Progression of unique records in NAPP US data set

- 250,000 occupational titles from Norway censuses 1801, 1865, 1875, 1900 and 1910 (Minnesota Population Center 2008; Norwegian Digital Archive 2008a, b; Norwegian Historical Data Centre 2008)
- 200,000 occupational titles from Dutch vital event records (Historical Sample of the Netherlands 2010)

Overall, the six occupational data sets analysed exhibit very similar patterns, once differences in data set size are accounted for. Figure 3.3 plots all six progression curves after normalising both the data set size and the total number of unique records.

The distribution of unique occupational titles throughout a data set appears to be strongly consistent, as shown by the fact that the curves cannot be easily distinguished. The curves are consistent, regardless of the origin and size of the data sets, and the order in which the occupational titles are processed. We tentatively conclude that:

- previously unseen occupational titles are likely to continue to arise in significant numbers as data accumulates during a long-running classification process; and
- there is some reason for optimism that we may be able to estimate future numbers of such unseen occupational titles at intermediate points during the project.

Figure 3.4 shows the same data as Fig. 3.3, with the y-axis values normalised to the overall size of each data set. This illustrates that the proportion of unique occupational titles within the data sets varies from less than 2 % (US) to nearly 20 % (Netherlands). To a large extent this variation corresponds with the different sizes of the data sets: the largest set has the smallest proportion of unique occupational titles. This indicates that the rate of accumulation of unique occupational

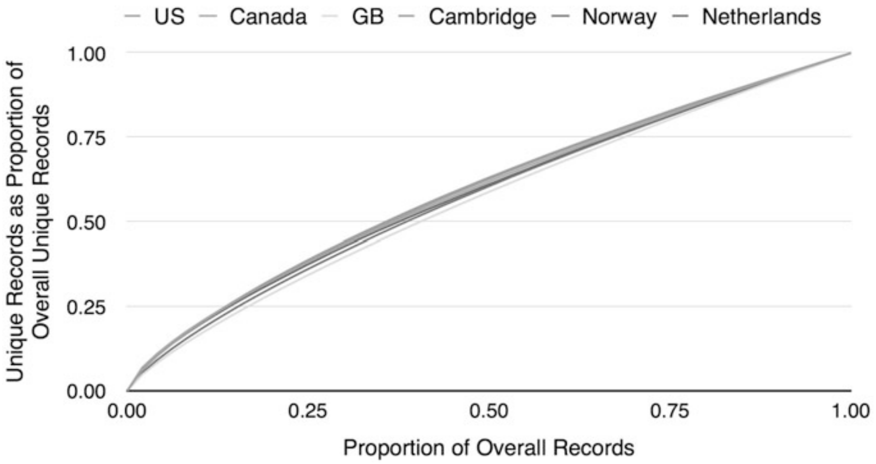


Fig. 3.3 Progression of unique records normalised to overall number of unique records

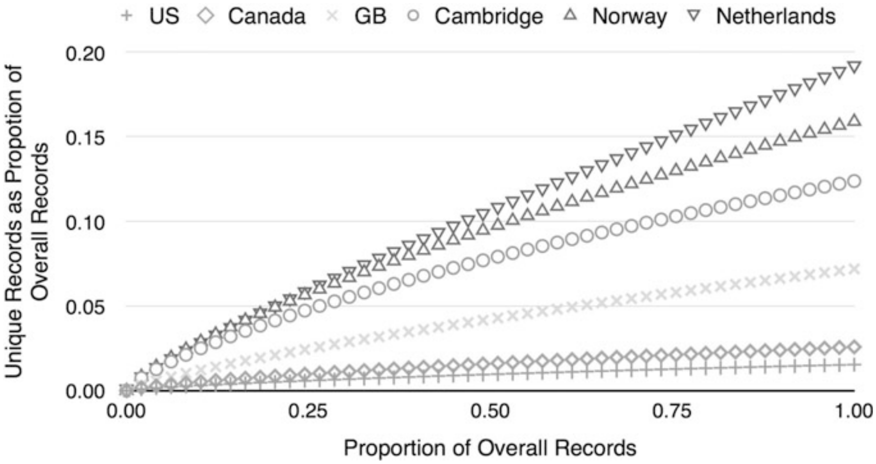


Fig. 3.4 Progression of unique records normalised to overall number of records

titles tends to continue falling as the data set grows. The Canada data set is different in this respect, with a relatively low proportion of unique occupational titles even though the data set is relatively small.

To investigate whether the observed patterns are particular to occupational titles, we performed the same analysis on two data sets containing ‘cause of death’ descriptions. Figure 3.5 analyses a set of cause of death records from Massachusetts in the period 1850–1912, provided by Susan Leonard (Leonard et al. 2012).

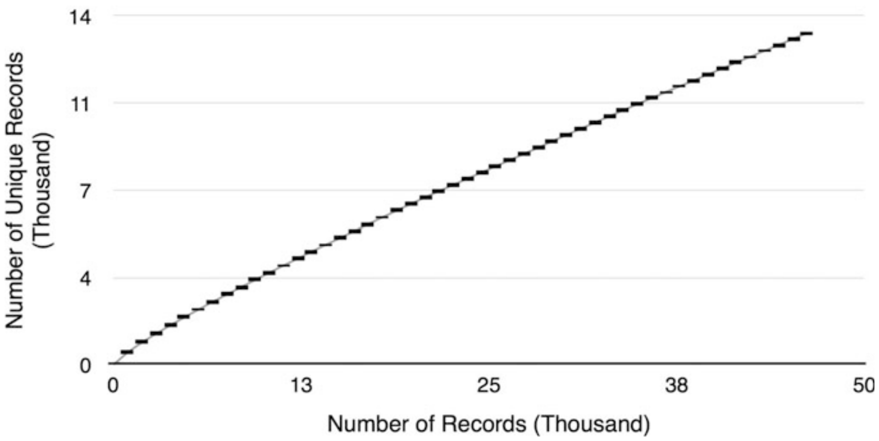


Fig. 3.5 Progression of unique records in Massachussetts cause of death data set

Figure 3.6 analyses a set of cause of death records from the UK in the period 1990–2012 (National Records Scotland 2014). Thus, very similar patterns were observed for occupations and causes of death.

The next step will be to investigate whether the unique record curve observed during the part of a classification process already completed can be extrapolated to provide a good estimate of the number likely to be encountered in the next phase. We plan to investigate whether previous work on estimating vocabulary size for historical authors can be applied to this problem (Efron and Thisted 1976).

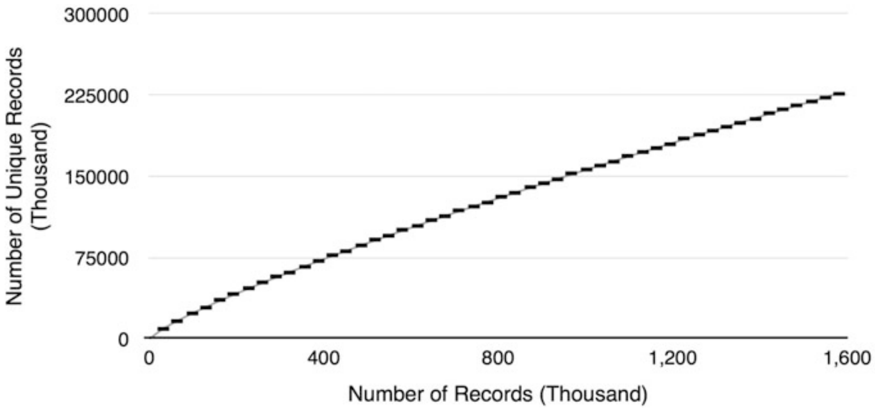


Fig. 3.6 Progression of unique records in UK cause of death data set

3.5.2 Future Work

Other directions for further development of this work include:

- a manual review of the occupational titles identified as incorrectly classified during evaluation using human-coded data, looking for any common patterns that might be candidates for additional cleaning steps;
- automatic selection of the most suitable classifier (whether individual or ensemble, and with or without spelling correction using various dictionaries), guided by experiments on samples from the data set;
- experimentation with other data cleaning measures such as spelling correction;
- experimentation with other string similarity metrics, and adding the ability for similarity-based classifiers to generate proxy confidence values, allowing them to be incorporated into confidence-weighted ensembles;
- investigation of the applicability of this work to the classification of text fragments from other domains of genealogical interest, for example the coding of family names to standard spellings.

3.6 Conclusion

The best automatic classification performance was achieved using the individual Naive Bayes classifier with cleaning of the data by correction of multiple-coded descriptions. This yielded micro-precision/recall of 75 %, macro-precision of 60.7 % and macro-recall of 65.4 %. This means that 75 % of the automatic classification decisions were correct. The lower macro-figures are due to poorer classification performance on a number of rare HISCO classes for which low numbers of occupational titles occurred in the data set. This poor performance is likely to be caused, at least to some extent, by the lack of training set examples of such classes adversely affecting the machine learning process.

It is not yet clear whether automatic classification with this level of performance, taken with any further improvements achievable through the planned further work, will be cost-effective for the task of occupation coding of the full Scottish data set. It is worth noting that the performance metrics have been calculated using the most specific classes of the HISCO classification. For some research uses, it may be sufficient to have occupational titles classified to more general levels of the HISCO hierarchy, and it is possible that classification performance will be better when assessed against these wider classes.

The performance of the ensemble approaches was disappointing, yielding no improvement over the individual Naive Bayes classifier. It is possible, though, that there is still benefit to be had from an ensemble approach if other individual classifiers with performance comparable to Naive Bayes can be found. Although the SGD classifier performed very poorly here, it did perform well on some data sets in our previous work on cause of death classification. This leads us to think that further investigation of its pathological behaviour here would be worthwhile.

In conclusion, the use of machine learning classifiers and ensembles appears to be a potentially promising method for coding large data sets. We are continuing to investigate how to improve classification performance, and how to automate the overall process of selecting a classifier, deciding training set size, running the classification and validating the results.

Acknowledgments This work was supported by the Economic and Social Research Council grant numbers ES/K00574X/1 (Digitising Scotland), ES/L007487/1 (Administrative Data Research Centre—Scotland).

This chapter is an expanded version of a paper presented at the International Workshop on Population Reconstruction in Amsterdam, February 2014 (Kirby et al. 2014). We thank the original anonymous referee, many workshop participants and the book editors, for their helpful comments.

We thank Richard Zijdeman for advice on historical coding and access to data from the Netherlands.

References

Apache Software Foundation. (2011). *Apache Mahout: Scalable machine learning and data mining*. http://mahout.apache.org/. Accessed 1 Nov 2014.

Bottero, W., & Prandy, K. (2001). Women's occupations and the social order in nineteenth century Britain. *Sociological Research Online*, 6(2).

Canadian Families Project. (2002). *National sample of the 1901 census of Canada*. Victoria: University of Victoria.

Carson, J. K., Kirby, G. N. C., Dearle, A., et al. (2013). Exploiting historical registers: Automatic methods for coding C19th and C20th cause of death descriptions to standard classifications. *New Techniques and Technologies for Statistics*, 598–607.

Darroch, G., & Ornstein, M. (1979). *Canadian historical social mobility project. National sample of the 1871 census of Canada* [computer file]. Toronto: York Institute for Social Research and Department of Sociology, York University.

Dietterich, T.G. (2000). Ensemble methods in machine learning. In: Multiple classifier systems. *Lecture notes in computer science*, Vol. 1857, (pp. 1–15). Heidelberg: Springer.

Dillon, L. (2008). *1881 Canadian census project, North Atlantic population project, and Minnesota population center*. National Sample of the 1881 Census of Canada (version 2.0). Montréal: Département de Démographie, Université de Montréal [distributor].

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika, 63*(3), 435–447.

HISCO. (2013). *HISCO tree of occupational groups*. http://historyofwork.iisg.nl/major.php. Accessed 1 Nov 2014.

Historical Sample of the Netherlands (HSN). (2010). *Data set life courses release* 2010.01.

Kirby, G. N. C., Carson, J. K., Dunlop, F. R. J., et al. (2014). Automatic methods for coding historical occupation descriptions to standard classifications. In: *Proceedings Workshop on Population Reconstruction*, Amsterdam, February 2014. International Institute of Social History.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings 14th International Joint Conference on Artificial Intelligence* (pp. 1137-1143). Burlington: Morgan Kaufmann Publishers Inc.

Komarek, P. (2004). *Logistic regression for data mining and high-dimensional classification*. PhD Thesis, Carnegie Mellon University.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. In: *Proceedings AAAI-92*, pp. 223–228.

Leonard, S. H., Anderton, D. L., & Swedlund, A. C. (2012). *Grammars of death*. University of Michigan/ICPSR. https://sites.google.com/a/umich.edu/grammars-of-death/home. Accessed 1 Nov 2014.

Minnesota Population Center. (2008). *North Atlantic population project: Complete count microdata*. Version 2.0 [Machine-readable database]. Minneapolis.

National Records Scotland. (2014). *Set of cause of death records.*

Ng, A. Y., & Jordan, M. I. (2001). On Discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In: *Proceedings Neural Information Processing Systems* (pp. 841–848).

Norwegian Digital Archive (The National Archive), Norwegian Historical Data Centre (University of Tromsø) and the Minnesota Population Center. (2008a). *National sample of the 1865 census of Norway*, Version 2.0., Tromsø.

Norwegian Digital Archive (The National Archive), Norwegian Historical Data Centre (University of Tromsø) and the Minnesota Population Center. (2008b). *National sample of the 1900 census of Norway*, Version 2.0. Tromsø.

Norwegian Historical Data Centre (University of Tromsø) and the Minnesota Population Center (2008). *National sample of the 1875 census of Norway*,Version 2.0. Tromsø.

Prandy, K., & Lambert, P. (2012). *CAMSIS: Bibliographic review*. http://www.camsis.stir.ac.uk/review.html. Accessed 1 Nov 2014.

Ruggles, S., et al. (2010). *Integrated public use microdata series: Version 5.0* [Machine-readable database]. Minneapolis: University of Minnesota.

Schürer, K., & Woollard, M. (2003). *National sample from the 1881 census of Great Britain* [computer file], Colchester: History Data Service, UK Data Archive [distributor].

Schürer, K., & Woollard, M. (2008). *National sample from the 1851 census of Great Britain* [computer file], Colchester: History Data Service, UK Data Archive [distributor].

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association, 88*(422), 486–494.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*(4), 427–437.

US Bureau of Labor Statistics. (2010). *Standard occupational classification (SOC) system*. http://www.bls.gov/soc/. Accessed 1 Nov 2014.

van Leeuwen, M. H. D., Maas, I., & Miles, A. (2002). *HISCO: Historical international standard classification of occupations*. Leuven: Leuven University Press.

Witten, I., & Eibe, F. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). Burlington: Morgan Kaufmann.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In: *Proceedings 14th International Conference on Machine Learning*, ACM (pp. 412–420).

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: *Proceedings 21st International Conference on Machine Learning*, ACM (pp. 116–123).

Chapter 4
Learning Name Variants from Inexact High-Confidence Matches

Gerrit Bloothooft and Marijn Schraagen

Abstract Name variants which differ more than a few characters can seriously hamper record linkage. A method is described by which variants of first names and surnames can be learned automatically from records that contain more information than needed for a true link decision. Post-processing and limited manual intervention (active learning) is unavoidable, however, to differentiate errors in the original and the digitised data from variants. The method is demonstrated on the basis of an analysis of 14.8 million records from the Dutch vital registration.

4.1 Introduction

In record linkage, the decision to make a link between two instances of information can be complicated by spelling variation, variants (translation, suffix variation, changes in order of name elements, etc.) and errors. A usual approach to cope with this kind of variation is to define a distance or similarity metric (at written or phonemic level) to describe the spelling difference between two names. If the difference between the names is less than some threshold, they are considered to be variants and may indicate the same person (Christen 2012). This approach has the limitation that (1) the same threshold is used for all names, while a name-dependent threshold may be more effective (although distance measures may incorporate this to some extent), and (2) the threshold is chosen arbitrarily or is at best decided upon by its overall effect: the linkage process should not produce too much overlinking,

G. Bloothooft (✉)
Utrecht Institute of Linguistics-OTS, Utrecht University, Trans 10, 3512 JK, Utrecht
The Netherlands
e-mail: g.bloothooft@uu.nl

M. Schraagen
Leiden Institute of Advanced Computer Science, Leiden University, Leiden
The Netherlands
e-mail: m.p.schraagen@liacs.leidenuniv.nl

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_4

i.e. too many false links. Although small variation in names can be identified in this way, larger variation, such as between *Jan* and *Johannes*, is usually beyond a threshold. On the other hand, small differences in names, like the surnames *Bos* and *Vos*, do not always imply a genuine variant. These observations indicate the need for a corpus which explicitly describes name variants that could have been used for the same person. Experts could help in the laborious task to construct such a corpus, but it would be efficient if these variants could be learned at least in part automatically from data.

There are circumstances where sources are rich enough to allow for record linkage while not using all available information. Names that are not needed in the matching process, may contain true variants (but errors as well). This chapter investigates the procedures needed to construct a corpus of true name variants in a largely automated way, applied to 14.8 million records from the nineteenth century vital registration of the Netherlands.

To derive name variant pairs, record links based on several elements or fields (e.g., the names of various people mentioned in a record) are examined. In case one of the fields differs between the records while the other elements are exactly equal, the differing field values are assumed to contain a name variant. After variant construction, post-processing using rules and heuristics takes place to remove erroneous variant pairs.

The chapter is structured as follows: Sect. 4.2 describes related work, Sect. 4.3 describes the source data, Sect. 4.4 describes the method to collect name variants, while Sect. 4.5 discusses the options to differentiate true variants from errors. In Sect. 4.6 results are presented and evaluated. The possibility to use name variants for clustering and name standardisation is explored in Sect. 4.7, including an extra iteration of the main name pair construction method. In Sect. 4.8 a comparison is made with a name variant corpus of FamilySearch, and Sect. 4.9 concludes.

4.2 Related Work

The basic name variant derivation procedure can be compared to corpus-based stemming of regular text using co-occurrence of terms in a document. An example is discussed in Xu and Croft (1998), where the basic assumption is that word variants that should be conflated will co-occur in a (typically small) text window. The approach of Xu and Croft is intended to address issues in rule-based and dictionary-based stemming. A text window in a document can be compared to a pair of linked records, in the sense that in both cases sufficient information is present to conflate variants. However, the construction of record links is non-trivial, which complicates the current approach. On the other hand, the structure of a record is given by the division into fields, in contrast to the structure of a natural language sentence. This reduces the need for complex co-occurrence statistics in the current approach.

The current method of extracting name variants from record matches is essentially a network approach, in which ambiguous nodes can be combined if the sets of connected nodes (in this case other names in a record) are similar (Malin 2005; Getoor 2007). However, the current dataset is represented as a simple network in which records are small, equally sized unconnected subgraphs, therefore reducing the need for elaborate graph traversal algorithms.

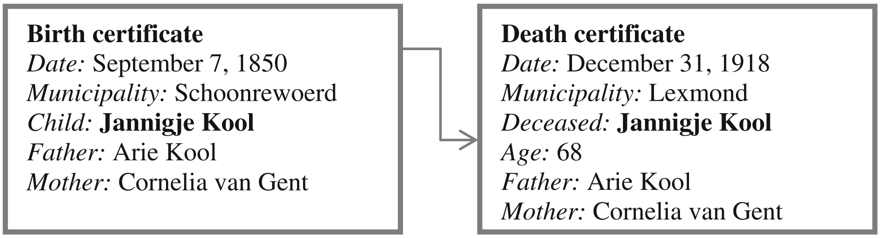
Name variant construction from data has been discussed by Driscoll (2013), who uses text patterns to search for name-nickname variants on web pages, combined with various morphological rules and matching conditions, partly automatically induced from the initial variant pairs. The results are promising, especially when all methods are combined using an automatically derived weight for each method. A key component of the current approach is however not used by Driscoll, which is the selection of a large amount of candidate pairs from record links. The overall quality of the candidates in this selection allows the rules applied in the current method to be less strict, which improves coverage without a large increase in error rate.

4.3 Material

The data used in the investigation is extracted from the Dutch *WieWasWie* (who was who) database (www.wiewaswie.nl, release November 2011 as Genlias). *WieWasWie* contains civil certificates from the Netherlands, for the events of birth, marriage and death, of which the registration started in 1811. Most documents originate from the nineteenth and early twentieth centuries. A record consists of the type of event, a serial number, a date and a place and information about the participants. The parents are listed for the main subject(s) of the document, i.e. the newborn child at birth, the bride and groom at marriage, and the deceased person at death, respectively. The documents do not contain identifiers for individuals and no links are provided between documents. The digitisation of the certificates is an ongoing process that is performed by volunteers. For the 2011 release it is estimated that key information from 30 % (4.1 million) of the birth certificates, 90 % (3.1 million) of the marriage certificates, and 65 % (7.6 million) of the death certificates has been made available. This concerns about 55 million references to individuals. These references include 101,830 different male first names, 128,800 different female first names, and 565,647 different surnames (all singular elements, if necessary derived from composite names).

4.4 Method and Variant Pair Construction

The three different types of certificates (birth, marriage and death) all contain information about individuals and their parents. This information can be used for matching, as illustrated in Fig. 4.1. The method described in this chapter uses the



Birth certificate
Date: September 7, 1850
Municipality: Schoonrewoerd
Child: **Jannigje Kool**
Father: Arie Kool
Mother: Cornelia van Gent

Death certificate
Date: December 31, 1918
Municipality: Lexmond
Deceased: **Jannigje Kool**
Age: 68
Father: Arie Kool
Mother: Cornelia van Gent

Fig. 4.1 Example match between birth and death certificates with *Jannigje Kool* as main person

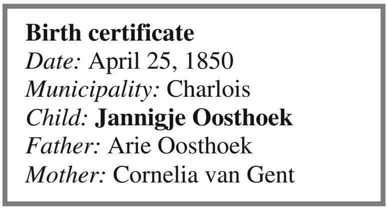
assumption that true record matches can be found using a subset of the available information, which enables construction of variant pairs based on the remaining information.

In the application of the method we required exact matching of the first name of the main person and equal year of birth (derived from age in marriage and death certificates, plus or minus one year). Moreover, three out of the four names of the mother and father should match exactly as well. Note that in the Netherlands women always keep their maiden name in the administration. The fourth name of the mother or the father was not part of the linkage decision and open to variation and, if differing between the two records, generated a name variant pair. This could concern a male (father) or female (mother) first name, or a surname (father or mother).

We tested whether the requirement of matching four out of five names plus year of birth was sufficient for accurate record linkage by selecting matches between birth and death certificates for which *all* five names and year of birth were available and matched, under the assumption that an exact match on all available names would generate only true matches in our dataset. Subsequently, one of the four names of the parents was ignored and it was counted in how many cases more than one link was generated. This was the case for only 85 out of 1,107,162 matches. We considered this as sufficient support for our assumption that three out of four equal names of parents were a sufficient condition for accurate linkage, as violation would generate only a few errors.

An example of a rare match where the condition did not hold is: 7 September 1850 birth of *Jannigje Kool* in Schoonrewoerd, from the parents *Arie Kool* and *Cornelia van Gent*, and her decease in 31 December 1918 in Lexmond, at 68 years of age with mention of the same parents (see Fig. 4.1). Competing is the birth of *Jannigje Oosthoek* on 25 April 1850 in Charlois, from the parents *Arie Oosthoek* and *Cornelia van Gent* (see Fig. 4.2). Although there are matches of the names *Jannigje*, *Arie*, *Cornelia*, and *van Gent*, and the year of birth 1850, this leads to the erroneous surname variant pair *Kool/Oosthoek*. The—disentangling—place of birth was not used in the matching decision, as this information is error prone, especially when mentioned in the death certificates.

A name can consist of several elements, such as the first name *Johan Willem Frederik*. Although we required identity of four out of five (full) names of a person



Birth certificate
Date: April 25, 1850
Municipality: Charlois
Child: **Jannigje Oosthoek**
Father: Arie Oosthoek
Mother: Cornelia van Gent

Fig. 4.2 Competing birth certificate for the birth certificate in Fig. 4.1

Table 4.1 Examples of variant pairs constructed from composite names

Name 1	Name 2	Differences	Name pair
Anna Christina Elizabeth	*Christiena Elizabeth*	Missing name, name variant	*Christina/Christiena*
Virgin Thomasa Franken	*Thomasa Virginia*	Reversed order, missing name, name variant	*Virgin/Virginia*
Adriana Agnita Cornelia	*Adriana Cornelia*	Missing name	None

and parents, in case of composite names, more single names were involved and thus provide stronger support for a true link. A considerable 50 % of the name pairs were accompanied by five or more identical single names in the comparison, instead of the four identical names minimally required.

The selection on differences in the fourth name of a parent over all combinations of birth, marriage and death records resulted in 897,426 name pairs. For composite names the difference could be caused by different name order, missing names and/or actual name variation. For these cases variants are identified using alignment of name elements based on minimal edit distance, see Table 4.1 for examples.

After this compositional analysis, which was also performed for surnames, pairs of single names remained. Since the order of names in a pair is unimportant, name pairs with opposite order were taken together. The results of this step are shown in Table 4.2. The most frequent name variant pairs, which have only minor spelling differences that mostly do not influence pronunciation, are also shown in Table 4.2.

Table 4.2 Variant construction results and examples

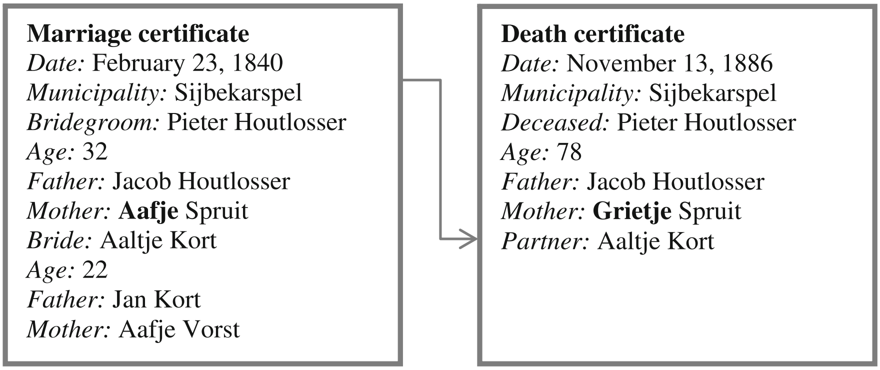
Female first name		Male first name		Surname	
Pairs	Tokens	Pairs	Tokens	Pairs	Tokens
48,684	246,519	31,885	183,050	177,258	374,901
Most frequent variant pairs					
Elisabeth/Elizabeth		*Johannes/Johannis*		*Jansen/Janssen*	
Willemina/Wilhelmina		*Jacob/Jakob*		*Bruin/Bruijn*	
Geertrui/Geertruij		*Arij/Arie*		*Ruijter/Ruyter*	

4.5 Variants and Errors

In this research we wish to construct a clean corpus of name variant pairs, but name errors complicate the process, also when the record links themselves are correct. Name errors can not only originate from the writing of the original certificates, but also from misreading or typing errors in the recent digitisation process, or result from violation of the assumption that four out of five equal names and equal year of birth describe a person uniquely (rare, but shown before). Where true name variants can replace each other in any condition and thus help record linkage under less favourable conditions, name errors should be recognised as such and not be propagated.

As an example of a registration error we consider *Pieter*, born in 1808 as son of *Jacob Houtlosser* and *Aafje Spruit*, as mentioned in the marriage certificate (see Fig. 4.3). But his death certificate mentions *Grietje Spruit* as mother, resulting in the erroneous first name variant pair *Aafje/Grietje*. Additional evidence that the records concern the same person comes from the partner name *Aaltje Kort*, mentioned in both certificates, and the correspondence in municipality (although place and partner information is not used in the matching process).

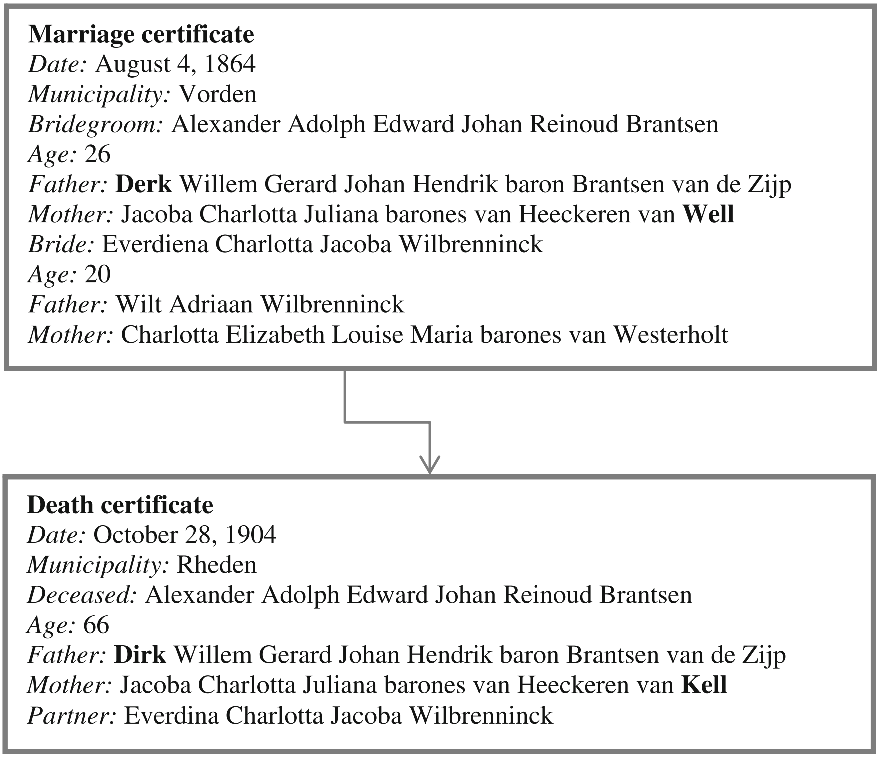
A distinction between a true variant (*Dirk/Derk*), and an error (*Dirk/Klaas*) is not at all easy to make. We chose for a definition of true variants as names that belong to the same lemma, while errors do not. A lemma [see, e.g., Bratley and Lusignan (1976)] is a usually etymologically based name from which by processes of pronunciation, suffixation, abbreviation, etc., derivate forms can be generated. These processes are very difficult to model or to predict and therefore it is hard, if not impossible to differentiate automatically between a true variant and an error. In many cases onomastic or linguistic expertise is required.



Marriage certificate
Date: February 23, 1840
Municipality: Sijbekarspel
Bridegroom: Pieter Houtlosser
Age: 32
Father: Jacob Houtlosser
Mother: **Aafje** Spruit
Bride: Aaltje Kort
Age: 22
Father: Jan Kort
Mother: Aafje Vorst

Death certificate
Date: November 13, 1886
Municipality: Sijbekarspel
Deceased: Pieter Houtlosser
Age: 78
Father: Jacob Houtlosser
Mother: **Grietje** Spruit
Partner: Aaltje Kort

Fig. 4.3 Record match resulting in the erroneous variant pair *Aafje/Grietje*



Marriage certificate
Date: August 4, 1864
Municipality: Vorden
Bridegroom: Alexander Adolph Edward Johan Reinoud Brantsen
Age: 26
Father: **Derk** Willem Gerard Johan Hendrik baron Brantsen van de Zijp
Mother: Jacoba Charlotta Juliana barones van Heeckeren van **Well**
Bride: Everdiena Charlotta Jacoba Wilbrenninck
Age: 20
Father: Wilt Adriaan Wilbrenninck
Mother: Charlotta Elizabeth Louise Maria barones van Westerholt

Death certificate
Date: October 28, 1904
Municipality: Rheden
Deceased: Alexander Adolph Edward Johan Reinoud Brantsen
Age: 66
Father: **Dirk** Willem Gerard Johan Hendrik baron Brantsen van de Zijp
Mother: Jacoba Charlotta Juliana barones van Heeckeren van **Kell**
Partner: Everdina Charlotta Jacoba Wilbrenninck

Fig. 4.4 Record match resulting in the correct variant pair *Derk/Dirk* and the erroneous variant pair *Well/Kell*. Note that the variant pair *Everdiena/Everdina* is not considered in the procedure

There can be substantial evidence that the same persons are concerned (for instance in case of long composite names), but this does not exclude errors. An example is shown in Fig. 4.4 where our onomastic knowledge tells us that *Dirk/Derk* is a genuine first name variant, while *Kell/Well* relates to miswriting, misreading or mistyping and should not be generalised beyond this single occurrence. Unfortunately, this differentiation between a true and erroneous name variant pair cannot be made automatically.

Also the frequency of a variant pair (or its probability) is of limited help. Both errors and variants can be rare or frequent. Frequent erroneous variants for male first names are for instance combinations of popular names, see Table 4.3. The use of rules and manual inspection (active learning) is unavoidable to make a distinction between variants and errors.

Table 4.3 Erroneous name pairs consisting of popular names

Name 1	Frequency (in millions)	Name 2	Frequency (in millions)	Interchange frequency
Jacob	0.50	*Jan*	2.39	331
Hendrik	1.11	*Jan*	2.39	212
Jacobus	0.39	*Johannes*	1.37	149
Willem	0.88	*Jan*	2.39	138
Gerrit	0.73	*Hendrik*	1.11	104
Gerrit	0.73	*Jan*	2.39	99
Willem	0.88	*Hendrik*	1.11	82
Gerrit	0.73	*Cornelis*	0.86	63
Klaas	0.33	*Jan*	2.39	60
Dirk	0.35	*Jan*	2.39	59

The frequency of the individual names in the WieWasWie 2011 corpus, as well as the frequency that name 1 is erroneously replaced by name 2 (or reversely) is given

4.5.1 Name Pair Cleaning

The name pairs resulting from the automatic analysis are post-processed in order to remove erroneous pairs. Three different methods have been applied, of which the first method uses an external manually compiled name lexicon, the second method developes and uses a corpus of non-variants, and the third method is based on manually designed variant classification rules (see diagram in Fig. 4.5). The methods are described in detail in the remainder of this section. In case of acceptance by the first method or rejection by the second method application of the third method was not needed. Additional manual review of a limited selection of variant pairs was applied to correct post-processing errors. The selection of pairs for review has been performed in an Active Learning setting (see Olsson (2009) for an overview, Sarawagi and Bhamidipaty (2002) for an application in nominal record linkage), considering pairs based on the frequencies of both the name pair and the individual names in the pair. The frequency values act as a confidence score which allows the algorithm to automatically single out pairs for which manual review is useful without the need to manually evaluate every single pair.

4.5.1.1 Using Name Dictionaries

Variants share a lemma and errors do not. The decision that names share a lemma can be based on expert onomastic knowledge, as laid down in name dictionaries. If available, the content of the dictionaries is usually much more limited than the name variation found in current resources. For the Netherlands, a dictionary of first names (van der Schaar 1964, first edition) is available which associates about 20,000 first names to 3737 gender-independent lemmas. This could be helpful as a starting

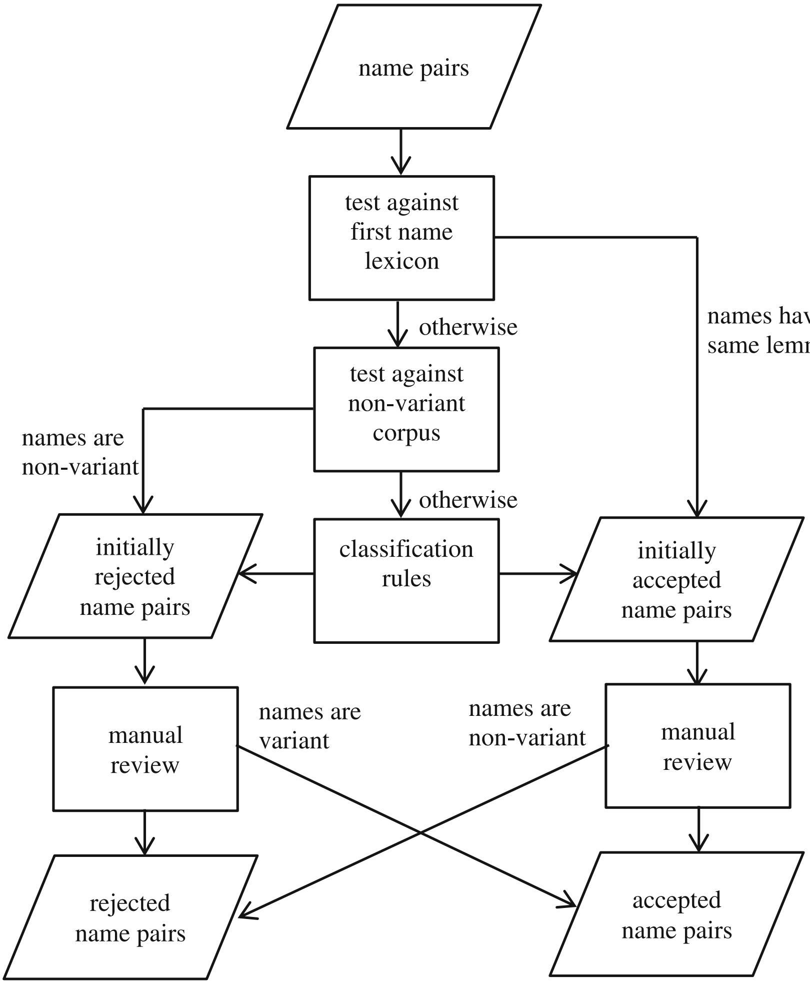


Fig. 4.5 Flow chart of post-processing of name pair variants

point for the identification of many more name variants, but there are a number of limitations. Many (abbreviated) names in the dictionary are associated to more than one lemma, especially short names. For instance, *Aai* with lemmas *Aai, Aalt, Adriaan*. Furthermore, lemmas can be too refined (given our observations of variation in practice), such as *Adagonda, Adelgonde* and *Aldegonde*. Sometimes association has subtle differentiation, such as *Nelie* with the lemma *Cornelis*, *Nelly* with the lemma *Cornelis* or *Petronius* (as *Petronella*), and *Nella* with lemma *Petronius*, which does not seem to conform to the use of names in practice either.

In our case, the dictionary has been used to accept first name variants that share a lemma, while making no decision on names that are associated to different lemmas.

A total of 3615 female and 2878 male first name pairs were accepted, which is about 5 % of all pairs found. The main gain of this approach is that we can accept name pair variants that differ so strongly that they would not make it through the rules we apply later. This avoids manual intervention for them. For surnames, a comparable dictionary is not electronically available for the Netherlands.

4.5.1.2 Data-Driven Identification of Erroneous First Name Pairs

A data-driven option to identify first name variant *errors* is based on the assumption that first name variants do not show up together in a composite name (Oosten 2008). This would imply that names that *do* show up in a composite name are not variants. From the first name *Anne Maria Helena* we may then conclude that *Anne*, *Maria* and *Helena* are no variants from each other.

This method was tested using all (possible composite) first names from the Dutch WieWasWie 2011 release. These names have 55 million tokens. From the composite first names in this set, all combinations of two names were determined, keeping the order of appearance from left to right in the composite name. This resulted in a non-variant-corpus of 907,660 pairs of first names, with 18 million tokens.

However, the dataset contains errors introduced in the digitisation phase. Patronymics [referring to the first name of the father, such as *Jansz*, short for *Jan's zoon*, son of *Jan,* cf. Anderson (2007)] and parts of the surname, or the whole surname, were sometimes included in the first name field. For example *Aagtje van Eck*, with *van Eck* as surname, is present as a first name, which results in the incorrect non-variant pairs (*Aagtje/van*), (*Aagtje/Eck*), (*van/Eck*). To exclude these errors, we required that name pairs should be seen in both orders, under the assumption that it is unlikely to find a patronymic or a surname before the first name.

Another problem was first name fields with descriptive content, such as *zoon van Geertruida* (*son of Geertruida,* 1 time) and *Aleida Geertruida van (Aleida Geertruida from,* 55 times), which resulted in the erroneous first name pair (*Geertruida/van*), seen in both orders. These name pairs were excluded by requiring a capital initial and a name length of at least three characters (which also excluded single initials). After this, a non-variant-corpus of 118,532 first name pairs resulted (only 13 % of the originally collected pairs), with 15 million tokens (83 %).

In conflict with the assumption of the approach, however, also true variant pairs show up jointly in a composite first name. Frequent examples were *Jan Johannes, Neeltje Cornelia, Arie Adrianus, Jannetje Johanna*, indicating that parents did not mind or even did not realise the common basis of both names. After removal of the pairs that have the same lemma in the dictionary of first names and a few manual corrections, the no-variant-corpus was held against the name variant set and resulted in the removal of 2458 female first name pairs and 2343 male first name pairs. The advantage of this approach is that we can exclude erroneous name pair variants that would pass the rules we apply in the subsequent step.

4.5.1.3 Rules That Accept Name Variant Pairs

If there were no errors in the source material, our method would not require additional cleaning methods. But since the source material is not error free, additional methods are needed, and the application of rules is one of them. Our rules can be much more relaxed, however, than rules that apply on any pair of names as they are used on a pre-selected corpus of name pairs.

Two sets of rules are applied: a first set of rules converts a name into a semi-phonetic form, while a second set of rules compares the differences between two names on the basis of Levenshtein distance and additional requirements that resemble the Jaro-Winkler distance measure (Winkler 1990). Both sets will be explained in this section.

In the past, the lack of spelling rules has promoted variation in spelling. Attempts can be made to apply rules on written forms that result in a version that has close correspondence to the original pronunciation. Since it is impossible to catch all spelling variation (especially under the presence of all kinds of errors), a limited but robust rule set was developed that converts names from Dutch sources into a semi-phonetic form (Bloothooft 1995). Semi-phonetic implies that although the coding is inspired by the conversion of written characters into speech sound symbols, no attempt has been made to arrive at a correct phonetic transcription, which is impossible under the presence of unpredictable writing or digitisation errors. This rule set resembles other approaches to phonetic encodings [Soundex (Russell 1918), Double Metaphone (Philips 2000), cf. also Dolby (1970)], but is tailored towards properties of the relation between spelling and pronunciation for Dutch.

Major rules are (1) symbol simplification by ignoring diacritics, (2) reducing all character replications to a single symbol, (3) reducing all vowel combinations to single symbols, (4) rules for resolving the ph, gu, ch and ck combinations, and (5) rules for the letters c, d, h, j, q, v, x, z. Examples are *Jannigje* > JANYGJE, *Cornelia* > KORNELYA, *Jozeph* > JOSEF. In further processing, this semi-phonetic form of a name was used.

A second set of heuristic rules was adopted that limits the acceptable differences between two names. A variant pair that complies with a rule was accepted. Major ingredients were the Levenshtein distance between the names, the name lengths, and number of identical (semi-phonetic) initials (at least one). These rules have some relationship to the Jaro-Winkler distance measure, but are more relaxed.

There is a considerable group of name pairs that result from (understandable) misreading of the initial. Frequent misreadings are found between the initials *T* and *F, P, J, S* or *K; F* and *P* or *J; I* and *J;* and *M* and *H, W,* or *Al.* The difficulty of misreading (at the digitisation phase) is that there is often a bias towards an (erroneous) existing name on the basis of the knowledge of the person who digitises (for instance, the first name pairs *Pietje/Tietje, Jannetje/Tannetje, Wessel/Hessel,* and the surname pairs *Tol/Pol, Meijden/Heijden, Noort/Voort*). If this misreading

Table 4.4 Heuristic rules containing thresholds for variant pair acceptance

Levenshtein distance	Length	Minimum length of shared prefix	Example
1	Shortest >4	1	*Joanna, Johanna*
2	Shortest >4	2	*Gerrit, Geurt*
3	Longest >5	3	*Annegien, Annigje*
4	Longest >7	4	*Laurentius, Laurijs*
5	Longest >8	4	*Franciscus, Frans*
Total length of pair minus Levenshtein distance >16		1	*Lingmandus, Luigmondus*

In addition to the six rules specified in this table, a more complex seventh rule on suffixes is explained in the text. Rules are applied both to the original and semi-phonetic form of a name

happens systematically, the resulting name confusion needs not even be rare. Automatic detection of them is difficult because the Levenshtein distance is small (only 1 because of the initial). Therefore we required by rule the same initial in the name pair, and more equal initial characters for more relaxed conditions of the Levenshtein distance between the names (at the semi-phonetic level, which already takes care of the major genuine spelling variation of the initials).

Rules are summarised in Table 4.4. These rules were applied to both the original and the semi-phonetic name form. If a variant pair passed a rule in either the original or semi-phonetic form, the pair was accepted. There was a final rule—applied to the semi-phonetic name form only—which required two identical initial characters, while the name ends in (any part of) the semi-phonetic suffixes TSJEN, TJEN, TYN, KJEN, KEN, KYN, YA, PJEN, PY or was empty. For instance: *Eva*/*Eefje* > EFA/EFJE > EF + A/EF + JE is accepted as variant pair.

From Table 4.4 it can be seen that variant pairs with a Levenshtein distance well over 2 can be accepted by a rule, which also holds for the additional suffix rule discussed above. A general threshold of 2 or 3 is common, the gain of the current method is in the conditional acceptance of a wider range of edit-distances.

It is impossible to fully automate the decision on the status of name variant pairs by rules. For instance, the genuine name pair *Willem/Guillaume* differs as a Dutch–French translation too much in spelling. Manual decisions, on the basis of expert knowledge, are unavoidable but should be kept to a minimum. An additional manual review was critical, and concentrated on true variants of low frequency and rejected variant pairs with a high frequency. If there was any doubt on the status of a variant pair, the name pair was not accepted. A manual decision could imply a rejection of a name pair that was accepted by rule, of acceptance of a name pair that did not pass the rules (for instance because the initials were not equal).

4.6 Results and Evaluation

A summary of the results of all phases in the cleaning process is presented in Table 4.5 for first names and in Table 4.6 for family names. For the accepted name variant pairs, the percentage with a certain Levenshtein distance is given in the tables as well, both for the original and the semi-phonetic form of the names. A Levenshtein distance equal or larger than 3 (usually too large to be accepted in straightforward record linkage as this generates abundant overlinking), is found—in original form—for 15.7 % of the female first names, 11.4 % for the male first names, and 7.0 % for the surnames (10.9, 8.3, 3.9 % for the semi-phonetic form, respectively). In terms of tokens the percentages are somewhat lower. This may be considered the gain of the method. As expected, the Levenshtein distance in the semi-phonetic form is lower for than in the orthographic form, but mainly for distances up to 2. Larger name pair differences originate in suffix variation or translation rather than in spelling differences for the same pronunciation.

Table 4.5 Overview of cleaning results for first name variant pairs

	Female first names		Male first names	
	Name pairs	Tokens	Name pairs	Tokens
Initial name pairs	48,684	246,519	31,886	183,050
Accepted by dictionary	3610	94,551	2877	90,761
Rejected as non-variant	2412	12,041	2289	6538
Rejected by rules	11,336	18,716	7077	10,079
Rejected manually	118	414	42	126
Accepted manually	1001	3917	563	2458
Total accepted	34,818	215,438	22,478	166,307
Total rejected	13,866	31,081	9408	16,743
Levenshtein distance (*original*)				
1	58 %	69 %	65 %	70 %
2	26 %	20 %	24 %	18 %
3	9 %	5 %	7 %	5 %
>3	7 %	6 %	4 %	7 %
Levenshtein distance (*semi-phonetic*)				
0	19 %	29 %	22 %	29 %
1	52 %	45 %	53 %	46 %
2	18 %	15 %	17 %	13 %
3	7 %	7 %	5 %	7 %
>3	4 %	4 %	3 %	5 %

The exclusion and acceptance mechanisms as described in Sect. 4.5.1 are detailed. For all accepted name variant pairs the percentage with a certain Levenshtein distance is given, both for original and semi-phonetic name forms

Table 4.6 Overview of the results of the various steps in cleaning the initial corpus of name pair variants for family names. Details as in Table 4.5

	Family names	
	Name pairs	Tokens
Initial name pairs	177,258	374,901
Accepted by dictionary		
Rejected as non-variant	103	199
Rejected by rules	56,694	79,079
Rejected manually	346	507
Accepted manually	783	2410
Total accepted	120,115	295,116
Total rejected	57,143	79,785
Levenshtein distance (*original*)		
1	69 %	77 %
2	24 %	19 %
3	5 %	3 %
>3	2 %	1 %
Levenshtein distance (*semi-phonetic*)		
0	29 %	44 %
1	53 %	45 %
2	14 %	8 %
3	4 %	2 %
>3	0.2 %	0.2 %

As mentioned in the previous section the heuristics used in the classification process resemble the well-known Jaro-Winkler similarity, as both methods compute similarity based on shared prefixes and number of edit operations relative to the length of the string. To compare both methods, the Jaro-Winkler similarity (which is expressed as a similarity value between 0 and 1) is computed for all candidate variant pairs of first names and surnames together that have been selected by the basic method outlined in Sect. 4.4. In Fig. 4.6 the amount of pairs is presented for different similarity values, using separate curves for pairs accepted or rejected by the joint post-processing methods. Both the similarity in the original names and the similarity in the semi-phonetic forms are shown in the graph.

Figure 4.6 shows that the two methods are indeed correlated: accepted pairs generally receive a higher Jaro-Winkler score than rejected pairs. The score at the intersection of the curves of accepted and rejected name pairs is around 0.85 and could be taken as a threshold. This value is consistent with those used in the literature (see e.g. de Vries et al. 2009). The area under the curve for rejected pairs >0.85 (20 % false acceptances) and <0.85 under the curve for accepted pairs (13 % false rejects) is the gain of the current post-processing methods over the application of the Jaro-Winkler similarity. The curves in Fig. 4.6 do not differ much for names in original and semi-phonetic form. This implies that the Jaro-Winkler similarity does not improve by application on the semi-phonetic name form.

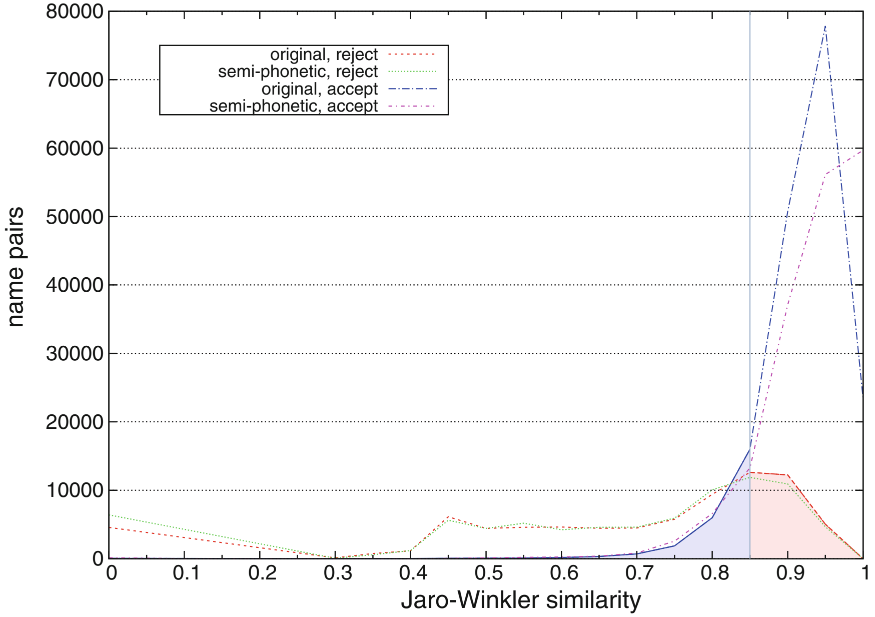


Fig. 4.6 Jaro-Winkler similarity for candidate variant pairs. Values are binned with interval 0.05 for readability. The vertical line at 0.85 is the optimal decision threshold for acceptance as similar names

In addition, it is of interest to consider the name pairs that are not accepted although they have a Levenshtein distance ≤2. These figures are not presented in Table 4.5, but amount to 39 % of the 13,866 erroneous female first name pairs, 49 % of the 9408 erroneous male first name pairs, and 43 % of the 57,143 erroneous surname pairs in original form, and 48, 42 and 48 % in their semi-phonetic form, respectively. If encountered in record linking, and considered by Levenshtein distance only, these names will be incorrectly accepted. This demonstrates the need for explicit knowledge of name variants.

A comparison of the name pair types and tokens shows that the rules mainly exclude rare name pairs as there are about only 1.5 times more tokens than types. On the other hand, first name pairs that were accepted because the names share the same lemma in a dictionary are frequent with on average about 30 times more tokens per pair. The latter variants are obviously well-known and made it to the dictionary.

Although we collected more surname variant pairs (120,115) than first name variant pairs (57,296 in all), the tokens of first name variant pairs were more frequent. On average, a first name variant pair was observed 6.7 times, while this was 2.4 times for surnames. This shows that there is much more variation in first names than in surnames.

The analysis of name pairs does not show how many different names are involved. This is shown in Table 4.7, together with the figures found in the full release (WieWasWie 2011). The number of singletons (i.e., name types with frequency 1) in both collections is presented as well, as they constitute 50 % of all first

Table 4.7 Number of different names (and singletons among them) in the accepted name pairs, and in the full WieWasWie corpus (release 2011)

	In accepted name pairs		WieWasWie 2011	
	All	Singletons	All	Singletons
Female first names	28,574	5766	128,800	63,132
Male first names	20,234	4048	101,830	51,000
Family names	129,929	22,361	565,647	225,389

names and 40 % of all surnames in the full corpus while they are present in 20 % of the accepted first name pairs, and in 15 % of the accepted surname pairs.

Given the constraints applied to arrive at accepted name pairs, the number of different names in the current analysis is relatively limited. Names might have been missed if they did not meet the required conditions, or if they were consistently written in the same way for any person and do not have a variant (the latter names will not present problems in record linkage). However, the selected names have a high coverage as will be shown in the next section. Coming to grips with the variation in these names can have a highly positive effect on record linkage.

4.7 Name Clusters and Standardisation

A corpus of true name variant pairs can be used to create clusters of names for which variant pairs are only found within a cluster. On this basis, yet unseen name variant pairs can be anticipated. We applied a non-standard clustering technique on the derived variant pairs which involved the following steps. Initially, for every name it was counted how many variants (types) were available. Based on the assumption that names with many variants likely constitute a cluster kernel, names were analysed in the order of the number of variants they had. The cluster procedure was applied on the semi-phonetic form of a name. For female first names, *Elisabeth* (including equal semi-phonetic forms) had most variants (148), for male first names this was *Hendrik* (69 variants), and for surnames *Tijssen* (50 variants).

As a first step, all variants of a name under investigation were added to the cluster. As a second step, for each of the names in the cluster it was analysed whether they had variants themselves that were not already in the cluster. If such a new variant name shared more than 60 % of the own variant *tokens* with names already in the cluster, the new name was added to the cluster. The value of 60 % was arbitrarily chosen on the basis of pilot analyses. A higher and more restrictive percentage would result in too many missed true variants in the cluster, while a lower and more permissive percentage would result in incorrect inclusion of variants. This process was continued until no new names could be added to the cluster. Subsequently, the next potential kernel name (with the then highest number of different variants) was analysed. The procedure continued until no names could be clustered anymore.

In the final phase of the procedure, variant pairs remained of which both names had no relation to other names and could be considered as mini-clusters. For the current analysis we left them out of consideration. For male first names 1221 clusters resulted comprising 16,487 names, for female first names there were 1530 clusters with 23,813 names, and for surnames 11,696 clusters with 93,793 names resulted. The largest clusters per name type were found for *Maria* (669 names), *Hendrik* (392) and the patronymic surname *Klaassen* (168). These figures are higher than the maximum number of variants found for an individual name since not all possible combinations of names in a cluster were present in the corpus. The cluster for the female first name *Elisabeth* is shown in Fig. 4.7 as an example.

In ideal clusters, the two names of a variant pair should both be members of the same cluster. This was the case for 91 % of the male first name pairs, 89 % of the female first name pairs and 85 % of the surname pairs. The remaining pairs could be erroneous and overlooked during the manual inspections, or the names were

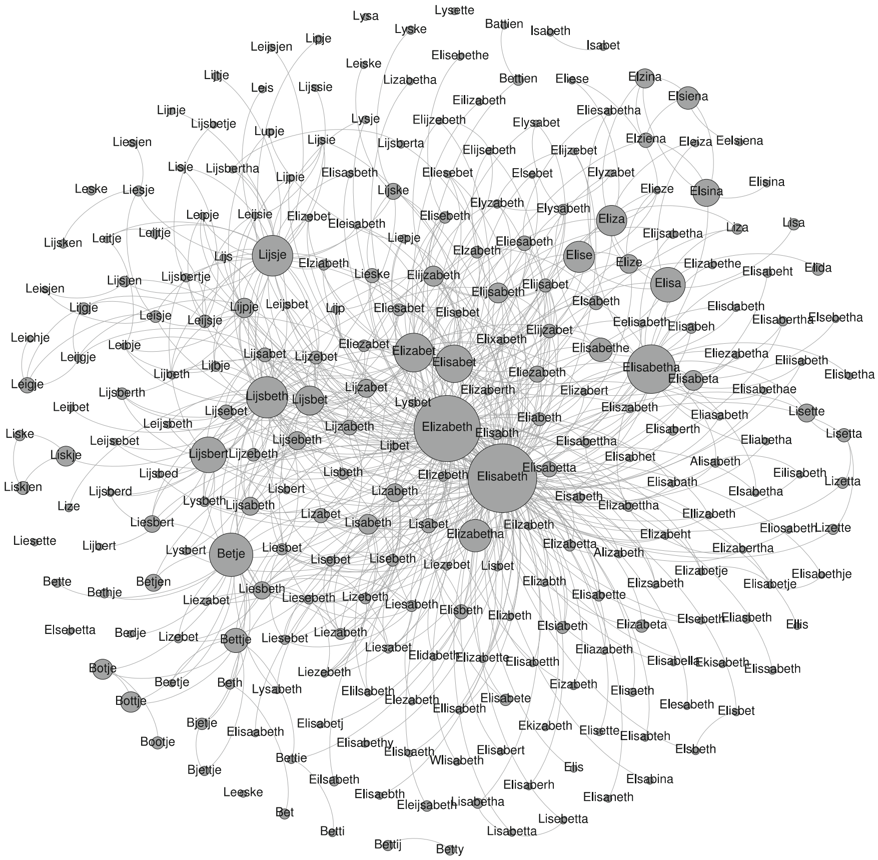


Fig. 4.7 Example cluster for *Elisabeth*. Edges denote variant pairs found in data. Node size is proportional to name frequency. Only names with a frequency ≥10 are shown

incorrectly associated in the cluster procedure. But there also can be genuine reasons that the names in a pair are not associated to the same cluster.

Names can reside in two closely related clusters although the current analysis did not provide sufficient evidence for such a merger, for instance the clusters *Egidius* and *Gilles*. Also, names can have more than one interpretation in terms of clusters, for instance *Louwis* as *Louis* or *Laurens*. The same holds for abbreviations, where the short form can be derived from several distinctive names. Productive are first names based on a suffix such as *Dina*, with possible derivations from *Alberdina, Berendina, Gerdina*, and so on. Whereas most names are associated to a single cluster, names in the latter categories are better associated to a group of clusters and analysed separately in record linkage procedures.

Once one has arrived at name clusters, it is a small step to name standardisation. The name in the cluster with the highest frequency was arbitrarily chosen as the standard name. This opens the possibility to perform a second round in discovering name variant pairs. Names in the original corpus were replaced by their standard if this could lead to four equal names of a person and parents, and a different fifth name. On this basis a total of 1,433,707 variant pairs (tokens) were collected, an increase of 60 % over the first round. A new analysis was performed on these pairs (while keeping all earlier manual decisions). This resulted in 26 % more accepted variant pairs (types) for both male and female first names, and 37 % more surname variant pairs.

After clustering of these variant pairs, 19,757 male first names were distributed in 1334 clusters, 28,509 female first names in 1694 clusters and 127.194 surnames in 15,114 clusters. The gain relative to the first round was 20 % for first names and 36 % for surnames, while the number of clusters increased less: 10 % for first names and 29 % for surnames. The ratio between the relative gain of newly created clusters and newly associated names is lower for first names than for surnames, indicating that for first names many new names were added as variant to existing clusters, while for surnames they created more new clusters.

If we use the clusters as a basis for the standardisation of names, about 20 % of all 101,830 male and 128,800 female first names get a standard, and 23 % of all 565,647 surnames. Whereas these percentages are rather low, the remaining names are rare and many are even singletons (hapaxes). In terms of tokens, the standardised first names describe 98.2 % of all 63 million element tokens and the standardised surnames 89.1 % of all 56.6 million element tokens. The number of standardised names can be extended by including non-standardised names with the same semi-phonetic form. This increases the number of standardised names by 63 %, but since these names are of low frequency the increase in the number of tokens is 0.2 % for first names and 1.5 % for surnames (to 98.4 and 87.4 % respectively).

4.8 Comparison to the FamilySearch Name Variant Corpus

The quality of the variant pairs derived above (referred to as the *test database*) can be estimated by a benchmark comparison with another, independent variant database. This section describes such a comparison with the name variant database of FamilySearch[1] which is the research department of the Church of Jesus Christ of Latter-day Saints (LDS, more commonly known as the Mormon Church). This database will be referred to as the *LDS Database*.

The LDS database is created as a by-product of genealogical research conducted by the LDS Church. Genealogies have been constructed from a large variety of sources, including census data, church records, court and inheritance records, land ownership records and migration records. The sources and resulting records have been reviewed by church clerks and linguists between the 1940s and mid-1980s in order to record name variation. This review has been mainly a manual process, based on general phonetic, syntactic and etymological guidelines with name variants resulting from genealogy research as a starting point. Source data originated mostly from North America, the British Isles (including Ireland) and continental Europe, but also some Central and South America and a small amount of sources from Asia. An estimated total of one billion name tokens has been used for the name variant database.

In order to be informative, the comparison setup needs to satisfy at least the following conditions:

1. The benchmark database and the test database have been constructed using the same definition of name variation.
2. The set of names contained in the test database is a subset of or equal to the set of names contained in the benchmark database.
3. The authority of the benchmark database is established.

If the first condition is not satisfied, a different classification due to a difference in definition is expected for an unknown number of name pairs. If the second condition is not satisfied then the accuracy of classification of name pairs containing names which are not present in the benchmark database cannot be established. If the third condition is not satisfied, differences in classification may be due to errors in the benchmark database instead of errors in the test database. In all three cases a comparison of name pair classifications is less informative. For the current comparison these three conditions will be discussed.

The main method described in this chapter has the aim to decide on variant pairs on the basis of true record matches without further requirements. However, the subsequent cleaning required a definition of a true variant pair:

[1] A web-based query interface is available on https://familysearch.org/stdfinder/NameStandardLookup.jsp

"A true *name variant pair* is a pair of names which can be traced to the same lemma". A similar notion of onomastic variation has been used in the LDS database, therefore, the first condition seems to be satisfied. However, the association of a name to a lemma may be uncertain or ambiguous and different interpretations may be used. If a name variant is based on a spelling error (reading error or typo), morphological or etymological information that could trace to the lemma may be lost. This may especially occur in short names, e.g., *Aatje*, which could be a variant of *Ada/Adriana* (morphological) but also could be a spelling error of *Aafje* or *Aaltje/Alida* which have different lemmas. Furthermore, the granularity at the lemma level can be different. In the LDS database the names *Gerrit* and *Geurt* (mentioned as variants in Table 4.4) are considered to belong to different name groups. Conversely the name group for *Sophie* in the LDS database contains *Fae, Feetje, Feye* which are remote or even unlikely variants that may be considered as belonging to different lemmas such as *Feie*. These issues indicate that the first condition of the benchmark procedure is not entirely satisfied.

In the LDS database many names from WieWasWie cannot be found (see Table 4.8) and vice versa, therefore the second condition is clearly violated.

Also, the authority of the benchmark is not fully clear. The procedures and guidelines used in the construction of the LDS database have not been documented in detail and manual decisions have influenced the database to a large extent. In the experience of the authors of the current chapter the overall quality of the LDS database is high, but a significant amount of classifications seems debatable or plain incorrect. The violation of all three conditions should be kept in mind in assessing the value of the benchmark validation. Obviously it would be preferable to use a benchmark database which does satisfy the conditions. Alternatives include *NameX* (commercial, namevariants.co.uk), *NamepediA* (community based, namepedia.org) or *JRC-Names* [named entities, see Steinberger et al. (2011)]. However, considering coverage, availability and technical accessibility, the LDS database is the most suitable for the current benchmark comparison.

Table 4.8 Results of a benchmark comparison with the name variant database of FamilySearch (LDS)

	Post-processing result	First names	%	Surnames	%
Total name pairs		80,570	100.0	177,258	100.0
Not in LDS		37,624	46.7	124,902	70.5
Present in LDS		42,946	53.3	52,356	29.5
Variant in LDS	Rejected	936	2.2	609	1.2
	Accepted	*18,675*	*43.5*	*12,414*	*23.7*
Non-variant in LDS	Rejected	*13,882*	*32.3*	*22,622*	*43.2*
	Accepted	9453	22.0	16,711	31.9

4.8.1 Comparison Results

The basic method described in Sect. 4.4 resulted in 257,828 name pairs, of which 80,417 pairs have been rejected by post-processing (see Tables 4.5 and 4.6). All pairs have been compared to the LDS database, of which the results are summarised in Table 4.8.

In Table 4.8 a name pair is considered not in LDS if either one or both names are missing from the LDS database. This category applies to around 47 % of the first name pairs and 71 % of the surname pairs. For these names the benchmark is unable to provide an indication of the accuracy of the basic algorithm or the post-processing procedure. This is partly caused by the presence of spelling errors in the test database and the coding of diacritic marks. However, also many valid names consisting of only basic characters are not present in the LDS database, mostly low-frequent names such as *Elijzebet* (frequency: 22), *Edcko* (frequency: 1 as part of a composite name) or *Ruighaven* (frequency: 12). More common names are missing from the LDS database as well, for example the surname *Paardekooper* (English: *horse merchant*, frequency: 1888) is not included while the variants *Paardekoper, Paardenkooper, Paerdekooper, Paerdekoper, Parrdekooper, Peerdekooper, Peerdekoper* are present. The omissions include several high-ranked names, such as *Pieters* as a surname (frequency: 38,005).

In Table 4.8 the relatively high values for variant-accepted and non-variant-rejected, as well as the low values for variants-rejected show a reasonable agreement between the benchmark and the test database (indicated by italic numbers in the table). However, the high values for non-variant-accepted show disagreement: according to the LDS database many of the names in these pairs belong to different name groups while the post-processing algorithms consider these names as valid variant pairs. This result could be interpreted as an indication that the current algorithms are too permissive. However, as noted above, the databases are subject to granularity differences which influence the classification. For the combination non-variant-accepted a high value indicates a more fine-grained clustering in the LDS database, which is consistent with limited manual browsing of the database.

In general it can be concluded that the amount of agreement is higher than the amount of disagreement, which means that both the LDS database and the approach of this chapter are capable of capturing a significant amount of person name variation.

4.9 Discussion

This research was based on a set of record matches with a very high confidence level. We focused on extraction of true name variants from these record matches to apply them later under less favourable conditions. Although the method automatically produced name variants, even with very large edit-distances, errors in the source data implied the extraction of erroneous variant pairs as well. The need for detection of these erroneous name pairs compromises the method, especially because manual inspection proved unavoidable. Nevertheless, the automatic selection, although not error free, assists enormously to identify true name variants from real data.

From Tables 4.5 and 4.6 it can be seen that the level of name errors that are present in this corpus concerns about 30 % of both the first name and surname variant pairs (and 9.2 % of the female, 13.0 % of the male first name pair tokens, and even 21.3 % of the surname pair tokens). These error levels may be worrying, but the reassuring observation is that they were detected by post-processing and evaluation procedures. In sources that are less rich in information on individuals, these errors cannot be traced that easily. In such cases it may only be hoped that more complex decision strategies (other than pair-wise comparison of records) can be developed, to perform accurate matching and error detection.

Part of the errors we identified are likely reading/transcription errors of the type *Pietje/Tietje*, in which *P* and *T* are confused, or typing errors like *Bos/Vos* with *B* and *V* as neighbouring keys. The additional problem with these errors is that they can result in existing names. Because we focused on name variants that have an onomastic basis, these pairs were labelled as errors. However, if we could estimate the likelihood of these errors, this could be incorporated in a linkage decision model (rather than requiring excess of information to be able to circumvent these reading errors).

Name variants can be summarised by clustering and name standardisation. Such a standardisation could realise a large efficiency gain in nominal record linkage, but can also be very helpful in search procedures. Whereas a considerable percentage of name variants indeed has names which belong to the same cluster, there are also variant pairs of which the names are associated to different clusters. This may indicate an erroneous name pair, but often it is an indication of ambiguity: names can be associated to more than one lemma. This is particularly true for first names which are derived from suffixes (for instance *Fien* from *Afien, Adolfien, Josefien, Rudolfien,* etc.) and for abbreviated names. Linkage and search procedures should then test all possible options for interpretation.

The method of using reliable record links to discover name variants is promising, but the process is complicated by the cleaning of errors in the data. This can only be partially performed by automatic procedures. Manual inspection and expert judgement, implemented in an active learning setting, is unavoidable. An explorative comparison of the results with the name variant corpus of FamilySearch revealed that many variants are not shared by both corpora, indicating the enormous scale of name variation (usually of low frequency). In name standardisation, choices for standards are not at all obvious, but influence the results of a comparison.

The final proof of the gain of the presented method is in application of the results in record linkage (through acceptance of names within a cluster of variants and the acceptance of specific name pair variants of which the names reside in different clusters). Such an evaluation requires a golden standard of linked records on which various linkage methods can be applied.

Acknowledgments This work is part of the research programme LINKS (LINKing System for historical family reconstruction, http://www.iisg.nl/hsn/projects/links.html), which is financed by the Netherlands Organisation for Scientific Research (NWO), grant 640.004.804.

References

Anderson, J. M. (2007). *The grammar of names*. Oxford: Oxford University Press.

Bloothooft, G. (1995). *Rules for semi-phonetic conversion of first names and family names*. Uil-OTS internal report (in Dutch).

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1, (Article 5).

Bratley, P., & Lusignan, S. (1976). Information processing in dictionary making: Some technical guidelines. *Computers and the Humanities*, *10*(3), 133–143.

Christen, P. (2012). *Data matching—Concepts and techniques for record linkage, entity resolution, and duplicate detection. Data-centric systems and applications*. Berlin: Springer.

Dolby, J. L. (1970). An algorithm for variable-length proper-name compression. *Journal of Library Automation, 3*(4), 257–275.

Driscoll, P. (2013). Computational methods for name normalization using hypocoristic personal name variants. In *Multi-source, multilingual information extraction and summarization* (pp. 73–91), Springer.

Malin, B. (2005). Unsupervised name disambiguation via social network similarity. In *Proceedings of the workshop on link analysis, counterterrorism, and security* (pp. 93–102).

Olsson, F. (2009). A literature survey of active machine learning in the context of natural language processing. *SICS Technical Report T, 2009*, 06.

Oosten, M. (2008). *Past names, family relation based on data from Genlias*, MSc thesis, LIACS, Leiden University (in Dutch).

Philips, L. (2000). The double metaphone search algorithm. *C/C++ Users Journal*, *18*(6), 38–43.

Russel, R. (1918). Index. US Patent 1261167.

Sarawagi, S., & Bhamidipaty, A. (2002). Interactive deduplication using active learning. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 269–278). ACM.

Schaar, J. van der (1964). *Woordenboek van voornamen*, Aula (since 1992 edited by D. Gerritzen).

Steinberger, R., Pouliquen, B., Kabadjov, M., Belyaeva, J., & van der Goot, E. (2011). JRC-NAMES: A freely available, highly multilingual named entity resource. In *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP)* (pp. 104–110).

Vries, T. de, Ke, H., Chawla, S., & Christen, P. (2009). Robust record linkage blocking using suffix arrays. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (ACM)* (pp. 305–314).

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)* (pp. 354–359).

Xu, J., & Croft, W. B. (1998). Corpus-based stemming using co-occurrence of word variants. *Transactions on Information Systems (TOIS), 16*(1), 61–81.

Part II
Record Linkage and Validation

Chapter 5
Advanced Record Linkage Methods and Privacy Aspects for Population Reconstruction—A Survey and Case Studies

Peter Christen, Dinusha Vatsalan and Zhichun Fu

Abstract Recent times have seen an increased interest into techniques that allow the linking of records across databases. The main challenges of record linkage are (1) scalability to the increasingly large databases common today; (2) accurate and efficient classification of compared records into matches and non-matches in the presence of variations and errors in the data; and (3) privacy issues that occur when the linking of records is based on sensitive personal information about individuals. The first challenge has been addressed by the development of scalable indexing techniques, the second through advanced classification techniques that either employ machine learning- or graph-based methods, and the third challenge is investigated by research into privacy-preserving record linkage (PPRL). In this chapter, we describe these major challenges of record linkage in the context of population reconstruction. We survey recent developments of advanced record linkage methods, discuss two real-world case studies, and provide directions for future research.

5.1 Introduction

In the past decade, record linkage has attracted much interest by researchers and practitioners from various domains, including national census, health and social science research, businesses, and crime and fraud detection (Christen 2012a; Herzog and Scheuren 2007; Naumann 2010; Talburt et al. 2011). Also known as

P. Christen (✉) · D. Vatsalan · Z. Fu
Research School of Computer Science, The Australian National University, Canberra
ACT 0200, Australia
e-mail: peter.christen@anu.edu.au

D. Vatsalan
e-mail: dinusha.vatsalan@anu.edu.au

Z. Fu
e-mail: sally.fu@anu.edu.au

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_5

data linkage, entity resolution, data matching, or duplicate detection, these techniques aim to identify and link all records that refer to the same real-world entities within a single or across several databases. In most applications, the entities under consideration are people, such as customers or patients.

The two areas where record linkage has traditionally been employed are national censuses (Winkler 2006) and the health domain (Kelman et al. 2002; Newcombe 1988). Most record linkage systems in these areas are based on the probabilistic record linkage approach developed by Newcombe and Kennedy (1962) and formalised by Fellegi and Sunter in 1969.

More recently, computer scientists have developed various techniques that allow the linking or deduplication of large databases with the aim to, for example clean customer records (Hernandez and Stolfo 1995) or identify fraudsters and criminals in financial and national security databases (Jonas and Harper 2006). Record linkage and deduplication techniques are also being employed to remove duplicate entries returned by search engines (Su et al. 2009), or to identify all bibliographic records of by the same author in publication databases (Lee et al. 2007).

Social scientists working in the area of demographics and genealogy have also employed record linkage techniques, commonly using historical census, or birth, death, and marriage (BDM) data (Fure 2000; Newton 2013; Quass and Starkey 2003; Reid et al. 2002; Ruggles 2002). The aim of such linkages is to identify and link not just individuals across two or more databases, but rather to create complete family trees over significant periods of time (Antonie et al. 2014a; Bloothooft 1995; Fu et al. 2014a). Such reconstructed (or reconstituted) family trees allow social scientists to investigate many aspects of past societies, such as changes in employment, mobility, fertility and morbidity, and even the genetic factors of certain diseases (Glasson et al. 2008).

Compared to contemporary data, the major challenges specific to the linking of historical data, which are based on census returns or BDM registers, are:

- The generally low levels of literacy of both census collectors and householders meant census items were often not recorded correctly. Dates of birth, and even ages, were commonly not known, and addresses were not clearly defined. There were no standard classifications of employment categories.
- Over time people moved, died, and were born, and so the structure of households and families changed significantly. Even if census returns are available for a full country, immigration and emigration mean a significant number of individuals simply ‘appear’ or ‘disappear’ without birth or death records. The influence of people’s movements is significantly worsened if only a small subset of census returns, like from a certain district or area, is available for research.
- Both given- and surnames often had strong local distributions. It was not uncommon for a large portion of a population to have one of a few common names.
- Only a small number of attributes were collected in many national censuses in the nineteenth century. For each individual they usually included the name, age, gender, relationship to the head of household, and occupation. Other data

sources, such as vital and parish registers (containing birth, baptism, death, and marriage records), can also provide rich sources of detailed information about families and their structures (Newton 2013; Reid et al. 2002).

- Historical documents are commonly hand-written and therefore have to be scanned and transcribed, either manually or automatically using optical character recognition techniques. These processes are likely to introduce further errors and variations into the data (Block and Star 1995).

Contemporary administrative and census databases are increasingly used for social science research. While present-day data are generally of higher quality and contain more detailed information, they pose their own set of challenges:

- As more information is being collected, today's databases not only become larger but they also contain more details about individuals, and they might also contain more complex types of data (such as text or multimedia documents). Linking very large databases poses significant computational challenges, as will be discussed in Sect. 5.2.
- The data collected are about people who are still alive, and therefore can contain sensitive information, for example about a person's health or their financial details. In today's 'Big Data' society, such information is highly valuable for organisations such as advertisers, insurers, financial institutions, and even governments, because it can facilitate for example specific individual targeting of advertisements, or the calculation of highly predictive credit risk scores (Siegel 2013). Privacy and confidentiality are especially of concern when records are linked across databases held by different organisations, as we will discuss in Sect. 5.3.

This chapter extends an earlier shorter workshop paper on the same topic (Christen 2014). In the following section, we provide a brief overview of advanced methods and techniques that have been developed in recent years. In Sect. 5.3 we discuss privacy issues relevant to record linkage and we summarise techniques that have been developed to facilitate linking databases across organisations without the need to reveal private or confidential information. In Sect. 5.4 we then illustrate, using two case studies, the issues discussed in Sects. 5.2 and 5.3. In Sect. 5.5 we present our view of important research directions for record linkage in the context of population reconstruction. We conclude this chapter in Sect. 5.6 with a summary of our findings. We also provide an extensive bibliography to relevant work.

5.2 Advanced Record Linkage Methods

A variety of techniques have been developed that allow the linking of large databases. The main areas of research have been to improve scalability to linking large databases, and to improve linkage quality using advanced classification techniques.

5.2.1 Scalable Indexing Techniques

When two databases are linked, each record from one database potentially has to be compared with all records from the other database. The vast majority of these comparisons will be between records that are not matches (i.e. refer to different entities). Indexing is the process of reducing this possibly very large number of record pairs that need to be compared in detail between databases by splitting each database into smaller sets of blocks or clusters, or by sorting the databases. The aim is to identify *candidate record pairs* from records in the same blocks or clusters that likely correspond to true matches, and that need to be compared in detail, generally using approximate string comparison functions (Christen 2012a).

The traditional blocking approach employs a *blocking criteria* (a single or set of attributes) to insert each record into one block (Fellegi and Sunter 1969). For example, if a ‘postcode’ attribute is used as blocking criteria then all records with postcode ‘2000’ are inserted into the same block. Only records within the same block are then compared with each other. The sorted neighbourhood approach (Hernandez and Stolfo 1995) sorts a database according to *sorting criteria* (usually a set of concatenated attributes) and then moves a sliding window over the sorted database. Only records that are within a certain window are compared with each other.

Many of the recently developed indexing techniques insert each record into more than one block, thereby aiming to overcome errors in attribute values (Christen 2012b). Overlapping clusters (called canopies), sorted suffix arrays, and q-gram-based indexing, are examples of such techniques. A different approach is to map records into a multi-dimensional space such that the distances between records are preserved (Jin et al. 2003). A multi-dimensional index data structure together with nearest-neighbour queries are then used to extract blocks of candidate records.

Adaptive techniques that, based on the characteristics of the data, dynamically modify the size of the window in the sorted neighbourhood method (Draisbach et al. 2012; Yan et al. 2007) or in suffix array-based indexing (de Vries et al. 2011) have recently shown to obtain blocks of higher quality. Other recent work has investigated indexing techniques for real-time record linkage, where a stream of query records is to be linked in sub-second time to a database of entity records (Christen et al. 2009; Ioannou et al. 2010; Ramadan et al. 2014). Related to real-time record linkage are approaches that allow for dynamic databases, where records are added, modified, or removed, on an ongoing basis (Dey et al. 2010; Ioannou et al. 2010).

While traditional indexing approaches require manual decisions about the choice of blocking criteria, several approaches have been proposed to learn optimal blocking criteria either using training data (pairs or groups of records known to be true matches or non-matches) (Bilenko et al. 2006; Michelson and Knoblock 2006), or more recently by exploring the distribution of attribute values in records and the similarities between them (Kejriwal and Miranker 2013). The aim of such learning techniques is to find blocking criteria that lead to small blocks which contain mostly matches only and overall have a high coverage of all matches (if they are known from training data). Generally, as will be discussed next, techniques that make use of true matches and non-matches obtain blocking results of higher quality compared to techniques that are only based on data distributions.

Only limited experimental evaluations have been conducted to compare the performance of indexing techniques. Christen (2012b) identified that none of 12 variations of six techniques outperformed all others when employed on several data sets, and that one of the most important factors for efficient and accurate indexing is the definition of an appropriate blocking criteria.

None of the indexing techniques discussed here is specific to a certain type of data, and therefore any can be used in the context of linking data for population reconstruction. However, given the often low quality especially of historical data, techniques should be applied that are able to cope with 'dirty' data and bring matching records together that likely contain errors and variations. To this end, techniques that insert each record into several blocks can be of advantage (at the cost of having to compare a larger number of candidate pairs), as can be techniques that incorporate domain expertise to guide the indexing process [for example by learning good blocking criteria (Bilenko et al. 2006; Kejriwal and Miranker 2013; Michelson and Knoblock 2006)].

5.2.2 Accurate Classification Techniques

The objective of record linkage classification is to decide if a pair or group of records is a *match* (assumed to refer to the same real-world entity) or a *non-match* (refer to different entities). In the traditional probabilistic record linkage approach (Fellegi and Sunter 1969), each compared record pair is classified independently into one of three classes (*matches*, *non-matches* and *potential matches*). The third class is those pairs or groups of records that require manual classification through a clerical review process (Christen 2012a).

Besides requiring an often time consuming manual clerical review step, this traditional approach has several other drawbacks. First, it assumes independence between attributes. Statisticians have investigated approaches that allow dependencies between some attributes to be modelled (Winkler 2006), and have achieved improved classification outcomes in some situations. Second, the estimation of the parameters needed for the probabilistic record linkage approach is a non-trivial undertaking and requires knowledge about the error rates in the databases to be linked (which is often difficult to obtain) (Herzog et al. 2007). Third, individual pair-wise classification can lead to a violation of the transitive closure property (if record pairs (a, b) and (a, c) are classified as matches, then pair (b, c) must also be a match).

Machine learning based approaches aim to overcome these deficiencies. They are either following a supervised learning approach, where training data in the form of known matching and non-matching record pairs are required (Elmagarmid et al.

2007), or they are based on unsupervised clustering techniques which group records according to their similarities (Naumann and Herschel 2010). While supervised approaches generally achieve higher linkage quality, their main drawback is the challenge of obtaining a large number of suitable training examples. Active learning techniques aim to overcome this drawback (Arasu et al. 2010; Bellare et al. 2012). They select a small number of difficult to classify record pairs and present these to a domain expert for manual classification, followed by a re-training of the classification model. This process is repeated until high enough linkage quality is obtained.

Several collective classification techniques for record linkage have recently been developed. Compared to the traditional classification of individual record pairs, based on a graph representation of the databases to be linked these techniques aim to find an overall optimal solution when assigning records to entities. Both Bhattacharya and Getoor (2007) and Kalashnikov and Mehrotra (2006) build a graph with records as nodes and relational and attribute similarities between them as edges. On the other hand, Dong et al. (2005) build a dependency graph where each attribute value pair is represented as a node that contains the similarity between the two values. An overall optimal classification is calculated in an unsupervised way by iteratively merging or splitting parts in such a graph into smaller sub-graphs, such that at the end of the process each sub-graph corresponds to an entity. A related technique is group linkage (On et al. 2007), where groups rather than individual records are considered and linked based on some form of group similarity.

Most experimental evaluations of these collective and group linkage techniques have been conducted using bibliographic databases, where different types of entities (authors, papers, venues, and affiliations) provide a rich and well-defined setting of relational information between entities. Compared to historical data, the quality of bibliographic data is generally high, but ambiguities occur, for example when non-standardised abbreviations of conferences or journals are recorded, only the initials of authors are given, or several authors have the same name and even work in the same research area. For two ambiguous author records, co-author similarities or having published in similar journals or conferences can provide the evidence needed to decide if the two records refer to the same author or not. The databases used to evaluate collective classification techniques generally contained less than one million records, and scalability of these techniques to very large databases has only been investigated recently (Rastogi et al. 2011).

Only limited work has been conducted in machine learning-based record linkage for population reconstruction. Antonie et al. (2014a, b) use a support vector machine classifier to link historical Canadian census data, while Efremova et al. (2015) use a linear scoring model to weight different similarity measures in the context of matching historical Dutch BDM records. These works highlight the successful application of supervised classification techniques for population reconstruction, but they also discuss the challenges in acquiring the required training data.

Fu et al. (2014a, 2011b, 2012) have recently investigated group linkage methods on historical census data by treating households as groups and combining pair-wise record linkage with household linkage. Their evaluation on UK census data showed a significant reduction in the number of multiple links (i.e. where a single record from one database is linked to several records in another database).

The unique structure between records within a family or household has only recently been explored for record linkage. While most personal details of people change over time, some aspects of the relationships between the members of a family or household keep constant even over long periods of time. For example, the age differences between two parents, and between parents and their children, do not change (assuming they are recorded accurately). As we will illustrate in Sect. 5.4.1, Fu et al. (2014b) recently proposed to build one graph per household using such time-invariant information as edge attributes, and they showed that such an approach can help to improve household matching in historical census data. Graph-based approaches can exploit such rich sources of structural information and allow the development of improved record linkage techniques in the context of population reconstruction.

5.3 Privacy Aspects in Record Linkage

Due to the lack of unique entity identifiers, record linkage is generally based on comparing partially identifying personal details of individuals, such as their names, addresses, dates of birth, and so on. When historical data are being linked then usually no privacy concerns are being raised, because these data do not contain any information about living individuals. However, as social science research increasingly requires the linking of contemporary databases obtained from diverse sources, privacy and confidentiality issues become crucially important. National census agencies are currently considering the use of anonymisation techniques to facilitate matching their databases with records sourced from public as well as private administrative data (Office for National Statistics 2013).

While a single database that contains the personal details of individuals can already contain sensitive information, linking records sourced for instance from government agencies with records from commercial databases can reveal information that is highly sensitive. For example, an individual's social security (unemployment) record linked with their financial details obtained from a bank database would be of high value for a credit rating agency. As recent events in the context of national security data leakages have shown (Edward Snowden's copying and releasing of thousands of top secret US government documents) (Toxen 2014), people are wary that their information is being collected by and shared across different organisations, especially if this is done by governments.

The linking of contemporary databases from diverse sources can allow studies at levels of detail and at scales otherwise not possible, and therefore safeguards must be in place to make sure no private or confidential information can be revealed. In the health domain, specific protocols (Churches 2003; Kelman et al. 2002) have been developed and are in use that split sensitive health data from the attributes used for the actual linkage. These protocols, however, still require a trusted third party to conduct the linkage using the actual personal details of individuals. Ideally, it should be possible to conduct record linkage without the need of any sensitive information to be exchanged between the parties that are involved in a record linkage project.

Researchers working in the area of 'privacy-preserving record linkage' (PPRL) are aiming to achieve this goal (Verykios and Christen 2013). Vatsalan et al. (2013) provide an extensive review and propose a taxonomy of current PPRL techniques, and they discuss research challenges and directions. The basic ideas of PPRL techniques are to (somehow) encode (or mask) the databases at their sources and to conduct the linkage using only these encoded data (i.e. no sensitive data are ever exchanged between parties). At the end of such a PPRL process, the database owners only learn which of their own records have a high similarity with certain records from the other database(s). The database owners can then negotiate the next steps, such as exchanging the values in certain attributes of the linked records, or sending selected attribute values to a third party (for example a researcher, as discussed in Sect. 5.4.2).

The two basic scenarios in PPRL are two- and three-party protocols. In the latter type, a linkage unit is conducting the actual linkage based on encoded data received from the two database owners. On the other hand, in two-party protocols the two database owners directly exchange encoded data between them. The advantage of two-party over three-party protocols is that they are more secure, as there is no possibility of collusion between one of the database owners and the linkage unit. However, two-party protocols are generally more complex in order to make sure that the two database owners cannot infer any sensitive information from each other during the PPRL process.

Research into PPRL started in the mid 1990s, and the developed techniques can be categorised into three generations (Vatsalan et al. 2013). The first only considered the exact matching of attribute values without revealing these values. These techniques basically convert attribute values into hash codes (bit-patterns of a certain length) using one-way hash algorithms such as SHA or MD5 (Schneier 1996), and then compare the generated hash codes in an exact fashion. These hash codes are secure in that having only access to a hash code makes it nearly impossible (with current computing techniques) to find the corresponding plain-text string in a reasonable amount of time. The major drawback of the first generation of PPRL techniques is that even a single character difference between attribute values results in completely different hash codes, and so only exact matching of values is possible. As data, especially personal details such as names and addresses, often contain variations and errors, exact matching does not work well in most practical linkage situations.

The second generation of PPRL techniques aimed to overcome this drawback by allowing for approximate matching. Approaches for secure edit-distance, Jaccard and overlap similarity, and Cosine distance have been developed, with several recent surveys providing comparative evaluations of such techniques (Durham et al. 2012; Karakasidis and Verykios 2010; Trepetin 2008; Vatsalan et al. 2013;

Verykios et al. 2009). A variety of techniques have been investigated, including Bloom filters (bit-arrays) (Schnell et al. 2009; Vatsalan and Christen 2012), phonetic encoding (such as Soundex or NYSIIS) (Karakasidis and Verykios 2009), random and public reference values (Karakasidis et al. 2011; Pang et al. 2009; Vatsalan et al. 2011), embedding spaces (into multi-dimensional spaces) (Scannapieco et al. 2007; Yakout and Atallah 2009), and secure multi-party computation (Atallah et al. 2003; Inan et al. 2008; Li et al. 2011; Ravikumar et al. 2004). The interested reader is referred to the above cited survey articles for details.

While allowing for approximate matching was a significant improvement for PPRL, the problem of scalability to linking large databases has only recently been considered in the third generation of PPRL techniques (Al-Lawati et al. 2005; Bonomi et al. 2012; Durham 2012; Inan et al. 2010; Karakasidis 2012; Karapiperis and Verykios 2014; Kuzu et al. 2013; Sehili et al. 2015; Vatsalan et al. 2013a). Different techniques have again been developed which combine traditional indexing techniques (Christen 2012b) with encoding, perturbation, or cryptographic approaches (Vatsalan et al. 2013). Thus far, only a few small comparative studies of such techniques have been published (Durham 2012; Vatsalan et al. 2013a, 2014). The issues involved in evaluating PPRL techniques have also received increased attention in recent times (Vatsalan et al. 2014).

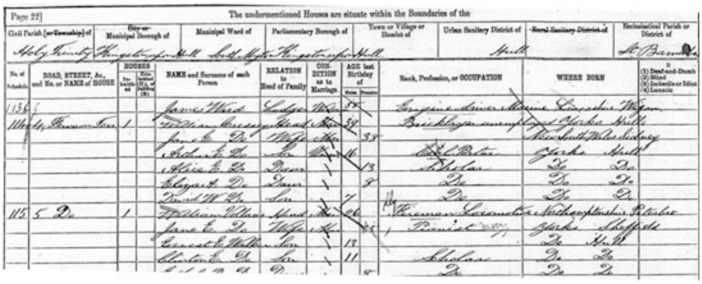
5.4 Case Studies

In this section we present two case studies with a focus on advanced record linkage techniques being employed to population reconstruction. The first study discusses the use of group and graph linking in the context of linking historical census data, while the second discusses approaches to preserving privacy when linking contemporary data from several sources from both private and public organisations.

5.4.1 Advanced Linking of Historical UK Census Data

Our case study uses historical census returns collected from the district of Rawtenstall, which in the nineteenth century was a small cotton textile manufacturing town in North-East Lancashire in the United Kingdom (UK). Currently released historical census data in this area were collected since 1851 in ten-year intervals. The original data were hand-filled census forms, which contain 12 attributes, that for each individual residing in a household include the address, full name, age, gender, occupation, place of birth, and their relationship to the head of the household.

These hand-filled census forms were transcribed manually onto enumerator’s returns sheets, and these sheets were subsequently scanned into digital form. Since the late 1990s, various organisations began transcribing these data from images into



Page 22] The undermentioned Houses are situate within the Boundaries of the

Civil Parish [~~or Township~~] of	~~City or~~ Municipal Borough of	Municipal Ward of	Parliamentary Borough of	Town or Village or Hamlet of	Urban Sanitary District of	~~Rural Sanitary District of~~	Ecclesiastical Parish or District of
Holy Trinity	Kingston upon Hull	South Myton	Kingston upon Hull		Hull		St Barnabas

No. of Schedule	ROAD, STREET, &c., and No. or NAME of HOUSE	HOUSES Inhabited	Uninhabited (U.), or Building (B.)	NAME and Surname of each Person	RELATION to Head of Family	CONDITION as to Marriage	AGE last Birthday of Males	Females	Rank, Profession, or OCCUPATION	WHERE BORN	(1) Deaf-and-Dumb (2) Blind (3) Imbecile or Idiot (4) Lunatic
113				James Ward	Lodger	Widr	58		Engine driver Marine	Lancashire Wigan	
114	4 Pearson Terr	1		William Bessey	Head	Mar	39		Bricklayer unemployed	Yorks Hull	
				Jane E. Do	Wife	Mar		38		New South Wales Sidney	
				Arthur E. Do	Son	Unm	16		Tel Porter	Yorks Hull	
				Alice E. Do	Daur			13	Scholar	Do Do	
				Elizabeth Do	Daur			9	Do	Do Do	
				David W. Do	Son		7		Do	Do Do	
115	5 Do	1		William Walton	Head	Mar	26		Fireman Locomotive	Northamptonshire Peterboro	
				Jane E. Do	Wife	Mar		25	Pianist	Yorks Sheffield	
				Ernest E. Walton	Son		13			Do Hull	
				Clinton E. Do	Son		11		Scholar	Do Do	
				[illegible]	Daur			8	Do	Do Do	

Fig. 5.1 Historical census sample

tabular form and stored them in spreadsheets where they could be examined by members of the public. A sample of a scanned image is shown in Fig. 5.1. Our collection consists of six data sets, with around 160,000 records in total, corresponding to the censuses from 1851 to 1901.

To link such historical census data, several key steps are necessary to calculate the similarities between records from the individual data sets (Christen 2012b). These steps include data cleaning and standardisation to improve data quality and make attribute values more consistent before comparison; blocking or indexing, as discussed in Sect. 5.2.1, to subdivide a data set into blocks so that records in a block are only being compared with other records in the same block in the comparison step; and finally the classification of the compared record pairs into matches and non-matches, as was discussed in Sect. 5.2.2.

The differences between traditional record linking methods and those based on group or graph methods are in the final classification step. Traditional approaches only perform linkage at the record-pair level, relying only on the output of record attribute similarities to classify record pairs. In practice, this strategy often faces difficulties because most historical data have significant data quality problems, and only limited details about people are available in historical (census) data that can be compared attribute-wise between records. Group- or graph-based methods, on the contrary, consider households (or families) as integral entities, and use the whole of household information to improve the effectiveness and accuracy of record linkage.

To illustrate how household information can help group- and graph-based record linkage, let us consider the following example. Table 5.1 shows a household with four people, consisting of the parents and two children, extracted from the 1871

Table 5.1 1871 household sample. The example record for *Sarah Ashworth* is highlighted in italics

ID	Address	Surname	Firstname	Relation_to_head	Sex	Age
25531	union street	ashworth	john	head	m	30
25532	union street	ashworth	alice	wife	f	28
25533	union street	ashworth	richard	son	m	4
25534	*union street*	*ashworth*	*sarah*	*daughter*	*f*	*2*

census return. The key attributes and their values for each member are displayed, with 'ID' being a unique record identifier.

When applying traditional pair-wise record linkage between 1871 and 1881 census returns, we can see that Sarah Ashworth (ID 25534) has two matched records in 1881. One (ID 12534 in Table 5.2) lived in the same address but with a wrong age, and the other (ID 20858 in Table 5.3) lived at a different address with the correct age. Based only on the attributes in these records, it is difficult to determine which is the correct Sarah. Most pair-wise record linkage approaches will take the match with ID 12534 to be the correct one because it has a higher similarity to the 1871 Sarah than the second option, because street addresses normally contain more distinguishing information than age. As example, the pair-wise linking method by Fu et al. (2014a) uses approximate string comparison functions (Christen 2012a) on the address and name attributes and absolute differences on the age attribute, and gives a total similarity score of 0.9 for the record pair with ID 25534 and ID 12534 of Sarah Ashworth, higher than the total similarity score between ID 25534 and ID 20858, which is 0.84. This shows that pair-wise record linkage is not always reliable.

If we take other household members into consideration, it is obvious that record with ID 20858 in Table 5.3 should be the true match for record with ID 25534 in Table 5.1. The reason is clear: the names and ages of Sarah's parents and her brother Richard in this 1881 household (with Richard abbreviated as 'rd' in 1881) also match the corresponding members in the 1871 household, while the other members of the household in Table 5.2 do not. Therefore, household information can greatly help the decision making so as to reduce the ambiguity that arises from the pair-wise linkage results.

Table 5.2 Wrongly matched 1881 household

ID	Address	Surname	Firstname	Relation_to_head	Sex	Age
12532	union street	ashworth	henry	head	m	48
12533	union street	ashworth	eruble	wife	f	47
12534	*union street*	*ashworth*	*sarah*	*daughter*	*f*	*18*
12535	union street	ashworth	john	son	m	12

Table 5.3 Correctly matched 1881 household

ID	Address	Surname	Firstname	Relation_to_head	Sex	Age
20855	whittle st	ashworth	john	head	m	40
20856	whittle st	ashworth	alice	wife	f	36
20857	whittle st	ashworth	rd.	son	m	14
20858	*whittle st*	*ashworth*	*sarah*	*daughter*	*f*	*12*
20859	whittle st	ashworth	john	son	m	8
20860	whittle st	ashworth	harold	son	m	3

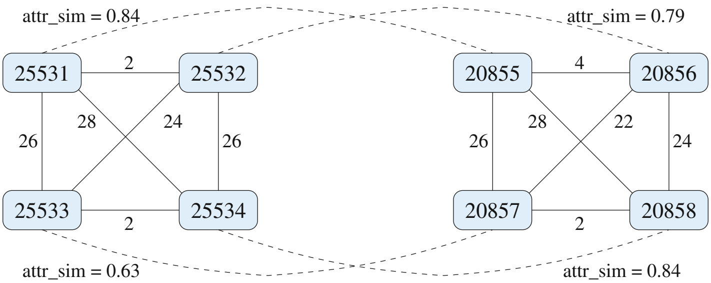
The key to utilising household information is how to model household members and their relationship. Group and graph linking are two methods aiming to solve this problem. Group linking (Fu et al. 2011b, 2012, 2014a; On et al. 2007) generates group similarity scores for each pair of households. Such household pair similarities are calculated in several steps. First, the number of household members in each household is counted. Then, the sum of the pair-wise record similarity scores between the household pairs is calculated. This sum is normalised by the number of distinct members in the two households being compared so as to generate the household similarity score. When two or more households are compared with a target household, the one with the highest household similarity is being matched.

For example, we know that the households in Tables 5.1 and 5.2 both have four members. Pair-wise linking results, again following the approach taken by Fu et al. (2014a), show that the similarity between records with ID 25531 and ID 12535 is 0.9, between ID 25532 and ID 12533 is 0.66, and between ID 25534 and 12534 is 0.9. Then, the group linking similarity score, using the bipartite similarity (Fu et al. 2011b), between this household pair is calculated as $(0.66 + 0.9 + 0.9)/(4 + 4 - 3) = 0.49$, i.e., the sum of record pair-wise similarities divided by the number of distinct members in these two households. The same calculation gives a group linking similarity score between the two households in Tables 5.1 and 5.3 of 0.78. Based on this it becomes clear that the households in Tables 5.1 and 5.3 are matched, and that Sarah Ashworth with ID 25534 and ID 20858 are matches.

Graph-based linking can be considered as an extension of the group linking step. Graph linking does not only consider the similarity of all record pairs in two households, it also takes structural information of households into consideration. While personal information, such as marital status, address and occupation, may change over time, surnames of women may change after marriage, and even ages may change due to different times of the year for census collection or input errors, some aspects of the relationships between household members generally remain unchanged. Such relationship aspects include, but are not limited to, age and generation difference, and role-pairs of two individuals in a household (Fu et al. 2014b). By incorporating such relationship aspects between household members into the linking model the linking accuracy can be improved.

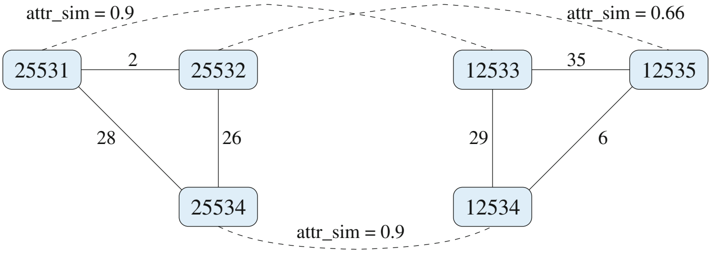
The graph method in (Fu et al. 2014b) treats members in a household as the vertices (nodes) of a graph, and uses edges to show the relational aspects between these vertices. The method first calculates record-pair similarities, which are used to find matched candidate record pairs with high similarities. These pairs are then used to connect the graphs of two candidate households. This transforms the household linking problem into a graph matching problem. The graph similarity score for each pair of households is calculated as the weighted sum of the vertex and edge similarities.

As an example of graph-based linkage using the three households from Tables 5.1, 5.2 and 5.3, Fig. 5.2 shows the graph generated between the 1871 and the correctly matched household, while Fig. 5.3 shows the graph generated between the 1871 and the wrongly matched household. Only the AGE attribute is used in



$$\mathit{avr\_node\_sim} = (0.84 + 0.79 + 0.63 + 0.84)/4 = 0.78$$
$$\mathit{avr\_edge\_sim} = (0.50 + 1.00 + 1.00 + 0.92 + 0.92 + 1.00)/6 = 0.89$$
$$\mathit{household\_sim} = \lambda \times \mathit{avr\_node\_sim} + (1 - \lambda) \times \mathit{avr\_edge\_sim} = \mathbf{0.84}$$

Fig. 5.2 Graph structure and similarity calculations of the two matched households from Tables 5.1 and 5.3. Edge values are absolute differences in AGE values, while the dotted lines show attribute similarities between records in the two households. We set the weighting parameter (Fu et al. 2014b) $\lambda = 0.5$



$$\mathit{avr\_node\_sim} = (0.90 + 0.66 + 0.90)/3 = 0.82$$
$$\mathit{avr\_edge\_sim} = (0.06 + 0.97 + 0.23)/3 = 0.42$$
$$\mathit{household\_sim} = \lambda \times \mathit{avr\_node\_sim} + (1 - \lambda) \times \mathit{avr\_edge\_sim} = \mathbf{0.62}$$

Fig. 5.3 Graph structure and similarity calculations of the two non-matched households from Tables 5.1 and 5.2. We again set the weighting parameter $\lambda = 0.5$

this example (in practice, the relationships between individuals would also be used). Only records that have high attribute similarities between the households are included in these graphs. The shown edge attribute values are age differences between records in the corresponding vertices, while the dotted lines between vertices in these household graphs correspond to the record attribute similarities calculated in the pair-wise linkage step. The edge (AGE) similarities are calculated as $age\_sim = 1.0 - abs(age\_diff)/max\_age$ (Christen 2012a).

When only the node similarities are considered, the similarity for the matched household pair is 0.78 for the households from Tables 5.1 and 5.3, which is lower than the similarity for the non-matched household pair (0.82), as shown in the calculations in Figs. 5.2 and 5.3, respectively. On the other hand, when the age relationships between household members are considered, a higher overall similarity is calculated for the matched household pair (0.84) compared to the non-matched pair (0.62), resulting in correctly matched households.

The results on the six Rawtenstall data sets show that the proposed methods significantly reduce the number of multiple record and household matches, with a more than 85 % reduction using either group or graph linking approaches (Fu et al. 2011b, 2014b).

5.4.2 Privacy-Preserving Record Linkage Across Several Organisations

Linking records across several databases held by different organisations, using the common identifiers that contain personal information, often involves privacy and confidentiality concerns of the individuals represented by the records in these databases (Vatsalan et al. 2013b). Generally, organisations are not allowed or willing to exchange such personal and sensitive information due to privacy and confidentiality concerns as well as government or business regulations.

As an example scenario, assume a demographer who aims to investigate how mortgage stress (having to pay large sums of money on a regular basis to repay a house) is affecting people from different ethnic backgrounds, and with different education and employment levels, with regard to their mental and physical health. This research will require data from financial institutions, as well as different government agencies (social security, health, and eduction), and potentially other private sector providers (such as health insurers). Neither of these parties is likely willing or allowed by law to provide their databases to the researcher. The researcher only requires access to some attributes of the records that are linked across all these databases, but not the actual identities of the individuals that were linked. However, personal details are needed to conduct the actual linkage due to the absence of unique identifiers across all the databases. As was discussed in Sect. 5.3, PPRL aims to address this problem.

Assume three databases from three different organisations, as shown in Tables 5.4, 5.5 and 5.6, need to be linked in order to identify the matching entities across these databases. A set of common personal identifiers, which are first_name / given_name, last_name /surname, and postcode, are used as quasi-identifiers (QIDs) for conducting the linkage. Exchanging the actual values of these QIDs is not possible in this scenario as it would compromise the privacy and confidentiality of the individuals represented in these databases. Therefore, the linkage has to be conducted on masked (encoded) versions of the QID values which have a specific functional relationship with the actual QID values (Vatsalan et al. 2014).

There have been various masking functions proposed in the literature, as reviewed in Vatsalan et al. 2013b. Bloom filter encoding is one masking approach

Table 5.4 Example bank database

ID	First_name	Last_name	DOB	Gender	Postcode	Loan_type	Period	Amount	Paid
6723	peter	robert	20.06.72	M	2617	Mortgage	20	350,000	130,000
8345	miller	roberts	11.10.79	M	2602	Personal	5	10,000	1,900
9241	amelia	millar	06.01.74	F	2415	Mortgage	30	475,000	154,250

Table 5.5 Example social security database

SSN	Title	Last_name	First_name	Age	Postcode	Employment	Income	Benefits	Payment
490814	Mrs	amilia	smith	39	2642	teacher	60,000	child care	45,000
581233	Mr	peter	roberts	42	2617	engineer	110,000	family tax	50,000
932389	Mr	william	smith	69	3205	retired	–	pension	35,000

Table 5.6 Example health database

PID	Surname	Given_name	Age	Postcode	Sex	Pressure	Stress	Last_visited	Reason_of_visit
P1209	robertt	peter	41	2617	m	140/90	high	25 days ago	chest pain
P4204	miller	amelia	39	2415	f	120/80	high	61 days ago	headache
P4894	sieman	jeff	30	2602	m	110/80	normal	15 days ago	checkup

that has widely been used in PPRL (Durham 2012; Ranbaduge et al. 2014; Schnell et al. 2009; Sehili et al. 2015; Vatsalan and Christen 2012, 2014). An example of Bloom filter encoding is illustrated in Fig. 5.4. The Bloom filter encoded QID values can then be compared using a set-based similarity function such as the Dice-coefficient (Schnell et al. 2009). The Dice-coefficient of P Bloom filters $(b_1, \ldots, b_P)$ is calculated as:

$$dice\_sim(b_1, \ldots, b_P) = \frac{P \times c}{\sum_{i=1}^{P} x_i} \tag{5.1}$$

where c is the number of common bit positions that are set to 1 in all P Bloom filters (common 1-bits), and x_i is the number of bit positions set to 1 in b_i (1-bits), $1 \leq i \leq P$.

As discussed in Sect. 5.2.1, comparing all pairs or sets of records is not scalable due to the resulting quadratic or exponential complexities, respectively (Vatsalan et al. 2013b). Generally, private blocking or indexing techniques (Durham 2012; Karakasidis 2012; Kuzu et al. 2013; Ranbaduge et al. 2014; Vatsalan et al. 2013a)

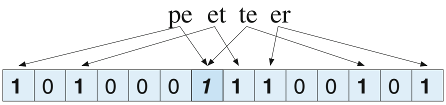


Fig. 5.4 Example Bloom filter encoding of value ‘peter’. The q-grams ($q = 2$) of ‘peter’ are hash-mapped into a Bloom filter of $l = 14$ bits using $k = 2$ hash functions

are used to reduce the number of comparisons that are required in PPRL. For example, applying Soundex-based phonetic blocking (Christen 2006) on the three example databases using the surname/last_name attribute as the blocking criteria results in blocks as shown in Tables 5.7, 5.8 and 5.9 with their encoded Bloom filters (made-up) using first_name/given_name, last_name/surname, and postcode as the QIDs.

Records are then compared with only the records from other databases that are in the same block. In the running example, comparing records (Bloom filters) in the blocking key (Soundex code) ‘r163’ using the Dice-coefficient similarity function and classifying records as matches that have a *dice_sim* (Eq. 5.1) of at least a minimum threshold $s_t = 0.8$, are shown in Figure 5.5. The records with ID 6723 from Table 5.4, SSN 581233 from Table 5.5, and PID P1209 from Table 5.6 are classified as corresponding to the same person as the similarity of these (masked) records is $dice\_sim = 0.86$ (≥0.8). Identifying matching records from subsets of databases (e.g. ID 9241 from Table 5.4 and PID P4204 from Table 5.6) is also an important problem in PPRL which requires further research.

Schnell et al. (2009) and Durham (2012) proposed to use a third party (linkage unit) to compare and classify the Bloom filters from two database owners.

Table 5.7 Records in the example bank database (Table 5.4) with their blocks, QIDs, and Bloom filter encodings. The records in block ‘r163’ are highlighted in italics

Block	Rec_ID	QID	Bloom filter
r163	*6723*	*peter,robert,2617*	*1 1 1 1 0 1 0 1 0 0 1 1 0 1 1 0 1 0 1 1*
r163	*8345*	*miller,roberts,2602*	*1 1 1 0 0 0 1 0 1 0 1 0 1 1 1 0 1 1 0 1*
m460	9241	amelia,millar,2415	1 1 0 0 1 0 1 1 0 0 0 1 1 0 0 1 1 0 0 1

Table 5.8 Records in the example social security database (Table 5.5) with their blocks, QIDs, and Bloom filter encodings. The record in block ‘r163’ is highlighted in italics

Block	Rec_ID	QID	Bloom filter
s530	490814	amilia,smith,2642	1 1 0 1 0 1 1 0 0 1 1 0 1 0 0 1 1 0 1 0
r163	*581233*	*peter,roberts,2617*	*1 1 1 1 0 1 0 1 1 0 1 1 1 0 1 0 1 0 1 1*
s530	932389	william,smith,3205	1 0 0 1 0 1 1 0 0 1 1 1 0 0 1 1 0 1 0 0

Table 5.9 Records in the example health database (Table 5.6) with their blocks, QIDs, and Bloom filter encodings. The record in block ‘r163’ is highlighted in italics

Block	Rec_ID	QID	Bloom filter
r163	*P1209*	*peter,robertt,2617*	*1 1 1 1 0 1 0 1 0 1 1 1 0 1 1 1 1 0 1 1*
m460	P4204	amelia,miller,2415	1 1 0 1 1 0 1 1 0 0 0 1 1 0 0 1 1 1 0 1
s550	P4894	jeff,sieman,2602	0 1 0 1 1 0 0 0 1 0 1 1 0 0 1 0 0 1 0 1

	Candidate set	Bloom filters	x and c	Similarity	Class
1.	(6723,	1 1 1 1 0 1 0 1 0 0 1 1 0 1 1 0 1 0 1 1	$x_1 = 13$	$\frac{3 \times 12}{(13+14+15)}$	Match
	581233,	1 1 1 1 0 1 0 1 1 0 1 1 1 0 1 0 1 0 1 1	$x_2 = 14$		
	P1209)	1 1 1 1 0 1 0 1 0 1 1 1 0 1 1 1 1 0 1 1	$x_3 = 15$		
		& 1 1 1 1 0 1 0 1 0 0 1 1 0 0 1 0 1 0 1 1	$c = 12$	$= 0.86$	
2.	(8345,	1 1 1 0 0 0 1 0 1 0 1 0 1 1 1 0 1 1 0 1	$x_1 = 12$	$\frac{3 \times 7}{(12+14+15)}$	Non-match
	581233,	1 1 1 1 0 1 0 1 1 0 1 1 1 0 1 0 1 0 1 1	$x_2 = 14$		
	P1209)	1 1 1 1 0 1 0 1 0 1 1 1 0 1 1 1 1 0 1 1	$x_3 = 15$		
		& 1 1 1 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 1	$c = 7$	$= 0.51$	

Fig. 5.5 Comparison and classification of Bloom filters in block 'r163' from Tables 5.7, 5.8 and 5.9

A privacy risk with using a linkage unit is the possible collusion between a party and the linkage unit with the aim to learn about data from the other parties (Vatsalan et al. 2013b). A two-party protocol (Vatsalan and Christen 2012) was later proposed where the database owners iteratively exchange selected bits from their Bloom filters and classify the record pairs without requiring a third party. Most work in PPRL so far support the linkage of two sources only. However, two novel approaches for multi-party PPRL for more than two databases based on Bloom filter encodings were recently proposed (Ranbaduge et al. 2014; Vatsalan and Christen 2014). One of the main challenges with multiple parties is the exponential increase in the number of record sets that potentially have to be compared.

In addition to Bloom filter encoding, several other masking functions, ranging from computationally expensive cryptographic techniques (Lindell and Pinkas 2009) to differential privacy (Dwork 2006), *k*-anonymity (Sweeney 2002), reference values (Pang et al. 2009), and noise addition techniques (Karakasidis et al. 2011) have been used in the literature to preserve privacy while allowing the linkage. Other privacy components that need to be considered in a PPRL project are encrypted communication among the parties using public/private key pairs, secure generation and exchange of keys, employee confidentiality agreements to reduce internal threats, as well as secure connections and servers to reduce external threats.

5.5 Research Directions

Most advanced record linkage techniques have been developed by computer science researchers. The focus of these techniques was not only on data that contain personal information, as is generally required for population reconstruction, but often on bibliographic records, or consumer product or business data. Based on existing techniques and approaches, the following research directions can be identified:

- A main open challenge is how collective and graph-based classification techniques, that have shown to be highly accurate, can be used on personal data such as those available in (historical) census and BDM databases. Compared to the bibliographic databases on which such techniques so far have been evaluated, much less relational structure is available in personal data. Specifically, the number of different entity types, and their relationships, are more limited.
- Only limited work has been conducted on how to incorporate temporal information into the linkage process, such as personal details like name and address values that can change over time (Chiang et al. 2014; Christen and Gayler 2013; Li et al. 2011). However, such changes, especially in address attributes, occur regularly and at significant rates.
- As in many applications no or only a limited amount of training data in the form of true matches and non-matches are available, further investigating active learning techniques (Arasu et al. 2010; Bellare et al. 2012), specifically in the context of population reconstruction, could lead to significant reduction in the manual efforts currently required with traditional record linkage approaches. Furthermore, visualising, for example multiple households or families that were linked over time, and highlighting ambiguities and conflicts in the obtained linkages, could help to both better understand problems in linkage algorithms, and also improve the selection and preparation of manual training examples.
- Related to the previous point, given the generally low quality of historical data, developing (semi-) automatic data cleaning and standardisation techniques (Fu et al. 2011a), based on approaches that learn the characteristics of data errors and variations, will significantly reduce the time consuming and cumbersome process of manual data cleaning that is still commonly required today. The requirements of training data of such learning algorithms should be minimised, by for example employing active learning (Arasu et al. 2010; Bellare et al. 2012) or bootstrapping approaches where increasingly accurate models are trained in an iterative fashion (Churches et al. 2002). Additionally, such learning techniques should also be transferable from one domain to another, or allow re-training with little (manual) effort.

With regard to PPRL, while significant advances have been achieved in this area, there are several open research questions that need to be solved in order to make PPRL practical (Christen et al. 2014):

- So far most PPRL techniques have only investigated the linking of two databases. However, as the example scenario in Sect. 5.4.2 has shown, in many real-world applications data from more than two sources need to be linked. Our recent work in multi-party PPRL (Ranbaduge et al. 2014; Vatsalan and Christen 2014) has highlighted the significant computational challenges when aiming to link data from several sources, as even when using sophisticated blocking techniques the number of candidate record sets to be compared increases exponentially with the number of parties involved. Besides these computational challenges, possible collusion between subsets of parties needs to be considered.

- Most existing PPRL techniques only employ a simple threshold-based classifier to classify record pairs into matches or non-matches. Only group linkage (Li et al. 2011) has been considered within a PPRL framework, but none of the other advanced collective and graph-based approaches discussed in Sect. 5.2.2 have so far been investigated for their applicability in PPRL. A major challenge for classification in PPRL is the use of training data for supervised learning approaches, because such data generally require access to actual sensitive attribute values.
- How to assess linkage quality and completeness has so far not been thoroughly investigated for PPRL. This is, however, a must-solve problem as otherwise it will not be possible to evaluate the efficiency and effectiveness of PPRL techniques in real-world applications, making these techniques non-practical.
- Unlike for measuring linkage performance and quality, where standard measurements, such as run-time, reduction ratio, pairs completeness, pairs quality, precision, recall, or accuracy can be used (Christen 2012a), there are currently no standards available for measuring privacy for PPRL. Different measures have been proposed and used (Vatsalan et al. 2013b, 2014), making the comparison of techniques difficult.
- Finally, no framework has been developed that allows the experimental comparison of different PPRL techniques with regard to their scalability, linkage quality, and privacy preservation. Ideally such a framework should allow researchers to easily ‘plug-in’ their algorithms. Related to this issue is the lack of standard test data sets, a problem that is not just specific to PPRL but to record linkage research in general (Christen 2012a; Köpcke and Rahm 2010). A possible alternative to using real-world data sets, which are difficult to obtain due to privacy and confidentiality reasons, is to use synthetic data that are generated based on the characteristics of real data (Christen and Vatsalan 2013).

Improved collaboration between domain experts, computer scientists and statisticians who work on the algorithmic aspects of record linkage is needed to obtain the best outcomes for the field of population reconstruction. Neither research area can work in isolation. While multidisciplinary research brings its own challenges, the importance of such applied research is now increasingly being recognised by research areas that traditionally have worked in isolation (Rudin and Wagstaff 2013).

5.6 Conclusions

As our society moves into the ‘Big Data’ era, tremendous opportunities arise for research in the social sciences to use large-scale population-based databases collected both by commercial organisations as well as government agencies. Compared to small controlled studies based on surveys and experimental set-ups, using large databases can help overcome sampling bias and potentially reduce costs. In an analogy to genomics and bioinformatics, Kum et al. (2013) recently proposed the notion of the 'social footprint' or 'social genome', and the field of 'population informatics' which deals with the collection, integration, and analysis of data about people gathered from many different domains, including healthcare, education, employment, finance, and so on. Reconstructing a population from such data, and enriching existing (census) data collections with such external data, will allow insights into many aspects of today's societal challenges.

National census agencies are also realising both the challenges and opportunities that matching their data with external, possibly commercial, databases can bring (Baffour et al. 2013; Office for National Statistics 2013). The acquisition of data from a variety of organisations is, however, a complicated process that involves negotiations with various partners. Privacy and confidentiality, as well as data quality issues, need to be considered carefully. As computers become more powerful, the computational challenges of linking large databases become less of an issue compared to non-technical challenges such as obtaining access to the data required for certain studies, or communication between researchers from different domains.

Nevertheless, research into techniques that allow efficient and effective population reconstruction based on data linked from a variety of sources will likely not only attract more interest from academia, but also from governments and private sector organisations. Understanding the structures and characteristics of populations, and how they change over time, becomes more valuable for organisations in an ever more competitive environment, where a better understanding of their data can give an organisation the competitive edge it needs to be successful (Siegel 2013).

Acknowledgments The authors would like to thank Mac Boot (The Australian National University) and Vassilios S. Verykios (Hellenic Open University) for their contributions to the work presented in this chapter.

References

Al-Lawati, A., Lee, D., & McDaniel, P. (2005). Blocking-aware private record linkage. In International Workshop on Information Quality in Information Systems (pp. 59–68). Baltimore.

Antonie, L., Inwood, K., Lizotte, D. J., & Ross, J. A. (2014a). Tracking people over time in 19th century Canada for longitudinal analysis. *Machine Learning, 95*, 129–146.

Antonie, L., Inwood, K., & Ross, A. (2014b). Dancing with dirty data: Problems in the extraction of life-course evidence from historical censuses. In *Population Reconstruction*.

Arasu, A., Götz, M., & Kaushik, R. (2010). On active learning of record matching packages. In ACM SIGMOD (pp. 783–794). Indianapolis.

Atallah, M. J., Kerschbaum, F., & Du, W. (2003). Secure and private sequence comparisons. In ACM Workshop on Privacy in the Electronic Society (pp. 39–44). Washington, DC.

Baffour, B., King, T., & Valente, P. (2013). The modern census: Evolution, examples and evaluation. *International Statistical Review, 81*(3), 407–425.

Bellare, K., Iyengar, S., Parameswaran, A. G., & Rastogi, V. (2012). Active sampling for entity matching. In ACM SIGKDD (pp. 1131–1139). Beijing.

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data, 1*(1), 5.

Bilenko, M., Kamath, B., & Mooney, R. J. (2006). Adaptive blocking: Learning to scale up record linkage. In IEEE ICDM (pp. 87–96). Hong Kong.

Block, W. C., & Star, D. L. (1995). Data entry and verification. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 28*(1), 63–65.

Bloothooft, G. (1995). Multi-source family reconstruction. *History and computing, 7*(2), 90–103.

Bonomi, L., Xiong, L., Chen, R., & Fung, B. (2012). Frequent grams based embedding for privacy preserving record linkage. In CIKM (pp. 1597–1601). Maui, Hawaii.

Chiang, Y. H., Doan, A., & Naughton, J. F. (2014). Tracking entities in the dynamic world: A fast algorithm for matching temporal records. *PVLDB, 7*(6).

Christen, P. (2006). A comparison of personal name matching: Techniques and practical issues. In Workshop on Mining Complex Data, held at IEEE ICDM. Hong Kong.

Christen, P. (2012a). *Data Matching—Concepts and techniques for record linkage, entity resolution, and duplicate detection. Data-centric systems and applications*. Berlin: Springer.

Christen, P. (2012b). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering, 24*(9), 1537–1555.

Christen, P. (2014). Advanced record linkage methods and privacy aspects for population reconstruction. In Population Reconstruction.

Christen, P., & Gayler, R.W. (2013). Adaptive temporal entity resolution on dynamic databases. In PAKDD (Vol. 7819, pp. 558–569). Gold Coast, Australia: Springer.

Christen, P., Gayler, R. W., & Hawking, D. (2009). Similarity-aware indexing for real-time entity resolution. In ACM CIKM (pp. 1565–1568). Hong Kong.

Christen, P., & Vatsalan, D. (2013). Flexible and extensible generation and corruption of personal data. In ACM CIKM (pp. 1165–1168). San Francisco.

Christen, P., Vatsalan, D., & Verykios, V. S. (2014). Challenges for privacy preservation in data integration. *ACM Journal Data and Information Quality, 5*(1–2), 4.

Churches, T. (2003). A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers. *BMC Med Res Methodol, 3*(1), 1.

Churches, T., Christen, P., Lim, K., & Zhu, J. X. (2002). Preparation of name and address data for record linkage using hidden Markov models. *BMC Med Inform Decis Mak, 2*, 9.

Dey, D., Mookerjee, V. S., & Liu, D. (2010). Efficient techniques for online record linkage. *IEEE Transactions on Knowledge and Data Engineering, 23*(3), 373–387.

de Vries, T., Ke, H., Chawla, S., & Christen, P. (2011). Robust record linkage blocking using suffix arrays and Bloom filters. *ACM Transactions on Knowledge Discovery from Data, 5*(2), 9.

Dong, X. L., Halevy, A., & Madhavan, J. (2005). Reference reconciliation in complex information spaces. In ACM SIGMOD (pp. 85–96). Baltimore.

Draisbach, U., Naumann, F., Szott, S., & Wonneberg, O. (2012). Adaptive windows for duplicate detection. In IEEE ICDE (pp. 1073–1083). Washington, DC.

Durham, E.A. (2012). A framework for accurate, efficient private record linkage. Ph.D. thesis, Faculty of the Graduate School of Vanderbilt University, Nashville, TN.

Durham, E. A., Xue, Y., Kantarcioglu, M., & Malin, B. (2012). Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. *Information Fusion, 13*(4), 245–259.

Dwork, C. (2006). Differential privacy. Automata, languages and programming (pp. 1–12).

Efremova, J., Ranjbar-Sahraei, B., Oliehoek, F. A., Calders, T., & Tuyls, K. (2015). A baseline method for genealogical entity resolution. In: G. Bloothooft, P. Christen, K. Mandemakers, M. Schraagen (Eds.), *Population reconstruction*. Berlin: Springer.

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering, 19*(1), 1–16.

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

Fu, Z., Boot, M., Christen, P., & Zhou, J. (2014a). Automatic record linkage of individuals and households in historical census data. *International Journal of Humanities and Arts Computing, 8*(2), 204–225.

Fu, Z., Christen, P., & Zhou, J. (2014b). A graph matching method for historical census household linkage. In PAKDD (Vol. 8443, pp. 485–496). Tainan, Taiwan: Springer.

Fu, Z., Christen, P., & Boot, M. (2011a). Automatic cleaning and linking of historical census data using household information. In Workshop on Domain Driven Data Mining, held at IEEE ICDM. Vancouver.

Fu, Z., Christen, P., & Boot, M. (2011b). A supervised learning and group linking method for historical census household linkage. In AusDM, CRPIT (Vol. 121). Ballarat, Australia.

Fu, Z., Zhou, J., Christen, P., & Boot, M. (2012) Multiple instance learning for group record linkage. In PAKDD (Vol. 7301, pp. 171–182). Kuala Lumpur, Malaysia: Springer.

Fure, E. (2000). Interactive record linkage: The cumulative construction of life courses. *Demographic Research, 3*(11), 3–11.

Glasson, E., De Klerk, N., Bass, J., Rosman, D., Palmer, L. J., & Holman, D. (2008). Cohort profile: The Western Australian family connections genealogical project. *International Journal of Epidemiology, 37*(1), 30–35.

Hernandez, M. A., & Stolfo, S. J. (1995). The merge/purge problem for large databases. In ACM SIGMOD (pp. 127–138). San Jose.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). Data quality and record linkage techniques. Berlin: Springer.

Inan, A., Kantarcioglu, M., Bertino, E., & Scannapieco, M. (2008). A hybrid approach to private record linkage. In IEEE ICDE (pp. 496–505). Cancun, Mexico.

Inan, A., Kantarcioglu, M., Ghinita, G., & Bertino, E. (2010). Private record matching using differential privacy. In EDBT (pp. 123–134). Lausanne, Switzerland.

Ioannou, E., Nejdl, W., Niederée, C., & Velegrakis, Y. (2010). On-the-fly entity-aware query processing in the presence of linkage. *VLDB Endowment, 3*(1), 429–438.

Jin, L., Li, C., & Mehrotra, S. (2003). Efficient record linkage in large data sets. In DASFAA (pp. 137–146). Tokyo.

Jonas, J., & Harper, J. (2006). *Effective counterterrorism and the limited role of predictive data mining*. Policy Analysis (584) (2006).

Kalashnikov, D. V., & Mehrotra, S. (2006). Domain-independent data cleaning via analysis of entity-relationship graph. *ACM Transactions on Database Systems, 31*(2), 716–767.

Karakasidis, A., & Verykios, V. S. (2009). Privacy preserving record linkage using phonetic codes. In Fourth Balkan Conference in Informatics, IEEE (pp. 101–106). Thessaloniki, Greece.

Karakasidis, A., & Verykios, V. S. (2010). Advances in privacy preserving record linkage. In E-activity and Innovative Technology, Advances in Applied Intelligence Technologies Book Series (pp. 22–34). IGI Global.

Karakasidis, A., & Verykios, V. S. (2012). Reference table based k-anonymous private blocking. In ACM Symposium on Applied Computing (pp. 859–864). Trento, Italy.

Karakasidis, A., Verykios, V. S., & Christen, P. (2011). Fake injection strategies for private phonetic matching. In International Workshop on Data Privacy Management. Leuven, Belgium.

Karapiperis, D., & Verykios, V. S. (2014). An LSH-based blocking approach with a homomorphic matching technique for privacy-preserving record linkage. *IEEE Transactions on Knowledge and Data Engineering*.

Kejriwal, M., & Miranker, D. P. (2013). An unsupervised algorithm for learning blocking schemes. In IEEE ICDM (pp. 340–349).

Kelman, C. W., Bass, J., & Holman, D. (2002). Research use of linked health data—A best practice protocol. *Aust NZ Journal of Public Health, 26*, 251–255.

Köpcke, H., & Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data and Knowledge Engineering, 69*(2), 197–210.

Kum, H. C., Krishnamurthy, A., Machanavajjhala, A., & Ahalt, S. (2013). Population informatics: Tapping the social genome to advance society: A vision for putting 'Big Data' to work for population informatics. *Computer, PP*(99).

Kuzu, M., Kantarcioglu, M., Inan, A., Bertino, E., Durham, E., & Malin, B. (2013). Efficient privacy-aware record integration. In EDBT (pp. 167–178). Genoa, Italy.

Lee, D., Kang, J., Mitra, P., Giles, C. L., & On, B. W. (2007). Are your citations clean? *Commununications of the ACM, 50*, 33–38.

Li, F., Chen, Y., Luo, B., Lee, D., & Liu, P. (2011). Privacy preserving group linkage. In SSDBM (Vol. 6809, pp. 432–450). Portland: Springer LNCS.

Li, P., Dong, X. L., Maurino, A., & Srivastava, D. (2011). Linking temporal records. *VLDB Endowment, 4*(11), 956–967.

Lindell, Y., & Pinkas, B. (2009). Secure multiparty computation for privacy-preserving data mining. *Journal of Privacy and Confidentiality, 1*(1), 5.

Michelson, M., & Knoblock, C. A. (2006). Learning blocking schemes for record linkage. In AAAI. Boston.

Naumann, F., & Herschel, M. (2010). An introduction to duplicate detection. *Synthesis Lectures on Data Management* (vol. 3). Morgan and Claypool Publishers.

Newcombe, H. B. (1988). *Handbook of record linkage: Methods for health and statistical studies, administration, and business*. New York: Oxford University Press Inc.

Newcombe, H. B., & Kennedy, J. M. (1962). Record linkage: making maximum use of the discriminating power of identifying information. *Communications of the ACM, 5*(11), 563–566.

Newton, G. (2013). Family reconstitution in an urban context: Some observations and methods. Technical Report, University of Cambridge, CWPESH No. 12.

Office for National Statistics. (2013). Beyond 2011 matching anonymous data. Methods and Policies Report M9.

On, B. W., Koudas, N., Lee, D., & Srivastava, D. (2007). Group linkage. In IEEE ICDE (pp. 496–505). Istanbul.

Pang, C., Gu, L., Hansen, D., & Maeder, A. (2009). Privacy-preserving fuzzy matching using a public reference table. *Intelligent Patient Management, 189*, 71–89.

Quass, D., & Starkey, P. (2003). Record linkage for genealogical databases. In ACM SIGKDD Workshop on Data Cleaning, Record Linkage and Object Consolidation (pp. 40–42). Washington DC.

Ramadan, B., Christen, P., & Liang, H. (2014). Dynamic sorted neighborhood indexing for real-time entity resolution. In ADC (Vol. 8506, pp. 1–12). Brisbane: Springer LNCS.

Ranbaduge, T., Christen, P., & Vatsalan, D. (2014). Tree based scalable indexing for multi-party privacy-preserving record linkage. In AusDM, CRPIT (Vol. 158). Brisbane, Australia.

Rastogi, V., Dalvi, N., & Garofalakis, M. (2011). *Large-scale collective entity matching. VLDB Endowment, 4*, 208–218.

Ravikumar, P., Cohen, W., & Fienberg, S. (2004). A secure protocol for computing string distance metrics. In Workshop on Privacy and Security Aspects of Data Mining held at IEEE ICDM (pp. 40–46). Brighton, UK.

Reid, A., Davies, R., & Garrett, E. (2002). Nineteenth-century scottish demography from linked censuses and civil registers: A'sets of related individuals' approach. *History and Computing, 14* (1–2), 61–86.

Rudin, C., & Wagstaff, K. L. (2013). Machine learning for science and society. *Machine Learning, 95*(1), 1–9.

Ruggles, S. (2002). Linking historical censuses: A new approach. *History and Computing, 14*(1–2), 213–224.

Scannapieco, M., Figotin, I., Bertino, E., & Elmagarmid, A. K. (2007). Privacy preserving schema and data matching. In ACM SIGMOD (pp. 653–664). Beijing.

Schneier, B. (1996). *Applied cryptography: Protocols, algorithms, and source code in C* (2nd ed.). New York: Wiley.

Schnell, R., Bachteler, T., & Reiher, J. (2009). Privacy-preserving record linkage using Bloom filters. *BioMed Central Medical Informatics and Decision Making, 9*(1), 41.

Sehili, Z., Kolb, L., Borgs, C., Schnell, R., & Rahm, E. (2015). Privacy preserving record linkage with PPJoin. In BTW Conference. Hamburg.

Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. New York: Wiley.

Su, W., Wang, J., & Lochovsky, F. H. (2009). Record matching over query results from multiple web databases. *IEEE Transactions on Knowledge and Data Engineering, 22*(4), 578–589.

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems, 10*(5), 557–570.

Talburt, J.R. (2011). Entity resolution and information quality. Morgan Kaufmann.

Toxen, B. (2014). The NSA and Snowden: Securing the all-seeing eye. *Communications of the ACM, 57*(5), 44–51.

Trepetin, S. (2008). Privacy-preserving string comparisons in record linkage systems: a review. *Information Security Journal: A Global Perspective, 17*(5), 253–266.

Vatsalan, D., & Christen, P. (2012). An iterative two-party protocol for scalable privacy-preserving record linkage. In AusDM, CRPIT (Vol. 134). Sydney, Australia.

Vatsalan, D., & Christen, P. (2014). Scalable privacy-preserving record linkage for multiple databases. In ACM CIKM. Shanghai.

Vatsalan, D., Christen, P., O'Keefe, C. M., & Verykios, V. S. (2014). An evaluation framework for privacy-preserving record linkage. *Journal of Privacy and Confidentiality, 6*(1), 3.

Vatsalan, D., Christen, P., & Verykios, V. S. (2011). An efficient two-party protocol for approximate matching in private record linkage. In AusDM, CRPIT (Vol. 121). Ballarat, Australia.

Vatsalan, D., Christen, P., & Verykios, V. S. (2013a). Efficient two-party private blocking based on sorted nearest neighborhood clustering. In ACM CIKM (pp. 1949–1958). San Francisco.

Vatsalan, D., Christen, P., & Verykios, V. S. (2013b). A taxonomy of privacy-preserving record linkage techniques. *Information Systems, 38*(6), 946–969.

Verykios, V. S., & Christen, P. (2013). Privacy-preserving record linkage. *Wiley Interdisciplinary reviews: Data Mining and Knowledge Discovery, 3*(5), 321–332.

Verykios, V. S., Karakasidis, A., & Mitrogiannis, V. K. (2009). Privacy preserving record linkage approaches. *International Journal of Data Mining, Modelling and Management, 1*(2), 206–221.

Winkler, W. E. (2006). Overview of record linkage and current research directions. Technical Report RR2006/02, US Bureau of the Census, Washington, DC.

Yakout, M., Atallah, M. J., & Elmagarmid, A. K. (2009). Efficient private record linkage. In IEEE ICDE (pp. 1283–1286). Shanghai.

Yan, S., Lee, D., Kan, M. Y., & Giles, C. L. (2007). Adaptive sorted neighborhood methods for efficient record linkage. In ACM/IEEE-CS joint conference on Digital Libraries (pp. 185–194). Vancouver.

Chapter 6
Reconstructing Historical Populations from Genealogical Data Files

Corry Gellatly

Abstract Over the past two decades, a huge number of historical documents have been digitised and made available online. At the same time, numerous software options and websites have encouraged people to conduct research into their family trees, leading to a surge in the availability of genealogical data. A major advantage of genealogical data, from a scientific research perspective, is that it combines information from many sources into a format that is structured by family relations and descendancy, which is very useful for studying the dynamics of population change over the generations. A critical issue for researchers who want to use genealogical data is how to assess the quality of the data and put in place measures to correct the errors that we find in it. In this chapter, I present some of the methods that are being used to filter, clean and aggregate genealogical data to create large datasets that may be used across a diverse range of academic research disciplines.

6.1 Introduction

There are examples of genealogical data being used to address research questions across a diverse range of disciplines, including, for example, historical demography (e.g. Zhao 1994), historical geography (e.g. Otterstrom and Bunker 2013), family history (e.g. Post et al. 1997), human fertility and natural selection (e.g. Moreau et al. 2011), heritability analyses (e.g. Gavrilov and Gavrilova 2001) and genetic genealogy (Larmuseau et al. 2013). However, genealogical data is arguably an under-used resource in scientific research, given the boom in genealogical research over the past couple of decades, which has been driven by the popularity of genealogy websites and the increasing availability of historical records online.

C. Gellatly (✉)
Department of History and Art History Utrecht University,
Drift 6, 3512 BS Utrecht, The Netherlands
e-mail: c.gellatly@uu.nl

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_6

It has been argued by some that because the majority of genealogical research is conducted by amateurs, it is too unreliable to be of much use for academic research purposes, but the validity of that argument has not been proven. It can alternatively be argued that the increasing mass of genealogical data being generated by both amateur and professional researchers represents a potential boon for academic research, because it is of a volume that could not realistically be gathered within the scope of one, or even several, academic research projects.

There is a huge potential benefit for historians and other researchers who are able to make use of the vast amounts of genealogical data that are in existence, because these effectively represent many thousands of hours of crowd-sourced work and a considerable source of knowledge about individual families.

The most widespread format for exchange of genealogical data over the last two decades has been the GEDCOM format. It is an open format with a simple lineage-linked structure, in which each record relates to either an individual or a family and relevant information, such as names, events, places, relationships and dates appear in a hierarchical structure below the record ID. The format was developed by the Church of Jesus Christ of Latter-day Saints, and the current standard specification is version 5.5, which was released in 1996.

There are a number of reasons why the GEDCOM format has become the dominant means for exchanging genealogical data:

- It provides a systematic and standardised way of structuring information about individuals, their families and life events.
- It satisfies the demand that exists for a common file format for exchanging genealogical data, because the ability to share family trees is a useful tool for genealogists to identify common connections and expand their family trees.
- The format has allowed for easy development of software applications that can read or export to the file type, because it is an open text-file format. It has been estimated that over 500 computer programs have been written that can export files in the GEDCOM format.[1]
- The format is flexible in terms of what can be added to a file, allowing users and software to easily make use of the format, even without having to conform to the correct standard specifications.

Despite the predominance of GEDCOM for exchange of genealogical data, there are a number of legitimate criticisms of the format, these are:

- The standard specification gives a list of tags that can be used to describe events; however, in many instances, software will use non-standard tags. This results in confusion or data loss when the file is read by a different software application. In such cases, the flexibility of the format is arguably problematic.
- There are no constraints on the data that may be entered under each tag. For example, it is possible to enter a date of birth from the future, or to enter, e.g.

[1] According to Louis Kessler, an expert on the GEDCOM format, speaking at Gaenovium 2014, a genealogy technology conference held on 7 October 2014 in Leiden, The Netherlands.

'<1900' or 'Born between 1900 and 1920' in a date field, which may not be understood by a different software application. It is often the case that place identifiers are ambiguous, so that confusion may arise about the geographical location. For example, if the place name is simply 'Washington', this could refer to the city or the state in the US, or the town or village in the UK.
- There are technical constraints with the standard format. An example is that multiple people cannot be linked to a single place, source citation or note, so these may have to be replicated within each file many times. In technical terms, the data format is not normalised, which leads to excessive replication of content and increased risk of data corruption.

The above problems with the GEDCOM format have almost certainly contributed to the poor quality of many genealogical files in existence, simply because the format does not inherently prevent errors. However, it is perfectly possible for GEDCOM files to contain accurate, well-structured and well-documented genealogical research, and they often do.

An important point about genealogical research conducted by amateurs is that it tends to include the knowledge that people naturally acquire about their own family history. This knowledge may typically extend no further than a few generations, but may be essential to construction of the genealogy in some instances. In this sense, use of amateur genealogical data for population research has some commonality with indirect techniques that are sometimes used for demographic estimation (United Nations 1983), because the data is not purely constructed from direct sources, such as marriage records, censuses, etc., but also photographs, letters, heirlooms and spoken knowledge passed down through a family.

There is an opportunity for those with an interest in reconstructing historical populations or following life history variables through generations to utilise the considerable amounts of genealogical data available, with the proviso that it must be assessed for probability of correctness. In this chapter, I detail the genealogical data collation methods that are being developed as part of an inter-disciplinary project involving historians and biologists, in which we are seeking to identify the social and biological determinants of the historical increase in life expectancy in Europe since the Early Modern period.[2] The methods build on those initially developed to study inter-generational patterns of human sex ratio variation, in which a genealogical database was compiled from >900 family trees (Gellatly 2009). This database includes GEDCOM files downloaded from several online repositories, but principally from a now defunct website: http://www.genealogyforum.com. The database contains mostly North American genealogies, though also some Western European genealogies and a small number of records from the rest of the world. This database has been geocoded and error checked, and is used as the basis for the methods, tests and analyses described here.

[2] Dutch NWO funded project conducted at University of Utrecht: Nature or nurture? A search for the institutional and biological determinants of life expectancy in Europe during the early modern period (276-53-008).

I briefly describe a series of steps for screening out substantially flawed genealogical data files, cleaning dates and geocoding place information. I then address the issue of how to extract research datasets that do not contain duplicate individuals. I describe tests of group-linking methods, which may be used to identify duplicate individuals based on family characteristics, such as number and sex of siblings and children, as well as on individual features, such as surname, year of birth, etc. A number of variations on this method are compared for their efficiency in identifying duplicate matches, and for their power—in terms of accuracy and data coverage. It is found that some of the group-linking methods have significant advantages in terms of identifying duplicates, particularly when taking into account problems with first names in genealogical data, in which nicknames, abbreviated names, middle names and actual correct given names are inconsistently used.

The methods described are being developed and made available via an open source web-based application called TreeChecker, available at the GitHub code repository: https://github.com/cgeltly/treechecker, with a wiki page at: http://www.treechecker.net/wiki. At present, the application is able to extract data from a variety of GEDCOM files, produce error reports and assist with error tracing. In time, the code will be developed to add extra capabilities, including the capacity to export data to new formats, such as RDF and XML, also the development of error tracing algorithms and the capability to match duplicates across multiple files. A website that will offer the opportunity for users to upload their GEDCOM files and run a detailed series of error checks is currently in the beta testing phase, and will be made available at http://www.treechecker.net, once testing is completed.

6.2 Building a Genealogical Database

There are some large GEDCOM files in existence that contain many thousands of individuals, but most contain far fewer. If we are seeking to understand historical trends and demographic phenomena, then the potential research questions that can be addressed with single GEDCOM files are limited. It is for this reason that we can aggregate individual files together to create large genealogical databases. In the process, we screen out poor quality files, carry out data cleaning and identify linkages between separate genealogies, which allows us to derive useful datasets from the database that can be used to address various research questions.

6.2.1 Screening Files

The process of screening involves the identification of errors and potential errors in GEDCOM files, in order to identify those that contain poor quality research or are incorrectly constructed. All errors need to be evaluated in context. Is it one of only a few errors in a large, otherwise good, genealogy? Is the error easily corrected? The following list of checks need to be carried out:

- A check for low mean number of offspring per family due to inclusion only of the author's direct lineage and exclusion of their ancestors' siblings. This type of bias in genealogical research excludes much of the familial information that we require, e.g. for reconstructing kin networks (Post et al. 1997) or following the heritability of traits such as sex ratio (Gellatly 2009). If the mean number of offspring per family is 1, then a file clearly contains a lineage. If it is close to 1, then further inspection may reveal branches of the tree where consecutive generations of individuals have one parent and no siblings.
- Age at death. If there are individuals older than 110 or younger than 0, this is a potential cause to exclude files, unless these are clear typo errors. For example, an individual with parents born in the 1930s and siblings born in 1961, 1962 and 1963 was likely to have been born in 1965 and not 965. In such clear-cut cases, the typo may be corrected, rather than the file being rejected.
- Events before birth. There are no events in the standard GEDCOM specification that occur to an individual before birth, so an event such as marriage, baptism or death occurring is an obvious error—possibly a typo or possibly an incorrect attribution of an individual to an event.
- Time between dates of birth precludes the possibility of a stated relationship between individuals. For example, if an individual becomes a parent younger than 13 years old, or if a woman is older than 55 at the birth of a child or a man older than 80, then this should flag a possible error.
- The spacing between births to a mother is less than one year.
- A relationship to an individual not listed in the family tree. A family may contain the ID of an offspring or parent, though the actual record of that person does not exist within the file. This indicates that the file is corrupted or has not been correctly built or parsed.
- Individuals have more than one mother or father. This may be an error, however, it may also indicate adoption because biological and adoptive parents may both be linked via the child 'CHIL' tag, whilst the adoption 'ADOP' tag is an event tag that tends to be used to indicate the date adoption occurred rather than the relationship itself. This is one of the more taxing problems with the GEDCOM format. The absence of an adoption tag on multiple parentages should flag up the file for possible exclusion or cleaning.
- Incestuous parentage. This is also a challenging problem, because there is no standard tag for it, whilst a brother and sister, father and daughter or mother and son having a child may be an error, rather than something that actually happened. An investigation of other events, e.g. marriages and deaths will often allow incest to be ruled out.
- The number of children under a family record does not match the number of individuals linked to that family as a child. This may indicate an error such as a duplicate entry for a child in the family record.
- Individuals are listed as offspring of one sex, but occur as parents of another sex. This is usually wrong and suggests the file should be excluded or the error cleaned. However, same sex marriages may use the 'HUSB' and 'WIFE' tags,

whilst children may be adopted. As with many potential errors, the context is important and some further investigation may be required.

- A good reason for excluding GEDCOM files is that people may trace their ancestry back so far that it is not credible. If an author traces their ancestry to before the Early Middle Ages (500–1000 AD) then it may be presumed that their methods were not particularly rigorous.

6.2.2 Cleaning

If, at the point of screening out poor quality files, it is recognised that some files may only be suffering from a small number of typo errors, then it may be worth retaining those files and cleaning the data.

Typically, dates are the 'cleanest' fields in GEDCOM files, because most genealogy software enforces some control over date format, although some do not check for impossible dates, e.g. 29th February in a non-leap year, or 31st June. There are functions within most programming languages and databases for identifying incorrect dates. In the MySQL database, for example, incorrect dates cannot be entered into a date formatted field and will be replaced with NULL. However, typos that do not result in incorrect dates still occur and these are problematic, because they alter the ages at which people die or become husbands, wives, parents, etc.

The process of correcting typos and other minor errors in genealogical data typically involves taking into account the context of the event to make a decision about the potential cause of the error and whether or not it can be resolved. The example in Fig. 6.1 shows a potential error in a man's year of birth (YOB). His YOB is 1857, but his mother was born in 1845, his father in 1847, his siblings in 1868, 1869, 1871 and 1873 and his wife in 1874, he was married in 1894 and the

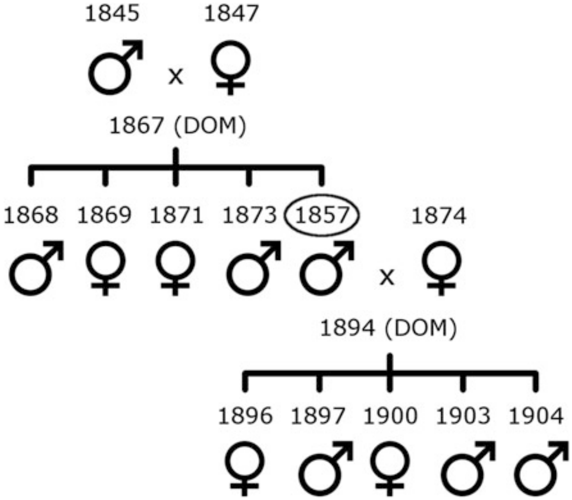


Fig. 6.1 Example of a potential transposition typo error for a year of birth (*circled*). The difference in age between this individual and his parents, siblings, spouse and children is in a normal range when his year of birth is 1875 instead of 1857

first of his children was born in 1896. If the YOB is correct, then he was born to a 10 year old mother, who then waited 11 years before having 4 more children, his wife was 17 years older than him and the first of his 5 children was born when he was 39. If, on the other hand, the YOB is a typo and the man was actually born in 1875, then he was born to a 28 year old mother who had the first of her 5 children at age 21, whilst his wife was 1 year younger and the first of his 5 children was also born when he was 21. On the balance of probabilities, it seems very likely that the YOB is a typo, and that it is simple transposition typo error, where the 5 and 7 have been entered in the incorrect order.

At present, the process of correcting typos requires a high degree of human oversight, to evaluate the potential error. The procedure is as follows:

- First, check other event dates, e.g. if the potential error is in the death event date then check for a burial date, if it is in the birth date then check for a baptism date. Also, check marriage age and whether it is a remarriage.
- Second, check dates of birth for parents, siblings, spouse and offspring. These can put the life of the individual in context, as with the example in Fig. 6.1.
- Third, conduct an internet search for the individual. In many cases, they may be recorded in a separate source, such as another genealogy, census, marriage record, gravestone, etc. The other sources may confirm a suspected typo error, or may point to an incorrect family connection. It may also become apparent that there are conflicting sources and wider difficulties in resolving information about the historical individual.

If, once the preceding checks have been carried out, the potential error remains ambiguous, or appears not to be a typo but instead due to an incorrect familial connection between individuals, for example, then the option to exclude the file from the database still remains.

It is typically a complex decision about how to resolve a potential error in genealogical data, but also a fairly laborious process to put the error in context and check alternative sources. As part of the TreeChecker project, work is underway to use algorithms that are able to examine the context of a potential error and to prompt the user with possible resolutions to it; so, for example, an algorithm will suggest a possible typo in a marriage date, or in mother’s date of birth, and also suggest what the correct date (or date range) might be. The dates suggested by the algorithm may also be used at this point for searching external data sources (e.g. marriage records, birth registrations, etc.) for the event in question, because the inclusion of a correct date can greatly improve the chances of locating an event.

6.2.3 Geocoding

The quality of geographical data in GEDCOM files varies considerably. According to the GEDCOM specification, place names should be entered in a comma separated hierarchy under the PLAC tag, in an order which roughly corresponds to:

Town, Region, State, Country

However, although the correct order is usually followed (i.e. smaller to larger place), churches, graveyards, etc. are often included and higher level information (i.e. country and/or region) is often missing. For example, a record may look like this:

Episcopal Church, Billings, Montana

when it should look like this:

Episcopal Church, Billings, Montana, USA

or perhaps this:

Billings (Episcopal Church), Montana, USA

The difficulty with this system of storing place names is that the standardisation is very weak. Latitude and longitude tags were introduced to the GEDCOM specification version 5.5.1, which allows for the storage of geo-coordinates and facilitates the use of accurate standardised geographical information. Although this is not found in the majority of available GEDCOM files, mapping and geocoding are increasingly being supported by family tree software, so this ought to improve. A painstaking approach, in which internet searches for place names were carried out, and records were all then sorted into countries, was applied to the genealogical database described here. However, the details of this method will not be elaborated on here. It is suggested that similar efforts in the future should make use of mapping software to generate geo-coordinates for places, because this produces standardised data, which is far more amenable to any number of analyses.

6.3 Dataset Extraction

The method described here for constructing a genealogical database involves bringing the data from a collection of GEDCOM files that have been screened for quality and undergone error cleaning and geocoding into a single database. This method retains the completeness of the original files; however, there is no process to deal with the issue of duplicates during the construction of the database. This issue may be dealt with at the point where a dataset is extracted from the database for analysis.

6.3.1 Identification of Duplicates

The appearance of duplicates within a database indicates where family trees overlap, which may in itself be of interest, but for most demographic analyses involving genealogical data, duplicates are problematic, because of their effect on statistical measures such as mean and standard deviation. A duplicated value will bias the mean in the direction of that value and will lead to a lower estimate of standard deviation, because there is no variance between two duplicated values. This is important for tests such as ANOVA (Analysis of variance), which evaluate the difference between means using the variance. If, for example, we were looking at changes in parental age over time, and there were more duplicates in earlier centuries, then failure to identify these duplicates could give misleading results for those centuries in comparison with later centuries.

The problem of duplicate identification is that of 'data matching' or 'record linkage', and is a problem that has been tackled in various other circumstances using a number of mathematical, computational and statistical techniques. Newcombe et al. (1959) first described the principles of probabilistic matching, which recognise that we can take into account factors that may influence the likelihood of a match between two separate records, such as the relative frequency of a particular surname within a sample—see Christen (2012) for a more detailed explanation. It is recognised that there is scope for employing probabilistic matching techniques within the context of genealogical data, because there will be a higher probability of a match between records when surnames are more common, or for more recent dates (because these are usually more prevalent than older dates in genealogies). However, this approach is not explored here, instead the focus is on those single variables (e.g. surname or date of birth) or group-linked variables (e.g. surname and date of birth) that give us the maximum power to identify duplicates and thereby increase our confidence that the datasets we extract from the genealogical database contain unique individuals and families.

In the absence of a 'perfect' method for identifying duplicates, there will always be a degree of error, whereby false duplicates are identified and real duplicates are not. Here I describe tests that were carried out to give an estimate of the accuracy and power of a number of duplicate identification methods, to give an idea of how useful the methods may be when applied to other genealogical databases.

Typographical and transcription errors are common to many types of records, but in the case of genealogical data, the problem of a person's name or an event date being entered differently in two separate files is compounded by the fact that names and dates may have been taken from different sources and recorded by researchers with different methods of transcription. For example, although there is a nickname tag 'NICK' in the GEDCOM format, it is quite common to find the nickname in brackets alongside the first name, whilst either abbreviated or entire first names may be used. Also, a date may be entered as '∼1876' or '>1654' or 'between 1701 and 1709', causing obvious difficulties for date matching.

Whilst various methods may be used to identify duplicates during dataset extraction, it is important to determine to what extent the research question allows for refinement of the data, before a particular method is chosen. If, for example, it is acceptable to only select individuals with an accurate date of birth (DOB) then this makes the task of removing duplicates easier, because this is unique to more individuals than year of birth (YOB). But, crucially, a record linking method that requires date of birth accurate to the day cannot determine if there are duplicates among those records that do not have this accuracy of detail, because it is not possible to conduct duplicate matching on null values. Having to exclude all records without an accurate DOB can result in a much reduced final dataset size and, often with fewer older records—as these are less likely to have marriage, birth, death and burial records that are accurate to the day.

In order to produce meaningful historical demographic estimates, we often want to gather datasets which contain a large number of individuals, for whom there is relevant information in relation to the matter in which we are interested, and at the time points and locations that we are interested in. It is not a simple matter to define exactly what size of sample we need, because this depends on a number of factors, including the population size of the town or region, the time range, the demographic group and the historical context. If, for example, we are interested in the relation between life expectancy and marriage age, this can be affected by factors such as war and migration, but can nevertheless vary significantly between individual couples. It might be that for a small village over a decade in the seventeenth century, 30 marriages is a reasonable sample. However, this would clearly not be sufficient to estimate marriage age in a city for a decade of the twentieth century. It is essential to clearly define the topic, scope and research questions, before designing the method required to extract useful datasets from aggregated genealogical data.

Group-linking methods have been described for various other types of data, e.g. census records (Fu et al. 2011) and publication records (Bhattacharya and Getoor 2007). The idea is that you can better understand whether two separate records refer to the same entity if you understand the relationships with other entities. So, if we know that Joe Bloggs has co-authored a number of books with James Smith, this raises our confidence that Joe Bloggs is the co-author of a book written by ‘J. Bloggs and J. Smith’. Similarly, if we have a Jane Smith and a Janet Smith, both born in 1850 in the same place, it is possible they are the same person (despite the first name variation), but we have much more confidence that they are the same person if they both married a Peter Jones in 1873. This approach has some similarity with the data mining approach developed by Ivie et al. (2007), which makes use of family relationships to identify whether more than one pedigree belongs to a single individual.

Although first names are obviously a very important way to distinguish between individuals in real life, there are several good reasons for avoiding the use of first names for duplicate identification in genealogical data: first of all, they are often abbreviated or swapped over with middle names; second, first name fields often contain nicknames or titles, e.g. ‘Mr.’ or ‘Dr.’ and descriptives, such as ‘Grandmother’ or ‘Baby boy’.

I will introduce several methods that do not use first names to identify duplicates, however, in the absence of a date of death, burial or marriage, or information such as number and sex of offspring, first names are extremely useful for distinguishing between same-sex twins and other same-sex multiple births, because these have the same surname, same number and sex of siblings and same date of birth (usually). In order to alleviate some of the problems with first names—mainly abbreviation and spelling variations—a truncated first name is used in some of the methods described, in which only the first three letters of first names are used for duplicate matching. The truncated first name values were nulled where the first name field was matched against a list of titles and descriptives, such as Mr., Dr., Baby, etc.

To identify matching family structures, two different numerical strings were constructed, one based on the number and sex of all full siblings of an individual, e.g. ‘0103’ (one brother and three sisters) and another detailing the number and sex of children born in each marriage, e.g. ’0402’ (four sons and two daughters). The surname and truncated first name values were nulled if they contained non-alphabetic characters (except whitespace and hyphens), whilst whitespace was removed from non-null values.

6.3.2 Testing Duplicate Identification Methods

6.3.2.1 Single Test File

In the first set of tests of the duplicate identification methods, a single GEDCOM file containing no duplicates was used, for the very reason that without duplicates we can test how many unique records are identified by the different duplicate matching methods and see how much data coverage is provided by each method. It is a US-based genealogy with birth years ranging from 1625 to 1994. In summary, this file contains:

- 2802 individuals
- 971 marriages/partnerships
- 681 conceptive relationships (the same man and woman producing one or more children)
- 1913 children

Although we know that the file contains no duplicates, we see that different methods for selecting data will, to varying degrees, incorrectly identify duplicates and will have varying levels of data coverage. In a typical GEDCOM file, a method based on YOB will cover more of the data than one based on DOB, but will have a higher chance of incorrectly identifying duplicates, because there is less uniqueness in YOBs. A group-linking approach can be used to increase the power of either DOB or YOB to identify duplicates within a genealogical database by combining these variables with others, such as surname and year of death (YOD) into a string that can be used to compare between records to identify duplicates, whilst other variables such as date of marriage (DOM) or year of marriage (YOM) can be combined with variables such as surnames of spouses and number and sex of children, to compare between families. In this context, a family is defined as that formed by the union of two people (spouses) and the children that belong to both

Table 6.1 A comparison of the power of different duplicate matching strings to identify unique records in a single test file, using single variables in the matching strings

Duplicate matching strings	Unique records (*a*)	Data coverage (*b*)	Individuals/ marriages (*c*)	Accuracy [*a*/*b*]*100 (*d*)	Power [*a*/*c*]**d*
Individual matching					
1. DOB (date of birth)	996	1040	2802	95.77	34.04
2. YOB (year of birth)	46	1357	2802	3.39	0.06
3. Surname	391	2790	2802	14.01	1.96
4. Siblings (no. male and female)	0	2802	2802	0.00	0.00
Family matching					
5. Children (no. male and female)	14	971	971	1.44	0.02
6. DOM (date of marriage)	284	292	971	97.26	28.45
7. YOM (year of marriage)	72	332	971	21.69	1.61

those people, which means that half-siblings will be found in different families, because they do not share both parents.

In Table 6.1 and 6.2, the results of using a number of different duplicate matching strings to identify duplicates in the test file are compared. The combination of unique records identified, data coverage and total records in the file are taken into account to give an estimation of the power of each method to identify duplicates in genealogical data.

It is seen from these tests of various duplicate matching strings (Tables 6.1 and 6.2) that the most powerful for identifying unique individuals is the combination of YOB + sex + surname + siblings + fn[3] (17), whilst the most powerful for identifying unique marriages is YOM + surnames + children (25). The estimation of power takes into account the data coverage of the method, hence, the reason that YOB and YOM are more useful than DOB and DOM when they are combined with other variables. In terms of accuracy (i.e. the percentage of unique records identified among the records covered) the (17) string is also the most accurate at identifying unique individuals at 99.7 %. The DOM + surnames (20) string, DOM + surnames + children (24) and YOM + surnames + children (25) strings were all 100 % accurate.

The results of this test are indicative and provide a relative approximation of the power of each method to identify unique cases and not to give false positive matches. It should be clear that the effectiveness of each method will to some extent depend on the character of the database to which it is applied. In a larger database, the number of duplicates identified by most methods is likely to increase, because there will be more matching dates and more matching names, and this may reduce the accuracy of some of the methods. In databases containing older records, we may see fewer records with a DOB or DOM (rather than YOB or YOM) and therefore lower data coverage for methods employing these variables.

It is important to recognise that there is an issue of censoring to take into account with the duplicate matching strings that are based on family structure (the numeric 'siblings' and 'children' strings) because the records of siblings or children may be missing, as opposed to a particular individual or couple not having any siblings or children. This is quite different to a year of birth or surname, when an individual must have had one, so it is clear whether or not it is missing. In the case of siblings, 31.7 % of individuals had no siblings. In the case of marriages, 29.8 % had no children. In isolation, the use of these family structure strings is not very useful, because they have very low power (Table 6.1) and suffer from the problem that they contain an unknown quantity of censored records. However, when the strings are included with other variables, we see that they can raise the accuracy of the duplicate matching. The YOM + surnames + children (25) string, for example, has the same data coverage as the YOM + surnames (21) string, but a higher accuracy rate, giving it a higher matching power (Table 6.2).

Uncertainty about whether the absence of siblings and children relates to missing information or people who never existed suggests that the family structure strings should be used with some caution; however, the fact that inclusion of these strings with other variables raises the accuracy of the duplicate matching string suggests

Table 6.2 A comparison of the power of different duplicate matching strings to identify unique records in a single test file, using multiple variables in the matching strings; fn[3] is the variable which includes the first 3 letters of the first name

Duplicate matching strings	Unique records (*a*)	Data coverage (*b*)	Individuals/ marriages (*c*)	Accuracy [*a*/*b*]*100 (*d*)	Power [*a*/*c*]**d*
Individual matching					
8. DOB + surname	1011	1039	2802	97.31	35.11
9. YOB + surname	1072	1356	2802	79.06	30.25
10. DOB + siblings	1014	1040	2802	97.50	35.28
11. YOB + siblings	950	1357	2802	70.01	23.74
12. DOB + sex + surname + sibs	996	1018	2802	97.84	34.78
13. YOB + sex + surname + sibs	1279	1320	2802	96.89	44.23
14. DOB + sex + surname + fn[3]	1014	1018	2802	99.61	36.05
15. YOB + sex + surname + fn[3]	1308	1320	2802	99.09	46.26
16. DOB + sex + surname + sibs + fn[3]	1014	1018	2802	99.61	36.05
17. YOB + sex + surname + sibs + fn[3]	1316	1320	2802	99.70	46.82
18. DOB + sex + surname + sibs + DOD	277	279	2802	99.28	9.81
19. YOB + sex + surname + sibs + YOD	413	417	2802	99.04	14.60
Family matching					
20. DOM + surnames	291	291	971	100.00	29.97
21. YOM + surnames	329	331	971	99.40	33.68
22. DOM + children	290	292	971	99.32	29.66
23. YOM + children	272	332	971	81.93	22.95
24. DOM + surnames + children	291	291	971	100.00	29.97
25. YOM + surnames + children	331	331	971	100.00	34.09

that their inclusion is warranted. Moreover, in cases where you are only interested in individuals with siblings or those marriages that produced children, the family structure strings may be very useful. They may also prove very useful in databases with high homogeneity of surnames, because they provide an additional means of distinguishing between individuals and marriages.

6.3.2.2 Genealogical Database (Multiple Test Files)

In the following sections, I apply some of the duplicate matching strings introduced in the previous section to the previously mentioned genealogical database, which contains data from 921 GEDCOM files. A number of datasets were extracted, in which duplicates were identified using the individual-based duplicate matching strings or the family-based duplicate matching strings. The usefulness of the different methods is evaluated.

6.3.2.3 Duplicate Matching of Individuals

Four different duplicate matching strings based on individual attributes were compared (Table 6.3). It is seen that much larger datasets were extracted using the YOB and YOD variables. The highest percentage of unique records compared to data coverage is for the YOB + sex + surname + sibs + YOD (19) string at 80 %, compared to 79 % for (18), 78 % for (17) and 76 % for (16). The use of the DOD and YOD variable has the advantage that first names do not have to be used, whilst the strings still have some power to distinguish between twins (when the assumption is made that twins died on different days [DOD] or years [YOD]).

A significant advantage of using YOB and YOD as opposed to DOB and DOD is that the datasets extracted contain, on average, individuals with earlier dates of birth and death (Table 6.4). This demonstrates that these methods are better suited to historical research which is seeking to look further back in time.

Table 6.3 A comparison of different duplicate matching strings with a genealogical database containing data from 921 GEDCOM files. The dataset size is the number of unique records identified plus one randomly selected copy of each duplicate record identified

Duplicate matching strings	Unique records (*a*)	Data coverage (*b*)	Individuals (*c*)	Duplicates removed	Dataset size
16. DOB + sex + surname + sibs + fn[3]	155,803	204,548	545,711	26,725	177,823
17. YOB + sex + surname + sibs + fn[3]	254,194	325,022	545,711	38,477	286,545
18. DOB + sex + surname + sibs + DOD	52,061	65,578	545,711	7242	58,336
19. YOB + sex + surname + sibs + YOD	89,881	112,022	545,711	11,826	100,196

Table 6.4 Average YOB and YOD for datasets extracted by the different methods

Duplicate matching strings	Mean YOB	Mean YOD	Dataset size
16. DOB + sex + surname + sibs + fn[3]	1888	1909	177,823
17. YOB + sex + surname + sibs + fn[3]	1862	1889	286,545
18. DOB + sex + surname + sibs + DOD	1853	1911	58,336
19. YOB + sex + surname + sibs + YOD	1831	1889	100,196

Table 6.5 A comparison of different duplicate matching strings in a genealogical database with data from 921 GEDCOM files. The dataset size is the number of unique records identified plus randomly selected copies of each duplicate record identified

Duplicate matching strings	Unique records (*a*)	Data coverage (*b*)	Duplicates removed	Total marriages (*c*)	Dataset size
20. DOM + surnames	47,539	62,406	8177	189,433	54,229
21. YOM + surnames	60,468	80,854	11,145	189,433	69,709
24. DOM + surnames + children	48,386	62,406	7674	189,433	54,732
25. YOM + surnames + children	61,899	80,854	10,282	189,433	70,572

6.3.2.4 Duplicate Matching of Families

Four different duplicate matching strings based on family attributes were compared (Table 6.5). The issue of multiple marriages is relevant here, because one marriage may have a date and another not, or a surname may be missing for one of the marriages. To deal with this problem, all marriages were excluded where dataset extraction resulted in a second (or first or third, etc.) marriage being dropped. However, this is not a vital step, it depends on the criteria for data selection and how important it is that second marriages are, or are not, missed out.

The highest percentage of unique records compared to data coverage is for the DOM + surnames + children (24) string at 77 %, compared to 76 % for (20) and (25) and 74 % for (21).

6.3.2.5 How Reliable Is the Duplicate Matching?

To evaluate the effectiveness of the duplicate matching, a closer inspection was carried out of those records that were matched when using the YOM + surnames (21) string that were not matched when using the DOM + surnames (20) string, but after excluding those records that were not matched on both simply because only a YOM was available and not a DOM.

There are 322 records (158*2 and 2*3 duplicates) that fall into this category. It is seen that 124 of these duplicate /triplicate pairs have distinctly different first names and different DOMs, indicating clearly that they are different couples and not duplicates. There were 32 pairs with a different DOM who were the same couple (and four where it was not possible to clearly determine). The indication from this analysis is that false matches on the YOM + surnames (21) string are about 1.6 %, whilst false matches on the DOM + surnames (20) string are about a quarter of that, i.e. 0.4 %. The lower percentage of false matches with the DOM + surnames (20) string may be expected, because DOM is a more accurate measure than YOM, but the fact that there are also false matches with this string shows that typographical and transcription errors are made in the entry of accurate dates. In terms of the estimates of the effectiveness of each string (Table 6.2), this supports the finding that the DOM + surnames (20) string has higher accuracy, although it has a lower power than the YOM + surnames (21) string.

6.4 Discussion

There are three major steps in combining multiple genealogical data files to construct research datasets, which I have detailed here:

The first step is to exclude genealogies that have numerous or major errors, because such errors are indicative of poorly or carelessly conducted research. There is always the option to make the filtering process more stringent, for example by excluding all files that are not geocoded with latitude and longitude coordinates, or by excluding any file that contains illegal characters within the name fields; however, very few files may meet very strict selection criteria, and it is not necessarily correct to assume that this would improve the usefulness of the datasets that are ultimately derived from the database.

The second step is to clean the data, which is a time-consuming task for large databases, because accurate automated procedures for error correction and geocoding are simply not available. The aim of the TreeChecker project is to report potential errors to users who upload their family tree files, so that they can go back and make corrections or improvements, perhaps even going back to their sources to check them again. In this way, it is hoped we can raise the standard of genealogical data from the ground up. In parallel, efforts to develop more automated systems for error checking, e.g. algorithms that can identify date typo errors, should make the process of cleaning genealogical data less labour intensive.

The third step is to extract a dataset for analysis with duplicates excluded. It has been shown that there are a number of ways to do this, but there is not one correct way. It is important to have in mind the research questions and type of analysis that will be carried out using the dataset, because these issues will determine the way in which the dataset should be extracted. The issue of censoring (i.e. missing information) has to be addressed before deciding on which type of duplicate matching string to use, because duplicate matching cannot be done on missing information. If, for example, duplicate matching is done on DOB or YOB, this may tend to be more often missing for women, which introduces a selection bias into your extracted dataset.

A comparative method to estimate the accuracy and power of different duplicate matching strings was demonstrated. The accuracy estimates were based on the number of records that were identified as unique from a straightforward matching operation on a string, as a function of the proportion of the data that was available to the operation (i.e. the proportion of data where the information required to construct the string was not missing). The power estimates also included the total number of individuals or families in the database as an input to the function. The comparison does not provide a completely objective assessment of the ability to remove duplicates, because it cannot account for errors, such as incorrectly entered dates, and it cannot account for ‘doppelganger’ type cases, where two people have a high number of attributes in common through pure coincidence. However, it is proposed that the accuracy and power estimations provide a comprehensive way to compare the effectiveness of the different duplicate matching strings for addressing the task in hand, which is to automatically remove duplicates from large genealogical databases.

The comparison between different matching strings also gives an indication of how duplication identification accuracy has to be traded off against data coverage to build larger or older datasets, for example by using YOM instead of DOM. It is seen, however, that inclusion of family structure strings (which are based either on the number and sex of an individual’s full siblings or the number and sex of children born in a marriage) can increase the accuracy of the duplicate identification process when using YOM instead of DOM, or YOB instead of DOB, to give more data coverage without necessarily losing duplicate identification power.

There are many combinations of variables that may be used to construct a group-linking string that can then be used to extract a research dataset with duplicates excluded. It may be true that the principle aim must be to objectively identify duplicate individuals, but there is always likely to be a degree of error, because information about the same individuals can differ between files. It is equally important that data extraction takes into account the issues that will be addressed in the analysis of the dataset. If censored data is going to be included in the analysis, then this has to be taken into account with the duplicate matching string. For example, if you want to determine whether there is a difference in longevity between marriage and death for those individuals for which you have birth dates and those which you do not (i.e. the censored group), then the duplicate matching string used for dataset extraction needs to exclude birth dates, e.g. a string based on DOM or YOM must be used.

A method that has not been covered here is that of using an external data source as a reference to verify aspects of a genealogical dataset. In certain circumstances this can be a very useful step for eliminating outliers and improving confidence in the data. An example is a study of the familial circumstances of centenarians in the US (Gavrilova and Gavrilov 2007), in which birth and death dates of centenarians in a database compiled from GEDCOM files were validated against the Death Master File of the US Social Security Administration. However, such an approach will of course depend on the availability of a suitable external reference source against which to validate your data.

References

Bhattacharya, I., & Getoor, L. (2007). Query-time entity resolution. *Journal of Artificial Intelligence Research, 30*, 621–657.

Christen, P. (2012). *Data matching*. Berlin: Springer. doi:10.1007/978-3-642-31164-2

Fu, Z., Christen, P., & Boot, M. (2011). A supervised learning and group linking method for historical census household linkage. In *Proceedings of the Ninth Australasian Data Mining Conference* (Vol. 121, pp. 153–162). Australian Computer Society, Inc.

Gavrilov, L. A. & Gavrilova, N. S. (2001). Biodemographic Study of Familial Determinants of Human Longevity. *Population: An English Selection*, *13*(1), 197–221.

Gavrilova, N. S., & Gavrilov, L. A. (2007). Search for predictors of exceptional human longevity. *North American Actuarial Journal, 11*(1), 49–67. doi:10.1080/10920277.2007.10597437

Gellatly, C. (2009). Trends in population sex ratios may be explained by changes in the frequencies of polymorphic alleles of a sex ratio gene. *Evolutionary Biology, 36*(2), 190–200. doi:10.1007/s11692-008-9046-3

Ivie, S., Pixton, B., & Giraud-Carrier, C. (2007). Metric-based data mining model for genealogical record linkage. In *IRI 2007, IEEE international Conference on Infomation Reuse and Integration.*

Larmuseau, M. H. D., Van Geystelen, A., van Oven, M., & Decorte, R. (2013). Genetic genealogy comes of age: Perspectives on the use of deep-rooted pedigrees in human population genetics. *American Journal of Physical Anthropology, 150*(4), 505–511. doi:10.1002/ajpa.22233

Moreau, C., Bhérer, C., Vézina, H., Jomphe, M., Labuda, D., & Excoffier, L. (2011). Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science, 334*(6059), 1148–1150. doi:10.1126/science.1212880

Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records: Computers can be used to extract "follow-up" statistics of families from files of routine records. *Science, 130*(3381), 954–959. doi:10.1126/science.130.3381.954

Otterstrom, S. M., & Bunker, B. E. (2013). Genealogy, migration, and the intertwined geographies of personal pasts. *Annals of the Association of American Geographers, 103*(3), 544–569. doi:10.1080/00045608.2012.700607

Post, W., van Poppel, F., van Imhoff, E., & Kruse, E. (1997). Reconstructing the extended kin-network in the Netherlands with genealogical data: methods, problems, and results. *Population Studies, 51*(3), 263–278. doi:10.1080/0032472031000150046

United Nations. (1983). *Manual X: Indirect techniques for demographic estimation*. United Nations Publication.

Zhao, Z. (1994). Demographic conditions and multi-generation households in Chinese history. Results from genealogical research and microsimulation. *Population Studies, 48*(3), 413–425. doi:10.1080/0032472031000147946

Chapter 7
Multi-Source Entity Resolution for Genealogical Data

Julia Efremova, Bijan Ranjbar-Sahraei, Hossein Rahmani, Frans A. Oliehoek, Toon Calders, Karl Tuyls and Gerhard Weiss

Abstract In this chapter, we study the application of existing entity resolution (ER) techniques on a real-world multi-source genealogical dataset. Our goal is to identify all persons involved in various notary acts and link them to their birth, marriage, and death certificates. We analyze the influence of additional ER features, such as name popularity, geographical distance, and co-reference information on the overall ER performance. We study two prediction models: regression trees and logistic regression. In order to evaluate the performance of the applied algorithms and to obtain a training set for learning the models we developed an interactive interface for getting feedback from human experts. We perform an empirical evaluation on the manually annotated dataset in terms of precision, recall, and F-score. We show that using name popularity, geographical distance together with co-reference information helps to significantly improve ER results.

J. Efremova (✉) · T. Calders
Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: i.efremova@tue.nl

T. Calders
e-mail: toon.calders@ulb.ac.be

B. Ranjbar-Sahraei · H. Rahmani · G. Weiss
Maastricht University, Maastricht, The Netherlands
e-mail: b.ranjbarsahraei@maastrichtuniversity.nl

H. Rahmani
e-mail: h.rahmani@maastrichtuniversity.nl

G. Weiss
e-mail: gerhard.weiss@maastrichtuniversity.nl

F.A. Oliehoek
University of Amsterdam, Amsterdam, The Netherlands
e-mail: f.a.oliehoek@uva.nl

T. Calders
Université Libre de Bruxelles, Brussels, Belgium

F.A. Oliehoek · K. Tuyls
University of Liverpool, Liverpool, UK
e-mail: k.tuyls@liverpool.ac.uk

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_7

7.1 Introduction

The process of integrating disparate data sources for understanding possible identity matches has been studied extensively in literature and is known under many different names such as Record Linkage (Bhattacharya and Getoor 2004; Schraagen and Kosters 2014), the Merge/Purge problem (Hernández and Stolfo 1995), Duplicate Detection (Christen 2012; Naumann and Herschel 2010), Hardening Soft Databases (Cohen et al. 2000), Reference Matching (McCallum et al. 2000), Object identification (Christen 2012), and Entity Resolution (Efremova et al. 2014; Getoor and Machanavajjhala 2013).

Gradually, Entity Resolution (ER) has become the first step of data analysis in many application domains, such as digital libraries, medical research, and social networks. Recently, ER has found its way into the genealogical domain as well (Ivie et al. 2007; Rahmani et al. 2014). In this domain, a real person entity could be mentioned many times, for instance in civil certificates such as birth, marriage, and death or in notary acts such as property transfer records and tax declarations. Usually, no common entity identifiers are available, therefore the real entities have to be identified based on alternative information (e.g., name, place, and date). All information presented in the corpus is distributed over different sources such as civil certificates and notary acts. As an example, consider a person named *Theodor Werners* born in *Erp* on *August 11th, 1861*. He got married to *Maria van der Hagen* in *1888*. *Maria Eugenia Johanna Werners* was their child, born in *Erp* in *October 1894*. Two years after the child's birth, they bought a house in *Breda*. *Theodor* died in *Breda* on *September 1st, 1926*. In our corpus, this information is spread over respectively the birth record of *Theodor*, a marriage certificate of *Theodor* and *Maria*, the birth certificate of their child, a notary act available in full text, and the death certificate for *Theodor*. All these documents do not contain personal identifiers, may contain name variations, or be available in full text only. Applying ER to such a problem poses many challenges such as name alternatives, misspellings, missing data, and redundant information.

Genealogical data contains a huge amount of inaccurate information and different types of ambiguities, therefore applying proper ER techniques for cleaning and integrating the reference extracted from different historical resources, has received much attention. Sweet et al. (2007) use an enhanced graph, based on genealogical record linkage, in order to decrease the amount of human effort in data enrichment. Schraagen and Hoogeboom (2011) predict record linkage potential in a family reconstruction graph by using the graph topology. Lawson (2006) uses a probabilistic record linkage approach for improving performance of information retrieval in genealogical research. Recently, Bhattacharya and Getoor (2007) propose a collective entity resolution approach where they use the relational information about references and combine it with similarity between common attributes. Christen (2012) describes in depth a variety of data matching techniques from a statistical perspective. He addresses main challenges in the overall data matching process including data preprocessing, name variations, indexing, record comparison and classification. The key application of information retrieval is also addressed by the work of Nuanmeesri and Baitiang (2008), in which they discussed the design and development of suitable techniques that can improve efficiency of a Genealogical Information Searching System. Singla and Domingos (2006) propose an integrated solution to the entity resolution problem based on Markov logic that combines first-order logic and probabilistic graphical models by attaching weights to first-order formulas.

The mentioned work in Genealogical ER mainly focus on linking references with *homogeneous structures* where the number of descriptive features and their types are identical in all references. In this chapter, in contrast, we are interested in applying ER to a real-world dataset with a *heterogeneous structure* where different references come from qualitatively different sources and references no longer have similar descriptive features. We refer to this problem as *ER* on multi-source data.

In particular, we are interested in performing multi-source ER on a database of historical records of a Dutch province called North Brabant. There are two types of sources in this dataset: “Civil Certificates” and “Notary Acts”. The former type has a structured form and contains three certificate types birth, marriage, and death certificates while the other type contains free-text historical documents indicating involvement of references in different formal activities such as property transfers, loans, wills, etc. We give the detailed description of the input source types in Sect. 7.2. To integrate these types of sources we, first, identify all the references involved in a given set of notary acts and then link the extracted references to their birth, marriage, and death certificates. This process faces many challenges such as ambiguity due to name alternatives, misspellings, missing data, or redundant information.

The remainder of this chapter is structured as follows. In Sect. 7.2, we describe our real-world collection of historical data. In Sect. 7.3, we discuss the general ER approach and its implementation to our data. The reference extraction approach is described in Sect. 7.4. The indexing techniques that we use to generate potential candidate record pairs we describe in Sect. 7.5. In Sect. 7.6, we introduce informative attributes of references, describe a computation of attribute similarities and then present a final classification of reference pairs. In Sect. 7.7, we introduce the tools developed for historians to label data and we show the evaluation of the obtained results. In our analysis we study the influence of the individual steps on the overall precision and recall. Section 7.8 offers a discussion about drawbacks and potential extensions of the proposed approach. Concluding remarks are included in Sect. 7.9.

7.2 Data Description and Problem Formulation

The genealogical data used in this chapter is provided by the Brabants Historisch Informatie Centrum (BHIC).[1] The data consists of two main sources. The first source, civil certificates, is comprised of the birth, marriage, and death certificates belonging

[1] http://www.bhic.nl/, the website of BHIC is available in Dutch only.

Table 7.1 Available features for each certificate type. PoB and PoD stand for place of birth and place of death respectively

Birth certificate	FirstName, LastName, Gender, BirthDate, PoB, FatherFirstName, FatherLastName, MotherFirstName, MotherLastName
Death certificate	FirstName, LastName, Gender, BirthDate, PoB, DeathDate, PoD, FatherFirstName, FatherLastName, MotherFirstName, MotherLastName, PartnerFirstName, PartnerLastName
Marriage certificate	GroomFirstName, GroomLastName, GroomAge, BrideFirstName, BrideLastName, BrideAge, GroomFatherFirstName, GroomFatherLastName, GroomMotherFirstName, GroomMotherLastName, BrideFatherFirstName, BrideFatherLastName, BrideMotherFirstName, BrideMotherLastName

to North Brabant, a province of the Netherlands, in the period 1811–1940. The level of detail of each certificate varies very much. Table 7.1 lists the descriptive features for each certificate type. As shown in Table 7.1, birth certificates include three individual references (i.e., child, father, and mother). Death certificates include four individual references (i.e., deceased, father, mother, and partner of deceased). Finally, marriage certificates include six references (i.e., groom, bride, and parents of each). Each mentioning of a person in each certificate is called a *reference*.

This database consists of around 1,900,000 certificates with around 7,500,000 references in total. The exact number of documents and details about the distribution between the different certificate types are provided in Table 7.2. Volunteers digitize scans of the original manuscripts and make them available in a database format. At this moment, the digitization work is the most complete for marriage and death certificates and the database continuously grows.

A sample civil certificate is shown in Table 7.3. We see that the certificate has a pre-formatted structure. We illustrate the certificate as it is presented in the database. Notice that although this record is structured, there may be inconsistencies in the way the fields have been completed. For instance, the field *gender* is filled as *zoon van*[2] instead of explicitly mentioning being *male* or *female*.

The second source, the dataset of notary acts, consists of around 234,000 free-text documents of North Brabant before 1920. These free-text documents include information about involvement people in different formal activities such as property transfers, loans, wills, etc. Notary acts are in a free-text format and not all details are mentioned in a structured way. They require additional Natural Language Processing (NLP) techniques to extract information such as person names from the text. According to the type of formal activity, the detailed information mentioned in each notary act varies very much. For instance, an inheritance act many person names and many relationships are mentioned, whereas in the purchase agreements usually only one person name is mentioned.

Table 7.4 shows statistical information about the dataset of notary acts.

[2] 'zoon van' is the Dutch term for 'son of'.

Table 7.2 Statistical information of civil certificates

Type	Number of documents
Birth certificate	345,046
Marriage certificate	391,273
Death certificate	1,042,558
Number of references	7,557,051

Table 7.3 An example of civil certificate showing birth data

Person name	Teodoor werners
Gender	zoon van
Place of birth	Erp
Date of birth	14-04-1861
Father name	Peter Werners
Father profession	Shopkeeper
Mother name	Anna Meij
Mother profession	–
Certificate ID	6453
Certificate place	Erp
Certificate date	16-04-1861

Table 7.4 Statistical information of notary acts

Description	Number of acts
Number of acts	234,259
Number of act types	88
Number of notary acts of type *'property transfer'*	23,275
Number of notary acts of type *'sale'*	17,016
Number of notary acts of type *'inheritance'*	12,335
Number of notary acts of type *'public sale'*	10,593
Number of notary acts of type *'obligation'*	9006

An example of a notary act is shown in Table 7.5 (the person names are in bold face type). The notary act also contains a short summary and details provided by volunteers: the date and the place of a document.

To integrate these two heterogeneous types of input sources we, first, extract all the references from the civil certificate. Second, we identify all the references involved in a given set of notary acts. Finally, we link the references mentioned in each notary act to the references extracted from civil certificates. Our main goal is to find all birth, marriage, and death certificates for every person mentioned in a notary act. We formalize the ER problem as follow. Let $\mathcal{R} = \mathcal{R}_N \cup \mathcal{R}_C$ denote the total set of references, where $\mathcal{R}_N = \{r_{n_i}\}_{i=1}^{k}$ and $\mathcal{R}_C = \{r_{c_j}\}_{j=1}^{l}$ are the sets of references extracted from notary acts and civil certificates respectively. Each reference r_{n_i} and r_{c_j} has a value for each attribute in $\mathcal{A} = \{a_i\}_{i=1}^{m}$. We aim to find a set of real-world

Table 7.5 An example of a notary act

Theodor Werners, burgemeester van Boekel en Erp, wonend te Boekel bekent schuldig te zijn aan gemeente Erp Fl. 200,–. Waarborg: woonhuis, tuin, erf, bouw- en weiland Dinther en bouw- wei- en hooiland te Boekel. Zijn vader **Peeter Werners**… (*Theodor Werners, mayor of Boekel and Erp, living in Boekel, admits to owe the township of Erp 200 guilders. Security: house, garden, yard, farmland, and pasture Dinther and farmland, pasture, and meadowland in Boekel. His father Peeter Werners…*)	
Text ID	100
Place	Boekel
Date	24-07-1896

entities $\varepsilon = \{e_i\}_{i=1}^{m}$ such that $e_i \subseteq \mathcal{R}$. The set of entities can be represented as a partitioning of the references, in which each partition corresponds to the set of all references that belong to the same entity. Every reference can belong to only one entity: $r \in e_i \wedge r \in e_j \Rightarrow i = j$. Then the ER problem can be defined as: $\forall r_{n_i}, r_{c_j} \in \mathcal{R} : \exists e' \in \mathrm{ER}(\mathcal{R}) : r_{n_i} \in e' \wedge r_{c_j} \in e'$ and vice versa. The objective is to determine whether $r_{n_i}, r_{c_j} \in \mathcal{R}$ are the same entity e' in the real world.

7.3 Entity Resolution for Genealogical Data

To apply ER to the multi-source collection of historical data we use the following steps: *data collection and preparation*, *indexing*, *similarity computation*, *learning algorithm and classification* (Christen 2012; Naumann and Herschel 2010). We illustrate the overall ER process in Fig. 7.1.

The first step is data collection and preparation, during which the raw data is collected from various sources, then cleaned and preprocessed. During this step we have to assure that all references have the same format (standardized date, null values, special characters, etc.) and extract all person references from civil certificates and notary acts. As discussed in Sect. 7.2, reference extraction from civil certificates requires data cleaning and standardization of null values. The notary acts, however, require more complicated preprocessing techniques to extract person names and other information from them. Dealing with the notary acts we use the natural language processing techniques and named entity recognition approaches (Chowdhury 2003; Nadeau and Sekine 2007) which we discuss in Sect. 7.4.

The second step of the ER process is data indexing and generation of candidate record pairs for further comparison. In order to avoid having to compare every reference in one source with every reference in another source, we split the

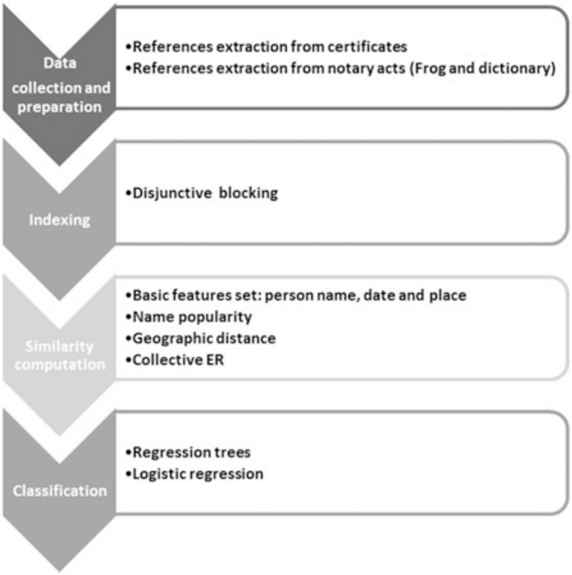

Data collection and preparation
•References extraction from certificates
•References extraction from notary acts (Frog and dictionary)
Indexing
•Disjunctive blocking
Similarity computation
•Basic features set: person name, date and place
•Name popularity
•Geographic distance
•Collective ER
Classification
•Regression trees
•Logistic regression

Fig. 7.1 The overall ER process

references into different partitions using an indexing technique. This partitioning allows us to reduce computational complexity by reducing the number of candidate record pairs. We discuss the applied indexing algorithm in Sect. 7.5.

The next step is the similarity computation step. The similarity score between two attributes, associated with two distinct references, is computed based on their types. We compare two attributes with type *String* using the *hybrid string similarity measure* described in (Efremova et al. 2014) and trained on the dataset of Dutch names from Meertens Instituut.[3] The hybrid measure combines using logistic regression (Sammut and Webb 2010) five string similarity functions: *Soundex* (SN), *Double Metaphone*, (DM), *IBMAlphaCode* (IA), *Levenshtein distance*, *Smith Waterman distance* (Elmagarmid et al. 2007; Naumann and Herschel 2010; Winkler 1995).

For attributes of type *Date* we calculate the similarity as the date difference in years. For every pair of references we compare essential attributes using an appropriate similarity measure. We discuss more about the features and methods for the similarity computation in Sect. 7.6.

The last step of the overall ER process is learning a model and classification. The score function computes the final similarity score between candidate record pairs using a supervised classification. Then pairs of references are classified into classes *Matched* or *non-Matched*, based on a threshold value of the score function. The classification step is described in Sect. 7.6.5.

[3] http://www.meertens.knaw.nl/nvb/

7.4 Data Collection and Preparation

We pre-process the data in order to extract references and other information from various sources. Civil certificates require a cleaning phase. There are many situations when person name in civil certificates is unknown. It happens, for instance, when child died at birth and did not receive a name. Then in birth and death certificates his name may be filled as: *onbekend,*[4] *niet vermeld,*[5] etc. We replace the common terms by *null* values.

Other fields in civil certificates also require standardization. We generalize the date of a document to its year, because it is specified in different formats: as a year, as an exact date or as text, for instance *[1861] Augustus*. We use regular expressions and consider the first four digits in the date field as the year of the document. The gender of a person is not always clearly mentioned. Instead of the direct specification of male or female, the gender may be given in a textual format, for instance using terms such as: *zoon van*, *zoontje van*,[6] etc. We standardize those values to an appropriate format.

After the cleaning phase civil certificates are ready for the reference extraction. Table 7.6 shows three sample references which are extracted from the civil certificate of Table 7.3.

To extract references from notary acts we apply the NLP tool Frog (Van den Bosch et al. 2007) which is a Dutch morpho-syntactic analyzer and dependency parser. The Frog tool extracts most of the names from notary acts, although some names are missed. To check the recall of name extraction we manually extracted names from randomly selected notary acts and compared to Frog results. Frog failed to identify 41 out of 166 manually extracted names. These missed references have a huge influence on the overall performance of our ER task, as there is no way to compensate for these missed references later on in the chain. Therefore, in addition to Frog name extraction we designed our own special-purpose name NLP rules.

The process of gathering the names from the notary acts hence proceeds in two steps. In the first step, **preprocessing**, some basic text polishing algorithms are run on the text, such as removing extra spaces and wrongly encoded symbols. Also punctuation is detected and checked, mistakes are corrected or reported for a manual inspection. In the second stage, which is **word labeling**, punctuation, the position of a word in the sentence, and dictionary information extracted from the structured data are used to label the words in the text as *person name*, *person name prefix*, *location name*, *location prefix*, *number*, *relation indicator*, *conjunction*, and *preposition*. This approach is iterative: for instance detecting a *location prefix* such as “te” (variant of “in” in Dutch used with locations) helps recognizing a following name as a location. Having labeled the words in the text, in the third stage, named **person name resolution**, the person references are extracted by considering every connected set of words labeled as “person name” and possibly a person name prefix

[4] ‘onbekend’ is the Dutch term for ‘unknown’.

[5] ‘niet vermeld’ is the Dutch term for ‘not mentioned’.

[6] ‘zoon van’ and ‘zoontje van’ are Dutch terms for ‘son of’.

Table 7.6 The references extracted from the sample civil certificate in Table 7.3

Ref_ID	Person name	Place	Date	Cert_ID
124358	Teodoor Werners	Erp	14-04-1861	6453
124359	Peter Werners	–	–	6453
124360	Anna Meij	–	–	6453

Table 7.7 Statistical information of reference extraction notary acts

Total number of extracted references	1,155,400
Minimum number of references in a notary act	1
Maximum number of references in a notary act	214
Average number of references in a notary act	5.7

Table 7.8 The references extracted from the sample notary act in Table 7.5

ref_ID	Person name	Place	Date	Text ID
94254	Theodor Werners	Boekel	24-07-1896	100
94255	Peeter Werners	Boekel	24-07-1896	100

in word sequence. Location entities usually follow a location prefix and contain a set of location names. The total number of extracted references and the general statistics about name extraction from notary acts is presented in Table 7.7.

As we see from Table 7.7 every notary act contains at least one reference. However, the number of person references per document varies a lot from only 1 to 214 references per document.

Returning to our example, using the NLP techniques described above, a sample person reference extracted form the notary act of Table 7.5 is shown in Table 7.8. The date and the place of the document are available in a short human-annotated summary of a notary acts and do not require an NLP extraction.

The data extracted from a notary act has only few features as compared to the structured data shown in Table 7.3.

7.5 Candidate Generation

It is computationally very expensive to compare every reference extracted from a notary act with every reference occurring in the civil certificates. Therefore, we use *indexing* to reduce the total number of potential candidate pairs, as this would require comparing $|\mathcal{R}_N| \times |\mathcal{R}_C|$ pairs. We do not compare every reference from a notary act with every reference from a certificate, but instead divide the references into buckets based on some basic characteristics, such as for instance the first four letters of the last name. Only references that fall into the same bucket will be compared. Obviously, the smaller the buckets, the faster we will be able to carry out all comparisons, but on the other hand, we may lose some pairs of references that refer to the same entity, because they accidentally get assigned to different buckets. To reduce this risk, we need to carefully select the characteristics on which we will decide the division into buckets, in order to optimize this trade-off.

In this work, we apply an adaptive blocking algorithm proposed by Bilenko et al. in (Bilenko 2006) which is based on learning an optimal set of disjunctions of blocking functions based on a labeled training set. To construct the set of predicates we use heads and tails of phonetic functions with variable size: 2, 3 or 4 characters for the heads, and 3 or 4 characters for the tails. There is a variety of phonetic functions. They use several rules to transform a name to a phonetic encoding. Some algorithms ignore all vowels and group the consonants, other algorithms analyze consonant combinations. For the experiments in this chapter, we construct an *indexing* using specified head and tales of the four phonetic functions: *Soundex*, *Double Metaphone*, *IBMAlphaCode,* and *New York State Identification and Intelligence System* (Christen 2006; Efremova et al. 2014. Table 7.9 shows an example of applied phonetic to encode imprecise names.

We analyze the performance of phonetic keys on the dataset of Dutch names as is described in Efremova et al. (2014). To index our data we use disjunctions of the following: $Head(Soundex, length = 4)$, $Head(DM, length = 4)$, $Head(NY, length = 4)$, $Tail(IA, length = 4)$. We apply the resulting formula to index first and last names in historical documents. That is, two references r_{n_i} and r_{c_j} will be compared if and only if they agree on at least one of these functions, hence the name “disjunctive blocking”. In this way, we can significantly reduce the number of candidates to be checked without losing too many true matches. Using different disjunctions of phonetic predicates helps us to reduce the number of candidate pairs to compare, however, some name variations can still occur in different partitions. In Sect. 7.7, we show that maximum achieved recall is above 92 % which is relatively high. The missed 8 % is partly because of the selected indexing approach and partly because of the name extraction phase. The first four letters of phonetic keys (e.g., first four letter of Soundex) are commonly used in literature for indexing purposes (Christen 2012). It is possible to use a less restrictive indexing strategy: only first

Table 7.9 An example of phonetic keys

Name	SN	DM	IA	NYSIIS
Theodoor	T600	TTR	0114	TADAR
Theodor	T600	TTR	0114	TADAR
Theodorus	T620	TTRS	0114	TADAR

Table 7.10 The coefficients of the logistic regression

(Intercept)	Name similarity	Place	Date	Name popularity	Geographical distance	Co-reference
−6.39	6.30	0.93	−0.01	−21.11	−1.45	2.93

letter of person names. However, this leads to a significant increase in the number of potential candidate pairs.

7.6 Feature Similarity Computation and Classification

One of the main challenges in multi-source ER is the lack of available information. It is virtually impossible to decide whether or not the person mentioned in a notary act is the same person as a person in a specific civil certificate, if there are more than 1000 other civil certificates that belong to persons with the same name in the same time period. For instance, it is much easier to find civil certificates that belong to *Bernardus Wijngaarden* whose name appears only few times in historical documents, than to find civil certificates that belong to *Theodor Werners* whose name appears much more often in the database. Therefore, in this section we, first, describe in detail informative features that we use to compare references, then we show how to compute a similarity for every feature and to classify a reference pair into *Matched* and *Non-Matched*.

7.6.1 ER Basic Features

We define a basic feature set $\mathcal{F} = \{f_1, \ldots, f_n\}$, which are used to compare pairs of potential candidate matches. These features can be obtained directly from one notary act and one civil certificate and do not require additional information. To construct a basic feature set $\mathcal{F}$ we use person *FullName*, *Date* (in years), and *Place*. Those attributes can all be extracted from a notary act. We use NLP techniques to extract person names as described in Sect. 7.4. Date and place of the document are specified by volunteers in a summary of the notary act. We compare the attribute *FullName* by a hybrid string similarity function (Efremova et al. 2014), the similarity between dates as the difference in years and the similarity between places as a Boolean value which is *true* when the two places in the pair of references have exactly the same name, and otherwise *false*. During the next step we extend the basic set of features and experiments by introducing additional attributes.

7.6.2 Considering Name Popularity

The person name is an important attribute in genealogical ER, however it is more difficult to find a certain match for a very common name than for an uncommon one. We think that the uncertainty caused by popular names is inevitable, and therefore aim at designing an algorithm to consider this important feature.

To compute the popularity of each name in the database, we make a list of full names using information from death certificates. We use only death certificates because they are more prevalent than the other types of certificates (i.e., birth and marriage). Under

FullName we consider the combination of *FirstName* and *LastName* of each person. We did not consider documents where first or last name were not filled. In the next step, for every full name we estimate its popularity as the fraction of name occurrence in death registers to the total number of death registers. In this way, we assign the lowest score to uncommon names and the highest score to the most popular ones.

We assign a name popularity value to a full name of a reference in a civil certificate. We do not compute name popularity of references extracted from notary acts because the name extraction using NER techniques is not always accurate. For instance, the name can be extracted with an extra symbol or an extra word like *'Theodor Werners'* or *'Theodor Werners te Erp'* which will not appear in the list. In the first case the name is extracted with an extra dot at the end and in the second case with the location prefix *te Erp*. If the name does not exist in the list, we assign popularity value 0. We extend the basic feature set $\mathcal{F}$ by adding name popularity as an extra feature: $\mathcal{F} \leftarrow \mathcal{F} \cup \{f_{\text{popular}}\}$.

We explore manual matches by humans on a manually annotated dataset described in Sect. 7.6.5. We are interested to see how often a match is assigned to popular names. Figure 7.2a shows the occurrence of every name matched manually in the overall collection of civil certificates. The highest values on the diagram belong to names such as: *Maria Janssen* (occurs 1,242 times in civil certificates), *Martinus Heijden* (962 times), *Johanna Martens* (900 times). It means that humans during the manual annotation identified only few matches that belong to very common names and most of manually annotated matches belong to relatively uncommon names. Name popularity information helps to improve the ER results compared to the basic set of features as discussed in Sect. 7.7.

7.6.3 Considering Geographical Distance

Although the historical documents belong only to North Brabant, which is relatively small, it is more likely to find a match between people from the same place than from different places in North Brabant that are farther apart. Therefore, we consider the geographical distance. We define the following three main groups based on geographical distances.

- intra city distance (from 0 to 5 km)
- inter villages distance (from 5 to 20 km)
- inter cities distance (more than 20 km)

For each place mentioned in the documents we define a spatial component: longitude and latitude (α, δ). We use the database of places provided by The Historical Sample of the Netherlands (HSN).[7] This database contains 7925 names of places in the Netherlands and their geographical coordinates. More details about

[7] http://www.iisg.nl/hsn/data/place-names.html

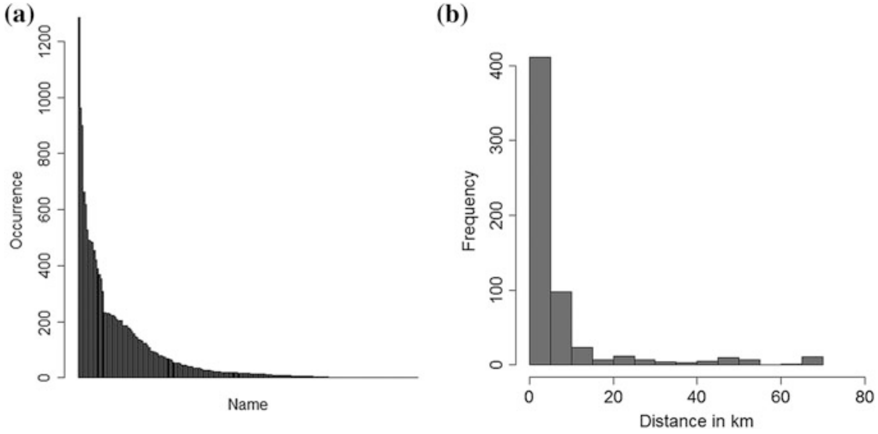


Fig. 7.2 Distributions of manual matched references. **a** Occurrence of person names that were manually matched. **b** Geographic distance between pairs of manually labeled references in km

the database of places can be found in Huijsmans (2013). Another way to retrieve geographical coordinates is to use the Google Geocoding API[8] with geo lookup functionality. However, the tool often confuses places that existed in the past with recent different more recent locations that have the same name. We calculate the geographical distance in kilometers for each pair of potential matches using the coordinates of the two places: (α_1, δ_1) and (α_2, δ_2) using Eq. 7.1 obtained from (Ramachandran et al. 2005):

$$\text{distance} = 2\mathcal{R} \cdot \arctan\left(\frac{\sqrt{\text{hav}(\theta)}}{\sqrt{1 - \text{hav}(\theta)}}\right), \tag{7.1}$$

where $\text{hav}(\theta) = \sin^2\left(\frac{\delta_1 - \delta_2}{2}\right) + \cos(\delta_1) \cdot \cos(\delta_2) \cdot \sin^2\left(\frac{\alpha_1 - \alpha_2}{2}\right)$ and $\mathcal{R}$ is the Earth radius.

We compute the geographic distance between two references for every candidate pair. To analyze how often humans are able to find a match between references from different places we made a distribution of geographical distances between two references in the manually annotated dataset as presented in Fig. 7.2b. Human annotators mainly find links between references that are from places not far apart. However, there are some references that were identified where the distance was up to 80 km within North Brabant. On the next step we convert geographic distances to defined groups and add this feature to the feature set: $\mathcal{F} \leftarrow \mathcal{F} \cup \{f_{\text{migration}}\}$. Adding the geographical distance helps slightly to improve results as it described in Sect. 7.7.

[8] https://developers.google.com/maps/documentation/geocoding/

7.6.4 Collective ER with Co-Occurrences of References

We carry out experiments with collective entity resolution (Bhattacharya and Getoor 2007; Štajner and Mladenić 2009) and take into account entity co-occurrence across the documents. All references within the same document are related to each other by a co-occurrence relationship. The co-occurrence relationship of references is widely used in ER and information retrieval. The idea behind it is that if entities often occur together, they are probably related to each other. We deal with the co-occurrence information by treating it as an additional feature for ER. For each pair of references (r_{n_i}, r_{c_j}) we construct the neighborhood sets $\mathrm{Nbr}(r_{n_i})$ and $\mathrm{Nbr}(r_{c_j})$ which include all co-occurred references of r_{n_i} and r_{c_j}, respectively. We perform pairwise *FullName* comparisons between all possible pairs of co-occurred references generated from $\mathrm{Nbr}(r_{n_i})$ and $\mathrm{Nbr}(r_{c_j})$.

Returning to our example, the neighborhood of the reference *'Theodor Werners'* extracted from a notary act contains one name $\mathrm{Nbr}(r_{n_i})$ = (Peeter Werners) and the neighborhood of the reference *'Teodoor Werners'* extracted from a civil certificate has two names $\mathrm{Nbr}(r_{c_j})$ = (Peter Werners, Anna Meij). We see that two neighborhoods have one similar name in common which has to be taken into account during the comparison of the references.

To compare *FullName* attributes of co-occurred references we use again the hybrid string similarity function described in Sect. 7.3. Then we assign the final similarity score as the highest similarity score between all possible pairwise comparisons.

Considering only the highest similarity score between $\mathrm{Nbr}(r_{n_i})$ and $\mathrm{Nbr}(r_{c_j})$ makes an algorithm to disregard that compared references may have more than one co-reference. However, finding at least one co-reference already helps us to improve the results significantly compared to the previous set of features as discussed in Sect. 7.7. Algorithm 1 demonstrates this approach.

Algorithm 1 Computation of reference co-occurrence

```
Input: A pair of references $(r_{n_i}, r_{c_j})$, a set of co-references $Nbr(r_{n_i})$ to $r_{n_i}$, a set of co-references $Nbr(r_{c_j})$ to $r_{c_j}$
Output: Computed co-occurrence information $f_{collective}(r_{n_i}, r_{c_j})$
$\mathcal{C} \leftarrow \emptyset$
for each co-reference $m$ in $Nbr(r_{n_i})$ do
    for each co-reference $n$ in $Nbr(r_{c_j})$ do
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{ComputeSim(m, n)\}$
    end for
end for
$f_{collective}(r_{n_i}, r_{c_j}) \leftarrow max(\mathcal{C})$
return $f_{collective}(r_{n_i}, r_{c_j})$
```

We add a collective feature based on the co-occurrence of references to the feature set: $\mathcal{F} \leftarrow \mathcal{F} \cup \{f_{\text{collective}}\}$.

7.6.5 *Classification*

The last step of the overall ER process is classification. Earlier in this section, we described informative attributes of reference pairs and appropriate attribute similarity metrics to compare them. However, to compute the overall similarity score of every reference pair we need to assign an appropriate weight to each attribute. This approach allows to estimate the final probability of each match using a score function. The score function computes the final similarity score between two references based on the results of single attribute comparisons. We learn the score function on a training dataset that we will discuss in detail in Sect. 7.7. After that, pairs of references are classified into *Matched* or *non-Matched* based on a threshold value.

There exists a variety of techniques from statistics, modeling, machine learning and data mining (Christen 2008; Florian et al. 2003) for designing a score function that combines individual similarity scores. We apply two predictive models. First, we use logistic regression (Sammut and Webb 2010) and calculate the score function as follows:

$$\mathrm{Score}(r_{n_i}, r_{c_j}) = \frac{1}{1 + \mathrm{e}^{w_0 + \sum_{l=1}^{k} w_l \cdot \mathrm{sim}\left(r_{n_i}.a_l, r_{c_j}.a_l\right)}} \tag{7.2}$$

where parameters w_0 to w_k are learned in the training phase.

The function $\mathrm{sim}(r_{n_i}.a_l, r_{c_j}.a_l)$ represents similarity measures of the attribute a_l between two arbitrary references r_{n_i} and r_{c_j}, while reference r_{n_i} and r_{c_j} have k attributes in common.

Additionally, we apply *Regression Trees* (Sammut and Webb 2010). The leaves of a tree represent class labels (*Matched* or *non-Matched*) whereas its nodes represent conjunctions of the features values.

7.7 Experiments and Results

The application of the multi-source ER approach and its evaluation on real-world data requires additional steps. The first step is the process of gathering expert opinions. This is a crucial requirement for the evaluation. Therefore, in this section first we present an interactive web-based interface which is used for getting input from humans. Then we elaborate on the application and the evaluation of the model.

We have two sets of experiments. **Experiment 1** is to obtain the performance results of ER algorithms on the manually annotated dataset. After the first experiment we select all *false-positive* (FP) matches that correspond to the maximal *F-score* value in order to evaluate to what extent they are really incorrect links or rather concern omissions in the human labeling. Given the extraneous nature of the labeling tasks it is indeed conceivable that human annotators may have missed a significant part of the links. Hence, in **Experiment 2** we evaluate new precision value after a manual review of false positive matches according to the prediction. In order to assess the performance of our results we apply the 10-fold cross-validation method on the entire ER approach.

7.7.1 Manual Labeling Phase

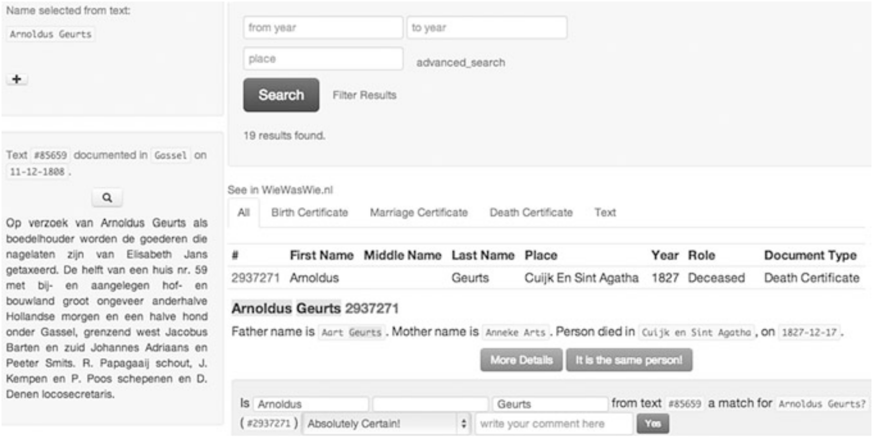
In order to generate adequate training/test set for the classification process, a web-based interactive tool was developed (Efremova et al. 2013) which allows historians to navigate through the structured and unstructured data, and label the matches they find between various references. This tool uses various programming tools for storage, exploration, and refinement of available data; it benefits from an intelligent searching engine, developed based on the Solr[9] enterprise search platform, with which historians can easily search through the dataset. Basically, the required data can be found via person name, location, date, and relationship types.

The developed Labeling tool, shown in Fig. 7.3, is very powerful and easy to use, which assists historians to link name-references mentioned in notary acts to name-references mentioned in civil certificates.

The time required to report a correct match between two name-references varies from a few seconds to probably hours of time, depending on how similar two references are (e.g., whether places, dates, ages, professions, and relatives match or not), and how easy it is to compare those two references. Consequently, the level of confidence in reporting a match varies. Therefore, the actions that historians take (e.g., which keywords they take and how fast they can recognize a match), and their level of confidence in reporting the match are all stored in the database. As a result, a rich benchmark is generated that includes the list of matches, the level of confidence, and the list of actions that historians search for before reporting the match.

We consider each pair of references labeled by a historian as an example of a positive match between two different sources of data. Due to insufficient information in a notary act, incomplete civil certificates or a very frequent person name, no matches might be found for some references. We assign a zero-matched status to such references. Using the developed tool we manually annotated 643 entity resolution decisions (matches between notary acts and civil certificates) from 82 randomly selected notary acts.

[9] http://lucene.apache.org/solr/



Name selected from text:

Arnoldus Geurts

Text #85659 documented in Gossel on 11-12-1808 .

Op verzoek van Arnoldus Geurts als boedelhouder worden de goederen die nagelaten zijn van Elisabeth Jans getaxeerd. De helft van een huis nr. 59 met bij- en aangelegen hof- en bouwland groot ongeveer anderhalve Hollandse morgen en een halve hond onder Gassel, grenzend west Jacobus Barten en zuid Johannes Adriaans en Peeter Smits. R. Papagaaij schout, J. Kempen en P. Poos schepenen en D. Denen locosecretaris.

from year | to year

place | advanced_search

19 results found.

See in WieWasWie.nl

#	First Name	Middle Name	Last Name	Place	Year	Role	Document Type
2937271	Arnoldus		Geurts	Cuijk En Sint Agatha	1827	Deceased	Death Certificate

Arnoldus Geurts 2937271

Father name is Aart Geurts . Mother name is Anneke Arts . Person died in Cuijk en Sint Agatha , on 1827-12-17 .

Is Arnoldus | | Geurts from text #85659 a match for Arnoldus Geurts?

(#2937271) Absolutely Certain! | write your comment here

Fig. 7.3 The developed web-based labeling tool for generating the required training/test dataset

7.7.2 Experiment 1: ER Before Manual Match Review

We evaluate the performance of the applied algorithms using the standard metrics precision, recall, and F-score. We compute the sets of True Positives (TP), False Positives (FP), and False Negatives (FN) as the correctly identified, incorrectly identified, and incorrectly rejected matches, respectively. In Fig. 7.4a, we show the achieved precision and recall values for different sets of features and for the two prediction modes: regression trees (RT) and logistic regression (LR). Figure 7.4b presents the evaluation of results in terms of F-score and threshold values. Table 7.11 shows the maximum F-score value and corresponding precision and recall.

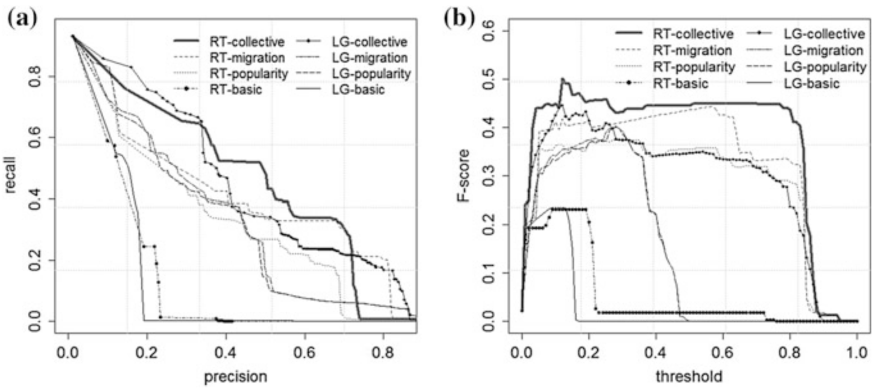


Fig. 7.4 Evaluation of ER quality using different feature sets. The label names: *basic*, *popularity*, *migration* and *collective* correspond to the respective set of features described in Sect. 7.6. **a** Precision versus recall. **b** F-score versus threshold

Table 7.11 The maximum F-score with corresponding precision and recall of different feature sets

Features	Logistic regression			Regression trees		
	Precision	Recall	Max F	Precision	Recall	Max F
Basic	0.161	0.445	0.236	0.218	0.246	0.231
Basic, name popularity	0.430	0.374	0.400	0.448	0.320	0.374
Basic, name popularity, geo distance	0.434	0.375	0.402	0.679	0.330	0.444
Basic, name popularity, geo distance, co-occur	0.338	0.653	0.445	0.486	0.518	**0.502**

As Fig. 7.4 and Table 7.11 show, the results improve significantly by adding the additional information. The basic set of features is clearly not sufficient for obtaining an appropriate performance level. Adding name popularity to the basic set of features almost doubles the maximum F-score. This can be explained as follows: it is a very difficult task to be certain in assigning a proper match among a huge amount of similar documents that belong to persons with the same name, so the final decision requires additional information and the overall score for matches of documents with popular names should be lowered. Adding a geographical distance to the feature set yields also a minor improvement (7.0 % for the RT and 0.2 % for LR). The last analyzed feature which improves the results significantly, is co-occurrence information. It increases the max F-score by 5.8 and 4.3 % for RT and LR respectively. To understand which features are more important we show the coefficients of the logistic regression in Table 7.10. These coefficients are applied to calculate the final similarity score using the function described in Eq. 7.2.

Overall, we compared the results of the two applied regression models. The highest F-score that we achieved is 0.502 by using the RT.

To show a computational complexity of applied overall ER approach we analyze a number of comparisons (candidate pairs) for every achieved level of precision and recall. The results are presented in Fig. 7.5 separately for LR and RT predictive models. The $\mathcal{Y}$ axes on the graph shows the total number of candidate pairs that need to be compared after applying an indexing technique described in Sect. 7.5. We see that to identify 643 manually annotated matches for references extracted from notary acts within a large collection of civil certificates we analyze more than 54,000 candidate pairs. This is much less than comparing each reference from notary act with every reference from civil certificates but this is still much larger than a number of true positive matches. The applied indexing strategy is not restrictive and generates among the true-positive matches a lot of extra candidate pairs to compare. Considering such a large amount of pairs we have recall value above 92 %. Then we use robust classifiers and the extended feature set that leads to promising results in distinguishing *Matched* pairs from a large amount of *Non-Matched* ones.

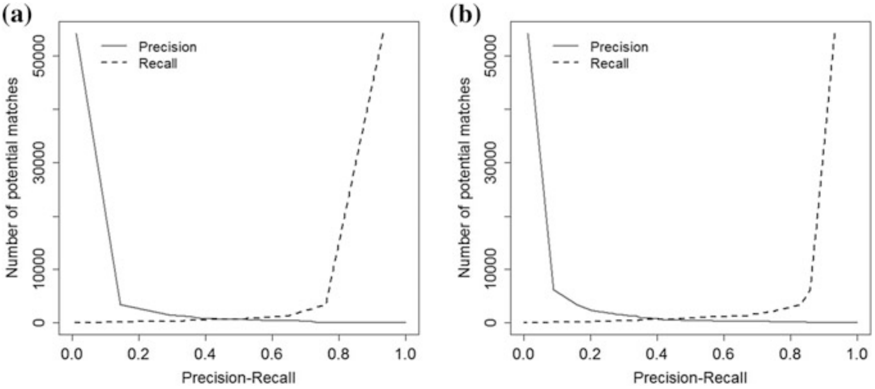


Fig. 7.5 Distributions of the number of potential candidate matches, and corresponding precision/recall values for two applied predictive models. **a** Using regression tree. **b** Using logistic regression

7.7.3 *Experiment 2: ER After Manual Match Review*

In this second experiment, we present the increase in precision after the manual cross-check of false positive matches which corresponds to the situation with the maximum *F-score* in Experiment 1. Experts manually review matches from the false positives, generated by the two prediction models (LR and RT). Table 7.12 presents recalculated precision results for each set of features using the logistic regression. We show the previous recall and optimal *F-score* values from Experiment 1 and compare two corresponding precision values: before and after manual matches review. The table shows that the initial accuracy has been greatly underestimated. After an additional review of matches that are positive according to the classifier, volunteers found that they missed 89 matches during the initial data annotation. To avoid boosting the recall artificially, we do not run a full set of experiments similar to the experiments described in Sect. 7.7.2. The cross-check of the false positive matches affects only the precision. Matches which were incorrectly rejected can not be identified during the manual review of the FN set. As we see from Table 7.12, for each set of features the precision is underestimated by 7 % on average.

Table 7.12 The improved precision in the Experiment 2 using the *Logistic Regression*

Features	$\max F_{\mathrm{exp1}}$	$\mathrm{Prec}_{\mathrm{exp1}}$	$\mathrm{Prec}_{\mathrm{exp2}}$	Δ_{prec}
Basic	0.236	0.161	0.218	0.075
Basic, name popularity	0.400	0.430	0.498	0.068
Basic, name popularity, geo distance	0.402	0.434	0.501	0.067
Basic, name popularity, geo distance, co-occur	0.445	0.338	0.413	0.075

Table 7.13 The improved precision in the Experiment 2 using the *Regression Tree*

Features	maxF_{exp1}	Prec$_{\text{exp1}}$	Prec$_{\text{exp2}}$	Δ_{prec}
Basic	0.231	0.218	0.289	0.071
Basic, name popularity	0.374	0.448	0.520	0.072
Basic, name popularity, geo distance	0.444	0.679	0.760	0.081
Basic, name popularity, geo distance, co-occur	0.502	0.486	0.626	**0.140**

Table 7.13 presents the results obtained using the Regression Trees. The precision is maximally improved by 14 %. The largest improvement corresponds to the extended set of features which includes the basic features and additional features such as name popularity, migration information, and reference co-occurrence. We see from the table that for each feature set the precision is increased after the manual review of FP matches. An additional review of the FP matches improves the precision evaluation. Nevertheless, the estimation of the precision value is very important for genealogical and population research. Therefore, we emphasize the precision calculation in this experiment.

7.7.4 Alternative Analysis

Since it is very hard to get the ground truth for our dataset we also run some alternative validations based on common sense. For instance, independently if we know which match is correct or not, when a person is matched to two birth certificates, one of the matches has to be wrong. In this subsection, we make an effort to evaluate our results using such common sense arguments. In Fig. 7.6 we show a detailed comparative analysis of the number of matches identified by humans and

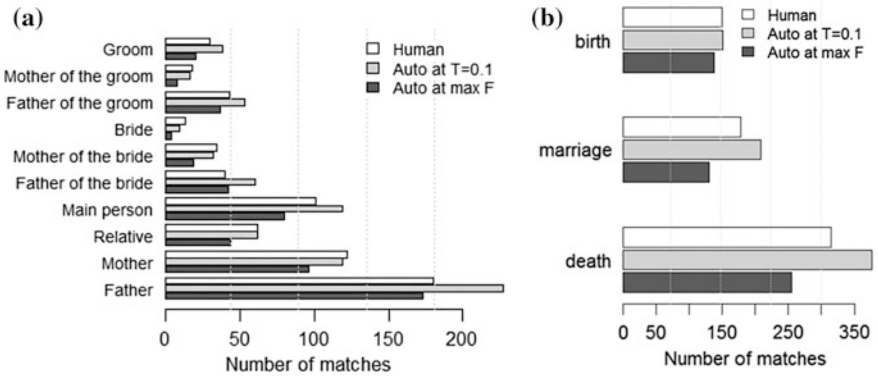


Fig. 7.6 The comparison of number of matches according to humans and automatic approaches for two threshold levels of RT score function. **a** For each person role in certificates. **b** For each certificate type

by the RT with the extended feature set (the best studied automatic ER method) with two selected threshold levels of score function at max F-Score and at the threshold level $T = 0.1$. We compare the number of matches for each type of certificate: birth, marriage, and death and for different role of people mentioned there. We see that the maximum number of matches are identified for death certificates by humans as well as by the automatic approach. This can be explained by the fact that the collection of death certificates is the most complete. We also see that for males (fathers or grooms) matches are found more often than for females. One reason for that is that males are mentioned more often than females in legal acts. However, the numbers of identified matches by humans and by the automatic approach is relatively similar.

7.8 Discussion

As can be seen from Sect. 7.7, the direct application of standard ER solutions to real-world multi-source genealogical dataset brings good results even though some space for the improvements is left. There are quite some differences between the civil certificates and the notary acts. On the one hand, some information which is available in the certificates, such as names of parents, is not always available in the notary acts. Furthermore, the available information in the notary acts is not fixed at all; depending on the type of the act there might be information about husband-wife relation or other family relations, while in other acts no family relations may have been mentioned. When evaluating the precision and recall of our approach we do not take into account what information may or may not be presented, but only assess the following criteria for each name that occurs in the notary act:

- Which links to certificates humans find for a name, where the algorithm has not reported them (i.e., recall)
- Which links to certificates humans do not find for a name, where the algorithm has reported (i.e., precision)

Since the labeling was not complete (due to the strenuous nature of this task) we additionally checked the top-links found by the humans in order to get an idea to what extent the accuracy figures were biased by the incomplete labeling.

Non-structural differences such as missing information may cause biases in the evaluation because the task becomes more difficult both for humans and computers. By the nature of our evaluation strategy, however, we try to counter this effect as much as possible.

Another challenge that we deal with is the lack of ground truth which makes it difficult to get reliable and high-quality evaluation. This problem is very common when dealing with real-world data (Alsaleh and van Oorschot 2013; Efremova et al. 2015).

In Fig. 7.7, using a Venn diagram, we demonstrate all possible intersections when a match is positive according to the absolute ground truth, the human

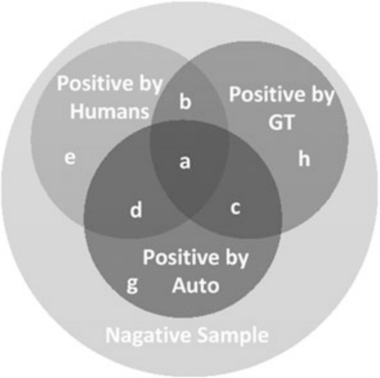


Fig. 7.7 A diagram of possible intersections between the ground truth, human judgment and the automatic ER approach. GT stands for a ground truth

judgment, and the baseline approach. Each circle in the diagram represents positive matches according to absolute ground truth, human judgment and the baseline approach. The closer human judgment agrees with the absolute ground truth, the more accurate is our evaluation.

In most machine learning approaches there is an implicit assumption that in the test data the absolute ground truth is known. In our diagram this would correspond to the cells labeled *e*, *c*, *d*, and *h* being empty, and hence the human judgment (green circle; i.e., the labels to which we have access) coincides exactly with the inaccessible ground truth (red circle). Given the nature of our problem, however, this is not at all true. On the one hand we calculate the perceived precision and recall as:

- perceived precision = $(a + d)/N$
- perceived recall = $(a + d)/(a + b + d + e)$, where $N = (a + c + d + g)$ represents the known number of positives by our classifier,

versus the real precision and recall:

- real precision = $(a + c)/N$
- real recall = $(a + c)/(a + b + \mathrm{c} + h)$.

Depending on the size of *c*, *d*, *e*, and *h* the differences may be significant. Therefore, we will now systematically analyze these 4 quantities and see how we can reduce their risk.

On the one hand we can be reasonable certain that the links labeled by humans are correct and hence cells *e* and *d* are probably small. On the other hand, however, cells *c* and *h* are likely very large given the arduousness of the task of labeling *all* matches. *c* (number of correct machine matches not found by humans) we control by running all seemingly false positives again by humans as explained in Sect. 7.7.3. There indeed we detected that there were several matches (7 % of the matched found by computer) not found by humans. In this way, we could reduce c and hence get accurate numbers for precision. Controlling *h*, on the other hand is much more difficult, as this concerns true matches not found by humans, nor by the machine. Even though we tried to reduce *h* as much as possible by reducing the number of notary acts, and requesting the human annotators to find all possible links for this reduced set of certificates, it is inevitable that a large part of true links go by unnoticed. This problem is to a large extent unsolvable and we tried to tackle it by the indirect, common sense-based evaluation in Sect. 7.7.4.

7.9 Conclusion

In this chapter, we studied the concept of ER in genealogical data research, where the data was provided from sources of different structure. We investigated the application of a number of existing ER techniques. Considering the multi-source characteristic of the data, classical ER techniques are difficult to apply due to the diverse types of data attributes and the lack of sufficient information. We focused our study on the extension of feature sets and on the analysis of the influence of name popularity, migration groups and co-reference information on the overall ER process. We showed that having inferred the name popularity, geographical distance together with the co-reference information helps to significantly improve ER results.

In order to assess the effectiveness of the applied ER approach and also to obtain a training dataset, an interactive web-based labeling tool was developed with which the human experts helped to manually identify the matches from an adequate sample of the whole data. The manually labeled matching was used for two purposes: obtaining training data and computing the evaluation metrics: precision, recall, and F-score. We cross-validated the overall ER process. Working with real-world data we had to deal with the lack of ground truth, which makes it difficult to get a reliable and high-quality evaluation. In the second experiment, we showed that experts missed a lot of true positives during the manual data annotation, therefore the precision in our results is underestimated.

The designed ER algorithm has some limitations. One of them is selected indexing strategy to generate candidate pairs. The disjunctions of partial phonetic keys helps us to achieve relatively high recall value, however as a part of our future work we want to exclude the blocking phase completely. One of a potential extensions is implementing a fuzzy name matching by using the *bit vectors* technique (Schraagen 2011) or Levenshtein automata (Schulz and Mihov 2002). In this case, we do not need to apply any data partitioning.

Another extension of the applied approach is to use more information from notary acts. It requires more advanced text processing techniques. For instance, inheritance notary acts contain information about many family relationships (parents, siblings, nephews, etc.) which should be taken into account during the ER.

We also work on more advanced ER techniques and want to improve the ER process by applying collective relational entity resolution (Getoor and Machanavajjhala 2012) where co-reference information is not processed as an additional attribute. Instead we want to apply more advanced graph-based techniques, taking into account that there may be multiple persons in the different acts and certificates that are co-referenced.

Another improvement concerns the applied predictive models. Instead of using logistic regression or regression trees that were proposed in the chapter we want to use Probabilistic Relational techniques, thus applying Probabilist Graphical Models to solve ER problem can be an appropriate next step.

In short, we proposed an ER approach which was capable of extracting various entities from multiple sources of information, some structured and some unstructured; the efficiency of the proposed approach was first improved by means of the labeled data provided by human experts, and afterward was evaluated in detail by human experts. The thorough evaluations of the work, showed good precision and recall, which is sufficient for some prosopographical and demographical researches, yet allows for various extensions. Thus, there is a potential for future work.

Acknowledgements The authors are grateful to the BHIC Center for the support in data gathering, data analysis and direction. In particular, we would like to thank Rien Wols and Anton Schuttelaars whose efforts were instrumental to this research and their patience and support appeared infinite. This research has been carried under Mining Social Structures from Genealogical Data (project no. 640.005.003) project, part of the CATCH program funded by the Netherlands Organization for Scientific Research (NWO).

References

Alsaleh, M., & van Oorschot, P. C. (2013). Evaluation in the absence of absolute ground truth: Toward reliable evaluation methodology for scan detectors. *International Journal of Information Security, 12*(2), 97–110.

Bhattacharya, I., & Getoor, L. (2004). Iterative record linkage for cleaning and integration. In *Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, DMKD'04* (pp. 11–18). USA: ACM.

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transaction on Knowledge Discovery from Data, 1*(1), 5.

Bilenko, M. (2006). Adaptive blocking: Learning to scale up record linkage. In *Proceedings of the 6th IEEE international conference on data mining ICDM-2006* (pp. 87–96). Piscataway: IEEE.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology, 37*(1), 51–89.

Christen, P. (2006). A comparison of personal name matching: Techniques and practical issues. In *Proceedings of the 'workshop on mining complex data' (MCD'06), held at IEEE ICDM'06* (pp. 290–294).

Christen, P. (2008). Automatic record linkage using seeded nearest neighbour and support vector machine classification. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '08* (pp. 151–159). USA: ACM.

Christen, P. (2012). *Data matching*. New York: Springer Publishing Company, Incorporated.

Cohen, W. W., Kautz, H. A., & McAllester, D. A. (2000). Hardening soft information sources. In R. Ramakrishnan, S. J. Stolfo, R. J. Bayardo & I. Parsa (Eds.) KDD (pp. 255–259). USA: ACM.

Efremova, J., Montes García, A., & Calders, T. (2015). Classification of historical notary acts with noisy labels. In *Proceedings of the 37th European conference on information retrieval, ECIR'15*. Vienna, Austria: Springer.

Efremova, J., Ranjbar-Sahraei, B., & Calders, T. (2014). A hybrid disambiguation measure for inaccurate cultural heritage data. In The 8th workshop on LaTeCH (pp. 47–55).

Efremova, J., Ranjbar-Sahraei, B., Oliehoek, F.A., Calders, T., & Tuyls, K. (2013). An interactive, web-based tool for genealogical entity resolution. In *25th Benelux Conference on Artificial Intelligence (BNAIC'13), The Netherlands*.

Efremova, J., Ranjbar-Sahraei, B., Oliehoek, F.A., Calders, T., & Tuyls, K. (2014). A baseline method for genealogical entity resolution. In *Proceedings of the Workshop on Population Reconstruction, Organized in the Framework of the LINKS Project*.

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering, 19*(1), 1–16.

Florian, R., Ittycheriah, A., Jing, H., & Zhang, T. (2003). Named entity recognition through classifier combination. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, CONLL '03* (Vol. 4, pp. 168–171). USA: Association for Computational Linguistics.

Getoor, L., & Machanavajjhala, A. (2012). Entity resolution: Theory, practice & open challenges. In *International Conference on Very Large Data Bases*.

Getoor, L., & Machanavajjhala, A. (2013). Entity resolution for big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1527–1527). USA: ACM.

Hernández, M. A., & Stolfo, S. J. (1995). The merge/purge problem for large databases. *SIGMOD Record, 24*(2), 127–138.

Huijsmans, D. (2013). Dataset historische Nederlandse toponiemen spatio-temporeel 1812–2012. In *IISG-LINKS*.

Ivie, S., Henry, G., Gatrell, H., & Giraud-Carrier, C. (2007). A metricbased machine learning approach to genealogical record linkage. In *Proceedings of the 7th Annual Workshop on Technology for Family History and Genealogical Research*.

Lawson, J. S. (2006). Record linkage techniques for improving online genealogical research using census index records. In *Proceedings of the Section on Survey Research Methods*.

McCallum, A., Nigam, K., & Ungar, L. H. (2000). Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 169–178). USA: ACM.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes, 30*(1), 3–26.

Naumann, F., & Herschel, M. (2010). *An introduction to duplicate detection*. San Rafael: Morgan and Claypool Publishers.

Nuanmeesri, S., & Baitiang, C. (2008). Genealogical information searching system. In *4th IEEE International Conference on Management of Innovation and Technolog*y, *ICMIT 2008* (pp. 1255–1259).

Rahmani, H., Ranjbar-Sahraei, B., Weiss, G., & Tuyls, K. (2014). Contextual entity resolution approach for genealogical data. In Workshop on knowledge discovery, data mining and machine learning.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America, 102*(44), 15942–15947.

Sammut, C., & Webb, G. I. (2010). *Encyclopedia of machine learning*. Berlin: Springer.

Schraagen, M. (2011). Complete coverage for approximate string matching in record linkage using bit vectors. In *ICTAI'11* (pp. 740–747).

Schraagen, M., & Hoogeboom, H. J. (2011). Predicting record linkage potential in a family reconstruction graph. In *23th Benelux Conference on Artificial Intelligence (BNAIC'11), Belgium.*

Schraagen, M., & Kosters, W. (2014). Record linkage using graph consistency. In *Machine learning and data mining in pattern recognition*. *Lecture Notes in Computer Science* (pp. 471–483). New York: Springer International Publishing

Schulz, K. U., & Mihov, S. (2002). Fast string correction with levenshtein automata. *International Journal of Document Analysis and Recognition (IJDAR), 5*(1), 67–85.

Singla, P., & Domingos, P. (2006). Entity resolution with markov logic. *Proceedings of the Sixth International Conference on Data Mining, ICDM'06* (pp. 572–582). USA: IEEE Computer Society.

Štajner, T., & Mladenić, D. (2009). Entity resolution in texts using statistical learning and ontologies. *Proceedings of the 4th Asian Conference on the Semantic Web, ASWC'09* (pp. 91–104). Berlin: Springer.

Sweet, C., Özyer, T., & Alhajj, R. (2007). Enhanced graph based genealogical record linkage. *Proceedings of the 3rd International Conference on Advanced Data Mining and Applications, ADMA'07* (pp. 476–487). Berlin: Springer.

Van den Bosch, A., Busser, B., Canisius, S., & Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In Computational linguistics in the Netherlands: Selected papers from the seventeenth CLIN meeting (pp. 99–114).

Winkler, W. E. (1995). Matching and record linkage. In *Business survey methods* (pp. 355–384). New York: Wiley.

Chapter 8
Record Linkage in the Historical Population Register for Norway

Gunnar Thorvaldsen, Trygve Andersen and Hilde L. Sommerseth

Abstract The Historical Population Register (HPR) of Norway aims to cover the country's population between 1800 and 1964 when the current Central Population Register (CPR) takes over. This may be feasible due to relatively complete church and other vital registers filling the gaps between the decennial censuses—In 1801 and from 1865 these censuses were nominative. Because of legal reasons with respect to privacy, a restricted access database will be constructed for the period ca. 1920 until 1964. We expect, however, that the software we have developed for automating record linkage in the open period until 1920 will also be applicable in the later period. This chapter focuses on the record linkage between the censuses and the church registers for the period 1800 until around 1920. We give special attention to database structure, the identification of individuals and challenges concerning record linkage. The potentially rich Nordic source material will become optimally accessible once the nominal records are linked in order to describe persons, families and places longitudinally with permanent ids for all persons and source entries. This has required the development of new linkage techniques combining both automatic and manual methods, which have already identified more than a million persons in two or more sources. Local databases show that we may expect linkage rates between two-thirds and 90 % for different periods and parts of the country. From an international perspective, there are no comparable open HPRs with the same countrywide coverage built by linking multiple source types. Thus, the national population registry of Norway will become a unique historical source for the last two centuries, to be used in many different multi-disciplinary research projects.

G. Thorvaldsen (✉) · T. Andersen · H.L. Sommerseth
Norwegian Historical Data Centre, University of Tromsø, Tromsø, Norway
e-mail: Gunnar.Thorvaldsen@uit.no

T. Andersen
e-mail: Trygve.Andersen@uit.no

H.L. Sommerseth
e-mail: Hilde.Sommerseth@uit.no

G. Thorvaldsen
Urals Federal University, Yekaterinburg, Russia

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_8

8.1 Background: Population Registers and Early Linking Methods

Currently, the Central Population Register (CPR) covers the population of Norway back to 1964. Earlier, local, card-based population registers covered an increasing number of municipalities from 1906, but these are not considered for computerization due to their volume and scattered archiving (Thorvaldsen 2008). Instead, the national Historic Population Register (HPR) is being built by linking the censuses and church record from 1800 until 1964. About a third of this source material has been transcribed, while the rest is being digitized. The HPR will become a unique historical source for the last two centuries and may be used in many different multi-disciplinary research projects. The potential inherent in the rich Nordic source material will be realized once the nominative records are linked together in order to describe persons, families and places longitudinally with permanent ids for all persons and source entries. This has required the development of new linkage techniques combining both automatic and manual methods, consisting of a composite of several established techniques combined with new methods that increase the linkage rates. Already more than a million persons have been identified in two or more sources.

Prior to the construction of the Historical Population Register, record linkage has been performed on local, nominative source materials, mainly censuses and ministerial records from localities, of which the transcription started in Norway around 1970. Even a bit earlier, family reconstitution was performed on selected parishes with manual methods, and later on these methods were adapted to link digitized records interactively after sorting them according to a number of criteria (Nygaard 1992). This process was still labour intensive and usually it was only possible to link a sample of the population in a local parish to subsets of nineteenth century censuses and church books. The first serious attempt with computer assisted semi-automated record linkage in 44 parishes around 1801 was performed by Jan Oldervoll and his students at the University of Bergen around 1980 (Engelsen 1983). A more automated approach has been used for linking the censuses of 1865, 1875 and 1900 for the province of Troms in the early 1990s (Thorvaldsen 1995, 2000). The database for the parish of Rendalen 1735–1900 was constructed on the basis of a manual family reconstitution in the late 1990s. Its use as a golden standard will be further explained in Sect. 8.5. In parallel, a system dedicated to interactive record linkage was constructed and used on two parishes outside Oslo for most of the nineteenth century (Fure 2000). This interactive process has been carried further with a semi-commercial, semi-automated system used to link censuses, church records, probate registers and other nominative sources in order to create layout for the printing of farm histories and genealogies in local community history books, called *Busetnadssoge* (Kjelland and Sørumgård 2012). Together with older software versions, this has been used to build linked, local population registers for about ten municipalities from the eighteenth century until the late twentieth century. For the last decades these build on digitized oral sources in addition to open written sources such as newspapers and family genealogies.

8.2 Sources

The keeping of church records in Norway spread slowly from 1623 with entries about baptism, burial and marriage (in chronological order) which became compulsory in 1680. It was not until 1812 that priests started to use printed forms with separate columns and defined headings, while the ministerial registers were organized by type of event with baptisms, marriages and burials in pre-defined sections. Høgsæt (1990) has found an undercount of 10–20 % in the eighteenth century records, particularly of children's burials. This, together with the fact that the earliest sources contain poor information on the ages, the names of married women and information about residence, motivates the decision to start the national HPR in 1801 with the first nominative census. It is realistic, however, to build local population registers for earlier periods.

The HPR will in principle link all openly accessible information about historical persons and their locations in Norway. It will be built mainly on church records and censuses, but may eventually include information from emigrant lists, newspaper notices, prison records and tombstones.[1] The HPR links information about the same persons appearing in several sources and their settlements. Combination of many sources will improve the HPR both by providing more information about people and places, and by making more reliable links. The period from 1735 to 1964 includes 9.7 million people and more than 37 million source events in census records, church records and other sources. From the period 1800 to 1960 some 10 million out of 30 million records in censuses and church records have been transcribed, and there are ambitious plans to transcribe the remainder of the church records until 1930 during the next few years (Eikvil et al. 2010).

The evolution of the population is closely linked to the development of settlements, communities and regions. It is an ambition of the HPR project to provide links to this information, which is nearly as dynamic as the population itself, involving the origin of smallholdings, the splitting of farms, urban growth, etc. A residence may be linked to several municipalities over time if administrative boundaries changed. The points of departure are the 1838, 1886 and 1950 farm tax registers. Census records can be connected to these, using farm name, place name, street name, title number, farm number, house number and street number.[2] Results from the project *Historical administrative boundaries* shows that 80–90 % of the farms can be linked to a unique title number. The dynamic farm register developed by the economist Kåre Bævre in this project can be linked to the HPR. Urban places

[1] A project to interpret newspapers with OCR-like methods has received separate funding from the Research Council. Genealogists have photographed and transcribed many thousand tombstones and added this information to a database.

[2] The cadastre from 1838 has been scanned and can hopefully be made available in machine-readable and searchable text form after being processed with OCR. Confer also http://www.dokpro.uio.no/cgi-bin/stad/matr50, http://www.rhd.uit.no/matrikkel/matrikkel1838.aspx and http://www.rhd.uit.no/indexeng.html

will require extensive new work due to the transition from the serial property numbers to street names and numbers in the towns. As will be discussed later, we rely on local and regional expertise to solve these puzzles correctly. Historian Arne Solli at the University of Bergen has done impressive work with the person and location information for cities in the BerGIS project for the period 1696–1906 and is seeking to extend this GIS to other urban areas (Solli 2006). Based on the coordinates for all main farms in the countryside and blocks of houses in the towns, it is possible to create illustrative and analytical graphics, such as displaying the migration patterns in a neighbourhood or the spreading of epidemics.

8.3 Data Processing and Database Structure

Several genealogical databases now use MediaWiki software to link and couple records via the Internet. *WeRelate.org* is the world's largest genealogy wiki with over 2.6 million people registered in the database. These are based on importing family trees via Gedcom files from genealogists. Since the database is not based directly on the source material, this creates problems with duplicates and database quality, and WeRelate administrators make substantial efforts to reduce redundancy (Quas 2011). We have experimented to use the WeRelate platform and to construct our own wiki version, but have decided that building a relational database gives a more flexible and efficient solution, while keeping much wiki-like functionality. The size of the HPR project makes this necessary. In the HPR we read and store all the events directly from the transcribed sources. This places great demands on the system as transparent, simple and designed to strengthen database quality over time rather than degenerating due to redundant duplicates which will pollute statistical use of a demographic database.

The HPR relational database will allow signed up participants to contribute with their expertise. Experience from other projects, such as the Local History Wiki illustrates how different contributors share knowledge on the basis of their particular expertise.[3] Topics include their own family and residence, geographic areas, occupations, ethnic groups, migrants, persons mentioned in historical literature. It is possible to describe the work within a particular theme in terms of project pages, inspiring users with different approaches and techniques to work on the data. The Norwegian Local History Institute is a partner which will link its Local History Wiki as an encyclopaedia with more comprehensive information about places as well-known persons found in the HPR.

The source entries have been transcribed verbatim, faithfully reproducing names, year of birth and other variables. The same person may have variant spellings of names, different age data and other incompatibilities in different sources. For the overall presentation of each person's data we choose the core data in which we have

[3] https://lokalhistoriewiki.no

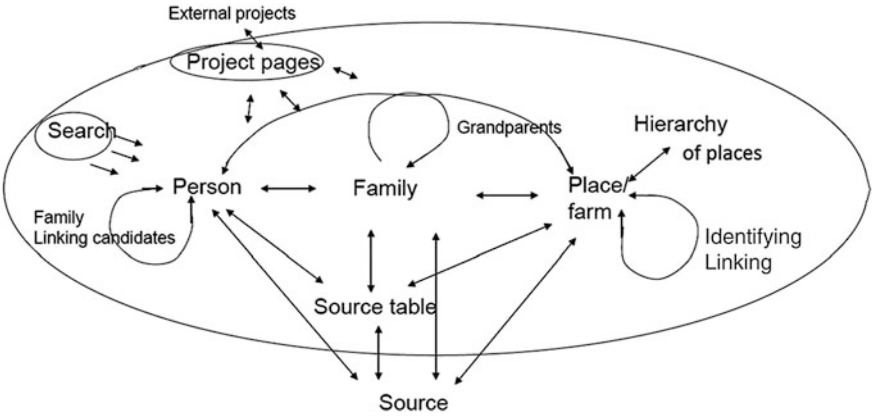


Fig. 8.1 One- and two-way pointers in the open HPR represented by *single* and *double arrows*. Project pages describe community history and local genealogy projects. The hierarchy of places is typically county-parish-farm-sub-farm which need to be linked internally and to the persons

most confidence based on the following priority rules: (1) selection by user during manual record linkage, (2) church records, where the older takes precedence over newer entries, (3) censuses, where recent takes precedence over older and (4) other sources, where recent takes precedence over older. Among church records the oldest will usually be the authoritative baptismal entry, whereas among the censuses the quality improved over time.

The main feature of the system is the establishment of a separate page for each person and family. A *person* page is of the type “Ole Hansen (21)”, which says that there are currently at least 21 occurrences of entries for persons named Ole Hansen registered in the HPR. These sequence numbers are only used in the places, families and persons where it is necessary to provide uniqueness. On this page users can collect all source information about the person, links to family members and even put a picture or write a biography. A *family* page can be entitled “Ole Hansen (21) and Maria Thorsdatter (8).” Their residence information is created from censuses and has pointers to other residents. Settlements and addresses are usually not established from the church records because these are less precise in their description of objects. The pages constructed from the source entries are merged/linked when we have sufficient certainty that it really is the same person, family or residence that is mentioned in various sources.

Figure 8.1 gives an overview of the entities in the HPR. An individual entity is created for each person’s occurrence and is linked to the relevant source entry. Family pages are created for each family from a marriage record or one parent with children. Family pages are merged when the parents’ records are linked. “Grandparents” is a special case of parent–child links exemplifying how basic family links can be nested to show information about extended families. In addition, there is a free-text page in order to describe peculiarities in the sources or any other topic related to linking or the population register. The HPR project will provide links to external sites about the project and links to relevant pages of the HPR, such as specific homes or people. There are also entries about each source used in the HPR. A place page for a farm or other types of residence may point to different municipalities in the different censuses because of municipality changes through time, and a family page may point to different place pages due to migration.

It is necessary to establish stable IDs for all persons in the HPR in order to have reliable references to individual occurrences of persons within the HPR for record linkage, and between the HPR and external databases. Today's social security numbers rely on stable and unique identification based on birth date, gender and additional digits. However, date of birth will often not be known precisely for historical persons and can hardly be used as part of the identifier. This is why the National Archives is establishing IDs for all *source entries* (called PKID) and all *properties* (eKIDs) for instance farms listed in the sources. We consider this as a reliable identification of each individual occurrence, referring each entity to a well-defined source entry.

The HPR uses a rule-based definition of each *individual's* ID, an unambiguous PID which is based on one of the PKIDs. Thus, the PID is defined as the PKID with highest priority among records linked to a particular person. This is a system that can handle revisions of the linking, as illustrated in Fig. 8.2. On the left side the PKIDs A and B are initially linked as PID A as A has highest priority. But when a third PKID (C) is linked to PKID B, the link between PKIDs A and B must be discarded. On the right side, PKIDs B and C are linked and priority is given to PKID B as personal identifier (PID) for the person whose records were linked. A separate table will contain the linkage history making it possible to see what links have been attempted unsuccessfully and what PIDs have been replaced due to the revision of links.

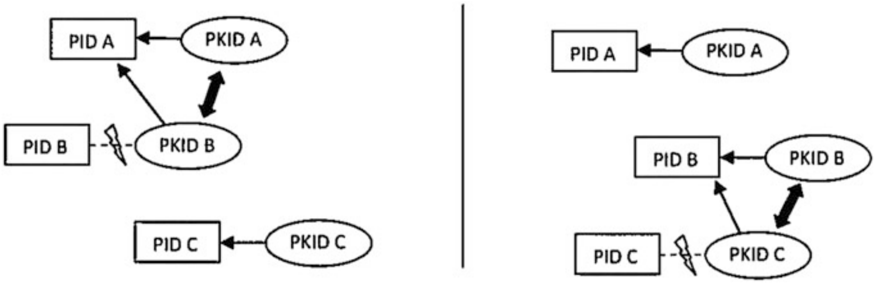


Fig. 8.2 Changing the person's ID (PID) when linking two person records (PKID). In the example on the left the PKID A has priority, and PKID B "inherits" the PID since the records refer to the same person (*thick arrow*). The example on the right shows that record B is instead linked to record C, and the previous link is broken. Record B's PKID becomes the PID because PKID C has lower priority

In the HPR we use the following source priorities to decide about the PID among the PKIDs when there are several to choose from:

1. Census 1910
2. Funerals/lists died before 1910
3. Births after 1910
4. Other censuses where recent censuses have priority over older ones
5. Lists of emigrants 1870–1930
6. Lists of immigrants from after 1910
7. Other entries in the parish registers where older entries (especially baptisms) have priority over later entries
8. Other sources where more recent entries have priority over older ones.

The first five categories are not overlapping, and the first three are presumably of relatively good quality. The census from 1910 will be crucial when linking the open HPR to the closed part of the population register from around 1920.

The results from the automatic linking described in Sect. 8.3 will be copied as lists of groups of PKIDs (the Ids described above) to be revised interactively. It is necessary that we include references to the sources and identify the individuals consistently, and the PKIDs fulfil both requirements. Thus, as soon as a new source is read into the HPR site, the straightforward links (on the two to three highest levels) are established automatically or semi-automatically. The tables will contain information about the quality of the links on the 0–10 scale (see Sect. 8.4.1) and linkage criteria. The web-based, interactive module of the HPR will receive suggestions for record linkage from many users, and such crowd-sourcing requires a standard set of rules with any conflicting links flagged and constant controls to verify consistency (see Sect. 8.3).

8.4 Source Dependent Record Linkage

The HPR will mainly rely on the input of data from censuses and church records, ensuring that we include all person records in these sources only once. There are some duplicate records also in the originals, primarily due to the combined use of de facto/de jure principles in the censuses and some entries of people who died away from their usual residence. These are far fewer and easier to remove by deduplication than all the redundant records which exist in databases built by combining collective genealogies, which are bound to contain more common ancestors as we move backwards in time.

Recognized relationships between spouses and other family members increase the likelihood for successful linkage, but parish registers and censuses provide information about relationship between persons in different ways. The census taker registered relationships on the level of family and household and these have been made explicit with pointers constructed through the cooperation of the North Atlantic Population Project (NAPP).[4] Parish registers provide information about relationship according to type of event. Marriage lists naturally inform about married couples and from 1820 onwards included the bride's and bridegroom's fathers. For the majority of baptism records we are able to relate explicitly the mother, father and child. The burial represents a more individualistic event, and provides a more difficult source to link. From 1877 the relations between a parent and the deceased child or between spouses are specified in a separate column, but fortunately the practice to register the father's or spouse's name started earlier in some parishes. Elsewhere, the burial lists until 1877 provide information only about the names, age and address of the deceased, which is often insufficient information for linking to other sources. Age can be missing or unreliable in the first phase of the HPR, and birth date is only found in the baptism lists during most of the nineteenth century. For the marriage and burial records a more thorough registration of birth dates started in 1877. Date of birth was introduced later in the censuses, for persons aged under two in 1891 and for all from 1910 onwards, while those older than two only had their birth year reported in the 1891 and 1900 censuses. If place of residence was registered along with first and last name in the parish register, and corresponds to the address in the census, reliable linkage is more likely, especially in the period around 1801 when information on birth place and other characteristics is often missing.

A person in the HPR should not have multiple fathers or mothers. Even if a person's ancestry cannot be linked because of a conflict between competing links to the father or mother, there is always hope that the parents' person records can be linked later when additional information becomes available. Thus, children and parents' potential person records are marked as candidates for linkage. Attempts to merge person entries manually where there is conflict between the candidate parents will result in a warning.

In spite of the fragmentary nature of the migration lists in the church books, the national scope of the project makes it possible to capture migration between different domestic areas. Migrants can usually be retrieved at both place of origin and destination, and special lists for migrating people who are not identified can be established. Similarly, it will be possible to rediscover the roots of most people of Norwegian descent living abroad in the HPR since the emigration lists from around 1870 with about one million records have been transcribed. Returnee emigrants is a bigger challenge, but many are listed explicitly in the 1910 and 1920 censuses, and from World War I onwards the lists of immigrants became more complete.

[4] https://www.nappdata.org

8.4.1 Record Linkage Principles

The main principles behind the automatic and manual linking in the open HPR are:

1. The HPR will build on many available sources of good quality. There should be two-way links, that is, from the HPR to the sources and from the source entries to the HPR.
2. We want the greatest possible openness and transparency. It should be possible to see who made the links and what criteria were applied.
3. We pursue the highest possible quality of links. All users are given opportunity to comment on the quality of the HPR.
4. All links will be marked with quality and linkage rule flags, and the representativeness of linked samples can be compared statistically with the population in the decennial censuses since 1801 in order to estimate bias in the linked part of the population.

These principles ensure unique source references and the combination of data sources ensures increasing data quality over time. The HPR functions as an index to the source entries instead of replacing them. With regard to citations and transparency, the open HPR will be different from the national historical population registry in Iceland, deCode Genetics, which is closed in such a way that people only have access to their own ancestry and is somewhat limited and intransparent with respect to source references and criteria for record linkage. The historical, longitudinal databases in Sweden (the Demographic Database, the Stockholm Roteman Archive, the Scanian Economic Demographic Database), and the Historical Sample of the Netherlands have a better basis for linking in their pre-linked, original sources, but so far cover only parts of the population of these countries (Mandemakers 2000; Thorvaldsen 1998).

In the HPR a distinction is made between *linking* and *coupling*. Whereas *linking* identifies and determines that the same person is referred to in at least two different source entries (e.g. two censuses or two records in a baptism list), *coupling* combines information about relations between persons in a specific event, e.g. in a baptism entry or in a census household list. Linkage and coupling are interrelated processes, since family information can be used to link persons and information from several sources may be necessary to decide which persons are related.

All links will get a quality flag on a scale from 0 to 10, according to guidelines following to what extent there are unique and equal characteristics about a person in two or more sources. These characteristics are first name, family name, gender, age or birth date and birthplace. Address and occupations can be used to obtain uniqueness, and identifying related persons across two sources strengthen the quality of the links. The following grading system is applied:

10: Completely secure link, fulfilling criterion 8 plus identifying the person as part of a family.

8: The same birth date and same or similar names indicated by a high name comparison score, as well as geographical origin on the farm or street level.

6: The same or similar name of both spouses and geographic affiliation.
4: Same or similar name, age and related to the same place.
1: A probable link, the records should likely be linked, but further information is required for a certain link.
0: Established as linkage candidate with reasons that specify the uncertainly.

Figure 8.3 illustrates the two competing goals when linking: both to link as many individual records as possible, and to avoid linking entries that do not belong together (Johansen 2002). The ideal is represented by the figure's origo at the bottom left, where all linked entries really belong together and no true links are omitted. This is hardly ever possible to achieve with historical records because of imprecise and missing information. The problem is that when we impose stricter rules in order to reduce the risk of linking records belonging to different individuals (horizontal axis), we easily rule out several links that really should have been included (vertical axis). Conversely, if we introduce more liberal rules in order to ensure that all potential links are realized (moving down on the vertical axis in order to minimize the number of "true negatives"), we may include more links that really should be discarded (introducing more "false positives" and thus moving towards the right end of the horizontal axis).

Johansen assessed that while getting close to the origo is ideal, acceptable linkage results lie between points A and B on the curve in the diagram, with few accepted erroneous links (A) and few excluded appropriate links (B). Linkage results along curve II he deemed unacceptable because this introduced too many false links in the database and too many correct links were omitted. His main point in the HPR context was that Danish and Norwegian source material is quite similar.

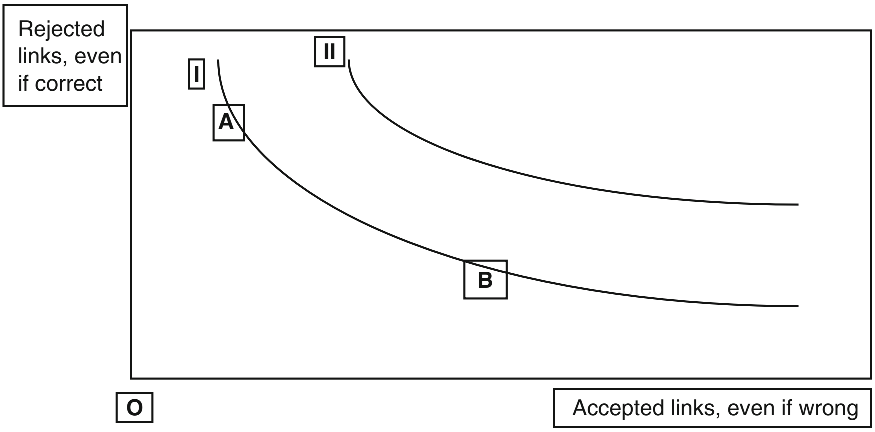


Fig. 8.3 The ratio of accepted links with errors ("false positives") and correct links that were rejected ("correct negatives") in sources from the early 1800s (*curve II*) and late 1800s (*curve I*). "*O*" marks the starting point of the *vertical* and *horizontal* axes—the origo. From Johansen (2002)

Starting with the second half of the nineteenth century, the source material has a precision that allows automatic linking along curve I, while sources from the earlier decades are so imprecise that they must be linked manually with auxiliary information as explained in the next paragraph. Fortunately, over the past decade we have developed techniques and methods that allow us to conclude differently, at least for substantial parts of the material.

First, we can now dynamically bring together variant spellings of the same name, both using standardized lists of names from the nineteenth century censuses developed by expert onomatologists, and algorithms for the comparison of strings (Alhaug 2011). Second, we have developed software which not only links individuals, but takes into account couples or entire families when linking. Both techniques are particularly valuable when dealing with a period when all information, including names and ages was imprecise or sometimes missing from the sources. By utilizing real-time name standardization and information on family relations, we can move a significant part of the population to be automatically identified in several sources from curve II to curve I in Fig. 8.3. Analysis of the 1801 census shows that about 80 % of women and 85 % of men were living with at least one other family member. Even if only a portion of these relations were stable over time, it still shows there is much potential in linking with group criteria. This is born out in our ongoing record linkage work where over a million persons have been linked in two or more sources, including many from the early nineteenth century.

In the next round, the problem of balancing the proportion of correct and incorrect links can be reduced further in the HPR by manual record linkage. The same quality and methods flag requirements apply to all linking, providing a quality indicator and justification for each link. Thus, researchers can make independent assessments and choose what links they will trust in their analyses. In summary, automatic record linkage has the great asset that documentation is inherent in the algorithms. The rules upon which the software is built can be spelt out in a database table, and references to this table can be a variable in the linked data set. This makes the links easier to trust, especially, when using material linked by others. However, there are decisions based on background knowledge and details in the sources which are not easy to automate: nick names used in certain families, knowledge about ancestors who cohabitated, names of neighbouring farms that were used interchangeably for cottars' places etc. Lists of property sales, probate registers and many other sources have been digitized only to a small degree, but have been used by genealogists to build their ancestral pedigrees. This wealth of information they can activate through record linkage crowd sourcing via the web.

Especially for the earliest period of the HPR we expect that a significant part of the linking and coupling will be done manually with contributions from genealogists and local historians volunteering. This is due to the simple structure of the source material with rather few variables, missing data and lack of consistency. It would be possible to automate much of this manual record linkage, but at a high cost. Typically, over 90 % of the investment in software would go into automating the 10 % special cases where human flexibility and special knowledge play a key role. A key element in this work is how to motivate many users to provide high quality links and how these contributions are monitored and how the quality is assessed. Most importantly, it must be easy in order to find the potential linkage candidates to search for all the people featured in a specific source, all families in a municipality, and all priests in Norway or other groups of individuals, families and places. Most rural municipalities in Norway have published local community history books which list the ancestries on the farms systematically. In addition, the volunteers can search for persons' records among the chronological events in the church books, all events on a farm over time or other criteria. If in doubt, they are enable to flag person records as linkage candidates and ask for the opinion from others to confirm or reject the link. Next, the automatic and manual links can be informed by checking the HPR against their personal databases or genealogies. National coverage provides a final solution to linkage problems: When only one unique candidate record for linkage remains the link has been ascertained by the elimination principle.

This interactive web-based system lets the user, after logging in, search one or several sources during a specific period for names, age or birthdate, birthplace occupation and place of residence. Once candidate records are displayed, the user can check what records have already been linked automatically. These can be modified, and new links added. Background knowledge about the persons from newspapers can be accessed, and a module with information about tombstones is planned. A beta version of the software is available for testing—so far only in Norwegian.[5]

8.4.2 Automatic Record Linkage: The Example of Lenvik Parish

The Norwegian Historical Data Centre has developed software for the automatic linking of married couples and other family members, also utilizing special algorithms for comparing names.[6] In connection with the project to celebrate the constitution of 1814 these programs link the 1801 census and the church records (baptism, burial and marriage) for the northern part of Norway (Bråthen 2011). Figure 8.4 displays the average number of links between source entries for all men who became fathers during the period 1799–1815 in Lenvik parish south of Tromsø. Information about both father and mother was used to link the baptisms. The mean number of births per year was 41.5, however with a wide variation from one year to another, with the extremes of 16 in 1811 and 65 in 1806.

[5] http://hbr2.nr.no/demo/avisproject/avisprosjekt.php

[6] Developed in PL/SQL, the scripting language for Oracle databases. Names can be compared efficiently in real-time with the built-in Jaro-Winkler string comparison algorithm, aiming for similarity levels of over 0.8.

As a result of automatic linking, a rather consistent result emerges when looking at the average number of about four birth list links over time, strengthening our confidence in the linking. In the baptism records the father is consistently linked to the child and the mother, and it is possible to keep track of the couple from one birth to the other. Approximately two-thirds of all parents in the baptism list were linked to the marriage list, which is a stable proportion over time. Both spouses' names are in both lists and the priests were consistent with respect to spelling etc. since they copied information from baptism or confirmation to the marriage protocol.

The census represents only a snapshot of the population in Lenvik parish in 1801, and we typically find persons in different stages of their life. Thus, marital status and place of residence for a couple in the baptism/marriage lists does not necessarily match with the registration in the census. As a consequence, we get good linkage results in the years close to the census year, with a gradual decrease towards the end of the period.

Linking to the burial list clearly was the major challenge. A significant proportion of the burials only provide the first and last name of the deceased, which is usually considered as inadequate linkage criteria. Regarding the entry of child deaths, the lists fortunately provide information about relatives, usually the name of the father. Thus, the burial linkages shown in Fig. 8.4 are in most cases links between a father's entry into a baptism list and the death of his child. Of all buried children under the age of two, approximately 50 % were found in the baptism list. Accordingly, the relatively high linkage rates must after all be understood against the background of the high fertility and child mortality rates in this period (Hubbard et al. 2002).

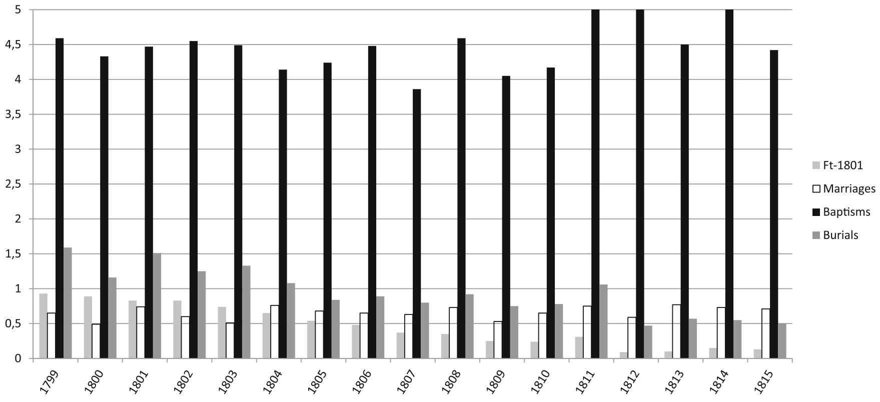


Fig. 8.4 Average number of links of fathers in between entries in the baptism list, the 1801 census (Ft-1801), the marriage list and the burial list. Lenvik parish in the period 1799–1815

About 40 % of all burials were not linked, and a large proportion of these were elderly men. They were typically registered without reference to relatives, often with first and last name only. From 1820 age information was usually included, but until the extensive revision of the church register that was undertaken by a royal resolution in 1877, providing a greater variety of data to link by, automatic and semi-automated techniques have clear limits. This is also visible when linking two censuses. Using semi-automated techniques, the 1865 and 1875 censuses were linked for parts of Troms province in Northern Norway. One-third of the inhabitants were not identified in both sources (Thorvaldsen 1995). In other words, we see a considerable scope for complimentary manual linking, employing the detailed knowledge of families and communities that exist in abundance in Norway. With the HPR's system for crowdsourcing via the Internet we activate this knowledge and resolve many of the problems individual investigators struggle with in vain.

8.5 A Ground Truth

To get an idea about the challenges we face and what kind of results can be expected from longitudinal databases at the local level we discuss the rather isolated parish of Rendalen on the border with Sweden. In Rendalen, church registers, censuses and other sources have been manually linked for the period from the local ministerial records that started in 1735 until 1950, when the last census before the one used to build the CPR was taken. The lowest linkage rates are in the eighteenth and early nineteenth century, before the regular, decennial nominative censuses started in 1866. As an isolated census without information about birthplace, the 1801 census increased the linkage rates more marginally. Of all the people that have been observed in the period 1815–1824, half were identified in both baptism, marriage and burial records, while another 20 % were identified in baptism and burial lists, but not in a marriage entry. In view of the high mortality in this period, we expect that a large number of people died unmarried (Bull 2006; Gjelseth 2000). Rendalen was a parish with little migration and the database is the result of thorough manual work on a small geographic area. We cannot expect equal coverage in municipalities with larger migration, but Rendalen can function as a "golden standard" or "ground truth" to evaluate the linking of sources in other parishes. Running the automatic record linkage software on the same source materials, is an interesting comparative exercise that still awaits completion.

Even after manual record linkage using the probate records, containing two generations as has been done for Rendalen parish, it is a challenge to reconstitute the population at a given time. These challenges are not diminished by the fact that mortality was high and that part of the population was geographically mobile during the turbulent times in the early nineteenth century. The high mortality meant that over one hundred marriages in Rendalen where recorded with widows or widowers from 1801 to 1815. To determine how these settled together with children from previous and new marriages over the following decades is no simple puzzle. The next nominative census was taken in 1866, but some help is rendered by the silver tax lists from 1816 and the farm tax lists from 1838—at least for the heads of households.

8.6 Enhanced Research Opportunities with the HPR

We know from previous local studies and international experience that an HPR can be used in a variety of local, regional and national studies and provide a basis for international comparisons. The HPR promotes international collaboration through the NAPP (Thorvaldsen 2011) and the European Historical Population Samples network (EHPS-net).[7] Our database employs algorithms developed by our partners at the Minnesota Population Center to encode the family structure automatically by creating location variables based on information about family position, sequence number in the household list, gender, age and last name. These pointer variables can be analysed together with other constructed and encoded variables (Sobek et al. 2011).

The HPR provides a new historical and social science understanding of the relevant periods. With longitudinal microdata we can study how family structure and social and geographical mobility changed longitudinally as opposed to the snapshots given in the censuses. In a medicine and health perspective, the HPR will be an important source for studies of the population's gene pool and for instance genetic diseases related to consanguinity (Surén et al. 2007). The National Institute of Health, one of our major partners, has a collection of bio-samples which can be linked to the HPR. The bibliographies maintained by the Demographic Database at Umeå University and the Minnesota Population Center in Minneapolis contain a host of further research topics which can be studied in more detail also in Norway.[8]

Within local history and genealogy it will be easier to place people's own family and community history into a broader context. In practical terms, it will be easier to identify the sources and more efficient to link to other people's work on their ancestries. It will always be possible to find more information, more sources and comparable life histories. But it will be less necessary to retrieve the same individuals and duplicate the same links, and we can instead complement the work of others. The aim is a database where we can follow individuals, families, farms, homes and other locations over time. Not least, the HPR will have a source critical function. Only by linking the sources on the individual level, will it be possible to spot and evaluate the many errors and inconsistencies in the basic source materials. It is not trivial to establish such a population registry, and this chapter describes

[7] https://www.nappdata.org and http://www.ehps-net.eu

[8] http://www.nappdata.org/napp/ and http://www.ddb.umu.se

some of the challenges we face and how they are solved. Record linkage will be done both with automated algorithms and interactively via the Internet. Thus, the HPR represents new technology for collaborating to link personal data.

Acknowledgments We are grateful for contributions from other team member, especially Lars Holden and Torkel Bråthen.

References

Alhaug, G. (2011). *10 001 navn. Norsk fornavnleksikon*. Oslo: Cappelen Damm.

Bråthen, T. R. (2011). Det norske folk i 1814. (The Norwegian people in 1814). *Slekt og Data, 4*, 44–45.

Bull, H. H. (2006). *Marriage decisions in a peasant society: The role of the family of origin with regard to adult children's choice of marriage partner and the timing of their marriage in Rendalen, Norway, 1750–1900.* (pp. 25–34). Unpublished doctoral dissertation. Oslo, Norway: University of Oslo. http://www.rhd.uit.no/nhdc/Chapter%203%20Hans%20Henrik%20Bull.pdf. Accessed 14 Dec 2014.

Eikvil, L., Holden, L., & Bævre, K. (2010). Automatiske metoder som hjelp til transkribering av historiske kilder. (Automatic methods in transcribing historical sources). Notat SAMBA/44/10, Norsk Regnesentral. http://www.rhd.uit.no/nhdc/HBR_notat_okt-2010.pdf. Accessed 12 Jan 2015.

Engelsen, R. (1983). Mortalitetsdebatten og sosiale skilnader i mortalitet. (The mortality debate and social differentials). *Historisk Tidsskrift 62*(2), 161–202. Abstract in English: http://rhd.uit.no/ht/ht62.html#2169. Accessed 18 Mar 2015.

Fure, E. (2000). Interactive record linkage: The cumulative construction of life courses. *Demographic Research.* doi: 10.4054/DemRes.2000.3.11 http://www.demographic-research.org/volumes/vol3/11/3-11.pdf. Accessed 18 Mar 2015.

Gjelseth, M. (2000). *Relasjonsdatabaser som verktøy i en historisk-demografisk studie*. (Relational databases as a tool for a historic-demographic study). Unpublished master's thesis for master's degree. Oslo, Norway: University of Oslo.

Hubbard, W. H., Pitkänen, K., Schlumbohm, J., Sogner, S., Thorvaldsen, G., & van Poppel, F. (2002). *Historical studies in mortality decline*, II. Hist.-Filos. Klasse Skrifter og avhandlinger Nr.3. Oslo: Novus Forlag in association with the Centre for Advanced Study, at the Norwegian Academy of Science and Letters.

Høgsæt, R. (1990). Begravelsesskikker og trosforestillinger i det gamle bondesamfunnet - en feilkilde når en bruker de eldste kirkebøkene til å studere dødelighet? (Burial customs and ideas of faith in the old agricultural community—a source of error when using the oldest parish registers to study mortality?) *Historisk Tidsskrift*, *69*, 130–145. Abstract in English: http://rhd.uit.no/ht/ht69.html#2591. Accessed 18 Mar 2015.

Johansen, H. C. (2002). Identifying people in the Danish Past. In H. Sandvik, K. Telste & G. Thorvaldsen (Eds.), *Pathways of the past: Essays in honour of Sølvi Sogner on her 70th anniversary 15 March 2002* (pp. 103–110). Oslo: Novus.

Kjelland, A., & Sørumgård O.M. (2012). Databases constructed by the Norwegian extended family reconstitution method as part of a national population register for Norway. Paper presented at ESSHC 2012 9th European Social Science History Conference Glasgow, April 2012. http://tilsett.hivolda.no/ak/FamilyReconstitutionDatabases_HPR.pdf. Accessed 22 June 2015.

Mandemakers, K. (2000). Historical sample of the Netherlands. In P. K. Hall, R. McCaa & G. Thorvaldsen (Eds.), *Handbook of international historical microdata for population research* (pp. 149–177). Minneapolis: Minnesota Population Center.

Nygaard, L. (1992). Name standardization in record linkage: An improved algorithmic strategy. *History and Computing, 4*, 63–74.

Quas, D. (2011). WeRelate: Suggestions/tweak to duplicates report. http://www.werelate.org/wiki/WeRelate:Suggestions/Tweak_to_Duplicates_Report. Accessed 4 Dec 2014.

Sobek, M. L. C., Flood, S., Hall, P. K., King, M. L., Ruggles, S., & Schroeder, M. (2011). Big data: Large-scale historical infrastructure from the Minnesota population center. *Historical Methods, 44*(2), 61–68.

Solli, A. (2006, August). *Urban space and household forms*. Paper presented at the eighth international conference on urban history, Urban Europe in Comparative Perspective, Stockholm, Sweden.

Surén, P., Grjibovski, A., & Stoltenberg, C. (2007). *Inngifte i Norge, Omfang og medisinske konsekvenser*, (Consanguineous marriage in Norway—prevalence and medical consequences), Folkehelseinstituttet. http://www.fhi.no/dokumenter/9b8f570dcd.pdf Accessed 20 Dec 2014.

Thorvaldsen, G. (1995). *Migrasjon i Troms i annen halvdel av 1800-tallet. En kvantitativ analyse av folketellingene 1865, 1875 og 1900.* (Migration in the province of Troms 1865–1900. A study based on the censuses). Unpublished doctoral dissertation, Registreringssentral for historiske data, University of Tromsø, Tromsø, Norway.

Thorvaldsen, G. (1998). Historical databases in Scandinavia. *The History of the family*. *An International Quarterly, 3*(3), 371–383.

Thorvaldsen, G. (2000). A constant flow of people? Migration in Northern Norway 1865–1900. *History and Computing, 11*(1–2), 45–59.

Thorvaldsen, G. (2008). Fra folketelling og kirkebøker til norsk befolkningsregister. (From censuses and church protocols to Norwegian population register). *Heimen, 45*, 341–359.

Thorvaldsen, G. (2011). Using NAPP census data to construct the historical population register for Norway. *Historical Methods, 44*(1), 37–47.

Chapter 9
Record Linkage in Medieval and Early Modern Text

Kleanthi Georgala, Benjamin van der Burgh, Marvin Meeng and Arno Knobbe

Abstract This chapter covers the topic of record linkage in historic texts, specifically documents from the Middle Ages and Early Modern period. The challenge of record linkage, in general, is to analyze large collections of data recording people, with the aim of recognizing links between these people, and deciding whether multiple mentions of people actually refer to one and the same person. The typical record linkage application, for example, involving birth and marriage certificates, deals with well-structured descriptions of people in terms of their first and last name, date and place of birth, and so on. In historic texts, however, specifically the medieval ones, people are not identified systematically, and one has to include a lot of the context of the occurrences in order to decide whether two descriptions actually refer to the same historic person. Here, we report on two recent projects, *ChartEx* and *Traces Through Time*, related to these challenges. We have been developing automatic techniques for recognizing links between documents, and determining the confidence that we have in the correctness of these links, based on the evidence provided in the text. Much of the work deals with the varied nature of the evidence, specifically with the role of first and last name being much more limited in the periods involved. We thus had to include identifying properties such as titles, professions, provenances, and family relationships, to determine confident

K. Georgala · B. van der Burgh · M. Meeng · A. Knobbe (✉)
Leiden Institute of Advanced Computer Science, Universiteit Leiden, Leiden, The Netherlands
e-mail: a.knobbe@kiminkii.com

K. Georgala
e-mail: k.georgala@liacs.leidenuniv.nl

B. van der Burgh
e-mail: b.van.der.burgh@liacs.leidenuniv.nl

M. Meeng
e-mail: m.meeng@liacs.leidenuniv.nl

A. Knobbe
Universiteit Utrecht, Utrecht, The Netherlands

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_9

links. This chapter describes the probabilistic record linkage system that was developed for this task, and presents a number of experiments on artificial data to test the workings of the system. Finally, we present some insightful examples of matches that our system was able to find in the Medieval and Early Modern data.

9.1 Introduction

Analyzing and linking collections of historic data is one of the central tasks for a historian. This involves identifying people and their roles in historic events, studying their progress through time, analyzing their relations with other individuals, and constructing their social networks based on the relations obtained from text. Furthermore, the attention drawn to handle large volumes of historic data has led to research into new ways of transforming the original data into large-scale digital archives.

Even though the digitized form of historic collections has served its purpose by helping historians, genealogists, and cultural organizations to have direct access to huge amounts of data, the challenges of linking large sources remain. Manual identification and linking of people, events, and places between large datasets is becoming impossible with growing collections, and fraught with human error. Meanwhile, parsing historic texts without the use of automatic tools limits the possibilities of linking simultaneously different data sources, which can often result in missing hidden and valuable information about individuals and events. Additionally, due to the lack of an automatic cross-reference procedure, the evaluation of the linkage procedure requires human effort that often fails in validating objectively the quality of the results.

Recently, the *ChartEx* project has been developing new ways of analyzing historic documents in an integrated fashion and reconstructing medieval social networks. Specifically, the project's aim was to develop tools to deal with medieval charters: records of legal transactions of property of all kinds (houses, workshops, fields, and meadows). The charters also describe the people who lived there and their relation to others. Long before records such as censuses or birth registers existed, charters were and still are the major resource for researching people, for tracing changes in communities over time and for finding ancestors.

As an extension of *ChartEx*, the project Traces Through Time (TTT) has a stronger focus on people, using an automatic and generalized procedure to link people across different data sources from the same period of time. The main goal of the project is to trace individuals and analyze their "stories" and progress, through large and diverse sources. The historic documents provide circumstantial evidence in order to recognize the characteristics of the people, such as their occupation, their place of origin, their role, and title in the society. This information provides a richer and more enhanced scope of the past: the organization of the society, the essential components of community, and gives a better insight into how changes were

Table 9.1 Projects and data sources

Project	Sources	Historic period	Locations	Language
ChartEx	various[a]	Medieval: twelfth–sixteenth century	England	Latin, trans. English
TTT	TNA[b]	Medieval: eleventh–fifteenth century	UK	Old English + Latin, trans. English
TTT	TNA[b]	Early modern: sixteenth–nineteenth century	UK	English

[a] http://www.chartex.org/

[b] http://www.nationalarchives.gov.uk/

applied throughout history. Additionally, TTT aims to explore the transactions between individuals that are highlighted in data, by constructing the social network of an entity, giving the opportunity to observe family and social relations between people. The challenging aspect of this work lies in the properties of the data: spelling errors, morphological variations, aliases, ambiguity in entity descriptions, linkage between large datasets that vary in format and language.

The basic characteristics of the data sources for *ChartEx* and TTT are described in more detail in Table 9.1. Both projects cover the medieval period. However, *ChartEx* includes several collections of *charters* from the twelfth to the sixteenth century, predominantly from England, whereas TTT mainly focuses on data sources from the eleventh to fifteenth century and covers a larger variety of content such as political, financial, juridical, and economic relations between people and the government. In addition, TTT covers the Early Modern era, from the sixteenth to nineteenth century. The sources include mainly secretariat documents that refer to political and economic relations, internal and external policies of the government, and juridical affairs between the state and people. The medieval datasets were originally written in old English or Latin. The documents that were collected from the Early Modern period were mostly written in more modern English compared to the medieval era. All datasets have been translated by scholars in twentieth century English and these translations were recently digitized.

Besides presenting the record linkage result achieved on these sets of data, we also aim to present our new probabilistic record linkage system, and experimentally evaluate it, both on the historic data as well as on artificial data. The problem with the historic data is that we do not have a ground truth, i.e., an unequivocal list of occurrences that refer to the same person, such that it is hard to evaluate the linkage process in an objective manner. As a result, also the evaluation of the correct setting of a number of parameters in the system is problematic. With the artificial data, our aim was to provide a ground truth of entities that are linked, such that the output of the system can be validated. We went to considerable lengths to make the artificial data mimic the actual data, for example, by using properties derived from the historic data to produce the artificial data. These properties, for example, include lists of statistics concerning first and last names of the period, and statistics about the use of various attributes to describe entities in text at the time, such as titles and towns people originated from. Additionally, we introduced misspellings, and accounted for the fact that some people (e.g., famous people) appear more often in texts than others. It should be noted that the artificial data in its strictest sense is not a golden standard. That would only be the case if we were to have a dataset of guaranteed matches, but in the absence of people from the era that are still alive, this is impossible to obtain. We believe that our realistic data provides the next best solution. In the experimental section, we first demonstrate how the realistic artificial data can be used to evaluate the linker, and analyze the use of various similarity measures that deal with spelling variations and typos.

9.2 Datasets

9.2.1 Medieval Period

The charters in the *ChartEx* collections record transactions of property. Typically, they contain fairly concise descriptions of the grantor and recipient of the transaction, some description of the property involved—often made more precise by mentioning the previous owners—and finally a list of witnesses. The following is an example of a charter from the Vicars Choral collection (pertaining to properties in the York, UK area):

> 408. Grant by Thomas son of Josce goldsmith and citizen of York to his younger son Jeremy of half his land lying in length from Petergate at the churchyard of St. Peter to houses of the prebend of Ampleford and in breadth from Steyngate to land which mag. Simon de Evesham inhabited; paying Thomas and his heirs 1d. or [a pair of] white gloves worth 1 d. at Christmas. Warranty. Seal.
> Witnesses: Geoffrey Gunwar, William de Gerford[b]y, chaplains, Robert de Farnham, Robert le Spicer, John le plastrer, Walter de Alna goldsmith, Nicholas Page, Thomas talliator, Hugh le bedel, John de Glouc', clerks, and others.
> January 1252 [1252/3]

Note that this charter primarily identifies two people, *Thomas, son of Josce, goldsmith and citizen of York* and *Jeremy*, his younger son (Fig. 9.1). The other person mentioned in the body text is *mag. Simon de Evesham* (mag. for magister, an academic title at the time) does not play a direct role, but is mentioned to specify a piece of land that is needed to identify the specific property being transferred here. The transaction relates to half of the land previously belonging to Thomas the goldsmith, and is mostly identified by how it is positioned in relation to other properties or landmarks, that were perhaps easier to identify at the time. Note that Petergate and Steynate are crossing streets that still exist in York as Petergate and Stonegate. The list of witnesses offers some clue as to the people involved, but does not play a major role in our work. Finally, the document is dated fairly accurately, but this is definitely not always the case, and differs from collection to collection, becoming common only after circa 1300.

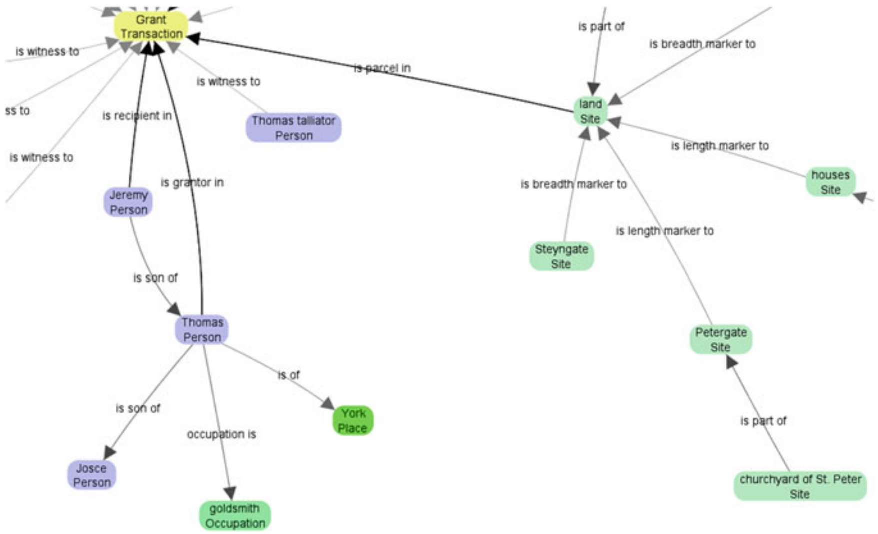


Fig. 9.1 Part of charter 408 from The Vicars Choral, in relational form. Note that this network does not yet involve record linkage across charters

One important thing to note here, is the fact that most information conveyed in a charter is in natural language, something which hinders direct interpretation of the documents and leaves room for ambiguity. For example, one could argue in this text that Josce is in fact the goldsmith and citizen of York, rather than Thomas. Additionally, there is no notion of registered land or geographic coordinates, nor do people have social security numbers, as one would expect in modern legal transactions. To make matters worse, there was no unified spelling of people and place names, such that a considerable level of flexibility will have to be assumed when matching names across documents. Also, the notion of last names was only slowly appearing in medieval England, such that people often are only identified by their first name. In many cases, people’s origin or profession served as last name, such as with William de Gerfordby or Robert le Spicer, but these “last names” did not serve as family names. Needless to say, the unequivocal matching of people and sites across charters is a challenge.

9.2.2 Early Modern Period

The Early Modern period datasets describe a variety of financial, economic, religious, and juridical relations between the state, or in general, governmental sectors and people. They are mostly secretariat documents, from the sixteenth to the nineteenth century, registering people and events that were performed in England,

Table 9.2 Calendar of state papers domestic: Elizabeth, 1581–1590

Date	Text
January 5	2. Information given by P.H. (indorsed B.B.) to Sir Francis Walsyngham. Journey of the Earl of Westmoreland to Rome, and his return to Flanders: he is compared with Campion. Departure of numerous Jesuits and seminaries from Rome for England. Intelligence of Wm. Smythe, Tyrrell, Mr. Gardynar, and Lady Foljambe. Landing of foreign troops in Ireland
January 10, London.	4. Thomas Norton to Sir Fr. Walsyngham. Sends the interrogatories, and thanks him for his letter. Has written twice about Sir G. Peckham and his several petitions, who desires leave to walk upon the leads near his chamber.
January 16	11. Relation of the proceeding in the opening of Parliament, "the "Queenes Highnes, with the L. L. and Bishoppes in Parliamente "robes, did ryde from Her Mats Pallaice at Whitehall to West"minster Churche, and there hearde a sermon"

Scotland, Ireland, and the British Colonies. In TTT, we are focusing on a particular dataset, the *Calendar of State Papers Domestic* mostly from the Elizabethan period that includes documents from 1547 to 1625.

The dataset is divided into eight volumes, where each volume consists of a set of documents in chronological order. Each document describes events that occurred during a month and is divided into sections, as illustrated in Table 9.2. We can observe that each section includes the date of the event *January 10* and regularly the particular location *London*, along with a unique identifier *4*.

The first sentence serves as a title, indicating the nature of the file, such as a letter (*January 5*) or proceedings of a parliamentarian meeting (*January 16*). The remaining text is a copy of the original document and is either copied explicitly (*January 5*) or written as being narrated from a third party (*January 10*, *London.*).

In Table 9.2, we can identify the entities such as *Sir Francis Walsyngham* that is being referenced either as *Sir Francis Walsyngham* or *Sir Fr. Walsyngham*. As you can see, the use of abbreviated fields in popular first names is a common characteristic among the official documents of the Early Modern period. For example, the occurrences *P.H.* or *B.B.* leave a lot of ambiguity when it comes to the true identity of the entity they reference. This can either indicate that the information is missing or that the first and last names were intentionally truncated to conceal their identity.

Observing the text, one can notice the concise sentences that are written in a very comprehensive manner. This writing style introduces many problems, especially in determining relations between individuals. For Section *January 5*, the text implies that the sender and the receiver share a close relation, whose nature cannot be defined explicitly. Additionally, we mostly notice links between people based on written communication and not family relations that were more common in the medieval period.

Furthermore, compared to the medieval documents described in the previous section, we observe a trend of using additional characteristics such as titles (*Sir*) and provenances (… *of Westmoreland*) in order to identify people, which decreases the ambiguity between entities. In the Early Modern era, people were described in a more unique and specific way.

9.3 Record Linkage

Throughout this chapter, we use the term record linkage to refer to the process of disambiguating references from one or several sources that refer to the same object. This problem is known in various fields under different names, such as "data matching" (Winkler 1995; Herzog et al. 2010), "entity resolution" (Singla and Domingos 2006; Bhattacharya and Getoor 2006), "deduplication" (Sarawagi and Bhamidipaty 2002; Christen 2012), and "reference matching" (McCallum et al. 2000). In this section, we will focus on the formalization of the concepts and techniques of record linkage in the context of historic data.

The very first notion of *record linkage* stems from Dunn (1946). He brought attention to the importance of being able to link records of various sources into a single document that he called the "Book of Life". Pages within the Book of Life for an individual describe his or her life from birth to death. In this respect, the goals of the Traces Through Time project are not that different, and indeed the basic approach that we use is very similar.

Within the documents, there are *occurrences* of names, be it first or last names, and other words that refer to a unique person. As we will see, these references can range from being very ambiguous (Mr. Johnson) to more or less uniquely identifying (King Richard I). It is the task of the record linkage system to deduce which of these references refer to the same person. The occurrences can be found using several techniques, such as simple syntactical parsing (e.g., a first name is optionally proceeded by a title and followed by a surname) or the more sophisticated Natural Language Processing (NLP) techniques (Nadeau and Sekine 2007; Ratinov and Roth 2009; Borkar et al. 2001). Since the emphasis in this chapter is not on the extraction of relevant information from text, we will assume that this step has already been performed. For simplicity, but without loss of generality, we can proceed to assume that the occurrences consist of a fixed number of attributes, although any value of a specific attribute might be missing if it is absent in a particular dataset. Figure 9.2 shows an example of an excerpt of a text and the corresponding occurrences.

The observation that two references indicate that the person in question is a king might increase our confidence that these references refer to the same person, since there are only few kings at any specific point in time. Therefore, the attribute *title* with value "king" can be considered highly informative. This is essentially the basis of our probabilistic record linker. It uses statistics, preferably computed on an already disambiguated dataset of the same time period, to compute a confidence score based on the probabilities associated with each of the specific values within an occurrence. If no such source of statistics is available one can be computed using the ambiguous references from the text itself, although this will introduce a bias

(a)

There once was a guy called King Henry of Scotland.

(b)

Title	First name	Last name	Provenance
King	Henry	–	Scotland
…	…	…	…

Fig. 9.2 **a** Example raw text. **b** The tabular format of occurrences extracted from the text

toward people that are mentioned more often in the text. We also allow for variations in names by use of a binary function s that decides whether two values are "similar enough" (Brizan and Tansel 2006).

$$s(v_i, v_j) = \begin{cases} 0 & \text{if} \quad d(v_i, v_j) > t \\ 1 & \text{if} \quad d(v_i, v_j) \leq t \end{cases} \tag{9.1}$$

where d can be any distance function and t is threshold value for that function.

Consider an $m \times n$ table T of occurrences with each row representing one of m occurrences with its respective values in its n columns, i.e., an occurrence o is defined as a tuple $o = (v_1, v_2, \ldots, v_n)$. Furthermore, we define a function that maps value pairs to their associated probability as follows:

$$f(v_i, v_j) = \begin{cases} 0 & \text{if} \quad s(v_i, v_j) = 0 \\ p(v_i) + p(v_j) & \text{if} \quad s(v_i, v_j) = 1 \wedge (\exists (p(v_i) \vee p(v_j))) \\ 0.0001 & \text{otherwise} \end{cases} \tag{9.2}$$

where $v_i, v_j \in V_x \in V$ for all $x \in 1, \ldots, n$ and $p(v_i)$ is simply the prior probability of v_i, if known and 0 otherwise. The function $p_{\text{link}}()$ now describes how likely such a combination of properties, shared by two occurrences o and o^*, is (assuming independence of the separate probabilities):

$$p_{\text{link}}(o, o^*) = \prod_{i=1}^{n} f(v_i, v_i^*) \tag{9.3}$$

Finally, a *confidence score* is computed as the reciprocal of the estimated number of people with such properties at the time:

$$\text{conf}(o, o^*) = \frac{1}{N \cdot p_{\text{link}}(o, o^*)} \tag{9.4}$$

Note that this confidence score depends on a proper estimate of the value of N: the number of people who can be expected to occur in the documents in question at the indicated time. $N \cdot p_{\text{link}}(o, o^*)$ is then an estimate of the number of people that can be expected to fit the properties shared by the two occurrences. Assuming a proper estimate of N, a confidence of 0.5 would then indicate that we are dealing with a reasonably reliable match, but we cannot be absolutely sure, since an estimated two people fit the description. A much lower value of 0.1 would indicate that about 10 people fit the description, so confusion between different people with the same properties is quite likely. Note that values above 1.0 can easily occur, although they would strictly indicate that less than one person is expected to have the specified properties. One should keep in mind here that the probability $p_{\text{link}}(o, o^*)$ can be arbitrarily small, if sufficiently restrictive evidence is provided (and matched). When the probability is, say, one in a million, and we assume a population of a thousand people, the estimated number of people with this property becomes 0.001. In this hypothetical situation, the confidence score is $\text{conf}(o, o^*) = 1000$. One could of course cap the confidence score at 1.0, if one was so inclined, but this removes the distinction between fairly confident and extremely confident.

In our experiments pertaining to the Early Modern data, we (somewhat arbitrarily) set $N = 10,000$. Although this value plays a large role in the confidence scores produced by the linker, note that a different value of N does not change the relative scoring of candidate matches. In other words, it is best not to interpret the confidence scores in an absolute sense, but rather as a *ranking* of candidate matches: a higher confidence indicates a more likely match between occurrences.

In the presence of errors and variations in names, it is apparent that the suggested procedure relies on both the similarity measure used and the quality of the statistics (Cohen et al. 2003; Köpcke and Rahm 2010). In Sect. 9.4 we will suggest several similarity measures.

9.4 Experiments

Before presenting a number of results obtained on the Medieval and Early Modern data, we first examine the record linker and how it deals with various degrees of information and uncertainty within that, for example, caused by spelling variation and typos. For this purpose, we have created an artificial dataset with ground truth.

9.4.1 Testing the Procedure

The investigation of the properties of our record linkage system requires a set of the unambiguous entities along with their corresponding occurrences. Additionally, the process of validating the linking results from the original datasets requires the presence of a group of experts that can cross-reference the outcome using specific in-domain history knowledge. We have produced an artificial dataset to evaluate our record linkage procedure, in order to study our system behavior in a controlled manner.

The artificial dataset is constructed to mimic the *Calendar of State Papers Domestic* dataset, from the Early Modern period. It consists of a set of 10,000 entities, from which our algorithm produced 20,000 occurrences according to a Zipf

Table 9.3 Frequency of attributes in occurrence

Attributes	Frequency (%)
Title	53
First name	50
Article	1.31
Last name	68
Role	12.7
Provenance	1.9

distribution (so on average two occurrences per entity, some with more occurrences, many with fewer). The data thus includes the notion that there might be a small set of important people whose occurrence in the documents is much more common.

Each entity was constructed using six fields: *title, first name, article, last name, role,* and *provenance*. In order to be consistent with the *Calendar of State Papers Domestic* data, the values for each field were drawn according to the distribution derived from the actual data. The creation of artificial occurrences was based on the idea of maintaining the structure and the characteristics of the original occurrences found in texts. In order to produce the sets of occurrences for our experiments, we used two techniques:

- Extracting one or more fields of the original occurrence, based on how often the field was assigned with value in the original set of occurrences (see Table 9.3).
- Replacing the value with its abbreviation (e.g., *Bart.* for *Bartholomew*), or by choosing to perform a spelling variation (e.g., *Goodrich* for *Godric*) on the initial value consulting a list of variants and trigram transformations, or random replacement of a letter. This variation was applied with a probability of 1.5 %.

Each dataset was accompanied by a set of statistics for each field, obtained by calculating the frequency of the unique values found in the artificial occurrences. Based on the observed presence of the various attributes in the original dataset, the following probabilities were applied to filling or leaving empty each attribute:

9.4.1.1 Similarity Measures

As mentioned in Sect. 9.3, one important aspect of our record linkage system is the decision criterion to establish if two occurrences describe the same entity. In order to determine the similarity between two occurrences, we have to compare the pair of occurrences by matching the values of one or more attributes. The performance of our record linkage procedure is influenced by the manner that various similarity measures (Brizan and Tansel 2006) determine whether two values of an attribute can be considered “similar” (Köpcke and Rahm 2010).

The most obvious similarity is the *Absolute Distance* function that considers two values similar if and only if their values are identical. However, we intentionally included spelling variations, grammatical errors, and name abbreviations in the artificial data. Therefore, we also consider other common similarity measures that define similarity in a different and more flexible manner. First, we consider *Soundex* (National Archives and Records Administration 2007), which is a phonetic algorithm that matches similar values with minor differences in spelling. Additionally, we include a set of similarity measures that belong to the edit-distance family of similarity measures:

- *Levenshtein* (*Leven*). The *Levenshtein* distance assumes that two values are similar if the number of single-character edits (insertions, deletions, or substitutions) required to transform the first value to the second value is less than a threshold *transp* (Levenshtein 1966).
- *Damerau–Levenshtein* (*DamLeven*), that is a modification of the *Levenshtein* distance that additionally defines as a possible single-character edit the transposition of two adjacent characters (Bard 2007).
- *Jaro–Winkler* that considers two values as similar if the weighted combination of the matching characters and transpositions is greater than a threshold *thrs* (Winkler 1990).

Finally, we include a similarity measure from the *QGrams* family of algorithms, which considers two values to be similar if the normalized sum of the equal trigram between the vectors of the two values is greater than a threshold *thrs*.

Since our algorithm incorporates four similarity measures with adjusted thresholds, our first experiment focuses on exploring their performance using different values for the parameters. For *Leven* and *DamLeven*, threshold *transp* is assigned values 1, 3, 5, 10, 20 and for *Jaro–Winkler* and *QGrams*, threshold *thrs* is assigned values 0.1, 0.3, 0.5, 0.7, 0.9, 1. The cost for each edit was set at 1.

The performance was measured by constructing Receiver Operating Characteristic (ROC) curves and by estimating the Area Under Curve (AUC) for each curve (Hanley and McNeil 1982). ROC is a graphical plot that is commonly used to illustrate the performance of a binary classifier. The curve is created by plotting the True Positive Rate (TPR = fraction of true matches correctly predicted) against the False Positive Rate (FPR = fraction of false matches incorrectly predicted). However, our record linker works as a ranker: it outputs a confidence score for each pair of occurrences using Eq. 9.4 and our system ranks those pairs in descending order. A ranker can be easily transformed into a classifier by setting a threshold value on the output of the system: every pair with confidence score above the threshold is considered a match (True Positive or False Positive) and every pair with confidence score below the threshold is considered a nonmatch. In our case, the set of threshold values is obtained using the set of unique confidence scores between occurrences. Therefore, for each threshold, we obtain the TPR and FPR and finally plot the ROC curve (Flach 2003).

Figures 9.3, 9.4, 9.5 and 9.6 illustrate the ROC curves for the four similarity measures, for different values of their thresholds, and Table 9.4 presents the associated AUC values for each curve.

From Figs. 9.3 and 9.4, we notice the similar performance of *Leven* and *DamLeven* for the different values of *transp*. As explained previously, *DamLeven* is

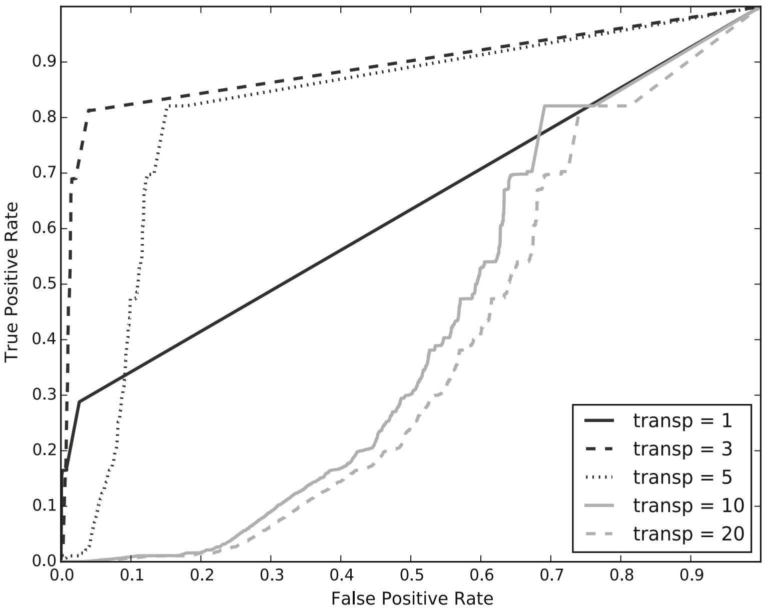


Fig. 9.3 ROC curve for *Levenshtein*

a variation of the *Leven* metric. Since the artificial data includes spelling variations, of which only few are actual transpositions between two adjacent characters, this similar performance is understandable.

Note that when allowing two values to be considered similar after performing at least five single-character edits, the record linkage algorithm falsely assumes that many occurrences reference the same person. On the other hand, if we over-restrict the number of allowed operations to one our record linker fails in capturing occurrences of the same entity, resulting in a decrease in the True Positive rate.

Continuing with our next similarity measure, our first observation about the experiments conducted for the *Jaro–Winkler* is that the amount of False Positives increased as the *thrs* gets assigned with a very low or a very high value. However, in contrast to *Leven* and *DamLeven*, over-restricting the *thrs* results in excluding a huge amount of falsely matched occurrences from the resulting set of pairs. The performance of the algorithm achieves the highest AUC score with *thrs* = 0.7, based on Table 9.4. For our final similarity measure, *QGrams*, the most noticeable observation is that the record linkage algorithm captures the most True Positive pairs only if the threshold is very small (*thrs* = 0.1). In other words, *QGrams* assigns a very low similarity score between similar attributes, especially if they include a spelling variation. Additionally, it allows more false positives to be present than any other edit-distance similarity metric.

After determining the optimal parameter setting for each of the similarity measures that incorporate a threshold, we also computed the performance for the

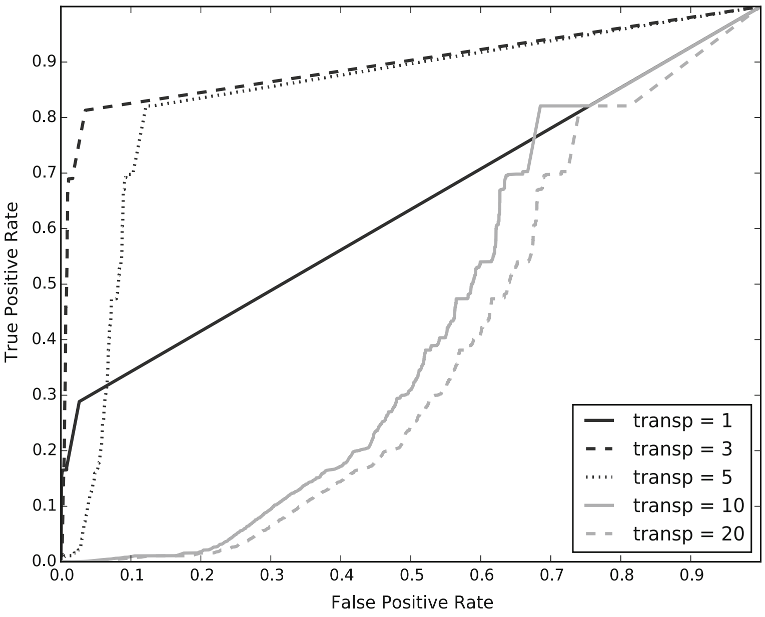


Fig. 9.4 ROC curve for *Damerau–Levenshtein*

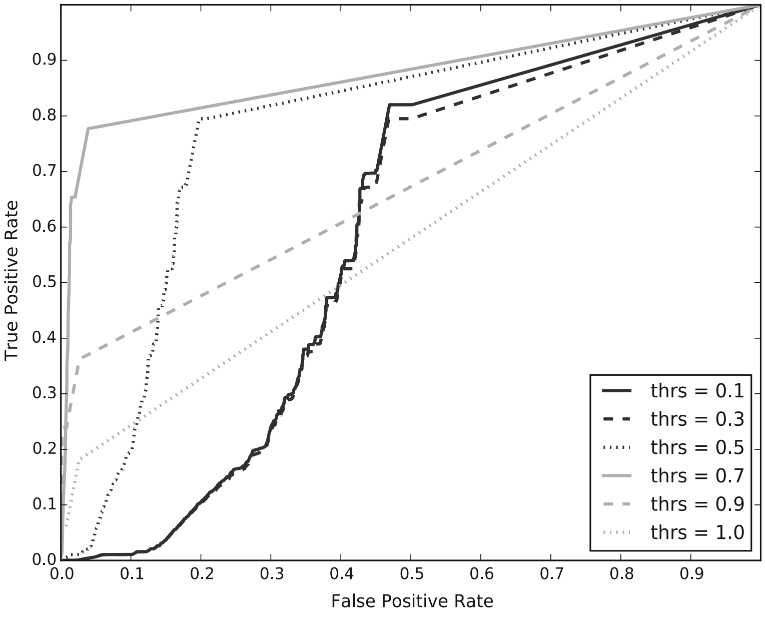


Fig. 9.5 ROC curve for *Jaro–Winkler*

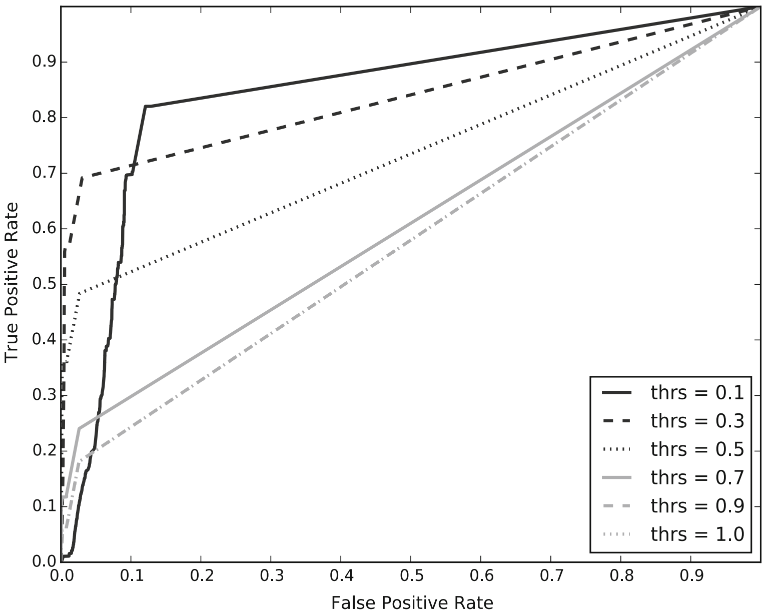


Fig. 9.6 ROC curve for *QGrams*

Table 9.4 AUC for each ROC curve of *Leven*, *DamLeven*, *Jaro–Winkler,* and *QGrams*

Method	AUC	Method	AUC
Leven, *transp* = 1	0.632433	*DamLeven*, *transp* = 1	0.63272
Leven, *transp* = 3	0.89223	*DamLeven*, *transp* = 3	0.89594
Leven, *transp* = 5	0.8157	*DamLeven*, *transp* = 5	0.83995
Leven, *transp* = 10	0.418235	*DamLeven*, *transp* = 10	0.42344
Leven, *transp* = 20	0.379297	*DamLeven*, *transp* = 20	0.3793
Jaro–Winkler, *thrs* = 0.1	0.58363	*QGrams*, *thrs* = 0.1	0.843617
Jaro–Winkler, *thrs* = 0.3	0.57395	*QGrams*, *thrs* = 0.3	0.836599
Jaro–Winkler, *thrs* = 0.5	0.7747	*QGrams*, *thrs* = 0.5	0.73247
Jaro–Winkler, *thrs* = 0.7	0.875145	*QGrams*, *thrs* = 0.7	0.608214
Jaro–Winkler, *thrs* = 0.9	0.67041	*QGrams*, *thrs* = 0.9	0.577945
Jaro–Winkler, *thrs* = 1	0.57781	*QGrams*, *thrs* = 1	0.57781

two remaining algorithms. Figure 9.7 gives an overview of the results and Table 9.5 summarizes the AUC scores, in order of performance.

As expected, *QGrams* receive the worst performance and the lowest AUC score among the edit-distance functions. The reason lies in the naive nature of the *QGrams* comparison strategy. *QGrams* match partially two values, without weighting the quality of the comparison. Therefore, it assumes that some common

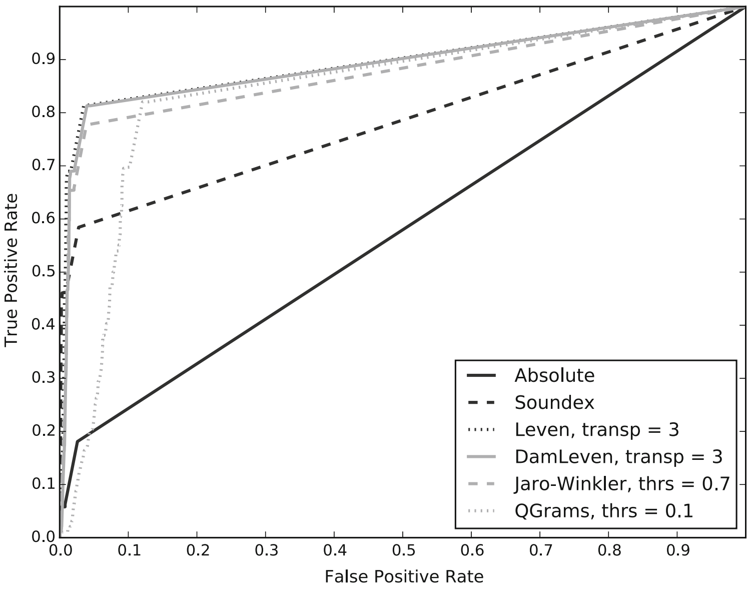


Fig. 9.7 ROC curves for all similarity measures

Table 9.5 AUC of the ROC curve for each similarity measure

Order	Method	AUC
1	*DamLeven*, *transp* = 3	0.8959
2	*Leven*, *transp* = 3	0.8922
3	*Jaro–Winkler*, *thrs* = 0.7	0.8751
4	*QGrams*, *thrs* = 0.1	0.8436
5	*Soundex*	0.7835
6	*Absolute distance*	0.5778

prefixes, suffixes, or infixes are enough evidence to assume that two occurrences can be references to the same entity with high confidence.

Regarding the performance of the other edit-distance measures, we notice that *DamLeven* and *Leven* outperform *Jaro–Winkler*. In contrast to the other edit-distance measures, *Jaro–Winkler*'s formula consists of adjusted weights only between the transformations and the common characters of two values, whereas *Leven* and *DamLeven* assign costs to each possible single-character operation.

For our final set of comparisons, observing the ROC curves of *Absolute Distance* and *Soundex*, we notice a nonsimilar performance of the two similarity measures. Both similarity measures perform string matching in a very similar manner. However, *Soundex* equates values if their phonetic representation is the same, covering cases of common misspellings. Therefore, the AUC score of *Soundex* is higher compared to *Absolute Distance*, which allows only equal values to be considered a match and receives the worst AUC score overall.

For our last set of experiments on the artificial data, we explore how our record linkage system can be influenced by the quality and quantity of information included in an occurrence. We conduct a set of experiments in which we add incrementally the probability of an attribute in the confidence estimation. As similarity measure we choose *DamLeven*, *transp* = 3, which based on our previous experiments had a very good performance in terms of AUC score. Additionally, *DamLeven* is widely used as a similarity measure for record linkage and entity resolution.

The set of artificial occurrences was constructed using attributes that were present in the original occurrences, whose values were filled using the probability of an attribute being assigned with a value in the set of the 60,000 occurrences (Table 9.3). In our initial experiment, we include only *first name* and *last name*, since they are the basic features for identifying a person. Then, we incrementally add *article, title, role,* and *provenance*. Table 9.6 includes the corresponding AUC scores.

We observe that adding the *article* attribute, the performance of our system does not change dramatically, since it is an attribute that appears the least often in data and it is less informative in describing an entity.

Eventually, the use of *title* seems to slightly boost the confidence of the record linker by increasing the AUC score and merely resulting in higher confidence scores between occurrences of the same entity. The use of titles provided extra information about the status of an entity and it was a very common custom to refer to people by their title. Furthermore, the extra information provided by *role* improves a bit further the linking process of our system. A very common problem in the original data is the use of titles to describe a person without any extra information about their first or last name. Therefore, this situation could result in great ambiguity between occurrences. Finally, the presence of *provenance* does not produce any improvement.

In conclusion, based on Table 9.6, the ability of our record linkage system to assign higher scores to occurrences of the same entity and identify occurrences that refer to the same person is influenced by adding attributes to the confidence estimation that are highly informative. The quality of information is based on the frequency of an attribute being assigned with a value and therefore, the *first name, last name, title* attributes provide enough evidence in order to resolve ambiguities

Table 9.6 AUC of the ROC curves for *DamLeven*, *transp* = 3

Attributes	AUC
First name, last name	0.8316
First name, last name, article	0.8316
First name, last name, article, title	0.8933
First name, last name, article, title, role	0.8959
First name, last name, article, title, role, provenance	0.8959

between occurrences. Additional attributes do provide additional AUC, but the benefit is not substantial.

9.4.2 ChartEx Data

For the Medieval data concerning charters, we had five separate collections at our disposal. Although these collections are typically fairly large, we work with more modest subsets of charters that were meticulously annotated by hand by historians. These subsets were specifically chosen for their known connectedness. Other than serving as a test set for our linkage approach, these annotated documents were also intended to serve as input to the Natural Language Processing (NLP) component of the project. The eventual goal was to automatically annotate the remaining charters in the vein of the manual annotation. Since the automatic annotation was incomplete and introduced a certain level of inaccuracy, we only present results on the manually annotated charters. The five collections are:

- The Vicars Choral (University of York), 125 charters manually annotated, English, 5000 charters (dated).
- Borthwick (Borthwick Institute, University of York), 55 charters manually annotated, English.
- DEEDS (University of Toronto), 49 charters manually annotated, Latin, over 10,000 charters.
- Wards2 (The National Archives, UK), 48 charters manually annotated, English, 7000 charters.
- Cluny (University of Columbia), 50 charters manually annotated, Latin, over 5000 charters (dated).

Before linking any actual occurrences, an initial analysis of the data was done to produce statistics, and to get a general understanding of the data. In the English documents, a total of 112 different first names occur, where we assume that different spellings are different names. Of these names, the gender of over 85 % could be inferred from the context in which they appeared (for example, *Thomas, son of Josce* implies Thomas is a male name). Of the names for which the gender was resolved, 36 % was female. It should be noted though that in absolute sense, women were much less mentioned than men, especially where ownership of property is concerned. In a ranking of names according to their frequency, the first female name (Margaret) appears at rank 15. Also, the common name John is over 17 times more common than Margaret. This medieval gender difference is also indicated by the occupation statistics, where the first clear female "occupation" (an annotation that was used somewhat liberally in *ChartEx*) is "widow" at rank 12, after clearly male occupations such as "yeoman" and "esquire".

We present a number of results from the Vicars Choral collection, which also featured in the introduction. Concerning the example in Sect. 9.2 of charter 408, and related charter 409, the following probabilities hold for *Thomas son of Josce,* *Goldsmith*, who appears in both charters. These probabilities were obtained by collecting statistics on names and professions from the five collections.

p(*Thomas*) = 0.12 (common name)
p(*Josce*) = 0.0015 (uncommon name)
p(*Goldsmith*) = 0.04 (common profession)

Note that *Thomas* itself is not a very informative name, but the fact that his father is called *Josce* makes up for this. Goldsmiths were fairly common in charters of the time, as they were typically fairly rich and acted as the medieval equivalent of bankers. Joining the individual statistics, we get the following confidence score, which indicates that the two occurrences actually fairly likely refer to the same person:

Thomas, son of Josce, Goldsmith, score: 0.9993

Another confident link is made between two occurrences of *Robert Warde*, who appears with identical spelling, and has as additional (linked) evidence:

is of York
is a merchant
his widow is named Beatrice (apparently Robert Warde was deceased at the time of writing)

Repeating the procedure for all candidate pairs in the collection, we find a reasonable number of confident matches (depening on one's choice of a lower bound on *conf*). In many cases, several occurrences within a charter can be linked to occurrences in other charters, such that certain charters actually become connected through several people mentioned simultaneously. As a result, a network of connections between charters starts to appear (see Fig. 9.8), which supports the historian in constructing a logical sequence of events, in the case of charters, of sales and inheritance of property in the Middle Ages.

9.4.3 Early Modern Data

For the Early Modern data, we use exactly the same set of attributes as for the artificial data. For identifying the occurrences in the natural text, we parse the documents using a combination of existing parsing tools and more sophisticated NLP techniques. Additionally, in order to identify the components of an occurrence, we assign part-of-speech (POS) tags to each token (Perrow and Barber 2006), consulting a set of early modern first names, last names, titles, and provenances from our own constructed lists and collections from TNA. The list of articles (*the, le*, etc.) is constructed manually observing the various articles that appeared in text. The attribute of *role* can be defined implicitly, by combining various POS-tags.

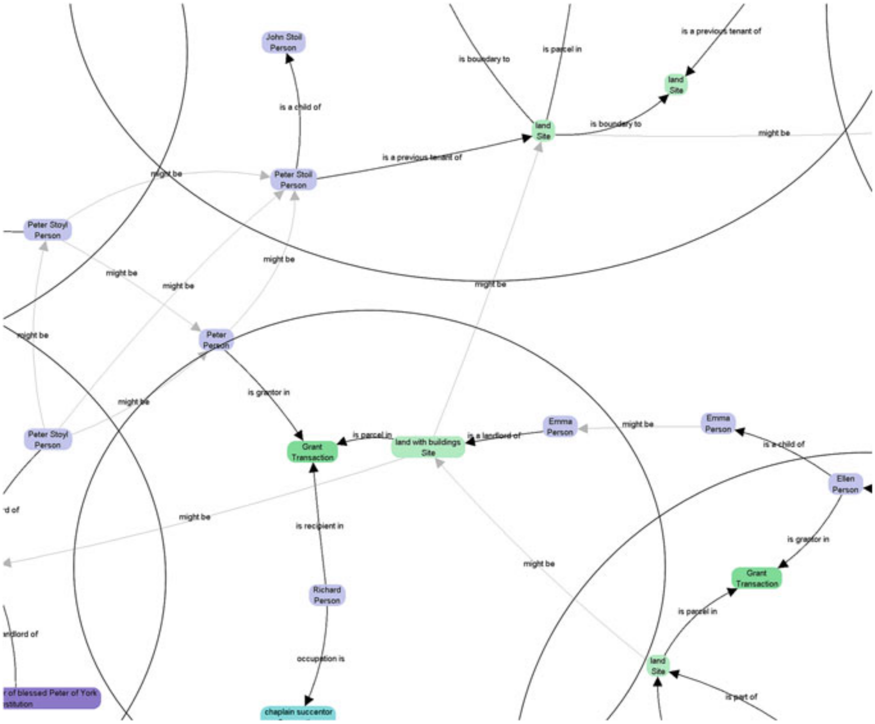


Fig. 9.8 Partial network generated from a subset of seven charters, known from the literature to describe related transactions. The ovals roughly cover each charter involved. *Gray lines* indicate hypothesized links between people in various roles in the charters. Note some of the spelling variations

As a result of the parsing procedure, we obtain a set of 60,000 occurrences from the text. Additionally, as described in Sect. 9.3, our record linkage procedure requires a set of statistics for each field in order to assign the final confidence score between two occurrences. The *Calendar of State Papers Domestic* dataset includes a partially deduplicated index of entities, from which we obtain statistics for each field.

In order to compute the confidence score between two occurrences, we used the statistics for each attribute obtained from the partially deduplicated index of the original data. In these experiments, we incorporated each of the similarity measures mentioned above, using the configuration that produced the highest performance in the artificial data.

The performance of our system could not be evaluated in the same manner as with the experiments on the artificial data, due to our lack of cross-reference techniques. Nevertheless, our intention behind these experiments is to explore the behavior of our algorithm in the presence of data from actual historic documents and observe how the different similarity measures influence the linkage process.

For our first example, we choose to observe the linkage results for a well-defined occurrence that consists of three common attributes (*title, first name, last name*): *Sir Wm Paget*. Each field includes relatively highly frequent values:

p(*title* = Sir) = 0.135
p(*firstname* = Wm) = 0.005065
p(*lastname* = Paget) = 0.0002445

Based on the linkage results, regardless of the similarity measure, our system assigns the high confidence score of 74.46 between *Sir Wm Paget* and other identical occurrences (*Sir Wm Paget*). Also, using the same configuration, our record linker links the occurrence of *Sir Wm Paget* with *Wm Paget, Sec of State,* and *Wm Paget* with the confidence score of 20.15. Even though the pair of occurrences shares only two common attributes with the exact same values, we notice that our system assigns between the two occurrences a lower confidence score. However, the matching results in obtaining additional information about the entity in description (*Wm Paget*), such as that it was addressed with the title Sir and his occupation was Sec of the State.

Also, the *Leven* and *DamLeven* distance functions identify the occurrence *Lord Paget* as a probable match to *Sir Wm Paget* with confidence 0.6456. The low confidence indicates that we cannot assume with high certainty that the two occurrences refer to the same entity since, *Lord Paget* is a title that could have been assigned to people with the same family name (*Paget*) over time.

Given the differences in the scores between *Sir Wm Paget* and the other similar occurrences, we notice that our system assigns a higher confidence between occurrences that share primarily the same subset of attributes, and especially it promotes pairs of occurrences whose values are also quite similar.

For our next example, we present the linking results for the occurrence *Robt Southwell Sheriff of Kent*. The special characteristic of this occurrence is that the text included only one occurrence with the exact same characteristics and many other occurrences with similar features:

p(*firstname* = Robt) = 0.00204
p(*lastname* = Southwell) = 0.0003365
p(*role* = SheriffofKent) = 0.000462

All of the similarity functions managed to link the pair *Robt Southwell Sheriff of Kent* with its identical occurrence *Robt Southwell Sheriff of Kent*, with very high confidence: 36360.645. Therefore, we can confirm that our record linker prioritizes in assigning higher confidence values between occurrences that share the same attributes with similar values. Additionally, the high value of the confidence score is a result of the high informative attributes values included in the occurrence.

Additionally, our system was able to match *Robt Southwell Sheriff of Kent* with the occurrence *Sheriff of Kent* with the low score of 0.108. However, even though the two occurrences share a common field with the same value, *Sheriff of Kent* is an ambiguous occurrence, because it includes no additional information about the first and the last name of the entity that it references to.

Observing the linkage results by using any edit-distance function as similarity measure, we notice that our system matches *Robt Southwell Sheriff of Kent* with the following occurrences:

Sir Ro Southwell, score: 5.8963
Sir Rob Southwell, score: 13.123
Sir Robt Southwell: 36.394

Using edit-distance functions to determine similarities between values, our system managed to recognize that the tokens *Ro* and *Rob* are common abbreviations of the token *Robt*.

Additionally, *Absolute Distance* or *Soundex* were also able to identify as potential match the pair *Robt Southwell Sheriff of Kent* and *Sir Robt Southwell*. However, in case of a system configuration that uses either these two methods as similarity functions, our linker matches only well-defined pairs of occurrences based mainly on how similar their matching attributes are, without considering the quality of information they can provided about the entity in question.

9.5 Conclusion

We have presented a system for record linkage over historic documents from various periods. Specifically, we have demonstrated its workings on Medieval and Early Modern data within two projects, *ChartEx* and the more recent Traces Through Time (a cooperation with The National Archives UK), which is still ongoing. In that respect, our efforts on the probabilistic record linker are work in progress.

The experiments demonstrate that our system is influenced by the choice of similarity measure. *Qgrams* is less sophisticated and very naive, therefore is not a good option for an edit-distance function. *AbsoluteDistance* and *Soundex* do not sufficiently support the real nature of data which is full of spelling variations, abbreviations, and typos, although reasonable results can be obtained with either choice. Edit-distance algorithms are more flexible and produce good results in realistic data because they compensate for the challenges mentioned. The method of evaluating the choice of similarity measure by means of ROC analysis appears very useful. This method can also be used to evaluate other choices or parameters in a record linkage system.

A particular contribution of this chapter is the realistic artificial dataset that was produced. It reflects many of the properties of the actual data, while offering the possibility of validation, something which is practically impossible for historic data of past centuries. Another possibility of the artificial data is that one can influence various aspects of the distributions, such that one can, for example, test the effect of adding more misspellings or introducing highly important people in the "population".

One important property of our probabilistic approach is the assumption that probabilities assigned to evidence (names, occupations, etc.) are independent, which is not necessarily realistic. Although the effect of this assumption is fairly moderate, we are considering an approach that accommodates for possible dependencies between attributes. Consider for example a dependency between first and last names. In the English data, one can assume that different ethnic background will produce different sets of both first and last names. For a person of Welsh decent, different probabilities hold for the first names. The same is true for the last name, such that the probabilities for the two attributes are not entirely independent. One approach that we are currently investigating is so-called *biclustering* over the matrix of joint probabilities of first and last names. Most likely, different ethnic groups and social classes will show up in the clusters of first and last names.

Another challenge in our current approach is the influence of occurrence on the statistics. Specifically, important people will occur more often in the text, such that statistics derived from the text itself will show a certain bias toward such people. We are working on an iterative approach that starts with the (somewhat) biased statistics, and then links the occurrences with the highest confidence. Once these safe links are established, the statistics are recomputed from the text, but the linked occurrences are no longer counted as separate cases, but rather as a single person. This process is then repeated until the statistics converge to a representation of the actual population, rather than the occurrences over this population as provided by the text.

References

Bard, G. V. (2007). Spelling-error tolerant, order-independent pass-phrases via the Damerau-Levenshtein string-edit distance metric. In L. Brankovic & C. Steketee (Eds.), *Fifth Australasian Information Security Workshop* (*Privacy Enhancing Technologies*) (*AISW*) (Vol. 68, pp. 117–124). Ballarat, Australia: CRPIT, ACS.

Bhattacharya, I., & Getoor, L. (2006). A latent dirichlet model for unsupervised entity resolution. In *Sixth SIAM Conference on Data Mining* (Vol. 5(7), pp. 47–58). Bethesda, MD, USA.

Borkar, V., Deshmukh, K., & Sarawagi, S. (2001). Automatic segmentation of text into structured records. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data* (pp. 175–186). New York, USA: ACM.

Brizan, D. G., & Tansel, A. U. (2006). A survey of entity resolution and record linkage methodologies. *Communications of the IIMA, 6*(3), 41–50.

Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering, 24*, 1537–1555.

Cohen, W., Ravikumar, P., & Fienberg, S. (2003). A comparison of string metrics for matching names and records. In *Proceedings of the Workshop on Data Cleaning and Object Consolidation at the International Conference on Knowledge Discovery and Data Mining, KDD* (pp. 73–78). Washington, USA.

Dunn, H. L. (1946). Record linkage. *American Journal of Public Health and the Nation's Health, 36*, 1412–1416.

Flach, P. A. (2003). The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *Proceedings Twentieth International Conference on Machine Learning* (*ICML'03*) (pp. 194–201). California, USA: AAAI Press.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29–36.

Herzog, T. N., Scheuren, F., & Winkler, W. E. (2010). Record linkage. In *Wiley Interdisciplinary Reviews: Computational Statistics* (pp. 535–543). New York, USA: Wiley.

Köpcke, H., & Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data & Knowledge Engineering, 69*, 197–210.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*, 707–710.

McCallum, A., Nigam, K., & Ungar, L. H. (2000). Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD '00*) (pp. 169–178). New York, NY, USA: ACM.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*, 3–26(24).

National Archives and Records Administration (2007). *The Soundex Indexing System*. http://www.archives.gov/research/census/soundex.html. Accessed May 30, 2007.

Perrow, M., & Barber, D. (2006). Tagging of name records for genealogical data browsing. In *Proceedings of the Sixth ACM/IEEE-CS Joint Conference on Digital Libraries* (*JCDL '06*) (pp. 316–325). New York, USA: ACM.

Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (*CoNLL '09*) (pp. 147–155). Stroudsburg, PA, USA: Association for Computational Linguistics.

Sarawagi, S., & Bhamidipaty, A. (2002). Interactive deduplication using active learning. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD '02*) (pp. 269–278). New York, USA: ACM.

Singla, P., & Domingos, P. (2006). Entity resolution with Markov logic. In *Proceedings of the Sixth International Conference on Data Mining* (*ICDM '06*) (pp. 572–582). Piscataway: IEEE.

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)* (pp. 354–359).

Winkler, W. E. (1995). Matching and record linkage. In *Business survey methods* (pp. 355–384). New York: Wiley.

Part III
Life Course Reconstruction

Chapter 10
Reconstructing Lifespans Through Historical Marriage Records of Barcelona from the Sixteenth and Seventeenth Centuries

Francisco Villavicencio, Joan Pau Jordà and Joana M. Pujadas-Mora

Abstract This chapter presents a methodology for reconstructing the lifespan of individuals through a nominal record linkage procedure using historical marriage records of Barcelona from the sixteenth and seventeenth centuries. The data are extracted from the Barcelona Historical Marriage Database (BHMD), a unique source that contains information about more than 600,000 marriages celebrated in both urban and rural parishes of the Barcelona area from over 450 years (1451–1905). We discuss the main characteristics of the database, the standardisation of the nominal information, the marriage linkage procedure and the reconstruction of the lifespans. Finally, we briefly introduce an application of Bayesian models to study adult mortality on the basis of the reconstructed lifespans.

10.1 Introduction

The census of Floridablanca, which was carried out in 1787, was the first modern census conducted in Spain and one of the most pioneering censuses in Europe at that time (Dopico and Rowland 1990; Simon Tarrés 1996). Prior to this point, enumerations of hearths and households have been conducted in Catalonia since the mid fourteenth century (Feliu 1999; Nadal and Giralt 2000). However, as they were carried out for tax and military purposes, they omitted certain social groups, excluded women, failed to record socio-demographic variables such as age or occupation and there was a lack of continuity which has hampered the study of

F. Villavicencio (✉)
Max Planck Institute for Demographic Research, Rostock, Germany
e-mail: villavicencio@demogr.mpg.de

J.P. Jordà · J.M. Pujadas-Mora
Centre for Demographic Studies, Autonomous University of Barcelona, Bellaterra, Spain
e-mail: jpjorda@ced.uab.es

J.M. Pujadas-Mora
e-mail: jpujades@ced.uab.es

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_10

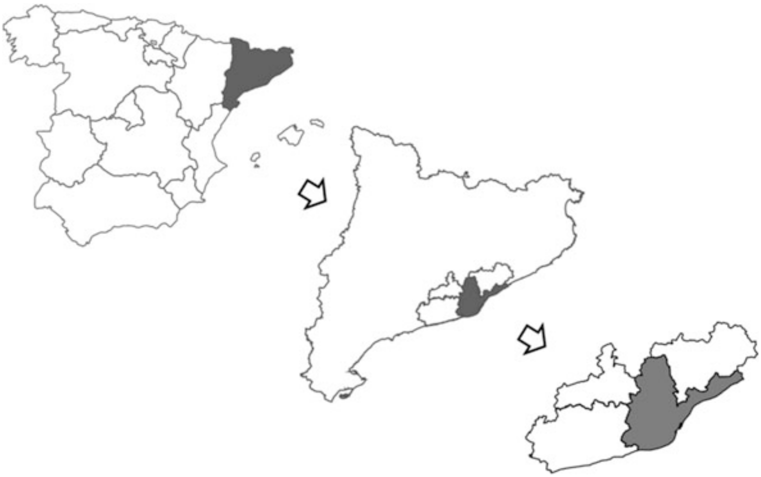


Fig. 10.1 Catalonia, Spain—The diocese of Barcelona—Main deanship of Barcelona

historical populations. To make matters worse, most of the parish registers that included vital events (*Quinque Libri*), and which had been compulsory since the Council of Trent (1545–1563), were not properly preserved due to the wars that ravaged Catalonia until the nineteenth century (Nadal and Giralt 2000).

In order to address this lack of specific demographic data, historical registers of marriage licenses from the diocese of Barcelona are a very rich source that offer interesting research opportunities. Between 1451 and 1905, those marriages were recorded in a set of 291 books conserved at the archive of the Barcelona cathedral and known as *Llibres d'Esposalles–Marriage License Books* (Baucells 2002). These books contain information about more than 600,000 marriages celebrated in over 250 urban and rural parishes of the Barcelona area (Fig. 10.1). The substantial volume of information available in the Marriage License Books was used to create the Barcelona Historical Marriage Database (BHMD), a unique database covering a period from over 450 years which was developed within the project *Five Centuries of Marriages* (Cabré and Pujadas-Mora 2011).

Each marriage record includes a range of information about the bridal couple, such as first names, surnames, occupations and places of residence. For some periods, additional information concerning the couples' parents is available, which allows for a nominal record linkage among marriage licenses.

The 32-year period from May 1597 until April 1629 (volumes 59–74) is of particular interest to researchers due to the quality and completeness of the data. For that period, the information in the marriage records about the parents of the spouses includes a remark about whether they were dead at the time of their children's marriage. Therefore, it is possible to have information on a given parent—his or her marriage, and the marriages of the offspring—while knowing whether he or she was alive at the latter events.

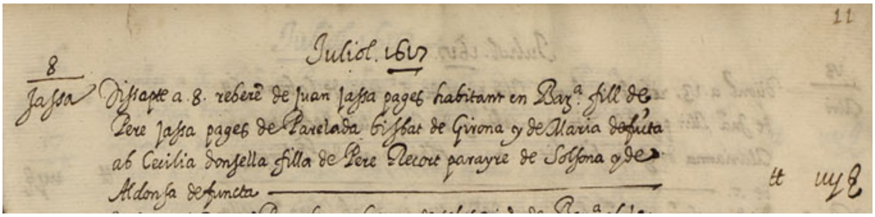
Even though the ages of the couples are unknown, once the marriage linkage is made, we are able to establish for each bride and groom some lower and upper bounds of birth based on the dates of the parents' marriages and the couple's own marriage, as well as some lower and upper bounds of death based on the last time the individual is observed alive, and the first time he or she is mentioned as dead in one of the children's marriages. In order to carry out this process, bridal couples from marriage records between 1573 and 1617 are used as the set of individuals who are expected to be found as parents of grooms or brides in marriage records corresponding to the 32-year period from 1597 to 1629 mentioned above. The final result is the reconstruction of the lifespans of individuals who married between 1573 and 1617 and who had children who married between 1597 and 1629. The whole period from 1573 to 1629 consists of 49,472 marriage licenses.

The reconstructed lifespans are being used in an ongoing research project in which adult age-specific mortality and life expectancy for data with unknown ages are estimated using Bayesian probabilistic models. A few previous studies have estimated life expectancy in Catalonia in the seventeenth century by reconstructing the population of some small Catalan parishes (Muñoz Pradas 1990; Torrents 1993). However, these studies were rather limited as they refer to small areas and a relatively small number of individuals.

10.2 Description of the Dataset

The first preserved volume of the Marriage License Books dates from 1451, but there are reasons to believe that there was a custom of recording marriages in the diocese of Barcelona in earlier times (Baucells 2002). According to Carreras Candi (1913), the origin of the books can be traced to a privilege given by Pope Benedict XIII (1328–1423) to the diocese for construction and subsequent maintenance of the Barcelona cathedral when he visited the city in September 1409.[1] The Pope granted the new cathedral the authority to levy a tax on all marriages celebrated in the diocese. These marriages were recorded in a centralised register until 1905, and each bridal couple had to pay a fee based on their socio-economic status. The seven-step fee scale went from *Amore Dei*—free of charge—for people who could not afford it or were exempted from payment, to a tax of 12 pounds for the high aristocracy (dukes, marquises, counts and viscounts). Different taxes applied to nobles, knights and lords; the urban oligarchy and medical doctors; and

[1] Benedict XIII, born Pedro Martínez de Luna and also known as *Papa Luna*, was an Aragonese nobleman who became Pope during the Western Schism (1378–1417).



11

Juliol. 1617

8

Jassa Dissapte a 8. reberē de Juan Jassa pages habitant en Barª. fill de
Pere Jassa pages de Peralada bisbat de Girona y de Maria defuncta
ab Cecilia donsella filla de Pere Recort parayre de Solsona y de
Aldonsa defuncta — lt iiij ß

Fig. 10.2 Example of a marriage license from July 1617 (volume 69): *Saturday the 8th we received from Juan Jassa peasant resident in Barcelona son of/Pere Jassa peasant from Peralada diocese of Girona and deceased Maria/with maiden Cecilia daughter of Pere Recort carder from Solsona and/deceased Aldonsa—4 shillings*

shopkeepers, royal notaries, merchants and masters of guilds. The vast majority of Catalan couples—about 90 %—paid four shillings (peasants, artisans and labourers).[2]

Each marriage record contains information about the tax paid, the exact date when it was paid—which was presumably close to the date of the marriage—the first name and surname of the groom and the first name of the bride. The occupation and the place of residence of the groom are usually also mentioned. However, the amount of additional information provided varies across time and different volumes: in the construction of the BHMD, more than 60 attributes were included for each marriage record. These attributes cover all the different types of information regarding the bridal couple and their relatives, at least once provided in the volumes. Examples of attributes which were sometimes but not always mentioned are nicknames, second surnames, or surnames and occupations of women.

For the 32-year period from 1597 to 1629, the standard information contained in each marriage record is structured as follows: [*date*] [*first name groom*] [*surname groom*] [*occupation groom*] [*residence groom*] [*martial status groom*] [*first name father groom*] [*surname father groom*] [*occupation father groom*] [*father groom dead or alive*] [*first name mother groom*] [*mother groom dead or alive*] [*first name bride*] [*marital status bride*] [*first name father bride*] [*surname father bride*] [*occupation father bride*] [*residence father bride*] [*father bride dead or alive*] [*first name mother bride*] [*mother bride dead or alive*] [*tax paid*]. The place of residence of the father of the groom is not generally mentioned and it is only included when it differs from the son's, as shown in the example of Fig. 10.2. If the place of residence differs from the origin of the individual, an additional attribute [*origin*] can be extracted. Furthermore, we assume that an individual was dead at the time of

[2] Taxes were paid in pounds and shillings, whereby one pound was equivalent to 20 shillings. The range of possible values was: *Amore Dei*, 4 shillings, 6 shillings, 12 shillings, 1 pound and 4 shillings, 2 pounds and 8 shillings, and 12 pounds. The currency and the social scale were not consistent across all of the volumes of the Marriage License Books (1451–1905), but those are the ones that correspond to the period of study of the present research (1573–1629).

the marriage only if the person's death is explicitly mentioned; when no mention of the person's death is made, we assume that the individual was alive.

Nevertheless, depending on the marital status of the bridal couple and their places of origin, the available information varies.

1. When the groom was not originally from Catalonia, information about his parents was not usually provided.
2. If the groom or the bride was widowed, information about his or her parents was also not provided. Nevertheless, detailed information was provided about the former spouse (name, occupation and place of residence) for widowed women, though not for widowed men.

Since it is unlikely that the parents of foreign grooms were married in the diocese of Barcelona, it would not be possible to use them for the marriage linkage even if they had been mentioned. The lack of information about the parents of widowed spouses is not a handicap, as those individuals had been married before, and that information should be available in a previous marriage record.

The detailed information about former husbands is attributable to the fact that women at that time were identified by their first names and their relationship to a man: to their fathers before marriage, to their husbands while married and to their former husbands when widowed (Sánchez Rubio and Testón Núñez 2012). Consequently, variables such as the surname or place of residence of the bride were rarely registered in the Marriage License Books, and they can only be inferred from the information about the referenced man.[3] This might also explain the greater availability of information about the parents of maiden brides compared to bachelor grooms, as analysed in a later section. In fact, over the 32-year period from 1597 to 1629, there are only 481 marriage records out of 30,633 (1.6 %) in which neither the parents of the bride nor the former husband of the bride are mentioned, and 60 % of those brides are identified as foreigners or orphans.

10.3 Standardisation of Nominal Information

Deciphering the information contained on the Marriage License Books can be challenging, as these marriages were recorded over a period of five centuries by different scribes with different handwriting. The fact that Catalan spelling rules were not defined until the early twentieth century (Fabra 1932) further hampers the comparability of data across time. Therefore, in the course of the construction of the BHMD, a process of standardisation was applied to the dataset, including the standardisation of occupations, geographical locations, first names and surnames.

[3] Under current Spanish surname customs, married women preserve their maiden surname, but this practise was not legally established until the late nineteenth century, when it became compulsory to keep the surnames of both the father and the mother (Art. 48 of the Spanish Civil Register Act from 1870).

Occupations were codified according the Historical International Standard Classification of Occupations (HISCO) code, while geographical locations were grouped according to the current Spanish postal code. In the case of foreign locations (i.e. France), a special code was assigned in each case. The standardisation of first names and surnames was more challenging due to their particular nature.

In Catalan, as in many languages, family names may have had several variants as a result of dialectal differences and foreign influences, as well as due to misspellings and phonetic transcriptions (Bas Vidal 1988). In most cases, the changes that appeared in Catalan surnames occurred for one of the following reasons: there was a deviation from a base name because of an augmentation, the use of a diminutive or a feminisation (*Pericàs*/*Pericot*/*Perica*); the surname was composed of two or more first names or surnames (*Bon+Fill*/*Bofill*, *Cap+Vila*/*Capdevila*) or a surname and an article (*La+Porta*/*Laporta*, *Sa+Font*/*Safont*); an *s* was added to the end of the surname; an archaic surname form persisted; the dialectal pronunciation varied (*Giral*/*Guiral*, *Bagué*/*Veguer*); or the name was misspelled (Moll 1959). The most common variations in the spelling of family names found in the Marriage License Books are generally attributable to one of several errors, including the removal of the final etymological *r* (*Ferrer*/*Farré*). Spelling changes may have also arisen from the confusion between certain letters which are phonetically very similar in Catalan: between *b* and *v* (*Rivera*/*Ribera*); between *s*, *c*, *ç*, *z* and *ss* (*Serra*/*Çerra*); between *y* and *i*; between *j* and *g*; or between *l* and *ll*. Additional reasons for spelling changes include the use of *c* instead of *ch* at the end of the word (*Benach*/*Benac*); and the use of *h* at the beginning of words that start with vowels (*Homs*/*Oms*).

A general discussion about the standardisation problems of historical data can be found in Bloothooft (1998), and there is also a large body of literature in which the difficulties in name matching even in modern times have been analysed (Christen 2012). In the present research, we have attempted to address these complications by carrying out a dictionary name-based standardisation of names to facilitate the nominal linkage procedure. All of the misspellings have been rectified according to current Catalan grammar rules, and all of the accents, articles, prepositions and conjunctions have been removed in order to keep only the root of each name. For the 1573–1629 period, 1,138 male and 1,081 female first names have been, respectively, grouped into 639 and 521 different standardised first names, with a reduction of 40 and 50 % of their variability. Accordingly, 30,637 surnames have been standardised into 12,108 different forms, reducing the variability by 60 %.

10.4 Marriage Linkage

A nominal record linkage is typically used to determine whether two different records belong to the same individual. However, in this case we are not comparing information about individuals, but about couples. Three key variables which are found in almost all records have been used to identify couples across marriage licenses: the first name of the groom, the first name of the bride and the surname of

Marriage 1		***Marriage 2***
First name of the groom	⟷	First name father groom, or First name father bride
First name of the bride	⟷	First name mother groom, or First name mother bride
Surname of the groom	⟷	Surname father groom, or Surname father bride

Fig. 10.3 Key variables used in the marriage linkage

the groom. There were only a few exceptions in which it was not possible to recover the information for these three variables: in some cases, the names were not recorded or they turned out to be illegible, while in other cases the original marriage record noted that the bridal couple wanted to remain anonymous.

The bridal couple of a marriage record (marriage 1) can be found as parents in the marriage records of their children (marriage 2). Thus, the first names of both the bride and the groom and the surname of the groom can later be found in the role of parents. This information provides us the key variables for record linkage, as shown in Fig. 10.3.

10.4.1 Completeness of the Key Variables Used in the Record Linkage

Before carrying out the marriage linkage, it is essential that we analyse the completeness of the key variables used in the nominal record linkage. Besides the names of the bridal couple, which are found in almost all records, the information about the parents deserves special attention.

As was already mentioned, the marriage records of widowed individuals do not offer information about their parents. This information is also generally missing for foreign grooms. In the 32-year period from May 1597 until April 1629, there are a total of 30,633 marriage records, 23,479 (76.6 %) of which involve bachelor grooms, and 23,214 (75.8 %) of which involve maiden brides. Table 10.1 summarises the available information about the parents of bachelor grooms and maiden brides for each of the key variables used in the marriage linkage. These counts only include cases in which the corresponding variables can be correctly recovered; cases in which the variables turn out to be illegible, or the parents are mentioned with incomplete names, are excluded.

As was mentioned above, we have more information about the parents of maiden brides than the parents of bachelor grooms because at that time a woman was always referenced in relationship to a man. Nevertheless, we can discern differences based on the origin of the groom. For example, 17,328 of the marriage records involve bachelor grooms who were originally from Catalonia (73.8 % of the records

Table 10.1 Available information about the parents of bachelor grooms and maiden brides. Volumes 59–74 of the marriage license books

Variable	Cases	Percentage
First name father groom	15,142	64.5
First name mother groom	14,659	62.4
Surname father groom	15,510	66.1
	Total bachelor grooms: 23,479	
First name father bride	21,341	91.9
First name mother bride	20,005	86.2
Surname father bride	22,500	96.9
	Total maiden brides: 23,214	

of bachelor grooms). Of these records, 14,988 (86.5 %) mention the groom's father's first name, 14,511 (83.7 %) note the groom's mother's first name and 15,355 (88.6 %) mention the groom's father's surname. As a result, if we restrict the analysis to those individuals for whom the information about their parents is complete—i.e. for whom all three of the key variables are well recorded—we have full information about the parents for 83.0 % (14,387) of Catalan bachelor grooms and 83.5 % (19,383) of maiden brides. Those are the records from the 32-year period from 1597 to 1629 which have been used in the marriage linkage.

10.4.2 The Linkage Algorithm

The marriage linkage is carried out by an algorithm that has been developed in the project *Five Centuries of Marriages*, and which is divided into two sections: (1) the record linkage itself, which has been implemented both in C++ (Stroustrup 2013) and R (R Core Team 2013) and (2) the data cleaning posterior to the initial link, which has been implemented in R.[4]

10.4.2.1 Nominal Record Linkage

The algorithm has been executed to find, separately, the potential parents of the grooms and the potential parents of the brides who married in the 32-year period from May 1597 until April 1629. The information about each parent is compared with the information about the bridal couples of previous marriages, and the string distance between each of the key variables analysed above is measured. This process is carried out using standardised names.

[4] The C++ version was developed by researchers from the Computer Vision Centre and the Centre for Demographic Studies (Barcelona, Spain) who created a software named *Busca Descendècies* (English translation: *Search Offspring*).

Following Roman legislation, in the fourteenth century the canon law established a minimum age at marriage of 12 years for women and 14 for men, and those limits were commonly accepted by the Catholic Church in the subsequent centuries (Gaudemet 1987). Therefore, assuming that all births befell within the marriage, these values have been adopted as the minimum time difference between the dates at marriage of an individual and of his or her potential parents. This has been the only restriction imposed while carrying out the marriage linkage.

Two different measures have been implemented in the process of nominal record linkage: the bag distance and a slightly modified Levenshtein edit distance. Let x and y be two different strings, the bag distance is the maximum between the number of elements of x that do not belong to y and the number of elements of y that do not belong not to x. Formally, let $\mathrm{ms}(x) = \{x_1, \ldots, x_n\}$ denote the set of symbols in x, then $\mathrm{dist}_{\mathrm{bag}}(x, y) = \max(|\mathrm{ms}(x) - \mathrm{ms}(y)|, |\mathrm{ms}(y) - \mathrm{ms}(x)|)$ (Bartolini et al. 2002). For instance, $\mathrm{dist}_{\mathrm{bag}}$ (‘mary’, ‘maria’) = max (|{‘*y*’}|, |{‘*i*’, ‘*a*’}| = 2.

By contrast, the Levenshtein distance measures the minimum number of operations necessary to transform one string into another with three possible operations: replacement, insertion and deletion of single characters. The procedure implemented to compute the Levenshtein distance is based on the method presented by Wagner and Fischer (1974), but with certain adjustments made to adapt it to Catalan grammar. Other string distances like the Jaro–Winkler which are commonly used for name matching (Christen 2012) were discarded because it was imperative that we control for the substitution of certain pairs of letters.

In the Levenshtein distance, each required operation for moving from one string to another generally has a cost equal to one, but our algorithm includes a system that assigns a lower cost to the substitution of letters that are phonetically very similar in Catalan; for example, *a* and *e*, *o* and *u*, or *ç* and *s*. The program also allows for the replacement of specific pairs of letters by a single letter (i.e. *ph* by *f*, *ll* by *l* or *ss* by *s*) with no additional cost. Moreover, a table of exceptions is used for surnames that are graphically very similar—only one or two letters changed—but etymologically different; for example, *Font* and *Pont* (English translation: *fountain* and *bridge*) or *Roca* and *Roda* (English translation: *rock* and *wheel*).

For both distances, a normalised similarity measure can be derived,

$$\mathrm{sim}(x, y) = 1 - \frac{\mathrm{dist}(x, y)}{\max(|y|, |x|)} \in [0, 1], \tag{10.1}$$

returning 1 if both strings are equal, and 0 if they are totally different (Christen 2012). This normalised similarity measure makes the distances between the strings of different length comparable, and can be used to accept or discard links according to a threshold. Nevertheless, in the marriage linkage procedure three strings—key variables—from each record are used, which leads us to a slightly different similarity index. Let *a* and *b* be two different marriage records; their total similarity is defined as

$$I(a,b) = 1 - \frac{1}{3}\sum_{i=1}^{3}\frac{\text{dist}(a_i,b_i)}{\max(|a_i|,|b_i|)} \in [0,1], \tag{10.2}$$

where a_i and b_i, $i = 1,2,3$ are the corresponding key variables from each record. Again, $I(a,b) = 1$ if each of the three pairs of strings are identical.

The bag distance has lower computational costs than the Levenshtein distance, and it has been proved that $\text{dist}_{\text{bag}}(x,y) \le \text{dist}_{\text{lev}}(x,y)$ for any pair of strings x and y, which makes the bag distance a useful tool for filtering out candidates before applying a more complex edit distance technique (Bartolini et al. 2002). This last equation implies that $\text{sim}_{\text{bag}}(x,y) \ge \text{sim}_{\text{lev}}(x,y)$, and consequently

$$I_{\text{bag}}(a,b) \ge I_{\text{lev}}(a,b). \tag{10.3}$$

Therefore, the marriage linkage procedure is divided into two steps:

1. First, the bag similarity index between two records is computed. If $I_{\text{bag}}(a,b) \ge 0.80$, the link passes to the next step; otherwise, it is discarded. Intuitively, this filter can be interpreted as showing that the key variables of both records share, on average, at least 80 % of their characters.
2. If $I_{\text{bag}}(a,b) \ge 0.80$ the Levenshtein similarity index between the two records is computed. The link is accepted only if $I_{\text{lev}}(a,b) \ge 0.90$.

Even if they are arbitrary, these thresholds offer a good balance between the number of links obtained and their quality, and they can be easily modified by the user each time the program is executed.[5]

While the variables needed to identify a couple are their first names and the surname of the male, some individuals had double first names, second surnames and nicknames that were also registered in the Marriage License Books. The algorithm has the ability to manage this kind of information. For example, our approach allows for the possibility that an individual whose first name was recorded as *Pere Joan* in a first marriage record appeared in a second marriage record as *Pere Joan*, *Joan Pere*, *Joan* or *Pere*. Accordingly, it has been established that moving from a single first name to a double first name or switching the order of a double first name has a lower cost than deleting, substituting or adding all of the necessary letters to move from one string to another. The same procedure is carried out for individuals who had two surnames.

Moreover, there are records in which nicknames that follow the individual's surname were included, and several cases have been identified in which the nickname of a father was passed to his son as a surname. These issues need to be taken

[5] Note that Eq. (10.3) is true only for the regular Levenshtein distance, but not for the slightly modified version used here. The selected thresholds were empirically tested to prevent from rejecting links in the first step that could be accepted in the second one.

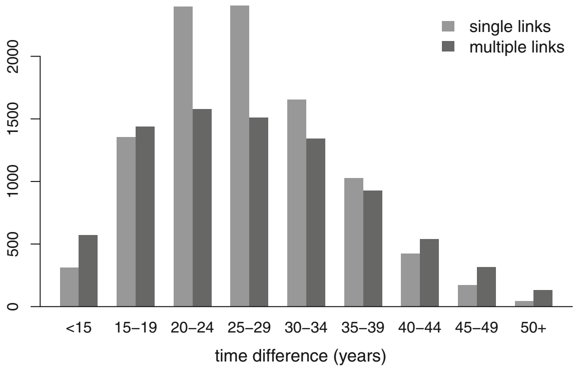


Fig. 10.4 Frequency of links according to the time difference between marriages

into account in the record linkage, as individuals with second surnames and nicknames might have appeared in different marriage records with those second surnames and nicknames as a surname, or vice versa. Thus, nicknames are treated for all purposes as second surnames in the marriage linkage.

10.4.2.2 Data Cleaning Following the Initial Link

The output of the linkage procedure consists of 18,147 links, of which 6,986 involved grooms and their potential parents and 11,161 involved brides. However, out of all of those links there are 8,359 (46.1 %) over-links with multiple candidates; these are cases in which more than one marriage of possible parents with similar names was found. The purpose of this section is to present a procedure to progressively discard the less likely links and reduce the amount of multiple links. Figure 10.4 shows the frequency of links according to the time difference between the marriages of the children and the marriages of their potential parents, broken down by single and multiple links.

Upper Limit of the Time Difference Between Marriages

Figure 10.4 shows that, for extreme values of the time difference between marriages, the number of multiple links exceeds the number of single links. So far, only a theoretical minimum time difference between marriages has been imposed for the acceptance or rejection of a link, but not a maximum value. If the upper 8 % of links are left out, an upper limit of 40 years for the time difference between marriages can be implicitly established, which is a reasonable limit based on the historical context. The exclusion of the 1,599 links corresponding to that 8 %—976 of multiple links and 623 of single links—reduces the amount of multiple links from 8,359 to 7,056 (42.6 % of the remaining 16,548 links).

Name Indexes

Standardised names are used to facilitate the nominal record linkage procedure and increase the number of plausible candidates. However, literal names—as they are written in the original marriage record—may contain information for selecting links among multiple candidates. To that end, a new similarity index like the one described in Eq. (10.2) may be computed, using the bag distance and literal names of the marriage records of each link.[6] Consequently, each link will have three similarity indexes: the one just described and the two computed in the marriage linkage with standardised names, one with the Levenshtein distance and the other with the bag distance.

Family Grouping

The nominal record linkage has been carried out by searching the marriage of the parents for each groom and bride married between 1597 and 1629. Reversely, all brides and grooms that refer to the same parents married between 1573 and 1617 can be considered as potential siblings. Inconsistencies may arise in cases in which the marriages of parents are linked to a child who might not have been theirs. For example, an individual might be listed as dead in the marriage record of one of his or her potential children, but be listed as alive in the marriage record of another potential child from a later date. Nonetheless, these inconsistencies represent a useful tool for discarding links and selecting the most probable link when there are multiple candidates, and can thus help to ensure the consistency and coherence of the dataset.

Three more attributes are used to compare the information among potential siblings: the occupation, the residence and the origin of the father. We did not want to compare the occupation and the residence of an individual stated in his marriage record, with the ones noted in the marriage records of his potential children, in order to avoid any bias to people who persist in the same area and in the same job (Ruggles 2002). However, we assume that the occupation and the places of residence of the fathers are less likely to change between the marriages of their children.

A different approach is given to the origin, which is a variable that is only mentioned if it differs from the place of residence. We consider that the origin should be the same during the life course of an individual. For example, if a groom was originally from France, his origin in his children's marriage certificate, if mentioned, needs also to be France. Therefore, we discard all links in which the origin of an individual stated in his marriage record does not coincide with the

[6] The purpose of this index is to obtain a measure of the string distance between the original literal names, but not to filter out candidates as in the linkage procedure. The bag distance is an optimal measure due to its lower computational costs, even though other string distances could have been used.

origin noted in the marriage records of his potential children. This allows to exclude 698 links and reduce the number of multiple links from 7,056 to 6,508 (41.1 % of the remaining 15,826 links).

Step-by-Step Selection of Links

Once all the necessary information for each link is recovered, a step-by-step selection of links is carried out in order to reduce the number of multiple links, based on the following:

1. First, all of the multiple links need to be identified and marked. The name indexes are compared to determine which links among multiple candidates have better similarity indexes.
2. All links are grouped by families. Then, comparing the information on the parents of each group of potential siblings, the less likely links are discarded.
3. After some of the links are excluded, the remaining multiple links should be identified again.

These three actions are iteratively repeated, changing at each time-step the criteria used on the selection of links. For example, in the first step, all multiple links which do not have the best name indexes and present inconsistencies regarding the information whether the parents are dead or alive, the occupation of the father and the place of residence (or origin) of the father, are excluded. In the next step, the conditions are slightly changed becoming more restrictive—as the less likely links have already been discarded—and then some other links are excluded. Table 10.2 summarises the results obtained and the criteria used at each iteration.

At the end of this iterative procedure, the number of multiple links was reduced to 1,111 (9.4 %) out of 11,839 links. It is interesting to note that this step-by-step process allowed to reduce the number of multiple links by 5,397, deleting only 3,987 links, a value which is 26 % lower. Nevertheless, there were some cases for which it was not possible to select the best link among multiple candidates. This problem was especially noticeable for links involving individuals with very common first names and surnames, and for whom no complementary information was available (i.e. occupation and residence of the father). Therefore, the 1,111 remaining multiple links were discarded. Moreover, in order to give consistency to the dataset, all of the links corresponding to parents that were listed as alive after having been listed as dead in a previous marriage record, were also excluded. As a result, the final output was composed by 10,131 single consistent links.

10.4.3 Evaluation of the Quality of the Links

So far, the marriage linkage has been carried out using an algorithm which finds the best links automatically. Moreover, we are assuming that single consistent links are

Table 10.2 Step-by-step selection of multiple links

Step[a]	Total links	Multiple links	Percentage	Links discarded	Multiple links reduction
0	15,826	6,508	41.1	0	0
1	15,662	6,304	40.3	164	204
2	15,365	5,930	38.6	297	374
3	15,277	5,818	38.1	88	112
4	15,166	5,664	37.3	111	154
5	14,922	5,325	35.7	244	339
6	14,773	5,112	34.6	149	213
7	14,655	4,972	33.9	118	140
8	14,422	4,681	32.5	233	291
9	14,364	4,603	32.0	58	78
10	14,207	4,412	31.1	157	191
11	13,937	4,079	29.3	270	333
12	13,805	3,908	28.3	132	171
13	13,525	3,488	25.8	280	420
14	13,274	3,182	24.0	251	306
15	13,148	3,013	22.9	126	169
16	12,738	2,505	19.7	410	508
17	12,151	1,488	12.2	587	1,017
18	11,839	1,111	9.4	312	377
			Total	3,987	5,397

[a] Exclusion criteria of multiple links at each step, comparing the information among links grouped in the same family: (1) not best name indexes and inconsistencies regarding the parent's deaths, the father's occupation and the father's residence or origin; (2) not best name indexes and inconsistencies regarding the parent's deaths and the father's occupation, residence or origin; (3) not best name indexes and inconsistencies regarding the parent's deaths; (4) not best name indexes and inconsistencies regarding the father or the mother's death, the father's occupation and the father's residence or origin; (5) not best name indexes and inconsistencies regarding the father or the mother's death and the father's occupation, residence or origin; (6) not best name indexes and inconsistencies regarding the father or the mother's death; for steps 7–12, the exclusion criteria are the same as the ones used in steps 1–6, but without the condition of not having the best name indexes; (13) worst name indexes and inconsistencies regarding the father's occupation, residence or origin; (14) not best name indexes and inconsistencies regarding the father's occupation, residence or origin; (15) inconsistencies regarding the father's occupation and the father's residence or origin; (16) inconsistencies regarding the father's occupation, residence or origin; (17) worst name indexes and (18) not best name indexes

all correct, which may not be the case. Therefore, to evaluate the quality of the final set of links, a manual review of a random sample is needed, according to the following criteria:

1. The names of the two records should match.
2. The spatial distance between the place where the marriages of each link took place and their geographical characteristics should follow a logical sequence according to the mobility trends in the historical context (Nadal and Giralt 2000; Simon Tarrés 1996).

3. The socio-occupational characteristics of the parents and their descendants and the amount of tax paid in each marriage should be plausible. Since the amount of tax paid was based on the socio-economic status of the bridal couple, the difference between the amount paid by the parents and by their children should not differ more than two tiers in the fee scale described in Sect. 10.2.

This qualitative analysis allowed us to validate 96 % of the links: 82 % were considered correct, 14 % turned out to be plausible but with some degree of uncertainty and 4 % were discarded as being clearly incorrect.

10.5 Reconstruction of the Lifespans

In the linkage procedure, the marriages of both singles and widow(er)s are considered; but in the reconstruction of the lifespans, only individuals who are identified in their first marriage are taken into account. As was explained in a previous section, because the former wives of widowed men who were remarrying are not mentioned, a man who was remarrying cannot be identified with his previous marriage. Therefore, the lifespan of a man might have been reconstructed twice—first based on information about the children of his first marriage, and again based on information about the children of his second marriage—and there is no way to know whether these lifespans belong to the same individual. We avoid this problem by considering only individuals whose first marriage has been identified.

Hence, the lifespans of 6,131 men and 5,872 women who first married between 1573 and 1617, and who had children who married between 1597 and 1629, have been reconstructed. This process has also been carried out through an automatic algorithm implemented in R using the output of the marriage linkage and considering all of the information about those individuals and their relatives, as shown in Fig. 10.5.

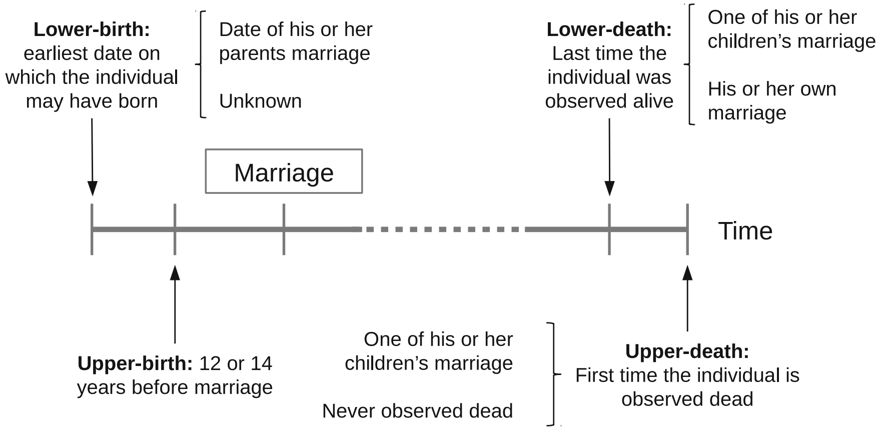

Lower-birth: earliest date on which the individual may have born
Date of his or her parents marriage
Unknown
Lower-death: Last time the individual was observed alive
One of his or her children's marriage
His or her own marriage
Marriage
Time
Upper-birth: 12 or 14 years before marriage
One of his or her children's marriage
Never observed dead
Upper-death: First time the individual is observed dead

Fig. 10.5 Reconstruction of the lifespan of each individual

Two different types of lifespans have been reconstructed:

1. Individuals for whom it has been possible to identify both the marriages of their parents and the marriages of their children. For those individuals, the lower and upper bounds of birth are better defined.
2. Individuals for whom it has only been possible to identify the marriages of their children.

In both cases, the individuals might be censored if they have never been observed as being dead, or they might be interval censored if they have been observed as being dead in the marriage record of at least one of their children.

10.6 Bayesian Model

The most immediate application of the reconstructed lifespans described above is in an ongoing research project in which adult age-specific mortality and life expectancy for data with unknown ages are estimated. The method is based on the Bayesian Survival Trajectory Analysis (BaSTA), a free, open-source R software package for estimating age-specific survival from capture–recapture/recovery data under a Bayesian framework which was originally designed to study the survival of wild animals with unknown ages (Colchero and Clark 2012; Colchero et al. 2012). The model has been extended for use in the analysis of this particular set of human historical data in which the ages of all the individuals are also unknown.

Bayesian methodology typically consists of three steps (Gelman et al. 2004): (1) a full probability model which includes some previous knowledge (*prior*) of the phenomenon of study must be specified; (2) knowledge about the unknown parameters based on the observed data must be updated, and a *posterior* distribution of the parameters is obtained; and (3) the adjustment of the model to the data must be evaluated. These three steps are usually repeated several times in an iterative procedure in order to obtain the best estimates.

All of the individuals from the Marriage License Books are left truncated at their age at marriage, whereby some of them are censored and others are interval censored. As the ages of all of the individuals are unknown, and because the dataset consists of marriage records, the model is conditioned on reaching the minimum age at marriage. Thus, the model uses a hierarchical approach that requires the implementation of a Markov Chain Monte Carlo (MCMC) algorithm to estimate mortality parameters and times of birth. At each iteration, the joint posterior is divided into two sections: (1) the estimation of survival parameters of a parametric mortality model and (2) the estimation of the unknown times of birth (i.e. the ages at the first detection). In order to assess the ergodicity to those estimations, the model requires us to run several parallel iterations of the MCMC algorithm.

10.7 Discussion

In this chapter, we have discussed a methodology for reconstructing the lifespans of individuals through a nominal record linkage procedure using historical marriage records from the sixteenth and seventeenth centuries. The data were extracted from the Barcelona Historical Marriage Database (BHMD), a unique source that covers a period from over 450 years (1451–1905) (Cabré and Pujadas-Mora 2011). The algorithm we implemented might, however, be exportable to other data sources with similar characteristics. The combination of two string distances—bag distance and Levenshtein distance—has been proved to be an efficient system for the nominal record linkage. In the data cleaning following the initial link, the selection criteria used at each step may be somewhat arbitrary, but there were selected in order to carry out a progressive exclusion of the less likely links taking into account the information available in each case. Certainly, while using this method some correct links may be discarded, especially in the last step of the selection procedure when it was not possible to select links among multiple candidates. However, the evaluation of the final set of links proved that our method is able to provide links of good quality.

The reconstruction of the lifespans of a population from the early modern period in a large area such as the diocese of Barcelona contributes to our understanding of the demographic behaviour of Catalonia in the pre-transitional era, and is by extension a useful tool for studying European population history of the sixteenth and seventeenth centuries. Moreover, the obtained links can be seen as very powerful tools for analysing issues such as the intergenerational transmission of social status, the diffusion and extinction of surnames and migration dynamics. The links can also be used to obtain indirect estimations of living standards and socio-economic indicators (Cabré and Pujadas-Mora 2011).

The reconstructed lifespans are being used in an ongoing research project in which mortality is estimated from incomplete data by applying Bayesian methods based on the BaSTA R package (Colchero et al. 2012). The development of methodologies for estimating mortality patterns from incomplete demographic data is a novel and rapidly growing research area. While applying methods which were originally designed for biodemographic studies is challenging, doing so opens up the possibility of adapting the model to similar demographic data from other countries and periods.

Acknowledgments This work is part of the Advanced Grant Project *Five Centuries of Marriages* (2011–2016) funded by the European Research Council (ERC 2010-AdG-20100407). The authors would like to express their gratitude to its principal investigator, Anna Cabré, for her support and guidance. We would also like to thank all of the members of the project, especially Miquel Amengual, Joan Mas, and Conchi Villar. Special mention goes to Sofia Aldabet, Fernando Colchero, Miriam Hils, Trifon Missov, and James Vaupel.

References

Bartolini, I., Ciaccia, P., & Patella, M. (2002). String matching with metric trees using an approximate distance. In *String processing and information retrieval* (pp. 271–283). Berlin: Springer.

Bas Vidal, J. (1988). *Els cognoms catalans i la seva història*. Barcelona: Cap Roig.

Baucells, J. (2002). Esposalles de l'arxiu de la catedral de Barcelona: un fons documental únic (1451–1905). *Butlletí del Servei d'Arxius, 35*, 1–2.

Bloothooft, G. (1998). Assessment of systems for nominal retrieval and historical record linkage. *Computers and the Humanities, 32*(1), 39–56.

Cabré, A., Pujadas-Mora, J. M. (2011). Five centuries of marriages (5CofM): A project of historical demography in the Barcelona area. *Papers de Demografia, 368*.

Carreras Candi, F. (1913). Les obres de la catedral de Barcelona (1298–1445). *Boletín de la Real Academia de Buenas Letras, 49*, 22–30.

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Heidelberg: Springer.

Colchero, F., & Clark, J. S. (2012). Bayesian inference on age-specific survival for censored and truncated data. *Journal of Animal Ecology, 81*, 139–249.

Colchero, F., Jones, O. R., & Rebke, M. (2012). BaSTA: An R package for Bayesian estimation of age-specific survival from incomplete mark-recapture/recovery data with covariates. *Methods in Ecology and Evolution, 3*, 466–470.

Dopico, F., & Rowland, R. (1990). Demografía del censo de Floridablanca. Una aproximación. *Revista de Historia Económica/Journal of Iberian and Latin American Economic History (Second Series), 8*(3), 591–618.

Fabra, P. (1932). *Diccionari general de la llengua catalana*. Barcelona: Llibreria Catalònia.

Feliu, G. (1999). La demografia baixmedieval catalana: estat de la qüestió i propostes de futur. *Revista d'Història Medieval, 10*, 13–43.

Gaudemet, J. (1987). *Le mariage en Occident: les moeurs et le droit*. Paris: Ed. du Cerf. English edition: Gaudemet, J. (2002). Marriage in the western world: Morals and the law (I. Roche, Trans.). Paris: University of Notre Dame Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman and Hall.

Moll, F. B. (1959). *Els llinatges catalans (Catalunya, País Valencià, Illes Balears): assaig de divulgació lingüística*. Palma de Mallorca: Editorial Moll.

Muñoz Pradas, F. (1990). Creixement demogràfic, mortalitat i nupcialitat al Penedès (segles XVII–XIX). *PhD. thesis, Autonomous University of Barcelona*.

Nadal, J., & Giralt, E. (2000). *Immigració i redreç demogràfic: els francesos a la Catalunya dels segles XVI i XVII*. Barcelona: Eumo Editorial.

R Core Team. (2013). *R: A language and environment for statistical computing. R Foundation for statistical computing*, Vienna. http://www.R-project.org/.

Ruggles, S. (2002). Linking historical censuses: A new approach. *History and Computing, 14*(1–2), 213–224.

Sánchez Rubio, R., & Testón Núñez, I. (2012). Situación y perspectiva de los estudios de antroponimia en la España moderna. In *L'Italia dei cognomi: l'antroponimia italiana nel quadro mediterraneo* (pp. 75–122). Nancy: The University Press.

Simon Tarrés, A. (1996). La població catalana a l'edat moderna. Deu estudis. *Monografies Manuscrits*. Bellaterra: Autonomous University of Barcelona.

Stroustrup, B. (2013). *The C++ programming language* (4th ed.). New York: Pearson Education.

Torrents, A. (1993). Transformacions demogràfiques en un municipi industrial català: Sant Pere de Riudebitlles (1608–1935). *Ph.D. thesis, University of Barcelona*.

Wagner, R. A., & Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM, 21*(1), 168–173.

Chapter 11
Dancing with Dirty Data: Problems in the Extraction of Life-Course Evidence from Historical Censuses

Luiza Antonie, Kris Inwood and J. Andrew Ross

Abstract This chapter builds on a recent use of SVM classification to create linked sets of Canadian 1871 and 1881 census records. The census data are imprecise and have limited granularity; many records share identical detail. In spite of these challenges, the SVM generates life-course information for large numbers of individuals with a low (3 %) false positive error rate. However, there is a higher incidence of error among apparent migrants when the true rate of migration is small. The linked data are broadly representative of the population with some underrepresentation of illiterates, young adults (especially unmarried women), older people (especially men), and married people of French origin. The new longitudinal data are of considerable research value but users must take into account these weaknesses.

11.1 Introduction

In this chapter, we explore strengths and weaknesses of a recent application of support vector machine (SVM) classification to Canadian historical census records. The classification identifies matched pairs of records from the 1871 and 1881 census. Each matched pair describes the same person and thus provides insight into the change in individual circumstances from one year to the next.

L. Antonie (✉)
School of Computer Science and Department of Economics,
University of Guelph, Guelph, ON, Canada
e-mail: luiza.antonie@gmail.com

K. Inwood · J.A. Ross
Department of History and Department of Economics,
University of Guelph, Guelph, ON, Canada
e-mail: kinwood@uoguelph.ca

J.A. Ross
e-mail: jaross@uoguelph.ca

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_11

The considerable importance of the North American census for historical research derives from its rich systematic detail and a paucity of alternate sources providing a comprehensive description of the population. Nowhere in North America was there an established church with a commitment to public vital registration. There was not even a dominant church whose records might serve that purpose, except perhaps in Utah or Quebec, and over time even their records became less comprehensive. Individual states and provinces gradually developed effective systems of vital registration but in both the United States and Canada the consistent national registration of births, marriages, and deaths emerged only in the twentieth century. Thus, it is in the absence of other sources that Canadian and American scholars turn to the nineteenth century censuses for population profiles and for the construction of longitudinal data that tracks individuals from census to census.

Since the 1980s, there have been significant advances in the method of linking records between censuses. A first wave of studies using manual techniques (Steckel 1988; Knights 1991; Ferrie 1996, 1999) has been followed by the use of machine-learning methodology (Ruggles 2006; Christen 2008; Goeken et al. 2011; Fu et al. 2014). The new approach is capable of generating in a near-automatic way large representative samples of longitudinal and even multigenerational data. The value of the new methodology for understanding historical populations makes it important to assess its strengths and weaknesses (Wisselgren et al. 2014).

In this chapter, we take the example of a recent application of SVM classification to historical Canadian historical census records (Antonie et al. 2013). A principal challenge for any attempt to track individuals from census to census is the relative imprecision of the 1871 and 1881 data. This requires careful adaptation of the classification methodology and some assessment of the quality of linking. We find that the linked data that we generate are reasonably representative of the population although care is needed, depending on research application, because some groups are harder to link: adolescents and young adults especially unmarried women, older people especially men, and married people of French origin.

An additional complication is that, while the overall error (i.e., false positive) rate is only 3 %, the records describing people who apparently migrate, or change categories in some other way, are difficult to interpret if the proportion migrating is small. These idiosyncrasies recommend some care in the research use of these valuable data.

11.2 Overview 1871–1881

We begin with an overview of the record-linking system described in greater detail elsewhere (Antonie et al. 2013). Our objective is to identify pairs of records that describe the same person in two different bodies of data: the 3.4 million records of the 1871 Canadian census (www.census1871.ca) and 4.3 million records of the 1881 census. The Church of Jesus Christ of Latter-day Saints created both databases;

the latter is housed at the Université de Montréal (www.genealogie.umontreal.ca/en). We construct records that follow individuals over time by comparing every 1871 record with each 1881 record, and then classify each comparison as a match or non-match. If we think a particular pair of records (one from 1871 and one from 1881) point to the same person, we accept them as a match.

The process requires us to compare, literally, millions of records in 1871 with millions of records in 1881 in order to establish which pairs are identical, i.e., describe the same person. The comparison is made using four personal attributes that should not change over time (last name, first name, gender, and birthplace) and two others that change in a predictable way (age and marital status). We do not use information about occupation, location, and household composition in order to avoid any bias to people who persist in the same area, in the same job or in the same family. The decision not to link with these characteristics reflects the sensitivity of hypothesis testing in history and the social science to bias (Ruggles 2006).[1]

The process has two computationally demanding steps. The first is to calculate how similar each 1871 record is to each 1881 record on each of the six characteristics. Then the system classifies each possible pairs of records as a match or non-match based on a score for their overall similarity. The classification is accomplished with a methodology, the SVM, used in a number of other classifications of historical census data (Christen 2008; Goeken et al. 2011; Richards et al. 2014). The classification software "learns" from a number of matches already confirmed as reliable on a case-by-case basis by expert genealogists. Without these "training data" the software would be unable to learn how to classify new pairs of records. We also use the individually prepared matches, or "true links", to assess accuracy.

An overview of the system is shown in Fig. 11.1. There are three main steps in the record linkage process. Step one consists of partitioning each census into smaller blocks to reduce the number of record pairs produced between the two censuses. Step two consists of comparing the records in each record-pair and creating a feature vector that contains information about how similar the records in the record-pair are to each other. In step three, the constructed record-pair feature vectors are labeled as matches or non-matches using a classifier that has learned from a training set constructed from both the 1871 and 1881 Canadian census data sets. During the comparison step, feature vectors are constructed for each record-pair (a, b) by comparing how similar the records attributes are to each other using various similarity measures. During the classification step, each feature vector is labeled as a match or non-match. The classification algorithm used in the classification step is a SVM classifier (Vapnik 1995). The SVM is trained on a labeled set of record-pair feature vectors constructed from the true links.

[1] The convention to understand and if possible avoid selection bias is part of the motivation for this paper. The consensus among social scientists on this point is sufficiently broad to recommend some reduction in linking accuracy in order to minimize bias, providing we also achieve a sufficiently small false positive link rate and a size of linked sample sufficiently large for hypothesis testing.

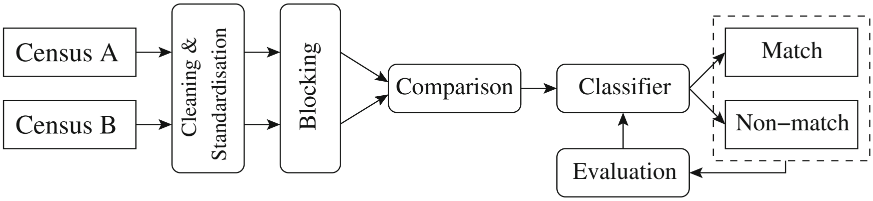


Fig. 11.1 Record linkage system

The record-pair feature vectors that are produced in the comparison step are given to the trained SVM classifier, from which they are labeled as positive or negative links. If the label for a feature vector is negative, the record-pair is seen as a non-match. If the label for a feature vector is positive, the record-pair is only seen as a match if each record is not found in another positive record-pair.

Four sets of true links are available to us: 8331 members of Ontario industrial proprietor families, 1759 residents of Logan Township, Ontario; 223 family members at St. James Presbyterian Church in Toronto and 1403 families of 300 Quebec City boys who were 10 years old in 1871.[2] The pairs of 1871 and 1881 records were established with additional information where available (e.g., church records in Toronto and Quebec City) although the chief criterion in all four collections was the census record of coresidence of other family members. Reliance on family context permits a high degree of confidence but biases the links toward those who persistently cohabit with the same family members. For example, we confirm the Logan and proprietor true links by (1) finding in both censuses at least one other household member (preferably two or more) with matching vital information, (2) making sure there is no significant contradictory information that makes a link improbable (for example, when one family member matches, but three others do not), and (3) determining that there is no other likely match in the 1881 Canadian census or the 1880 U.S. census.[3]

We have considerable confidence in the accuracy of the true links. They represent a useful diversity of population although, admittedly, they are not demographically representative insofar as they describe people living in the same family, or part of the same family in both years. This creates a bias to young children and married couples. Single people and those who became single over the decade (for example children leaving home) are underrepresented. Fortunately, even if the true links are not demographically representative, they still reflect the imprecision of

[2] The proprietors were linked in preparation for Inwood and Reid (2001). The Logan records were linked in preparation for Baskerville (2015). The St. James links were generated by Andrew Hinson for his doctoral dissertation (2010). The Quebec City links were made by the project *Population et histoire sociale de la ville de Québec* (www.phsvq.cieq.ulaval.ca) and kindly provided to us by Marc St-Hilaire.

[3] We check the United States census as well, because in this period Canadians could and did migrate to the United States.

information and name duplication needed to train the linkage system. Thus, our system will take this biased set of links and use it to produce new links that are less biased, more demographically representative and therefore more useful.

We use Ontario's high-performance computing grid SHARCNET (www.sharcnet.ca) because hundreds of millions of calculations are needed to compare name, age, place of birth, etc., and then to classify each pair of records as a match or non-match. Simply calculating similarities between millions of 1871 records and millions of 1881 records would require almost one year of continuous operation by a single processor.[4] Even running the system in parallel, however, a single run of the linkage system would be impractical without efficient code written in C, blocking to reduce the number of similarity comparisons and thresholding to remove some records from consideration.[5] We block by birthplace, marital status (allowing for obvious changes), first letter of surname, and our own first name groups (designed to allow for nicknames, diminutives, and unusual spelling variation). Similarity between pairs of names is assessed using the edit distance, Jaro-Winkler and double metaphone algorithms (Philips 2000; Winkler 2006). Similarity between birth years is assessed using a log-linear decay function. A description of the features used for linking and their similarity measures is given in Appendix Table 11.12.

Of course, many 1871 records cannot be matched because the individual died before 1881, left the country, or reported information differently in the 2 years. Nevertheless, the most common reason for failing to identify a match is not an inability to find someone with the same characteristics 10 years later. Rather, the biggest problem is that too many 1881 records have more or less the same characteristics as an 1871 record, and so produce multiple links. In such cases, we cannot identify which of the multiple links is correct. An example of records afflicted by the problem of "multiples" is given in Appendix Table 11.13.

The severity of the problem of multiples is clear from the distribution of outcomes for 1871 records, as reported in Table 11.1. About one-quarter are successfully linked in the sense that one 1871 record is classified as a match to only one 1881 record, and the 1881 record is matched to only one 1871 record. Another group comprising about one-quarter of the records cannot be linked with sufficient confidence to any 1881 record.[6] The largest group, 54 % of all 1871 records, consists of multiples. A multiple is an 1871 record that is either linked to more than one 1881 record or is part of a group of 1871 records linked to a single 1881 record,

[4] Computing similarity between all possible pairs of the 3 million and 4 million records on 8 string-based features with a single processor would require 343 days. Classifying each pair is additional.

[5] Blocking reduces the number of calculations. For example, we do not compare similarities between surnames beginning with different letters. Thresholding sets aside pairs of records that are sufficiently dissimilar that there is no prospect of being classified as a match.

[6] Thresholding and blocking remove 28 % of the 1871 records from consideration. A genealogical expert would be able to link some of these records but our automated system is less flexible. Table 11.1 reports the outcome of records submitted to the classification system.

Table 11.1 Outcome for 1871 census records in the classification system

	No. of records	Share
One-to-one links	550,726	0.215
No links returned	611,702	0.238
Many-to-one-and One-to-many links (multiples)	1,397,915	0.545

or both. Nevertheless, the system does report unique links for 550,000 people enumerated in 1871. This scale of longitudinal data is more than sufficient for most analysis providing these links are of sufficient quality. A careful assessment of these links is therefore needed.

11.3 The Level and Sources of Error Among the 1871–1881 Linked Records

In this section, we assess the level and sources of error among linked records. We begin with a general discussion of census data and their characteristics that make it difficult to link a substantial share of the records. This provides some context for the outcomes reported in the previous section. Next we assess the representativeness of the linked records using tabular descriptions and logistic analysis of the relative likelihood of linking different kinds of records. Finally, we point out that regardless of the overall error rate being low, a high proportion of the errors manifest themselves as individuals who have changed their location. This inflates the number of people who appear to have moved and complicates use of the linked data for migration analysis.

Our first question is if the system pairs up the right 1871 and 1881 records. Two kinds of mistakes are possible: an 1871 record can be linked to the wrong 1881 record, and an 1881 record can be paired to the wrong 1871 record. We assess the propensity for both errors by examining if the classification system has managed to identify correctly our true links, pairs of 1871–1881 records already linked with care by experts independent of the classification system. The fate of true links in the classification system indicates a combined incidence for both kinds of error of 3 %.[7] Is this a large or small number? We know that census data are in general somewhat imprecise. 3 % is similar to the rate for other sources of error in the North American historical censuses (Hacker 2013; Knights 1969; Parkerson 1991).

We might ask the same question of the 21 % rate of unique linking (Table 11.1). Is that high or low? Here is it useful to recognize that 30 % of our true links have surnames that differ by one or more letters and 20 % of the true links have name differences so large (edit distance > 0.15) that our classifier cannot find them. If the pattern of surname reporting in population is the same as in our true links, a full

[7] 3 % is the false positive rate on Ontario true links using a fivefold cross-validation method.

Table 11.2 Summary of probable limitations to potential link success

Records available	Loss of records	Reason and authority
100 %		
	20 %	Surname imprecision (true link analysis)
80 %		
	10 %	Age, birthplace, forename imprecision (true links)
70 %		
	5 %	Underenumeration estimate (Hacker 2013)
65 %		
	10 %	Emigration estimate (Emery et al. 2007)
55 %		
	10 %	Death estimate (Bourbeau et al. 1997)
45 %		Estimate of records available to be linked

20 % of the 1871 records cannot be linked for this reason alone. Imprecision in age, birth place, and first name reporting likely raises the "cannot link" share to at least 30 %. We also know that 10 % or more of the population would have died during the 1870s, and another 10 % would have emigrated. Another, smaller proportion may have been missed by enumerators.[8] Thus, we estimate, admittedly very roughly, that we are unlikely to be able to link more than 40–50 % of the records. We summarize the likely limitations to link success in Table 11.2. These estimates are of necessity somewhat speculative approximations.

The reason we achieve 21 % rather than 40–50 % is related to the reasons why there are any mistakes at all. Every time we cannot find the right person (for whatever reason), we are at risk of identifying the wrong person because of the widespread repetition of names, even among people with the same age, birthplace, and marital status. Multiple people who share a common set of characteristics are challenging in complicated ways. First, if a number of people have roughly similar characteristics (i.e., similar name, age, and birthplace), the system cannot distinguish among them, since a link cannot be accepted unless it is unique. Second, if the correct person reports age or name imprecisely, or if a woman changes her name at marriage, an incorrect person with similar characteristics might be selected in place of the correct one. In the first case no link is identified; in the second an incorrect link is made. A related problem arises if the correct person dies or emigrates before the next census, and therefore is not present in 1881. In this case, again, we are at risk of mistakenly selecting someone else with a similar combination of name, age, and birthplace.

Problems of this nature are more severe to the extent that names are common or that some kinds of people report their characteristics imprecisely. The imprecision means that occasionally we will connect together the wrong pair of records.

[8] Underenumeration in the nineteenth century the U.S. censuses is estimated to be about 5 % (Hacker 2013).

A second and perhaps more pervasive effect of imprecise reporting is to force a broadening of the tolerance for declaring a match. For example, we may accept any 1881 age between 28 and 32 for someone who reported 20 years in 1871 because someone is as likely to be 1–2 years off as to be exact in both years. Broadening tolerance, however, aggravates the problem of multiple links.

Classifying any data must strike a balance between broadening tolerance to avoid mistakes from a presumption of undue precision and; on the other hand, diminishing unique links by expanding the pool of multiples. It is particularly challenging to strike the right balance with our data because of their intrinsic imprecision. Many people did not remember their age or even their birth place correctly. The spelling of names varied a great deal. Enumerators who record information on the census manuscript page and volunteers who transcribe that information into a digital framework also made mistakes. In the face of this data imprecision, a combination of 21 % unique links and 3 % false positive errors (i.e., 3 % of the 21 %) reflects a successful balance of tolerances for linking characteristics. More importantly, the linked data are sufficient to identify and test hypotheses about broad patterns at the level of an entire population or large subpopulations.

11.3.1 Representativeness of the 1871–1881-Linked Records

Another way to assess our linked or matched data is to consider if they were broadly representative of the broader population. From the outset, we can anticipate reasons why linked records may be slightly atypical. We are more likely to link people with less common names and people who report their personal detail with greater precision and consistency. These biases are trivial unless they lead to other biases of greater analytical import.

In order to assess the implication of these and other biases, we compare the age and ethnicity of linked 1871–1881 records with the entire population in 1871. Here we use a subset of the linked records for which additional characteristics are available because they are part of a specially constructed 5 % representative sample. One effect is immediately apparent in Table 11.3: we link a much lower proportion of adolescents and young adults (15–25 years) than other groups. Young people are harder to link because they were of an age to move away from the family home, to start a new life, and to some extent reinvent themselves by reporting different characteristics. A propensity for women to change surname as they marry, of course, is an extreme example that leaves us with a noticeably smaller number of linked records for women aged 15–25 years.[9] The record-linking process is most

[9] We estimate, for example, that 40–45 % of single 15-year-old women in 1871 entered marriage during the following 10 years (comparing the number of single 15-year olds in 1871 with the number of married 25-year olds in 1881). The estimate is an approximation for only one birth year, however it suffices to indicate the scale of difficulty in linking young women. Only 2–3 % of women married someone with the same surname or retained their own surname in marriage.

Table 11.3 Age distribution of 1871 population and linked women and men

Age in 1871	Women		Men	
	Pop.	Linked	Pop.	Linked
0–14	0.38	0.40	0.39	0.38
15–25	0.24	0.16	0.22	0.19
26–55	0.31	0.37	0.30	0.35
56 and over	0.08	0.07	0.09	0.08

Source Canada, Census, 1871, 5 % microdata sample constructed at the University of Guelph http://census1871.ca (ignoring records for which age is missing). The linked records are generated by the People-in-Motion record-linking system (www.people-in-motion.ca) as described in Antonie et al. (2013)

successful for young children and the middle-aged, presumably because their information was reported more consistently over time. People over the age of 55 in 1871 are more difficult to identify in 1881 for a different reason—they were less likely to be alive in the latter year.

Interestingly, there is no bias to a more effective linking of the native born than of immigrants (Table 11.4). The foreign-born share of linked records is exactly the same as the foreign-born share of the population in 1871. The same is true for individual countries of birth (admittedly those born in England are overrepresented in the linked sample). This implies, unexpectedly, the linkage rate for immigrants is

Table 11.4 Distribution by nativity and ethnicity in the population and in linked records

	Pop.	Linked
Birthplace		
Foreign-born	0.19	0.20
England	0.04	0.06
Scotland	0.04	0.03
Ireland	0.06	0.06
Germany	0.01	0.01
U.S.	0.02	0.03
Canadian-born	0.81	0.80
Ontario	0.33	0.29
Quebec	0.29	0.30
Origin or ethnicity		
French	0.32	0.27
English/Welsh	0.20	0.27
Irish	0.25	0.23
Scottish	0.14	0.12
Continental Euro.	0.06	0.09
North American	0.01	0.003
African	0.01	0.005
Other	0.01	0.01

Source as Table 11.3

Table 11.5 Logit analysis (odds ratio) of 1871 records being linked uniquely, i.e., to a single 1881 record

		Married		Single/widowed	
		Women	Men	Women	Men
Male	1.18***				
Single	0.60***				
21–25	0.86***	0.85***	0.85***	0.71***	0.91**
>55	0.81***	0.87***	0.79***	0.99	0.65***
Fr. orig.	0.82***	0.72***	0.85***	0.91*	0.98
Illiterate	0.79***	0.67***	0.84***	0.81***	0.92
N	95,760	29,372	30,581	18,341	17,466

Note Full regression detail is available from the authors
*indicates that the coefficient differs significantly from 1.0 at 10 % confidence level
**indicates that the coefficient differs significantly from 1.0 at 5 % confidence level
***indicates that the coefficient differs significantly from 1.0 at 1 % confidence level

comparable to that of Canadian-born.[10] The linked records also mimic the population share of those born in the two largest provinces Quebec and Ontario.

There is some variance, however, with different ethnicities. Here we use the Canadian census category of "origin" as a measure of ethnicity. The information in Table 11.4 indicates a distribution of ethnicities roughly matching that of the population, with two important exceptions: fewer French-origin people are linked while the English origin are linked more successfully. The underrepresentation of people who report a French origin is notable.

11.3.2 The Likelihood of Establishing a Unique Link for Different Kinds of Records

We further investigate sources of linking bias in a logistic regression that considers the influence on being linked on age, sex, marital status, literacy, and if the individual reports a French origin.[11] The hazard, or odds ratios, reported in Table 11.5 indicate the contribution of each characteristic to the likelihood of being linked after controlling for other influences. A deviation from 1.0 indicates the size and direction of the effect; a number less/more than 1.0 indicates the odds of being linked for this category is less/greater than average. For example, in the first column

[10] This is unexpected because place of birth is reported more precisely for native born, to the level of province, in contrast to immigrants who simply report a country of birth. As well, immigrants or anyone moving a long distance has more scope for imprecise reporting of age, name, etc., than does someone living in the same location as his parents and family friends.

[11] We restrict the age categories being considered in this section because literacy is only available for people aged 21 or more years.

1.18 for men indicates they are 18 % more likely to be linked. The 0.60 reported for singles implies that they are 40 % less likely to be linked.

The odds ratios reported in the first column add to what we can learn from the previous tables. Men and married people are more likely to be linked; young adults and older people are less likely. These patterns conform to expectations. Singles are harder to link because they were more likely to change circumstances as they married (and of course most women changed their names). Younger adults were more likely to reinvent themselves as they left their parents' home. Some of them left the population entirely through emigration elsewhere in North America. Older adults were more likely to leave the population through death. We also see that people unable to read were less likely to be linked, as also for those reporting a French origin. The former is unsurprising. People lacking an ability to read probably reported their information with reduced precision. The French effect is more difficult to explain.

Partitioning the sample into married versus singles and men versus women allows more precise estimation of the age, ethnicity, and literacy effects (columns 2–5 in Table 11.4). For all groups, the youngest and oldest were less likely to be linked, but the effect was greatest for younger women (because of name-changing) and older men (because their 10-year survival rate was lower).[12] The French and illiteracy disadvantage is larger for women and for married people; the reason for these differentials is not immediately obvious. We do learn that the French effect is independent of literacy levels and age structure.

Records that are not linked fall into one of two groups: (i) we do not find even one good match in 1881 or (ii) we cannot identify the correct link because there are too many close possible matches.[13] We can estimate odds ratios for these effects separately (Tables 11.6 and 11.7). The odds of not finding of any match at all are large for older adults but this is offset by a smaller risk of losing sight of the correct match in a sea of multiple possibilities. In contrast, the younger adults are not at risk of being underlinked (Table 11.6) but they (especially single women) suffer a great deal from the problem of multiples (Table 11.7). For people reporting a French origin, the bias against finding a unique link arises primarily because of the failure to find even one possible link (similar to the older adults).

The challenge of finding unique links for the French-origin population leads us to estimate the odds of linking within this population. Table 11.8 reports the odds of finding at least one link. The pattern of odds ratios is very close to that of the general population (Table 11.6) with one exception. The impact of illiteracy on the odds ratio disappears for married men and becomes slightly stronger for married women.

[12] Similar patterns are observed if we abandon the restriction to people with 21 years of age or more. Literacy is unavailable for those under 21 but other effects are robust to the age restriction.

[13] For clarity, we highlight that the failure to find even one potential match can occur two ways: if the 1871 record is removed during the initial filtering or if it survives the filter but the classifier does not recognize any 1881 records with sufficient similarity.

Table 11.6 Odds ratios for finding at least one link for each record

	Married		Single/widowed	
	Women	Men	Women	Men
21–25 years	0.97	0.99	1.11***	1.16***
>55 years	0.60***	0.68***	0.51***	0.42***
Fr. origin	0.75***	0.78***	0.71***	0.74***
Illiterate	0.85***	0.91***	1.11**	1.01
N	29,372	30,581	18,341	17,466

Significance levels as in Table 11.5

Table 11.7 Odds ratios for finding only one link among the linked records

	Married		Single/widowed	
	Women	Men	Women	Men
21–25 years	0.83***	0.83***	0.62***	0.80***
>55 years	1.29***	1.11**	1.83***	1.22**
Fr. origin	0.82***	0.98	1.19***	1.22***
Iliterate	0.69***	0.87***	0.73***	0.88*
N	15,561	16,402	7,718	8,835

Significance levels as in Table 11.5

Table 11.8 Odds ratios for finding at least one link, French origin only

	Married		Single/widowed	
	Women	Men	Women	Men
21–25	1.08	1.13*	1.25***	1.16**
>55	0.65***	0.60***	0.49***	0.41***
Illiterate	0.83***	1.02	1.10*	1.11
N	9172	9670	6440	4527

Significance levels as in Table 11.5

Interestingly, although levels of illiteracy were higher in the French-origin population, and they are less likely to be linked, literacy patterns apparently did not contribute to the link bias (with the exception of married women).

Decomposing the link bias into two stages has not helped a great deal to understand the underlinking of older adults, people of French origin, and married people who cannot read. For these groups, we know only that we are less likely to find even one good link. Why that is the case remains unclear. The two-stage approach does help, however, with younger adults and singles who cannot read. We learn that there is a better than average prospect of finding a match for these groups (Table 11.6). Indeed, the problem is that we find too many good matches and in consequence cannot discriminate amongst them (Table 11.7). Any strategy for disambiguation of multiples might be especially helpful for the young adults.

We conclude that although the linked records are roughly representative of the 1871 population by birthplace and by major age and sex categories, there is some bias.

It is easy to see why younger single women and older men are less likely to be linked. Reweighting the linked sample by demographic category is an easy way to limit the impact of this bias in any analysis of the linked records.

There is a small but noticeable effect of illiteracy on the odds of being linked. The few people who described themselves as being unable to read were less likely to be linked. This must be kept in mind for any social or economic analysis using the linked sample. Fortunately, only a small share of the population was unable to read (about 10 % of young adults and 20 % of those aged 55 years or more).

There remains a mystery about the difficulty of linking people of French origin. This group comprises nearly one-third of the population. One possible explanation is that the quality of enumeration was influenced by language. Lower quality enumeration of the French-descended population might imply less precise or less consistent information that, in turn, would be more difficult to link. There is no reason, however, to think the census was undertaken less carefully in Francophone districts. A Quebec intellectual headed the Census Bureau in 1871, regional directors were drawn from the respective jurisdictions and most enumerators in French-speaking areas were themselves Francophone (Curtis 2000; Inwood and Kennedy 2012). Admittedly, any francophones relocating to English Canada were at greater risk of name misspelling.[14]

Dillon (2006) suggests (a) that the relatively small pool of French names increases the incidence of multiple links and makes it harder to isolate a unique match and (b) that the transcription of the 1881 census was weaker for French names. Both effects are plausible. Another possible influence is faster emigration of the French-descended population during the 1870s (Emery et al. 2007). Differential emigration and perhaps mortality would explain at least some part of the 25 % lower odds of finding at least one link for French-origin men and women (penultimate row of Table 11.5).

11.3.3 Error Rates Among Movers Versus Stayers

Linked or longitudinal census data are often used to describe and analyze mobility —both social and geographical. Elsewhere, we consider the broad patterns of occupational mobility during the 1870s (Antonie et al. 2015). Here we consider error rates among those who change location in order to assess the usefulness of these data for the study of migration. Since there are insufficient "movers" within our true links to support a comprehensive assessment along the lines reported in Sect. 11.3, we revert to a simpler strategy of checking if the individual links are

[14] The Jaro-Winkler and edit distance similarity measures are not phonetic and carry no obvious bias against recognizing similarities in the French language. Our third similarity measure, double metaphone, is phonetic but has been designed to minimize bias against languages other than English.

"credible" or not. This differs from the earlier evaluation insofar as we do not begin with secure knowledge acquired independently of the linkage system. Rather, we assess select links produced by the system in a way that relies to a large extent on the continued coresidence of other family members.

This process differs in principle from the generation of true links (above). Here we do not attempt to identify which 1881 record, if any, represents the same person as the 1871 record. That would require a broad investigation of all possible 1881 matches. The current process is more restricted and much less costly. We ask if the 1881 match recommended by the system has coresident family members who resemble those of the 1871 record using structured criteria (see Appendix 1). There might be a number of 1881 records with similar coresidents, but these are not checked. Rather, we assess the "credibility" of the one record selected by our linkage system.

This process is imprecise to the extent that we ignore other possible matches that, if examined, might reduce confidence in our results. Clearly, this implies a bias in favor of accepting matches recommended by the system. There are other sources of imprecision. For example, we use the coresidence in 1881 of people who would not be expected to be absent (given what we know from the 1871 family) as evidence undermining credibility.[15] And yet, families change for good reason; it is entirely plausible that family configuration changes and thereby creates the appearance of contradictory information. In these situations we may have a bias against acceptance of the correct match. Another complication is that we can assess the credibility of only those matches who have coresident family members in both years. We can say little about the credibility of links involving people who live alone or with non-family members in one or other year.

Although the process is particular in these ways, it provides an economical but plausible check on all kinds of linked pairs, with no obvious bias between different kinds of records. We use the method to compare people who appear to have changed provinces and those who do not. The distribution of linked pairs between interprovincial movers and stayers is reported in Table 11.9. Two verification assistants, independently, have checked each linked pair. Any differences are adjudicated; we report only those pairs on which there is consensus after adjudication.

We report our assessment of a random selection of links in Table 11.10.[16] The summary indicates a large difference between the movers and the stayers. Links for those people who stayed in place are highly credible; 83 % of the stayers are deemed credible (A and B categories) and only 5 % look to be incorrect (category D). In contrast, nearly half (45 %) of the linked pairs involving a change of province are incorrect.[17] The difference is dramatic, and invites explanation.

[15] Category D in Appendix 1.

[16] We examine a random selection of linked pairs for both movers and stayers.

[17] Although this example has only 39 movers, a larger sample of 1363 movers checked with a slightly different method had a comparable 42 % being deemed unlikely links.

Table 11.9 Distribution of 1871–1881 linked pairs by gender and interprovincial movers versus stayers

	Female	Male	All
No. of linked pairs	247,663	303,030	550,726
No. of links with change in province	8037	9848	17,910
Apparent movers as a share of all links	0.032	0.032	0.033

Table 11.10 Individual assessment of linked pairs implying movement between provinces from 1871 to 1881

	Movers	Stayers
Number of records checked	39	1787
Share assessed highly credible (A)	0.46	0.76
Share assessed credible (B)	0.05	0.09
Share that cannot be confirmed (C)	0.15	0.10
Share assessed likely incorrect (D)	0.33	0.05

NB: Here we report linked 1871–1881 pairs for which two independent assessments agree after adjudication. "Movers" are records that imply a change in province of residence. The assessment categories are described in Appendix 1.

One reason for errors among the reported movers is that some proportion of 1871 records cannot be linked properly. For example some individuals died or left the country before 1881, or were present and overlooked by enumerators, or were enumerated in 1881 with some misstatement of personal information.[18] Situations like these prevent the system from making a correct link.[19] Further, as noted above, when the correct link is not available, the system may identify incorrectly someone else with similar personal characteristics. For example, a 48-year-old woman named Joanna Munroe who in 1871 was enumerated in Southampton, New Brunswick, was linked in 1881 to Jane Munroe, a 58-year-old from Lingan, Nova Scotia. While all the linkage criteria match very well (only the first name is off), different coresident families make it clear they are different people. The linkage error is attributable to the 1881 Jane being enumerated as Jessie (a Scottish nickname for Jane) in 1871, and the fact that the 1871 Joanne Munro had likely died by 1881. Because location is not used for linking, mistaken 1881 links like these will have a wide geographical distribution. If the mistaken link is in another province the system can generate a "phantom mover".[20]

[18] Socially marginal groups such as aboriginal, African-descendants or Chinese are more likely to be enumerated with substantial imprecision (Reid 1995; Fryxell et al. 2015).

[19] The most careful, genealogical-like researchers seldom manage to surpass an 80 % rate of linking from one Canadian census to another, for exactly these reasons. See Darroch (2015), Baskerville (2015) and Olson (2015).

[20] Ron Goeken at the Minnesota Population Center first suggested this interpretation of the relationship between geography and errors in linking.

Table 11.11 Simulated share of observed state changes that are correct

True rate of state changes						
	0.03	0.05	0.1	0.2	0.3	0.5
# possible states						
2	0.47	0.34	0.21	0.12	0.08	0.05
3	0.54	0.41	0.26	0.15	0.10	0.07
4	0.57	0.44	0.29	0.17	0.12	0.07
5	0.58	0.46	0.30	0.17	0.12	0.08

Notes The simulation is based on the idea that a characteristic for a record has a number of possible states, e.g., for the locational characteristic, there might be two locations, or three locations, etc. For a consideration of interprovincial movement, there are four possible states in 1871, corresponding to the four provinces. We assume records are distributed equally across all possible states, i.e., if there are 2 states, they are 50 and 50 %. If 4 states, each has 25 %. This is like assuming the four provinces in 1871 are of equal population size. Further assume that any mistake in linking is random with respect to states/locations (e.g., if there only two locations, any "mistake" will be in the same place in 1881 as in 1871 half of the time). The other half of the time the mistake will register as a change of state. If there are three states, the mistakes will appear as a change of state two-thirds of the time. Some correct links also appear as a change of state since some people really do change provinces. We predict the likely number of true and phantom movers under these simple assumptions, and report the phantom share of reported movers in Table 11.11. Formally, the table is generated as

WM (wrong movers) = P (population size) * FPR (false positive rate) * $(S - 1)/S$; S is number of states
CM (correct movers) = P (population size) * TPR (true positive rate) * TRS (true rate of state change)
Phantom movers rate = WM/(WM+CM)
True movers rate = CM/(WM+CM)

This phenomenon complicates use of the data because people who really did not move but were linked incorrectly contaminate the evidence of movement. Indeed, if there are few genuine interprovincial movers, as in the 1870s (Baskerville 2015) then a large share of the apparent movers may be mistaken, and the overall level of mobility is exaggerated significantly. The overall error rate is still 5 % or less, but *among the reported movers* the proportion of mistakes can be much higher.

A simple simulation in Table 11.11 illustrates that an uncomfortably large proportion of the apparent movers will be mistakes if the true extent of movement is less than 15 %. Changes in religion, occupation, etc., will have a similar problem. The implication is that analysis of change by a small proportion of the population will be subject to more uncertainty than is suggested by the overall error rate of 5 %.[21] In practice, of course, the severity of this complication depends a great deal on particular circumstances, as illustrated in Table 11.11.

[21] This is independent of how well the system links people who really did move; the problem is not the quality of data describing true movers. That said, movers were disproportionately young adults who generally are more challenging to link. For this reason, the system may generate a higher rate of error among true movers. The only way to assess this possibility would be to generate more true links than currently are available.

11.4 Summary and Observations

The application of machine-learning systems to historical censuses generates useful data describing people at different points in their lives (Ruggles 2006). The method is especially important for jurisdictions that lack comprehensive church or public vital registration and must depend on the census for understandings of population-wide experience. The new longitudinal source provides, for the first time, large-scale and near-representative life-course information about nineteenth century Canadians. This is an important and very welcome development.

The nature of the source and underlying population does not allow us to link every record. Nevertheless, as we demonstrate with the 1871 and 1881 Canadian censuses, it is possible to generate samples large enough for most historical and social science research. The overall quality of the data, as reflected in a low rate of false positive links, is excellent. A carefully designed system brings the false positive rate down to an acceptable range, circa 3 % on independently verified links.

We assess the extent of bias or representativeness of the linked pairs by examining unconditional means and with logistic analysis of the propensity to link. We find that birthplaces are reasonably representative of the population. The linking method is slightly more successful for immigrants born in the British Isles but otherwise it roughly replicates the proportions of the population born in Canada versus immigrants and in one province versus another. People who were unable to read are noticeably more difficult to link, but they account for a small share of the population. Older men and young adults are more difficult to link than people at other ages. The former reflects differential mortality at advanced ages; the latter probably reflects change accompanying the departure of children from a family home. A near universal tendency for women to change their surname at marriage is the largest single complication in this vein.

A lower rate of linking people who report a French origin in 1871 is more puzzling. There is no reason to think that the enumeration of Francophone communities was in any way inferior. Logistic analysis rejects the hypothesis that ethnic differences in literacy are responsible. Literacy matters, but it does not explain the ethnic differential in linking. Breaking the process into two stages, identification of at least one promising match and discrimination among multiple possibilities, points to the first stage as especially challenging for the French-origin records. Again, however, there is no reason to think blocking or the use of similarity measures in the first stage carries a bias against French language names. Further investigation of similarity algorithms for French names may prove useful.[22]

[22] It is worth noting that the Canadian census category of 'origin' is itself obscure. People were asked their 'origin' in the sense of ancestry or ethnicity, but as best we know no instructions were made available about how to identify in the event of mixed ancestry. There is likely to have been some discretion in the self-identification of origin. An improved understanding of this process may help us to understand why French origin Canadians are more difficult to link.

Analysis of the odds of linking shows that the underrepresentation of French-origin population is pronounced only for married people (and is especially large for married women). Until this problem is better understood, it would be prudent in most research to reweight observations to correct the underrepresentation of French-origin married couples.

One final problem, a higher rate of mistaken links among those who appear to move between provinces, is easier to understand. The bias arises because a linking error for any reason is likely to generate the appearance of geographic relocation. The problem of phantom migrants looms large when the true rate of moving is low. In principle we might mitigate the effect by adjusting standard errors for hypothesis testing, but in practice this is difficult because we do not know the true rate of moving independent of the analysis. As a practical matter, therefore, when the reported rate of changing category falls below 15 %, it would be prudent to verify the intrinsic credibility of linked pairs implying a change of state. Admittedly, verification is only possible for those who continued to live with the same family members. Thus, even after a process of verification, linked data cannot be used to analyze the relationship between family evolution and migration if the rate of reported movement is small.

Our assessment of the linked records identifies specific limitations notwithstanding their excellent quality overall. Some problems are small enough to ignore (impact of illiteracy, small deviations in birthplace composition). Others require a simple reweighting to compensate for underrepresentation (younger single women, older men, French-origin married couples). The clustering of errors among movers when only a small proportion appears to move requires more caution and where possible manual verification. These are manageable problems, which further research and improvements to the record-linking system may reduce further.

Our experience linking historical census data indicates that, for this case at least, an optimal application of machine-learning methodology takes account of the quality of underlying data. Of course, this is only one case study. Nevertheless, if our experience were to be replicated elsewhere, it would be useful practice for computing science researchers to take data characteristics into account in their application of otherwise standard machine-learning methods. There is a comparable lesson for social science and historical users. If the issues encountered with linking the Canadian data were to obtain elsewhere, social scientists and historians would find it useful to assess and accommodate data quality issues that arise from the intersection of sophisticated machine-learning methodology and sometimes messy historical data.

Acknowledgments The authors are grateful for financial support from the Canadian Foundation for Innovation, Ontario Ministry of Research and Innovation, Social Sciences and Humanities Research Council, Google, Sharcnet and the University of Guelph. We would also like to thank our genealogical collaborators, the Ontario Genealogical Society, Ontario GenWeb and Family Search.

Appendix 1: Protocol for Checking Automatically Generated Links

We check the reliability of links in order to prepare Table 11.10 and assess the relative "movers" and "stayers." This process differs from that of determining true links insofar as (i) we do not rule out the possibility of other, equally plausible matches and (ii) we cannot bring to bear any insight from the independent study of some community or subset of the population. Checking involves two independent experts assessing a link without reference to each other's decision (blind double-checking). Each link is assessed based on the household information in the two census years, as well as the consistency of information, and then assessed with a quality letter grade. The basic question being asked and answered is the common genealogical query: Is this the same person in both records? (Tables 11.12 and 11.13)

Table 11.12 Description of linking features

Original attribute	Type	Similarity measure(s)	Feature score
Last name	String	Edit distance (ED): the minimum number of single letter edit operations needed to convert string A into string B	Float [0–1]
		Jaro-Winkler (JW): calculated based on the number of common characters, character transpositions and string length between two strings, giving preference to strings that share a common prefix	
		Double metaphone (DM1, DM2): transforms strings into their corresponding phonetic representation, creating a primary and secondary representation on which edit distance is applied	
First name	String	See above	Float [0–1]
Age	Integer	$$F(x) = \begin{cases} 1 & \text{if } x \in 0, 1, 2 \\ 1-1/x & \text{if } x \in [3,10] \\ 0 & \text{otherwise} \end{cases}$$	Float [0–1]
Gender	Binary	Exact match	Binary (0,1)
Birthplace	Categorical	Exact match	Binary (0,1)
Marital Status	Categorical	Rule based 1 if valid status change (ex. single to married) 0 otherwise	Binary (0,1)

Note Blocking techniques are applied on three different attributes to reduce the number of record-pairs being compared. These attributes are a name-code based on the first name, the first letter of the last name and birthplace. This means that a record-pair is considered for comparison only if the two records reside in the same name-code and last name block of their respective censuses, and their birthplaces match

Table 11.13 An example of census records with similar attributes

Surname	Forename	Age	BPL	Marital status
1871 Census				
Barns	Mary	11	15030	Single
Barns	Mary	9	15030	Single
Barns	Mary	8	15030	Single
Barns	Mary	12	15030	Single
Barns	Mary	10	15030	Single
Barns	Mary	10	15030	Single
1881 Census				
Barns	Mary	20	15030	Single
Barns	Mary	22	15030	Single

BPL = birthplace

In addition to the grading, experts provide reasons for their decisions by recording the answers to certain questions. This information (a) helps us refine our linkage system, (b) allows us to compare decision-making between coders and ensure consistency, and (c) possibly change quality grades in future without having to revisit the links manually.

1.1 Links: Primary and Subsidiary

A primary link is the one that the system linked using six linking variables (First Name (FN), Last Name (LN), Age, Marital Status (MS), Birthplace (BPL), and Sex), and this kind of link is the one that we are interested in giving a link quality assessment. In the course of checking the primary links, we may also see other people who link up. These we call a subsidiary link and are usually a household member of the primary link whom we have determined with good confidence is the same person in both years. It may be a spouse, sibling, or child or parent, or even servant.

1.2 Deciding on Quality

The six linking variables used by the automated linkage system to generate the primary link are likely to be very consistent, and so not very useful for distinguishing false positives by themselves (although commonness of surnames could be a consideration). Accordingly, in order to verify a link, checkers consider the household/family context as well as other personal fields (of primary and provisional subsidiary links) and also to assess whether or not they appear consistent.

1.3 The Questions to Ask

- Household/family Context—does the family have some of the same members in both years? Are family member details (FN, LN, Age, MS, BPL, Sex, Origin, Religion, etc.) consistent?
  - Does the spouse match? (Could the linked person have remarried?)
  - How many children match by name and age? (exclude those who were born in census period, i.e., those aged under 10)
  - Are there any other family members that are the same? (e.g., parents, servants)
  - Does the household transition make sense—deaths, leaving family to start new family, etc.
  - Are children on the same age ladder?
- Is birthplace consistent?
- Is ethnic Origin consistent? (children's origins sometimes change to follow one or other of the parents)
- Is Religion consistent, or show a likely transition (i.e., more likely between Protestant denominations than between Catholic and Protestant)?
- In some circumstances, contradictions in other fields may also give a reason to look more closely at a link (e.g., an unlikely occupational change, or an eastward (as opposed to westward) long distance (i.e., interprovincial) move).

The link checking protocol may include one or more of the above questions in explicit form as fields to be filled out, and these are usually designed to require notation only in cases where the answer is unexpected. In addition, there is a Comments field, in which checkers can indicate other difference in the information given for the same person in the two censuses.

1.4 The Link Quality Typology

The assessment of link quality is a holistic summary of the answers to these questions, with the primary consideration being the matching of family members, although contradiction of information is considered. The qualities are:

A = Two or more family members match

- With no major contradictory information (such as children appearing in one census that were not there ten years before)
- In some cases, neighboring families can be used to make an A (see CM and CF below)
- e.g., Spouse and child

- In very rare cases an A can be achieved with fewer than two subsidiary links if there is certainty it is the same person (e.g., in a case where a man has no family by next census (wife dead and children grown up) and he has moved in with neighbors/other family that are evident in both census but whose records not linkable because they are not in the link set in the first census year.)

B = One family member matches

- e.g., spouse or child
- With no major contradictory information

C = Possible match but no family in one year to confirm against

- e.g., single man in rooming house in 1871, or a man in barracks in 1881
- Information otherwise very consistent

CM = Single to married man with new family by next census

- MS will change from single to married, and children (if any) will be below the age of 10.
- e.g., a single man in 1871 (on his own or in a family) got married and started his own family by 1881.
- When possible, we check if family members are in neighboring households—in this case CM might be upgraded to an A.
- If the man is a widow in the first year, and married the next, then it is a CB.

CF = Single to Married woman with new family by next census (Rare)

- MS will change from single to married, and all children will be below the age of 10.
- e.g., a single woman in 1871 (on her own or in a family) got married and started a own family by 1881.
- Some women did keep their own names in some cases (French and Scottish), but in most cases single to married women with the same surname will be bad matches (D) (linking criteria may even prevent a link in the first place). Therefore, this code is used only when there is very good evidence it is the same person (e.g., she retains her maiden name in the married family (husband has different surname); or there is evidence she married a man with the same name (possibly a neighbor); or she has been enumerated with the same (birth) family in both years).
- When possible, we check if family members are in neighboring households—in this case CF might be upgraded to an A.

CB = Possible match, but for less common reasons

- These are possible matches where families do not match, but links may be possible. Examples of these are:
  - Widow/Widowers—A older married man or woman with family is alone by the next census and marital status has changed to widowed/divorced/separated (or possible married/spouse absent)
  - A single person who has joined a different family to work as a servant in 1881
  - Spinster/Bachelors who change families
  - A man with only a wife in both years, but wife's name might change, however all other information about her stays the same, and they still have the same neighbors.
  - If the man is a widow in the first year, and (re)married the next.

D = evidence of wrong match

- e.g., families are different, and/or there is significant contradictory information.

1.5 Evaluation/Arbitration

When checkers disagree on the quality of a link or whether a newID should be assigned, the records are either reevaluated by the checkers or arbitrated by a third party for a final decision.

References

Antonie, L., Inwood, K., Lizotte, D., & Ross, J. A. (2013). Tracking people over time in 19th century Canada. *Machine Learning, 96*(1) (S1), 129–146.

Antonie, L., Baskerville, P., Inwood, K., & Ross, J. A. (2015). Change amid continuity in Canadian work patterns during the 1870s. In P. Baskerville & K. Inwood (Eds.), *Lives in transition: longitudinal research from historical sources* (pp. 120–140). Kingston: McGill-Queen's University Press.

Baskerville, P. (2015). Wilson benson revisited: Movement and persistence in rural Perth County, Ontario, 1871–1881. In P. Baskerville & K. Inwood (Eds.), *Lives in transition: longitudinal research from historical sources* (pp. 141–164). Kingston: McGill–Queen's University Press.

Bourbeau, R., Légaré, J., & Édmond, V. (1997). *New birth cohort life tables for Canada and Quebec, 1801–1991*. Ottawa: Statistics Canada.

Christen, P. (2008). Automatic record linkage using seeded nearest neighbour and support vector machine classification. *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 151–159.

Curtis, B. (2000). *The politics of population: State formation, statistics, and the census of Canada, 1840–1875*. Toronto: University of Toronto Press.

Darroch. G. (2015). Lives in motion: Revisiting the 'Agricultural Ladder' in 1860s Ontario, a study of linked microdata. In P. Baskerville & K. Inwood (Eds.), *Lives in transition: longitudinal research from historical sources* (pp. 93–119). Kingston: McGill–Queen's University Press.

Dillon, L. (2006). Challenges and opportunities for census linkage in the French and English Canadian context. *History and Computing, 14*(1–2), 185–212.

Emery, H., Inwood, K., & Thille, H. (2007). Hecksher-Ohlin in Canada: New estimates of regional wages and land prices. *Australian Economic History Review, 47*(1), 22–48.

Ferrie, J. P. (1996). A new sample of males linked from the public use micro sample of the 1850 U.S. federal census of population to the 1860 U.S. Federal census manuscript schedules. *Historical Methods, 29*, 141–156.

Ferrie, J. P. (1999). *'Yamkees Now': European immigrants in the antebellum U.S., 1840–1860*. New York: Oxford University Press.

Fryxell, A., Inwood, K., & Van Tassel, A. (2015). Aboriginal and mixed race men in the Canadian expeditionary force 1914–1918. In P. Baskerville & K. Inwood (Eds.), *Lives in transition: Longitudinal research from historical sources* (pp. 254–273). Kingston: McGill–Queen's University Press.

Fu, Z., Boot, M., Christen, P., & Zhou, J. (2014). Automatic record linkage of individuals and households in historical census data. *International Journal of Humanities and Arts Computing, 8*(2), 204–225.

Goeken, R., Huynh, L., Lenius, T., & Vick, R. (2011). New methods of census record linking. *Historical Methods, 44*(1), 7–14.

Hacker, D. (2013). New estimates of census coverage in the United States, 1850–1930. *Social Science History, 37*(1), 71–101.

Hinson, A. (2010). Migrant scots in a British City: Toronto's scottish community, 1881–1911. Ph. D. Dissertation University of Guelph.

Inwood, K., & Kennedy, G. (2012). A new prosopography: The enumerators of the 1891 census in Ontario. *Historical Methods, 45*, 65–77.

Inwood, K., & Reid, R. (2001). Gender and occupational identity in a Canadian census. *Historical Methods, 32*(2), 57–70.

Knights, P. R. (1969). A method for estimating census under-enumeration. *Historical Methods Newsletter, 3*(1), 5–8.

Knights, P. R. (1991). *Yankee destinies: The lives of ordinary nineteenth-century bostonians*. Chapel Hill: University of North Carolina Press.

Olson, S. (2015). Ladders of mobility in a fast-growing industrial city: Two by two, and twenty years later. In P. Baskerville & K. Inwood (Eds.), *Lives in transition: Longitudinal research from historical sources* (pp. 189–210). Kingston: McGill–Queen's University Press.

Parkerson, D. (1991). Comments on the underenumeration of the U.S. census, 1850–1880. *Social Science History, 15*(4), 509–515.

Philips, L. (2000). The double metaphone search algorithm. *C/C ++ Users Journal, 18*, 38–43.

Reid, R. (1995). The 1871 United States census and black underenumeration. *Histoire sociale/Social History, 28*, 487–499.

Richards, L., Antonie, L., Areibi, S., Grewal, G., Inwood, K., & Ross, J. A. (2014). Comparing classifiers in historical census linkage. *Data Integration and Applications Workshop, in Conjunction with IEEE ICDM 2014*.

Ruggles, S. (2006). Linking historical censuses: A new approach. *History and Computing, 14*(1–2), 213–224.

Steckel, R. H. (1988). The health and mortality of women and children, 1850–1860. *The Journal of Economic History, 48*(2), 333–345.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Heidelberg: Springer.

Winkler, W. E. (2006). Overview of record linkage and current research directions. *Statistical Research Division Report*. U.S. Census.

Wisselgren, M. J., Edvinsson, S., Berggren, M., & Larsson, M. (2014). Testing methods of record linkage on swedish censuses. *Historical Methods, 47*(3), 138–151.

Chapter 12
Using the Canadian Censuses of 1852 and 1881 for Automatic Data Linkage: A Case Study of Intergenerational Social Mobility

Catalina Torres and Lisa Y. Dillon

Abstract This chapter discusses the issues of missing and uncertain data in the Canadian census sample of 1852 within the context of automatic linkage with the complete census of 1881. The resulting linked sample from these two censuses was created to provide an opportunity to study intergenerational social mobility in Canada between fathers (in 1852) and sons (1881). We discuss the accuracy and representativeness of the automatically generated links and show how the use of marriage registers can be helpful in order to verify the results of the automatic linkage. Our verifications suggest that most of the links are accurate. However, the linked sample is not representative of some subgroups of the studied population, since some attributes favoured while others hindered the fact of being automatically linked from 1852 to 1881. Finally, based on our efforts of manual linkage between the BALSAC marriage registers and the automatically linked census sample for the verification of the latter, we present some considerations about the great research potential of linking census and parish register data in Quebec.

12.1 Introduction

Between 2004 and 2006, the Programme de recherche en démographie historique (PRDH)[1] created a 20 % sample of the first nominal census of Canada in the nineteenth century: the census of 1852. The quality of this census has been criticized by some researchers. For example, Gagan (1974) described the "lack of

[1] Research programme in historical demography, Université de Montréal.

C. Torres (✉) · L.Y. Dillon
Programme de recherche en démographie historique (PRDH), Université de Montréal,
Montreal, Canada
e-mail: catalina.torres@umontreal.ca

L.Y. Dillon
e-mail: ly.dillon@umontreal.ca

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_12

consistency" of this census in making reference to the irregular quality of the 1852 census manuscripts. A more recent critique made by Curtis (2001) concerns the combination of the approaches *de jure* and *de facto* in the taking of the 1852 census. By this combination of approaches, the Canadian population of 1852 could be overestimated. Dillon and Joubert (2012), who have examined the 20 % sample of the 1852 census in the light of those critiques, suggest that the remarks made by Gagan and Curtis regarding the quality of this census concern a minority of the observations. Thus, the 20 % sample of the 1852 census offers unique opportunities to broaden our knowledge about the Canadian population of the mid-nineteenth century, particularly the rural population (Dillon and Joubert 2012).

Both the census sample of 1852 and the 100 % database of the 1881 Canadian census constitute rich sources of information about the Canadian population of the mid- and late-nineteenth century. For example, both sources contain valuable socio-economic variables and provide information at the individual level, making these data suitable for record linkage.[2] For instance, by linking individuals (e.g. the boys of a certain age) from 1852 to 1881, phenomena such as the intergenerational social mobility can be studied through a comparison of the occupation of the fathers (in 1852) and the sons (in 1881).[3]

This chapter analyzes an automatically linked sample from the Canadian censuses of 1852 (20 % sample) and 1881 (complete census). Since the aim of this sample is to provide opportunities to study the intergenerational social mobility between fathers (in 1852) and sons (in 1881), the linkage efforts were concentrated on a limited subgroup of the population, namely the boys aged from 0 to 15 years, living mainly in a rural area in the provinces of Ontario or Quebec in 1852. In total, our linked sample contains information about 4226 individuals linked from 1852 to 1881. This linked sample was created for exploratory purposes in the framework of the international project *Mining Microdata: Economic Opportunity and Spatial Mobility in Britain, Canada and the United States, 1850–1911.*[4] The two census data sources were provided by the PRDH. The linkage between both data sets was performed in the Historical Data Research Unit (HDRU) at the University of Guelph, while the Mining Microdata project is pursuing a parallel linkage effort of the 1852 and 1881 censuses headquartered at the Minnesota Population Center.

In order to analyze this linked sample, we start with a brief description of the linkage technique, followed by a discussion of the results of the automatic linkage. This discussion includes some considerations of the factors that affect the linkage

[2] The PRDH has lengthy experience with the record linkage of Quebec parish registers, and more recently undertook linkage of a sample of the 1871 Canadian census to the 1881 census. Our current effort to link the 1852 and 1881 Canadian censuses is funded by the international project *Mining Microdata: Economic Opportunity and Spatial Mobility in Britain, Canada and the United States, 1850–1911,* Digging Into Data Challenge, ESRC/NSF/CRSH.

[3] In both censuses, the only information about the socio-economic status of the individuals is the occupation.

[4] This project aims to contribute to the discussion about the social and geographical mobility in North America and in Great Britain in the late nineteenth and early twentieth centuries.

success, such as mortality and emigration. Indeed, linking individuals from 1852 to 1881 is a substantial challenge, since the larger the interval of time between the two observations, the more the individuals could be lost and become impossible to link due to those factors. Following this discussion, we present some analyses of the accuracy and representativeness of the automatically generated links. In this stage, we will show how the use of the BALSAC marriage registers (Balsac fichier de population 2013) can help assessing the validity of those links. Finally, we will discuss the utility of linking census data in order to study intergenerational social mobility between fathers and sons from 1852 to 1881.

12.2 Linkage Technique and Results

The automatic linkage procedure employed to generate our linked sample from 1852 to 1881 is very similar to that explained by Antonie et al. (2015, this volume) regarding the linkage between the Canadian censuses of 1871 and 1881. This procedure is based on the individual attributes that should not change over time, such as first and last name,[5] gender and place of birth. Other characteristics that change over time but in a predictable way, namely age and the marital status, are also used in the linkage procedure. We limited our record linkage criteria to this range of attributes to avoid biasing the sample in favour of stable individuals; this kind of bias could occur if variables which change over time, such place of residence and occupation, were used to link cases.

The linkage procedure—explained in more detail by Antonie et al. (2015)—starts with data cleaning and standardization. Once the variables previously mentioned are cleaned and standardized, the observations of the two data sets can be compared in order to find the 1881 record that corresponds to each individual in 1852. In order to reduce the number of comparisons, blocking by some characteristics is useful, for example by birthplace at the country or provincial level. This means, for instance, that boys born in Canada according to the 1852 census are compared only with men born in Canada according to the 1881 census. Similarity scores for the comparison of each attribute between two individuals as well as global score are generated. Based on the similarity scores, the link is accepted when there is only one candidate who passes a certain threshold (one-to-one approach). The candidates are chosen by a support vector machine (SVM) programme, which uses training data—a previously generated set of manual links made by genealogist experts—as a guide for the acceptable links.

[5] Since for our purposes we were interested in linking men only, we are not faced to the problem of changing name at marriage. This was usual among women and more frequent among some subgroups of the population than others.

Table 12.1 Survivors by age group, males aged 0–19 living in rural Quebec or rural Ontario in 1852

Age group (1852)	N (1852)	Survivors in 1881 (%)
0–4	18,492	73.6
5–9	18,015	80.6
10–14	15,425	78.9
15–19	13,638	75.6

Sources Canadian census of 1852 (20 % sample)

Despite the fact that some differences in name spelling between 1852 and 1881 are tolerated, as well as some discrepancies between the expected and the observed age, the linkage process briefly described above advantages individuals with more accurate information on both censuses. However, since the one-to-one approach reduces the quantity of false links, this linkage procedure should favour the precision and the representativeness of the linked sample (Roberts 2012).

With 4226 individuals linked from 1852 to 1881 and taking into account mortality, emigration and imprecise data, we estimate that the linkage rate is about 15 %. First, based on the period life tables of Boubeau et al. (1997), we estimate that around 25 % of the 57,023 boys who composed the initial population in 1852 died before 1881 (Table 12.1).

The method to calculate the proportion of survivors by age group in 1881 is based on the methods presented by Boubeau et al. (1997). The mortality quotients by age group and year for the male population of Canada come from their period life tables.

Regarding emigration, we consider that around 15 % of our initial subpopulation of boys could not be linked because of emigration to the United States. Contrary to the mortality calculations, the latter estimate is not based on quotients because there are no emigration quotients by age and sex for the Canadian population during the period 1852–1881. In automatically linking individuals from 1871 to 1881, Antonie et al. (2015) estimated that the percentage of linkage failure due to emigration is about 10 %. In our case, we can reasonably establish that the percentage of individuals who could not be linked because of emigration is higher because the American censuses of 1860, 1870 and 1880 suggest that immigration of men born in Canada between 1835 and 1854 (i.e. the approximate birth cohort of our subpopulation of interest) was particularly important during the 1860s. Table 12.2 shows that the immigration of males born in Canada between 1835 and 1854 seems

Table 12.2 Number of males born in Canada between 1835 and 1854 enumerated on at least one U.S. census between 1860 and 1880 by age cohort. For 1860 and 1870: samples (weights applied); 1880: complete census

Census year	Age cohort	N
1860	6–25	59,600
1870	16–35	135,600
1880	26–45	157,200

Source IPUMS U.S. censuses of 1860, 1870 and 1880 (Ruggles et al. 2010) to have been particularly important during the 1860s: in the 1870 census, the number of those Canadians is more than two times higher than the corresponding number in the census of 1860. In the light of these numbers, our estimate of 15 % is probably conservative.

Finally, regarding the accuracy of the information stated in the census manuscripts, Antonie et al. (2015) estimate that around 10 % of the observations could not be linked between 1871 and 1881 because of imprecise information regarding age, first name and place of birth. In our case, we can expect a higher percentage of linkage failure due to imprecise information, since persons in 1852 likely had lower literacy than persons in 1871, suggesting a greater possibility of imprecise declarations in 1852 compared to 1871. Moreover, we have to consider not only the imprecision of the information stated on the census manuscripts, but also the inaccuracies arising from the transcription of the names from the original manuscripts. Our estimate of linkage failure due to inaccurate information is 12 %.

Thus, with the three estimates regarding linkage failure due to mortality, emigration and imprecise information, we can establish that the percentage of boys in the initial subpopulation who it would be impossible to link to the 1881 census might be as high as 50 %. Our final linkage rate was 15 %, yielding 4226 linked cases. This rate must be interpreted in the light of this linkage failure estimation. Although a linkage rate of 15 % is low compared to other studies using similar linkage techniques (e.g. Antonie et al. 2015; Long 2005), one has to consider that an intervening span of 29 years between observations is quite long, which increases the chances of linkage failure due to mortality and emigration. In order to enhance the linkage rate, we could have used an intermediate census between 1852 and 1881, e.g. the 1871 census sample: because of the shorter time between observations, a linkage between 1852–1871–1881 would probably have resulted in a higher linkage rate. However, as the source for 1852, the 1871 data source is a census sample. Linking from sample to sample implies more uncertainty in the accuracy of the links, since the true link for an individual could be someone excluded from the sample. For this reason, “each linked pair of censuses must include at least one complete enumeration” (Roberts 2012, p 7). In addition, over and above the number of the links, it is the accuracy and the representativeness of the links that we are interested in.

12.3 Accuracy and Representativeness of the Linked Sample

An assessment of the accuracy and representativeness of the 4226 1852–1881 linked cases based solely on the 1852 and 1881 censuses is challenging. The main difficulty arises from the fact that we do not have a complete enumeration of the population in 1852 (the 1852 source is a sample of 20 %) as we do for 1881. A 100 % index of the 1852 census does exist, which includes first and last names, age, gender, place of birth and place of residence. However, unlike the 1881 microdata, this index is not clean for the moment and cannot yet be used for linkage projects. Furthermore, this index omits occupations, making the study of intergenerational social mobility impossible. The fact that our data source for 1852 is a sample that hinders the validation of the automatically generated links. For a given linked individual from 1852 to 1881, one potential match in 1881 for the corresponding individual in 1852 was found, and vice versa. However, the correct link for the 1881 individual may not be the one found in the 20 % sample but someone else who is not included in this sample. The linkage with the wrong 1852 individual might have been accepted because both the correct (not included in the 1852 sample) and the false (included) links have very similar personal information. Had the other "true", individual been included in the 20 % sample, the linkage had probably not been accepted because there would be more than one single candidate.

We can still verify the quality of the automatic links by drawing upon other historical data sources. Thanks to the availability of marriage registers for a subgroup of the population—namely Catholics who celebrated their marriage(s) in a Quebec parish between 1852 and 1881—, we could verify the accuracy of some of the automatically generated links via manual linkage at the individual level between the automatically linked sample (1852–1881) and the corresponding marriage registers.

Figure 12.1 provides an example of the basic principle used to assess the validity of a link between the censuses of 1852 and 1881: the automatically linked sons (Alfred and Isidore Poitras) lived with their father and mother or stepmother in 1852 (Narcisse Poitras and Thérèse Pelletier) and with their respective spouses in 1881 (Philomène Etu and Virginie Jalbert). In the marriage registers, we find the names of the new couple (the same names we would expect to find in the 1881 census) along with the names of the parents of the linked son (also the same names that we would expect to find in the 1852 census). As clearly indicated in Fig. 12.1, the only information available to establish the validity of this link are the first and last names of the son, his parents (in the 1852 census and in the marriage registers) and his wife (in the 1881 census and in the marriage registers).

We looked for the marriage certificates of 533 individuals who had been automatically linked from 1852 to 1881. However, among these individuals, 130 (24 %) could not be verified for one of the reasons indicated in Table 12.3. Thus, we could check the validity of the linkage for 403 individuals. Our verification of the automatically generated links using the marriage registers as illustrated in Fig. 12.1 suggests that the automatic linkage has produced satisfactory results: among the sample of 403 verifiable links, 73 % are accurate (Table 12.3). However, the percentage of false links is also considerable (27 % or 20 % when considering all 533 verified cases). Moreover, considering that the majority of the population in Quebec married—i.e. that a marriage certificate should exist for most individuals—, the fact that 107 (20 %) out of 533 cases were not verifiable suggests the need of further research in order to understand the high percentage of missing records.

Table 12.3 shows that to validate a link it is necessary to find information that demonstrates the connection between the 1852 and the 1881 censuses. This proof

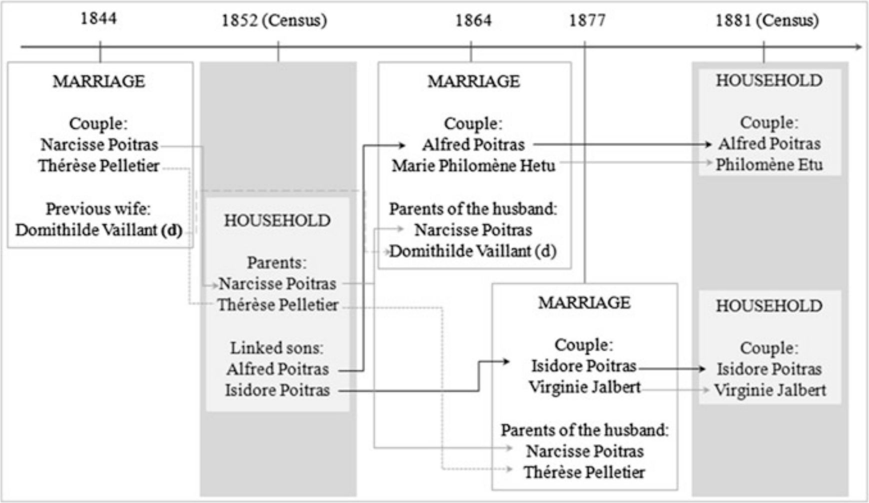


Fig. 12.1 Example of link validation via manual linkage between the 1852–1881 linked panel and the BALSAC mariage registers for Quebec. *Sources* BALSAC marriage registers and 1852–1881 linked sample. The letter "(*d*)" for "deceased" indicates the survival status of an individual at the moment of the marriage. In our example, Domithilde Vaillant died before the marriage of her son, Alfred. She is indicated as deceased on the marriage register of her son. We located the marriage certificate of the second marriage of Narcisse, Alfred's father, where we see the name of his second wife, Thérèse, who is observed in 1852. This same certificate also confirms the name of Alfred's mother, Domithilde

Table 12.3 Validation of automatically generated links between 1852 and 1881 using the BALSAC marriage registers, Catholic population of Quebec

Results	*N*	%[a]
Failure	*109*	**27.05**
Different spouse according to the 1881 census (1852 ⟶ M ↛ 1881)	62	
Another individual with the same name is married with the woman that appears as spouse in 1881 (1852 ↛ M ⟶ 1881)	47	
Success	*294*	**72.95**
The linked individual has the same spouse in 1881 (1852 ⟶ M ⟶ 1881)	264	
The linked individual remained single or became a widower between 1852 and 1881. Some family members who were present in 1852 are still present in 1881 (1852 ⟶ 1881)	30	
Non-verifiable	*130*	
No marriage register found and no other family members linked	107	
The linked individual did not live with his parents in 1852	18	
More than one marriage register found (common names)	5	
Total	*533*	
Total (verifiable)	*403*	

Sources 1852–1881 linked sample and BALSAC marriage registers
[a]% of verifiable cases; *M*: marriage register; 1852 and 1881: censuses

could be via the marriage register, as shown in Fig. 12.1, or through the presence of other family members in both census years, as shown in the "success" section of Table 12.3. Failures can be identified when the chain 1852-M-1881 is broken. Finally, in some cases, there is not enough information to establish whether a link is a failure or a success; these are the "non-verifiable" cases. This verification exercise suggests that most of the automatic links from 1852 to 1881 are accurate. However, as previously pointed out, our verification of the automatic links using the marriage registers is limited to Catholic individuals who married in a Quebec parish.

To explore the representativeness of the 4226 automatically generated links more generally, we present Table 12.4, which shows the results of a logistic regression analysis on the predictors of being automatically linked from 1852 to 1881. For this regression, the independent variables are: having a common name, living in a frontier district in 1852 (i.e. close to the border with the United States), being a farmer's son, going to school in 1852, the age group in 1852, the place of birth, the type of residence in 1852 and the fact of living in a household where the head was a labourer in 1852.

The results of the logistic regression suggest that the linkage was favoured by some characteristics, such as being older than 5 years in 1852, being a farmer's son, attend to school in 1852 and being born in Quebec or in England. Apart from the results concerning the place of birth, the results relating to each of the other attributes are as expected. Indeed, considering that in the context of our study mortality is considerably high in the first years of life, the proportion of survivors in 1881 (among the boys aged 0–15 in 1852) should be the lowest among those aged 0–5 years in 1852. This is what we have observed in our mortality and survival estimations (Table 12.1): 73.6 % of the boys aged 0–5 years in 1852 are expected to have survived until 1881 whereas the corresponding percentage for the boys who were in older age groups in 1852 is higher.

As to being a farmer's son, some studies have stressed the fact that the geographical stability of farmers is favourable for data linkage (e.g. Gagnon and Bohnert 2012; Dillon 2002). Moreover, farmers might have had lower mortality levels compared to individuals in other socio-economic groups (Gagnon et al. 2011). For our purposes, both the lower mortality and the geographic stability suggest that the likelihood of finding the corresponding record in 1881 is higher among farmers than among individuals in other socio-economic groups.

Concerning school attendance in 1852, it is possible that school-attending boys lived in households with other educated household members who in turn were more likely to provide accurate information on the census, favouring the linkage of that boy. We note, however, that the information about school attendance provided in the 1852 census manuscripts is quite limited: according to the 1852 census enumerator instructions, "By the words "attending school", not only those actually attending school at the time, but those who usually attend during some or any portion of the year, are meant to be included" (Gagan 1974, p. 360). Thus, our education variable identifies individuals who went to school at any time of the year, regardless of the amount of time spent at school. Questions also remain about the subgroup of boys who did not report school attendance. School attendance might

Table 12.4 Logistic regression: probability of being automatically linked, boys aged 0–15 years in the 20 % sample of the 1852 Canadian census

Variables	Odds ratio	
Common name[a]	0.315	***
Residence in a frontier district	0.925	*
Age group		
0–5 (*ref*)		
6–10	1.084	†
11–15	1.180	***
Farmer's son	1.272	***
School attendance in 1852	1.133	**
Birth place		
Canada (province not specified) (*ref*)		
Quebec	1.230	***
Ontario	1.052	
England	1.547	***
Ireland	0.787	**
Scotland	0.828	
Other	1.127	
Unknown	0.127	***
Type of place of residence		
Rural (*ref*)		
Village	1.052	
City	1.088	
Labourer household head	0.935	
N = 57023		

Source 1852–1881 linked sample

***p < 0.001, **p < 0.01, *p < 0.05, †p < 0.1

[a] We identified the 10 most frequent family names in each province among boys aged 0–15 years in the 20 % sample of the 1852 census. The top ten family names among these boys are (1) in Quebec: Coté, Tremblay, Gagnon, Roy, Morin, Ouellet, Gauthier, Boucher, Belanger and Demers and (2) in Ontario: Smith, McDonald, Campbell, Brown, Miller, Johnson, Scott, Wilson, Thompson and Taylor. In Quebec, 6.2 % of the boys of the subpopulation of interest have one of these common names whereas the corresponding percentage for the boys of Ontario is 5.7 %

have been lower among boys whose parents were labourers (Thernstrom 1973). However, Table 12.4 indicates that living in a household where the head was a labourer in 1852 does not significantly affect the chances of being automatically linked.

Regarding the place of birth, we observe that the reference category is being born in "Canada" (province not specified). In a separate exercise (not presented here) we looked at the distribution of the subpopulation of interest and of the linked sample by birthplace. We noticed that the share of individuals born in Canada (whether in Quebec, in Ontario, in "Canada" or in other parts of Canada) is higher among the latter (90 %) than among the former (84 %). This implies that the fact of being native Canadian increased the chances of being automatically linked. In particular—and surprisingly—, the probability of being automatically linked seems to have increased significantly with the fact of being born in Quebec (odds ratio significant at the 0.1 % level): the proportion of individuals born in Quebec is indeed higher in the linked sample (34 %) than in the corresponding group of boys in the population in 1852 (28 %). As previously mentioned, this result is rather surprising, since the greater homogeneity of last names among French-Canadians compared to Canadians of other origins should have diminished the chances of linking individuals born in Quebec (who were mainly of French-Canadian origin). In Canada, the stock of French surnames is indeed more limited than the stock of English origin surnames, since the French-Canadian population is descended from basically 10,000 French immigrants who arrived in Quebec before the 1760s. In contrast, the regular immigration to Canada of people from the British Isles during the nineteenth century nourished the pool of English last names (Charbonneau et al. 2000). This implies that the likelihood of finding more than one individual with the same name might be higher among French than among English Canadians. In our case, the share of boys of the subpopulation of interest in 1852 who had one of the common last names indicated in Table 12.4 is higher among those who lived in Quebec (6.2 %) than among those who lived in Ontario (5.7 %). Moreover, it has been suggested that, compared to individuals of English origin, French-Canadians were more often in lower socio-economic strata and had a lower school participation—for example in the city of Montreal (Gauvreau and Olson 2008). Thus, if French-Canadians were less educated and if they had more homogeneous last names compared to other Canadians, one would expect them to be under-represented on the linked sample. In their analysis of an automatically linked sample between the Canadian censuses of 1871 and 1881, Antonie et al. (2015) mention that married French-Canadians are among the under-represented groups. The linkage technique used in their study is the same as the one employed in the creation of our linked sample from 1852 to 1881. Thus, the reason why boys born in Quebec—who are mostly of French-Canadian origin—were favoured in the automatic linkage procedure between 1852 and 1881 is not completely clear yet. One possible explanation that needs further research is that, due to language and cultural barriers, French-Canadians in our subpopulation of interest might have been less likely to emigrate to the United States compared with their English counterparts.

An additional consideration about the place of birth concerning the accuracy of the data is worth mentioning here. As stated earlier, the reference category of this variable in Table 12.4 is being born in “Canada” (province not specified). In the 1852 census, most native Canadians provided rather vague information about their place of birth, since they indicated only their country of birth without specifying the province. Moreover, contrary to the 1881 census, the 1852 census did not include a specific question about the origin of the individuals. In our case, 42 % of the linked individuals from 1852 to 1881 were born in “Canada” (province not specified). The corresponding percentage within the subpopulation of interest is 42.4 %. By visualizing the 1852 census manuscripts, we could identify that some enumerators wrote a letter, “b” or “f”, next to the mention “Canada” as place of birth. In the PRDH, we recently discovered that these letters could be an indicator of the origin of the individuals born in Canada: the letter “f” indicates the French-Canadian origin of an individual whether a “b” might indicate the British (or English) origin of a person. Indeed, in almost all cases, the letters “f” and “b” that accompany the mention “Canada” as place of birth correspond to individuals whose last name is of French or of English origin, respectively. Moreover, the content of several census pages suggests that the enumerators did sometimes fill the column Place of Birth with mentions relative to the cultural origin of the individual, such as “French-Canadian”, “Br. Canadian”, “Irish”, etc.

In order to take the previous considerations into account, we created two new birthplace codes in the 20 % sample of the 1852 census, namely “Canada French” and “Canada English” (province of birth not specified in both cases). In the entire 20 % sample of the 1852 census, the proportion of individuals identified with “f” or with “b” as well as with the corresponding birthplace codes is about 10 % and 1 %, respectively. In particular, the code “Canada English”—which corresponds to mentions where there is a “b” accompanying the place of birth or where the place of birth includes an indication of the British or English origin of an individual—aims to correct a previous interpretation of the letter “b” regarding the place of birth. Initially, strings such as “Canada b” and “b Canada” were coded as born in Quebec, as the “b” was probably associated with “Bas” (prior to 1841, Quebec was known as *Bas Canada*, which means Lower Canada). However, we have considered the possibility that the letter “b” may not always mean “Bas”, e.g. when the enumerator was Anglophone. We believe that the new codes better document the mentions inscribed on the census manuscripts, indicating the cases in which a specific province of birth cannot be attributed, and the cases in which “f” and “b” suggest the cultural origin of an individual rather than a place of birth. For linkage purposes, these codes could be useful in the stage of manual verification of automatically linked individuals. For example, as previously mentioned, the 1881 census contains a direct question about the cultural origin of the individuals.[6] Thus, during the stage of manual verification of the links from 1852 to 1881, the indicator of cultural origin based on the letters “f” and “b” could be compared with the corresponding answer in the 1881 census.

Back to the discussion about the representativeness of the linked sample, we have so far treated of the factors that might have increased the chances of being automatically linked from 1852 to 1881. However, some other characteristics seem to have diminished the chances of being automatically linked. Such characteristics are the fact of having a common name, of living in a frontier district and of being

[6] According to the instructions to the enumerators of the 1881 census “Origin is to be scrupulously entered, as given by the person questioned; in the manner shown in the specimen schedule, by the words English, Irish, Scotch, African, Indian, German, French, and so forth” (Department of Agriculture (Census Branch), 1881, p. 30).

born in Ireland. These attributes have an odds ratio significantly lower than 1 (Table 12.4). Regarding the fact of having a common name, we did a separate analysis (not presented here) in order to know whether having a common name was associated with some especial socio-economic characteristics. This analysis suggests that the main differences between the boys who had a common name (i.e. one of the surnames indicated in Table 12.4) and those who did not are the fact of living in a household where the head was a farmer in 1852 and the fact of being born in Quebec: having a common name is more frequent among farmers and among individuals born in Quebec. Since individuals born in Quebec and farmers are overrepresented in our linked sample, we can say that the fact of having a common name did not introduce bias in our linked sample. This bias would have occurred if the individuals with more common names had been under-represented in the linked sample, which is not our case.

As to being born in Ireland, this characteristic seems to have significantly diminished the chances of being linked: the percentage of individuals born in Ireland is indeed higher among the subgroup of boys in the population in 1852 (4.4 %) than in our linked sample (3.5 %). Some studies suggest that the Irish living in North America during the second half of the nineteenth century were overrepresented among the manual labourers, who constituted the lowest and more vulnerable socio-economic group, especially in the cities (Gaurvreau and Olson 2008; Katz 1975; Thernstrom 1973). In our case, more than 25 % of the boys born in Ireland in our subpopulation of interest lived in a household where the head was a labourer in 1852. The corresponding percentages for the boys of other origins vary between 10 and 20 %. Despite that in our subpopulation of interest in 1852 boys born in Ireland lived more frequently in the house of a labourer compared to boys of other origins, we observe in Table 12.4 that the negative impact of being born in Ireland on the chances of being automatically linked persists even after controlling for the fact of living in a household where the head was a labourer in 1852 (which is not significant). We note that among the boys born in Ireland who were recorded in the Canadian census of 1852, some might have been migrants who fled from their native country due to the potato famine. These immigrants were particularly vulnerable and lived in unstable conditions in North America (Crowley et al. 2012). Thus, it is possible that mortality was higher among these boys, diminishing their probability of being present in 1881.

In short, despite that the linkage technique aims to increase the validity and the representativeness of the links, the previous analyses suggest that our linked sample from 1852 to 1881 is not representative of some subgroups of the population of interest. On the one hand, being an immigrant (particularly from Ireland), having a common name and living in a district close to the border with the United States are characteristics that decreased the chances of being automatically linked. On the other hand, being older than 5 years in 1852, being native (especially form Quebec), being a farmer's son and attending to school are characteristics that increased the chances of being linked.

One very important issue about representativeness which we have not yet discussed concerns the population by type of place of residence in 1852. Table 12.4

shows that, for the purposes of the automatic linkage, there is no significant difference between living in a rural place, in a village or in a city in 1852. Indeed, the distribution of the linked individuals by type of place of residence in 1852 is very similar to that of the boys in the subpopulation of interest in 1852: both lived mainly in a rural area (92 %), 2.5 % lived in a small village (up to 2999 inhabitants), 1.5 % lived in a big village (3000 or more inhabitants) and 4 % lived in a city.[7] Thus, our linked sample is representative of the subpopulation of interest regarding the type of the place of residence in 1852. However, the subpopulation of interest is composed by individuals who are included in the 20 % sample of the 1852 census, which is affected by the absence of one third of the records: the census manuscripts covering 34 % of the population disappeared before being microfilmed (Dillon and Joubert 2012). Most of those missing manuscripts contained the records of the urban population: in Ontario (Upper Canada), the records of the cities of Toronto, Kingston, London and the big district of Simcoe are missing; in Quebec (Lower Canada), the records of Montreal are lost, except for those of the neighbourhood of St. Louis.

According to Dillon and Joubert (2012), 9.3 % of the population enumerated in the provinces of Ontario and Quebec in 1852 lived in the cities of Montreal, Toronto, Quebec, Hamilton, Kingston, Bytown and London. Due to the loss of most of the urban manuscripts mentioned above, the corresponding percentage in the 20 % sample of the 1852 census is of 4.8 %. Fortunately, the data of some cities, namely Bytown (Ottawa), Hamilton and Quebec, has been preserved. This data can be used in order to increase the representation of the urban population in the 20 % sample of the 1852 census: by attributing weights to the population living in those cities in 1852, some aspects of the urban population of mid-nineteenth century Canada can be analyzed. A weight variable is already available on the 20 % sample. This variable is based on the distribution of the population by type of place of residence according to the volume of aggregated statistics of the 1852 census (Board of Registration and Statistics 1853). It gives more weight to the population living in the cities of Quebec, Hamilton and Bytown in 1852, so that they constitute 9.3 % (instead of 4.8 %) of the total population of the two provinces. For example, each one of the individuals included in the 20 % sample who lived in Hamilton or in Bytown in 1852 has a weight of 2.8. The corresponding weight for their counterparts living in Quebec city in 1852 has a value of 1.7.

Despite the availability of some urban data, and even if the majority of the Canadian population in the mid-nineteenth century was rural, the missing records are problematic for any linkage procedure at the individual level, since most of the people living in the big cities in 1852 will be excluded. Attributing weights is not an optimal solution to this problem, since the population living in the biggest cities, i.e. Toronto and Montreal, very likely differed in some aspects from the population living in smaller cities, such as Hamilton, Bytown or Quebec. For example, the

[7] According to the type of place indicated in the aggregated volume of the 1852 census (Board of Registration and Statistics 1853).

ethnocultural composition and the economic opportunities—which are factors that have an impact on the intergenerational social mobility—seem to have been different in the biggest cities, on the one hand, and in the smaller cities, on the other hand. According to the preserved volume of aggregated statistics of the 1852 census (Board of Registration and Statistics 1853) the share of individuals of French-Canadian origin was higher in Quebec city (58.3 %) than in Montreal (45 %) in 1852. In turn, the share of the people from Ireland and from Scotland was more important in Montreal (25.8 %) than in Quebec city (16.6 %). In the Ontarian cities, the share of French-Canadians was minimal in Toronto and Hamilton, whereas in Bytown one-quarter of the population was of French-Canadian origin. Regarding the economic opportunities, looking at the occupational distribution of men aged 18–65 years in the complete census database of 1881 gives us an idea of the economic opportunities in the same five cities. In 1881, the share of merchants, manufacturers and professionals was more important in Montreal and Toronto (around 23 %) than in Ottawa, Hamilton and Quebec city (around 17 %). The opportunity to have an occupation on the manual skilled sector seems to have been the highest in the city of Hamilton, where 44 % of the men aged 18–65 years were skilled workers in 1881 (the corresponding share in Montreal and Toronto was around 35 and 38 %, respectively). In Ottawa, the share of men employed in white collar occupations was the highest (around 17 %), whereas the corresponding part in the other four cities was around 10 %. Though these observations are based on the occupational distribution of 1881, they provide an idea of the differences that might have existed among the five cities compared regarding the development of certain economic sectors as well as the occupational opportunities for the individuals living in those cities.

The previous considerations suggest that Hamilton, Bytown and Quebec city differed in some aspects from Toronto and Montreal. For this reason, using the weights provided in the 20 % sample of the 1852 census in order to increase the representation of the urban population needs caution in our interpretations. Moreover, for automatic linkage purposes, the absence of most of the data of the population living in the biggest cities in 1852 means that the urban individuals available for linkage will represent only certain selected cities, leaving Toronto and Montreal under-represented. Thus, when using the 20 % sample of the 1852 census, one should consider whether the missing urban data constitutes a problem to the analysis of a phenomenon of interest. For example, the missing urban data should not be problematic in the analysis of the rural exodus, which could be studied by linking the census sample of 1852 with the complete enumeration of 1881. It should be kept in mind that the urban population of Canada in 1852 constituted only 10 % of the population, and many of the missing manuscripts are distributed across communities of varying sizes.

12.4 The Use of the 1852 Census to Study the Intergenerational Social Mobility

Despite the difficulties mentioned above regarding the use of the Canadian census of 1852, this source of data includes valuable information about the population who lived in the provinces of Ontario or Quebec at the mid-nineteenth century. The 20 % sample can be used to analyze several aspects of this population, since it includes information about family composition and coresidence[8] as well as about the socio-economic conditions of individuals and households. For instance, the 1852 census includes questions about the occupation and the type of dwelling (e.g. log house, stone house, shanty, etc.). Moreover, compared to other data sources (e.g. the parish registers), the census data provides more details about the composition and the socio-economic characteristics of households. In our case, the aim of linking individuals from 1852 to 1881 was to have an opportunity to study the intergenerational social mobility between fathers (in 1852) and sons (in 1881). Using linked census data is favourable to study this phenomenon, since the information about the socio-economic characteristics of the family of origin in 1852 is of great interest when one wants to analyze the intergenerational social mobility and the occupational attainment of the sons in 1881. Moreover, despite the long intervening time span of 29 years between the linked censuses, linking the sons (aged 0–15 years in 1852) from 1852 to 1881 increases the chances of observing the father and the son living together in 1852. This coresidence is essential in order to have information about the father in 1852.

The use of the 1852 census is also supported by the subject of study. If one is interested in studying the Canadian population of the mid-nineteenth century, the use of the 20 % sample of the 1852 census is appropriate when the difficulties associated with the data do not hinder the subject of study. For instance, if one wants to study the rural exodus, the 20 % sample is appropriate, whereas it would not be the case if the subject of study were the urban life in Canada in the mid-nineteenth century. In our case, we are interested in the intergenerational social mobility at the beginning of the industrialisation in Canada. For this purpose, we need to observe the father (at the mid-nineteenth century) and the son (some decades later) when they were adults in order to compare their occupations at similar points in their careers. Some researchers (e.g. Prandy and Bottero 2000; de Sève and Bouchard 1998; Delger and Kok 1998; Van Poppel et al. 1998) have criticized

[8] The 20 % sample of the 1852 census includes the variable “Household number” but not “Family number”. Thus, we can identify who lived with whom (in the same household) but not who belonged to which family in 1852. In order to have an idea of the different families that lived together, the 20 % sample includes some variables that aim to identify the relationship between individuals living in the same household. For example, the constructed variable CANREL indicates the relationship with respect to the household head (e.g. “wife of head”, “child of head”, “parent of head”, “other kin of head”, and “undetermined”). The variables MOMLOC and POPLOC indicate, within each household, the position (in order of enumeration) of the mother and the father of an individual, respectively.

the use of marriage registers as only source of data in the study of the intergenerational social mobility: “Studies using marriage records are obviously comparing, for the most part, fathers (and/or fathers in law) who are nearing the end of their working lives with sons who are at a fairly early stage in theirs” (Prandy and Bottero 2000, p. 4). This difference of age between fathers and sons increases the risk of overestimating downwards the social mobility. Thus, by linking individuals from different censuses, this risk can be reduced, since the comparison of fathers and sons can be made at more similar points in their respective occupational careers.

However, for the purposes of studying the intergenerational social mobility, the use of the 1852 census implies that we have to take care with our interpretations, since the mobility observed does not include the sons of the fathers who lived in the biggest cities (Toronto and Montreal) in 1852. This means that the observed mobility would concern mostly the rural population, which constituted the majority (90 %) of the Canadian population in 1852 (Dillon and Joubert 2012). Moreover, the occupation is the only variable available in the 1852 census that can be used in order to have an idea of the social status of an individual. The Canadian censuses started to include more questions about employment—e.g. the employment status (employer/employee) and the fact of earning a salary—only in 1891 (Baskerville 2000). Thus, before the 1891 census, the only information available in the personal censuses regarding the social status of the individuals is their occupation. Other types of census questionnaires containing economic information existed already before 1891. Some examples are the agricultural schedule of 1852 and the industrial return of 1871 (LAC 2014). However, most of the information contained in those questionnaires is not immediately available for research, since the manuscripts that have been preserved have not been transcribed yet (with the exception of a sample of the agricultural and the industrial return of 1871).

In short, despite the difficulties associated with the use of the Canadian census of 1852, this source of data provides valuable information about the socio-economic conditions as well as the household composition of the population living in Ontario or in Quebec at the mid-nineteenth century. This data is suitable for the study of certain phenomena as well as for data linking with other nominative sources of data, such as the complete 1881 Canadian census.

12.5 Conclusion and Discussion

This chapter has briefly described and discussed some aspects about an automatically generated linked sample from the Canadian censuses of 1852 and 1881. The linked sample, composed by males aged 0–15 years in 1852, was created with the specific purpose of studying the intergenerational social mobility between fathers and sons at the beginning of the industrialisation in Canada. In order to analyze this linked sample, we have briefly described the technique by which it was generated. This technique of automatic linkage, based on the individual attributes that should not change over time (or that should change in a predictable way), aims to favour the representativeness and the accuracy of the linked sample by reducing the number of false links. We have presented some analyses about the accuracy and the representativeness of our linked sample from 1852 to 1881: on the one hand, our verification of the automatic linkage via the use of marriage registers suggests that most of the links are accurate. On the other hand, our analyses regarding the representativeness of the linked sample suggest that some attributes favoured while others hindered the fact of being automatically linked from 1852 to 1881: immigrants (particularly from Ireland), individuals with common names and those living close to the border with the United States had fewer chances to be automatically linked, while the native (especially from Quebec), older than 5 years in 1852, farmer's sons and attending to school had more chances to be linked. Thus, the linked sample is not representative of some subgroups of the population (here, the "population" is composed by the boys aged 0–15 years in 1852 who are included in the 20 % sample). The identification of the under-represented groups is important in order to be careful in the interpretations of a study using this linked sample.

As a final consideration, we would like to put emphasis on the great research potential of linking censuses and parish registers. In our case, we used the BALSAC marriage registers in order to verify the accuracy of the automatically generated links between the censuses of 1852 and 1881. Besides being appropriate for this purpose, the marriage registers could be used to add more information about the linked individuals from 1852 to 1881. Indeed, though the automatically linked sample 1852–1881 provides valuable information about the composition and the socio-economic conditions of individuals and households in 1852 and 1881, it does not include information about the demographic events, such as the age at marriage of the individuals. This information could be added by linking the marriage registers to the linked sample between censuses, such as illustrated in Fig. 12.1. Such an approach would not exclude individuals who did not marry during a determined interval of time, since their information would be available on the censuses. We note, however, that in our case, the BALSAC marriage registers are limited to the catholic individuals who married in a parish in the province of Quebec while the linked sample from 1852 to 1881 includes individuals with other religious affiliations who married outside the province of Quebec between 1852 and 1881.

Our exercise of manual linkage between the automatically linked sample (1852–1881) and the BALSAC marriage registers is not currently connected to any other project in Quebec. However, researchers could profit from the linkage between these registers and the Canadian censuses, since both data sources contain information at the individual level and provide information about complementary aspects such as family composition and socio-economic conditions (the censuses) and the demographic events (the parish registers).

References

Antonie, L., Inwood, K., & Ross, A. J. (2015). Dancing with dirty data : Problems in the extraction of life-course evidence from historical censuses. *This book*, chapter 10.

Balsac, fichier de population (accessed 2013) [online]. http://balsac.uqac.ca/. The BALSAC marriage registers were accessed via the website of the *Registre de la Population du Québec Ancien* (RPQA): http://www.prdh.umontreal.ca/rpqa/.

Baskerville, P. (2000). Displaying the working class: The 1901 census of canada. *Historical Methods, 33*(4), 229–234.

Bord of registration and statistics. (1853). *Census of the Canadas* (pp. 1851–1352). Quebec: J. Lovell.

Boubeau, R., Légaré, J., & Émond, V. (1997). *Nouvelles tables de mortalité par génération au Canada et au Québec, 1801–1991*. Ottawa: Statistics Canada, Demography division.

Charbonneau, H., Desjardins, B., Légaré, J., & Denis, H. (2000). The population of the St. Lawrence Valley, 1608–1760. In M. R. Haines & R. H. Steckel (Eds.), *A population history of North America* (pp. 99–142). New York: Cambridge University Press.

Crowley, J., Murphy, M., Roche, C., & Smyth, W. (Eds.). (2012). *The scattering, in Atlas of the Great Irish Famine*. New York: New York University Press.

Curtis, B. (2001). *The politics of population: State formation, statistics, and the census of Canada, 1840–1875*. Toronto: University of Toronto Press.

De Sève, M., & Bouchard, G. (1998). Long-term intergenerational mobility in Quebec (1851–1951): The emergence of a new social fluidity regime. *Canadian Journal of Sociology/Cahiers Canadiens de Sociologie, 23*(4), 349–367.

Delger, H., & Kok, J. (1998). Success or selection? The effect of migration on occupational mobility in a Dutch province, 1840–1950. *Historical Methods, 13*(3–4), 289–322.

Dillon, L. Y. (2002). Challenges and opportunities for census record linkage in the French and English Canadian context. *History and Computing, 14*(1–2), 185–212.

Dillon, L. Y., & Joubert, K. (2012). Dans les pas des recenseurs: une analyse critique des dimensions géographiques et familiales du recensement canadien de 1852. *Cahiers québécois de démographie, 41*(2), 299–339.

Gagan, D. (1974). Enumerator's instructions for the Census of Canada, 1852 and 1861. *Social History, 7*(14), 355–365.

Gagnon, A., & Bohnert, N. (2012). Early life socioeconomic conditions in rural areas and old-age mortality in twentieth-century Quebec. *Social Science and Medicine, 75*, 1497–1504.

Gagnon, A., Tremblay, M., Vézina, H., & Seabrook, J. A. (2011). Once were farmers: Occupation, social mobility, and mortality during industrialization in Saguenay-Lac-Saint-Jean, Quebec 1840–1971. *Socioeconomic Inequalities and Mortality, 48*(3), 429–440.

Gauvreau, D., & Olson, S. (2008). Mobilité sociale dans une ville industrielle nord-americaine: Montréal, 1880–1900. *Annales de démographie historique, 115*(1), 89–114.

Katz, M. B. (1975). Transiency and Social Mobility. *The people of Hamilton, Canada West: Family and class in a mid-nineteenth-century city* (pp. 94–175). Cambridge: Harvard University Press.

Library and Archives Canada (LAC). (2014). Censuses. Online: http://www.bac-lac.gc.ca/eng/census/Pages/census.aspx.

Long, J. (2005). Rural-urban migration and socioeconomic mobility in Victorian Britain. *The Journal of Economic History, 65*(1), 1–35.

Prandy, K., & Bottero, W. (2000). Reproduction within and between generations. *Historical Methods, 33*(1), 4–15.

Roberts, E. (2012). Mining microdata: Economic opportunity and spatial mobility in Britain, Canada and the United States, 1850–1911. In *Digging into data challenge, National Science Foundation, Economic and Social Research Council (UK), and Social Science and Humanities Research Council (Canada).* Project description.

Ruggles, S., Alexander, J. T., Genadek, K., Goeken, R., Schroeder, M. B., & Sobek, M. (2010). Integrated public use microdata series: Version 5.0 [machine-readable database]. Minneapolis : University of Minnesota.

Thernstrom, S. (1973). *Poverty and progress: Social mobility in a nineteenth century city*. New York: Atheneum.

Van Poppel, F., de Jong, J., & Liefbroer, A. C. (1998). The effects of paternal mortality on sons' social mobility: A nineteenth-century example. *Historical Methods, 31*(3), 101–112.

Chapter 13
Introducing 'Movers' into Community Reconstructions: Linking Civil Registers of Vital Events to Local and National Census Data: A Scottish Experiment

Eilidh Garrett and Alice Reid

Abstract The release of national, individual-level census data for Scotland, via the Integrated Census Microdata (I-CeM) project undertaken at the University of Essex, makes it possible to identify the number of Scotland's residents by their county and parish of birth on each census night from 1851 to 1901. This chapter uses the anonymous I-CeM data for 1871, alongside individual, nominal census data from the 1881 census of Scotland, and details from the civil registers of births, marriages and deaths on the Isle of Skye to 'follow' all individuals born on the island during the 1860s and 1870s to their entries in the 1881 census. This allows the number of migrants to be gauged, and those moving within their home country to be distinguished from those who emigrated. During the linkage process a number of biases became evident, and the implications of these for record linkage and demographic history are discussed.

13.1 Introduction

For over a decade the authors have been studying the population history of the Isle of Skye, which lies off Scotland's north-west coast (e.g. Reid et al. 2002/2006; Garrett and Davies 2003; Garrett 2006; Blaikie et al. 2005; Davies and Garrett 2005; Galley et al. 2011).[1] We have compared and contrasted the demographic

[1] This work has been funded by a number of research awards: 'Determining the demography of Victorian Scotland through record linkage' ESRC award RES-000-23-0128; 'Doctors, Deaths, Diagnoses and Data: a comparative study of the medical certification of cause of death in nineteenth century Scotland' Wellcome Trust award (Ref 082200/Z/07/Z); Scottish Health Informatics Programme (SHIP) grant (Ref WT086113).

E. Garrett (✉)
University of St Andrews, St Andrews, UK
e-mail: eilidh.garrett@btinternet.com

A. Reid
University of Cambridge, Cambridge, UK
e-mail: amr1001@cam.ac.uk

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_13

experience of this very rural community with the demographic history of the lowland town of Kilmarnock in Scotland's 'Central Belt'. This work resulted in a better understanding of the differences between the demographic regimes operating in urban and rural areas of Britain during the second half of the nineteenth century. For both communities our work has involved linking together births, marriages and deaths recorded in the civil registers to reconstruct family groups and then linking these groups to the censuses of 1861, 1871, 1881, 1891 and 1901.[2] With the linked data we have been able to examine not only the demographic histories of our communities but also aspects of their medical, social and economic histories.

The record linkage undertaken in our studies is not 'family reconstitution' in the classic sense (Wrigley et al. 1997; Newton 2011). Reconstitution is usually undertaken using parish records and involves tracing individuals from their birth, or baptism, to their death, or burial, often via at least one marriage. A few studies have considered contiguous parishes (Perkyn 1999), but usually the focus has been on just one parish. Only data from individuals who are known to remain in the parish are included when demographic rates are calculated from a 'reconstitution' study. This is done to reduce any biases which might be introduced into the measurements by the inclusion of incomplete, or censored, life histories. It can, however, skew perceptions of the demographic experience of cohorts as a whole because, as Ruggles (1992, p. 507) has pointed out, 'migration can bias estimates of such measures as mean age at marriage and life expectancy, even if age-specific demographic rates of migrants and non-migrants (within a cohort) were identical'. Because the date of an individual's migration out of their birth parish is seldom known in a reconstitution study, it is difficult to gauge whether that person left before the average age of marriage (and may therefore have married at that age, or even earlier) or left the parish, unmarried, long after the age by which most of their peers had wed. In either case the estimate of the 'average age at marriage' of the 'stayers' (and probably the 'movers' too) has been 'biased' away from the figure for the cohort as a whole.

As Ruggles (1992, p. 507) notes 'most of the concern about the effects of the exclusion of migrants (from record linkage studies) has focused on the question whether demographic behaviour of migrants and non-migrants was similar, or not'. This has proven a very tough question to answer, certainly in the context of Great Britain, because it is virtually impossible to follow those who have moved away from their parish of birth when the only sources available are parish registers (Clark and Souden 1987).

We refer to the type of record linkage undertaken for Skye and Kilmarnock as 'family reconstruction'. The data only covered a 40 year period so it was not possible to follow all individuals born on Skye to their marriage or death, but the censuses provided independent evidence of their continued presence on the island at a particular date. Individuals and their families could therefore be followed from

[2] Special permission to access the civil register data was granted by the General Register Office for Scotland (now National Records of Scotland) with conditions attached.

one decade to the next. Those in-migrants to the island who were enumerated in the census could be identified and, if appropriate, incorporated into measures of demographic behaviour. This meant that the demographic rates calculated included a greater proportion of the population than would have been true of a classic 'reconstitution' study, albeit for shorter time periods. The census data allowed 'numbers at risk' to be calculated more accurately as individuals could be censored at the date of the last census in which they were observed rather than at the date of the last event in which they were involved. Nevertheless, the questions surrounding migration remained. Different demographic behaviour among those who moved away was still a potential source of bias.

Computer-readable, individual-level, census datasets covering the entire populations of England, Wales and Scotland in the second half of the nineteenth century are now available. As the censuses record each person's parish and county of birth it has become increasingly possible to follow an individual from a particular location in one census as they moved elsewhere before being enumerated in a succeeding census. Migrants leaving particular communities can be traced to see what life held in store for them, and whether their life history played out differently from those of their siblings or neighbours who remained in the original community.

This chapter thus reports on an 'experiment'; a first attempt to link civil register data, local census data and national census data across Scotland for a whole community. This experiment was designed to demonstrate what may be gleaned from such an exercise concerning migrants and migration. It also aimed to show that linking beyond the geographic limits of a community may be of benefit to those undertaking record linkage in other contexts, by illustrating why certain links might be 'missed'.

The chapter is arranged into a further five sections. Section 13.2 introduces the data sources and Sect. 13.3 considers the strategies used to link them. Section 13.4 reports the results of an exercise in which two cohorts of children born on Skye were linked to the 1881 census returns of Scotland. Section 13.5 then considers the variation in rates of success achieved when linking different groups before we present our final conclusions in Sect. 13.6.

13.2 The Data

In 1861 the Isle of Skye comprised seven parishes: Bracadale, Duirnish, Kilmuir, Portree, Sleat, Snizort and Strath.[3] In the course of our previous work we had transcribed the complete contents of the civil registers and the census enumerators' books covering these parishes for the whole of the period between census day 1861 and census day 1901. Details of every individual's name, sex, age, relationship to the head of the household, occupation and county and parish of birth were extracted

[3] The parish of Portree included the population of Raasay, a neighbouring island.

from each census. In the civil registers each entry gave the forename and surname of the individual or individuals being born, marrying or dying, the forename and surname of their father, and the forename and maiden surname of their mother. From 1860 onward Scottish birth certificates also gave the date and place of the parents' marriage (Sinclair 2000, p. 41). When a child was born outside marriage either just its mother's name was recorded, or the names of both parents were given but no date of marriage. In both cases the word 'illegitimate' was usually entered on the certificate. The date and place of each birth or death were given, as were the date and place of each marriage.

In addition to the details in the civil registers and the censuses for Skye, two different datasets of individual-level census data for the whole of Scotland were used. The first of these was the individual records of the 1871 census of Scotland recently made available by the Integrated Census Microdata (I-CeM) project based at the University of Essex. Unfortunately, commercial constraints mean that the data currently available for research do not include names and addresses so nominal record linkage data is not yet possible using these data.[4] However, the numbers of people who claimed to have been born on Skye can be counted, by age, sex and their location in 1871.[5] The 1881 census of Scotland is also available from I-CeM, but we chose to use a version previously deposited at the UK Data Archive by the Church of the Latter Day Saints, and later enhanced by Schürer and Woollard.[6] This version was preferred because it includes the forenames and surnames of each enumerated individual, allowing the linkage of individuals rather than just the identification of those who were 'Skye born'. The 'Skye born' living elsewhere in Scotland were extracted from the 1881 census along with the other members of their household.

As well as the lack of names for 'Skye born' individuals not resident on the island in 1871, a further limitation of the data for our 'experiment' was the fact that only individuals living in Scotland could be identified in the 1871 and 1881 censuses; those who had moved beyond Scotland's borders could not be followed. Even those who moved to England and Wales could not be identified in either

[4] The I-CeM data can be accessed via the UK Data Archive at http://icem.data-archive.ac.uk http://www.essex.ac.uk/history/research/icem. Documentation detailing the dataset and the access arrangements can be found at: http://www.essex.ac.uk/history/research/ICeM/documentation.html. Other census years and censuses for England and Wales are also available.

[5] This was done using the online Nesstar system (see http://icem-nesstar.data-archive.ac.uk/). Had we downloaded the individual records from I-CeM, it might have been possible to identify the names of the individuals from the images held by FindMyPast, I-CeM's commercial partner, but this was deemed too expensive and time consuming for an exploratory project.

[6] This dataset is available to the research community via the UK Data Archive (dataset SN 4178). We also made use of the CD Rom version created by the Church of Jesus Christ of Latter Day Saints (1999).

census because the instructions for the English and Welsh censuses made it clear that individuals born elsewhere were only to return to the country of their birth. This means that it is seldom possible to determine exactly where Scots migrating 'south of the border' had been born.[7]

13.3 Record Linkage Methods and Issues

13.3.1 Record Linkage on the Isle of Skye

The study attempted to trace all members of two birth cohorts born on Skye: an 1860s cohort—all individuals born on or after census day 1861 but before the date of the 1871 census (7 April 1861 to 1 April 1871)—and an 1870s cohort—all those born between the 1871 census and the 1881 census (2 April 1871 to 2 April 1881)—either to their death on the island before the 1881 census or to their entry in that census, wherever they were living in Scotland.

Our record linkage was based on names, relationships and ages using a 'sets of related individuals' approach as outlined in Reid et al. 2002/2006. Birthplaces were used as an additional piece of information in the linkage process. As part of the linkage strategy the names recorded in the censuses and civil registers were standardised against a set of 'name dictionaries' which were built up as the linkage progressed. The dictionaries helped to minimise the effect of misspellings or transcription errors on both forenames and surnames. They also helped to contend with the fact that the population of Skye was predominantly Gaelic speaking. Names on official documents, such as the vital registers and the census, were expected to appear in an 'English' form, but the Anglicised version of a Gaelic name could vary from source to source, over space or with time. On Skye, for example, the Gaelic name *Mòrag* never appeared in the records. It was most often translated as *Marion* but could appear as *Sarah*, depending on who recorded it. It was noticeable that boys born on Skye and registered as *Ewen* or *Ewan* (the Gaelic version *Eòghann* was never used) were subsequently found on the mainland in the 1881 census all recorded as *Hugh*. We used Morgan (1989) as a guide to Gaelic names and their spelling.

Perhaps because of this 'translation' issue, but also because of local naming practices, the name pool on Skye was relatively small. In 1862, for example, six boys born on the island were named *Donald McInnes*. Two had fathers called *Angus*, and two had fathers called *Lachlan*. Three had mothers named *Catherine*. It

[7] Some Scottish-born individuals in England and Wales nevertheless gave more detailed birthplaces. We have not used these in the current study as they are too few and are probably unrepresentative.

was also not uncommon for living siblings to be given the same forename (Galley et al. 2011a, b). To avoid confusion the population made extensive use of nick names and diminutives such as *Effie* for *Euphemia* and *Alex* or *Sandy* for *Alexander*. The name dictionaries helped us to link the 'formal' name given on a birth certificate with the 'everyday' name which was often reported in the census.

13.3.2 Record Linkage Between Skye and the Rest of Scotland

As a first step in the linkage individuals giving 'Skye', or one of its seven parishes, as their place of birth were identified in the '1881 census' database and the 1871 I-CeM data for the whole of Scotland. 'Skye born' individuals aged 0–9 were identified in 1871, and those aged 0–9 and 10–19 in 1881. Variations in the spelling of the parish names were allowed for as much as possible, but individuals may have been missed because their birthplace was misreported, misspelt or mistranscribed.

To make links between individuals on Skye and those elsewhere in Scotland, the same methodology was applied as had been used on the island. The combination of the child's and parents' names, child's age and birthplace were used to match the child's birth on Skye to an individual elsewhere in Scotland in 1881, when individuals' names were available.

The links between Skye and the rest of Scotland were constrained to some extent. We lacked the resources to seek out and obtain the death or marriage certificates of any members of the 1860s and 1870s birth cohorts who had migrated from Skye. The death certificates would have given details of the deceased's parents and therefore helped to distinguish one young person from another with the same name. The marriage certificates would have provided a bride's maiden name, increasing the chance that her census entry and her birth would be linked.

In some respects issues related to names also constrained the linkage process. The name dictionaries built up during work on Skye did not always include the spellings or forms of the names found in the 1881 census transcripts for the rest of Scotland. Migrants may have changed their names themselves, or the change may have been made by others, either deliberately or by mistake.

Conversely, the identification of one child from a sibling group as a result of a distinctive name often helped to identify the others. This can bias the links to some degree toward children living with at least one other sibling.

Given that links were made on the basis of the combination of the forenames and surnames of three individuals; a child and his or her father and mother, it was much more difficult to follow illegitimate children in the records than it was to trace their legitimate peers. This problem was exacerbated by naming issues amongst illegitimate children. Both on Skye and elsewhere in Scotland, some illegitimate children adopted their father's surname after their birth was registered using their mother's surname.[8] Both illegitimate and legitimate children occasionally took the surname of a step-father who had married their mother. Thus, although 91 % of legitimate children born in the 1860s could be linked to a death pre-1871 or to the 1871 census, links could only be made for 72 % of illegitimate children.

By 1881 many of the 1860s birth cohort were in their late teens and living away from their parents. This made it much more difficult to link their census entry to their birth entry, or vice versa, particularly as Skye's small name pool meant that there could be multiple individuals of a similar age with the same name. If two or more of the six *Donald McInnes* born in 1862, mentioned above, had been identified in 1881 employed as servants in different households, with none of their kin resident, it would be impossible to differentiate which birth linked to each *Donald*. Each birth or census entry would be considered 'unaccounted for' as a confirmed link could not be made.

As children grew older their ages became more prone to inaccuracy. In some households the ages of several children within a family were all 'adjusted' by the same amount when reported in a particular census, suggesting that at least one parent had lost track of the ages of their offspring. More distant relatives, or an employer, would have been even less sure of a youngster's exact age. The young people themselves may have 'massaged' their ages when seeking work or somewhere to live in order to be eligible for consideration. When ages could not be confirmed through triangulation with other key variables, the number of competing matches often increased, reducing the chance that a link would be made.

Birthplaces, as well as ages, could be misreported in the census. Some children on Skye stated that they had been born elsewhere in Scotland but nevertheless their birth could be identified on the island. Unfortunately, those who reported that they were born on Skye, but were, in fact, born elsewhere were more difficult to identify. They contributed to the number of census entries unlinked to a birth. Birthplaces, as reported in the census, thus need to be treated with caution. We refer to them as 'alleged birthplaces' and place 'Skye born' and 'not Skye born' in quotation marks to indicate this.

Before moving on from discussion of the linkage process to a discussion of the results, it should be noted that the combination of census and civil register data allows linkage to be conducted in two directions: either forward or backward in time. The data in each census provide 'a snapshot' of a family whereas, when linked, the events in the civil registers are more akin to a 'movie'. The two sources give quite different perspectives on the individuals being recorded. As becomes evident below, the direction of the linkage being undertaken yields rather different sets of links: an important fact that those reconstructing populations need to recognise when designing their linkage strategies.

[8] The great majority of illegitimate children took their father's surname if their parents subsequently married, but some seem to have adopted their father's surname without a marriage taking place.

13.4 Linkage Results

Skye had 19,605 residents recorded in the census of 1861. The community was in decline, however, and only 17,684 individuals were enumerated in the 1881 census: a fall of just under 10 %. Nevertheless, over these two decades the island saw natural growth (an excess of births over deaths) of 3445 persons. Therefore, the island actually lost more than 23 % of its original population over the course of the 20 years. Between the censuses of 1861 and 1881 there were 9778 births in total on the island. The 4906 born between the 1861 and 1871 censuses form our 1860s cohort, and the 4872 births which occurred between the censuses of 1871 and 1881 form the 1870s cohort.

There were, of course, individuals aged 0–9 in the 1871 census and 0–19 in the 1881 census who had moved to Skye since their birth. Their birthplaces provide some indication of the origins of in-migrants to Skye.[9] There were also a small number of individuals, apparently born between 1861 and 1881, who died on Skye between the censuses, for whom no birth could be traced. These too were assumed to be in-migrants, but very little—apart from their age at death and their father's occupation—is known about them. In this chapter the focus is on out-migrants rather than in-migrants.

The results of the linkage exercises are laid out in five tables. Table 13.1 shows the linkage between the 1860s birth cohort and the 1871 census, and Table 13.2 shows the linkage between the 1870s birth cohort and the 1881 census. In Panel A in each of the tables the births of the cohort on Skye are taken as a starting point and linked forward. In each case births were first linked to those dying on Skye and those seen, either on the island or elsewhere in Scotland, in the following census. Individuals who were not linked, who will be referred to as being 'unaccounted for', had either left Scotland, died after leaving Skye or were still on Skye, the link between their birth and the census having been missed. Table 13.1 shows that amongst the 1860s cohort almost 15 % had died on the island by 1871, just shy of 74 % remained on the island, 5 % had moved elsewhere in Scotland and a little over 6 % were 'unaccounted for'. Among the 1870s cohort shown in Table 13.2 the equivalent percentages were similar, being 13.4, 78.1, 3.7 and 4.9, respectively.

Panel B in both Tables 13.1 and 13.2 takes a different perspective and links backwards from the relevant census of Scotland to the births.[10] Those who were 'Skye born' and aged 0–9 in each census were linked back to a birth on Skye. This was possible for all those observed in the 1881 census, but only for those enumerated on Skye in 1871, as names were not available for those living elsewhere in Scotland.

[9] There were 624 individuals aged 0–19 on Skye in 1881 who reported that they had not been born on the island. Of these 167 had been born elsewhere in Inverness-shire, and 90 in neighbouring Ross-shire. A further 112 had been born in Lanarkshire, the vast majority in Glasgow.

[10] Those both born and enumerated on Skye were corrected for errors in age and birthplace. It is likely that similar errors occurred among those enumerated elsewhere, but the tables show that these errors largely cancelled out so are unlikely to lead to significant bias. There were, of course, children enumerated on Skye who were 'not Skye born'. These do not form part of the analysis although they are shown in the tables for completeness.

Table 13.1 Population accounting: linking between the births of the 1860s birth cohort on Skye, deaths before 1871 and individuals aged 0–9 years in the 1871 census[a]

PANEL A: LINKING FORWARD from births			
		N	%
A.	BIRTHS registered on Skye 1861–1871[a]	4906	100.0
B.	Linked to a death in Skye death registers 1861–1871[a]	724	14.8
C.	Linked to an entry in the 1871 census of Skye	3627	73.9
D.	Accounted for on Skye by 1871 (=B+C)	4351	88.7
E.	Not linked to a death or a 1871 Skye census entry (=A−D)	555	11.3
F.	‘Skye born’ aged 0–9 elsewhere in Scotland in 1871 census	247	5.0
G.	Not identified on Skye or elsewhere in Scotland in 1871 (=E−F)	308	6.3
		N	%
H.	DEATHS of those aged 0–9 registered on Skye 1861–1871	1012	100.0
I.	Linked to a 1861–1871 birth on Skye	724	71.5
J.	Individual observed in the 1861 census: not linked to a 1861–1871 birth on Skye	237	23.4
K.	Individual not observed in 1861 census; but should have been given age in death register	12	1.2
L.	Individual not linked to a Skye birth but age indicates born 1861–1871: in-migrant	39	3.9

PANEL B: LINKING BACKWARD from the 1871 census					
		N	%	*N*	*N*
		‘Skye born’	of ‘Skye born’ in Scotland	not ‘Sky born’	total
M.	Those aged 0–9 enumerated on Skye	3818	94.1	256	4074
N.	Enumerated on Skye ‘aged 10+’ but in birth register 1861–1871	+41			
O.	Enumerated on Skye ‘aged 0–9’ but also enumerated in 1861 census	−82			
P.	Enumerated on Skye aged 0–9 as ‘not Skye born’ but linked to a 1861–1871 birth on Skye	+32		−32	
Q.	Adjusted *N* aged 0–9 on Skye 1871 =(M+N+O+P)	3809	93.9	224[b]	4033
	Q1. ‘Skye born’ aged 0–9 on Skye linked to a birth	3627	89.4		
	Q2. ‘Skye born’ aged 0–9 on Skye *not* linked to a birth	182	4.5		
R.	‘Skye born’ aged 0–9 elsewhere in Scotland in 1871 census	247	6.1		
S.	All ‘Skye born’ aged 0–9 in Scotland	4056	100.0		

(continued)

Table 13.1 (continued)

PANEL C: ACCOUNTING for the 1860s birth cohort on Skye: in the civil registers and 1871 census

		N	%
T.	Births on Skye, 1860s cohort (A)	4906	100.0
U.	Survivors from 1860s cohort on Skye in 1871 (Q)	3809	77.6
V.	Deaths to 1860s cohort on Skye (I)	724	14.8
	Death rate amongst stayers (=*V/*(*U+V*) * *1000*) = *160 per 1000 births*		
W.	Survivors from 1860s cohort elsewhere in Scotland, 1871 (F, R)	247	5.0
X.	Births from 1860s cohort unaccounted for in 1871(includes deaths off Skye and migrants leaving Scotland) (=T−(U+V+W))	126	2.6
Y.	Total migrants leaving Skye amongst 1860s cohort (=W+X)	373	7.6
	If 'stayer' mortality applies to 'movers' (=*Y**(*V/U+V*)):		
	Y1. Estimated N of deaths amongst migrants leaving Skye	*60*	*1.2*
	Y2. N of survivors from 1860s cohort in Scotland (W)	*247*	*5.0*
	Y3. Estimated N of migrants from 1860s cohort surviving outside Scotland in 1871 (=*Y*−(*Y1*+*Y2*))	*66*	*1.3*
Z.	Assumed in-migrants aged 0–9 to Skye 1861–1871 (=L + 'Not Skye born' in Q)	263	

Note

[a]Births and deaths were taken from census day 1861 to the day before the 1871 census: from 7 April 1861 to 1 April 1871

[b]224 'Not Skye born'; considered to be in-migrants to the island

Panel B, Table 13.1 shows that 89 % of 'Skye born' 0–9 year olds were living on Skye and could be traced to a birth. A further 4.5 % were on the island but could not be linked to a birth and 6 % of the cohort were living elsewhere in Scotland. The picture is similar in 1881 (Panel B, Table 13.2): 92 % of the 'Skye born' 0–9 year olds in Scotland were living on the island and could be linked to a birth. A further 4 % were living on the island but could not be linked to a birth and 4 % were living off the island. Because the names of the latter individuals are known they can be linked to their births; 83 % (3.6 % of the cohort) could be linked to a birth on Skye. In all, out of the cohort of 'Skye born' 0–9 year olds surviving in Scotland in 1881, 4.6 % could not be linked to a birth on the island.

The possibility of missed links in our Skye data is problematic. Any missed links when linking forward will be classed as 'unaccounted for'. An accurate assessment of out-migration from Skye needs to take account of mortality amongst those migrating and of emigration from Scotland. A combination of forward and backward linking, as laid out in Panel C of Tables 13.1 and 13.2, allows this. Missing links on Skye are dealt with by assuming that the 'Skye born' seen in the census but not linked to a birth were indeed 'Skye born', and that errors in age and birthplace more or less cancelled each other out (as seems likely from Tables 13.1 and 13.2). It was also assumed that the death rate among those leaving Skye was the same as that amongst those remaining on the island. This is probably an overestimation, as it is likely that few children would leave the island as infants, so most young migrants will have survived the first, most lethal, year of life. Even if they were subject to

Table 13.2 Population accounting: linking between the births of the 1870s birth cohort on Skye, deaths before 1881 and individuals aged 0–9 years in the 1881 census[a]

PANEL A: LINKING FORWARD from births

		N	%
A.	BIRTHS registered on Skye 1871–1881[a]	4872	100.0
B.	Linked to a death in Skye death registers 1871–1881[a]	652	13.4
C.	Linked to an entry in the 1881 census of Skye	3803	78.0
D.	Accounted for on Skye by 1881 (=B+C)	4455	91.4
E.	Not linked to a death or a 1881 Skye census entry (=A−D)	417	8.6
F.	‘Skye born’ aged 0–9 elsewhere in Scotland in 1881 census	178	3.7
G.	Not identified on Skye or elsewhere in Scotland in 1881 (=E−F)	239	4.9
		N	%
H.	DEATHS of those aged <10 registered on Skye 1871–1881	812	100.0
I.	Linked to a 1871–1881 birth on Skye	652	80.3
J.	Individual observed in the 1871 census: not linked to a 1871–1881 birth on Skye	108	13.3
K.	Individual not observed in 1871 census; but should have been given age in death register	23	2.8
L.	Individual not linked to a Skye birth and age indicates born 1871–1881: in-migrant	30	3.6

PANEL B: LINKING BACKWARD from the 1881 census

		N	%	*N*	*N*
		‘Skye born’	of ‘Skye born’ in Scotland	not ‘Skye born’	total
M.	Those aged 0–9 enumerated on Skye	3910	94.3	322	4232
N.	Enumerated ‘aged 10+’ but in birth register 1871–1881	+59			
O.	Enumerated on Skye ‘aged 0–9’ in 1881 but also enumerated in the 1871 census	−39		*−1*	*−40*
P.	Enumerated on Skye aged 0–9 as ‘not Skye born’ but linked to a 1871–1881 birth on Skye	+39		−39	0
Q.	Adjusted *N* aged 0–9 on Skye 1881 =(M+N+O+P)	3969	95.7	282[b]	4251
	Q1. ‘Skye born’ aged 0–9 on Skye linked to a birth	*3807*	*91.8*		*3807*
	Q2. ‘Skye born’ aged 0–9 on Skye *not* linked to a birth	*162*	*3.9*		*444*
R.	‘Skye born’ aged 0–9 elsewhere in Scotland in 1881 census	178	4.3		178
	R1. Linked to a birth on Skye	*147*	*3.6*		
	R2. Not linked to a birth on Skye	*31*	*0.7*		
S.	all ‘Skye born’ aged 0–9 in Scotland (=Q+R)	4147	100.0		

(continued)

Table 13.2 (continued)

PANEL C: ACCOUNTING for the 1870s birth cohort on Skye: in the civil registers and 1881 census

		N	%
T.	Births on Skye, 1870s cohort (A)	4872	100.0
U.	Survivors from 1870s cohort on Skye in 1871 (Q)	3969	81.5
V.	Deaths to 1870s cohort on Skye (I)	652	13.4
	Death rate amongst stayers (=*V/*(*U*+*V*) * *1000*) = *141 per 1000 births*		
W.	Survivors from 1870s cohort elsewhere in Scotland, 1871 (F, R)	178	3.6
X.	Births from 1870s cohort unaccounted for in 1881(includes deaths off Skye and migrants leaving Scotland) (=T−(U+V+W))	73	1.5
Y.	migrants leaving Skye amongst 1860s cohort (=W+X)	251	5.1
	If 'stayer' mortality applies to 'movers' (=*Y**(*V/U*+*V*)):		
	Y1. Estimated N of deaths amongst migrants leaving Skye	*35*	*0.7*
	Y2. N of survivors from 1870s cohort elsewhere in Scotland (W)	*178*	*3.6*
	Y3. Estimated N of migrants from 1870s cohort surviving outside Scotland in 1881 (=*Y*−(*Y1*+*Y2*))	*38*	*0.8*
Z.	Assumed in-migrants aged 0–9 to Skye 1871–1881 =(L + 'Not Skye Born' in Q)	312	

Note

[a]Births and deaths were taken from census day 1871 to the day before the 1881 census: from 2 April 1871 to 2 April 1881

[b]282 'Not Skye born'; considered to be in-migrants to the island

much higher mortality rates at later ages than the members of their cohort who remained on Skye, a greater proportion of 'movers' than 'stayers' would have survived from birth to age 10.[11] Subtracting the number of deaths on Skye, the estimate of deaths off Skye, and the number of survivors both on and off Skye from the number of births in the cohort gives an estimate of the number of cohort members who had left Scotland. Under these assumptions 1.3 % of the 1860s birth cohort were alive overseas in 1871 and 0.8 % of the 1870s birth cohort survived outside Scotland in 1881. Thus, in total, 7.6 % of the 1860s birth cohort had left the island by 1871, and 5.1 % of the 1870s birth cohort had left by 1881. The loss of these individuals was, however, balanced to some degree by migration onto Skye of children aged 0–9, although some of these subsequently died before they could be recorded in the next census. At least 249 of the children aged 0–9 living on Skye on census day 1871 had arrived on the island since the 1861 census, and a minimum of 312 0–9 year olds arrived between the 1871 and 1881 censuses. Even if young adults were leaving the island, people with young children appear to have found Skye attractive.

In both the 1871 and 1881 censuses approximately 60 % of 0–9 year olds living on Skye, but not born on the island were living in a household headed by their

[11] Further work is required before the degree of overestimation of mortality amongst migrants which arises from our assumption of parity with 'stayers' can be gauged.

parent and a further 25 %, or thereabouts, were living in a household headed by their grandparent, suggesting that children under the age of 10 were unlikely to migrate in the absence of a relative. It was more common for those in the teenage years to venture away from home on their own. Table 13.3 shows the results of an attempt to link the children from the 1860s cohort who survived to be enumerated in the 1871 census to their census entries in 1881. The names of those enumerated on Skye in 1871 were known but only the number of those enumerated elsewhere in Scotland. Panel A of Table 13.3 shows forward linkage from the 1871 census to the 1881 census, and distinguishes between those individuals on Skye in 1871 who had been linked to a Skye birth, and those on Skye and those enumerated elsewhere in Scotland who could not be linked to a birth. Of those on Skye in 1871 who had been linked to a Skye birth, roughly 70 % were identified on Skye in 1881; just under 5 % had died on the island, approximately 9 % were found elsewhere in Scotland and 16 % were ‘unaccounted for’. Individuals for whom a birth-to-1871 link had not been made were much harder to link to 1881. Only 36 % of the individuals on Skye in 1871 who could not be linked to a birth could be identified in 1881, and, of course, as we did not have the names of the ‘Skye born’ elsewhere in Scotland in 1871, they could not be linked forward to 1881.

Panel B in Table 13.3 shows the results of linking backward from the 1881 census to the 1871 census and thence to the 1860s births. Of the 3747 individuals enumerated in Scotland in 1881 who were ‘aged 10–19’ and ‘Skye born’, 87 % were recorded on Skye and 13 % elsewhere. Of those enumerated on Skye, 81 % could be linked back to Skye’s 1871 census and 76 % could be linked to that census *and* to a birth. A further 2 % of individuals could not be found in 1871, but were linked back to a birth. This left 523 young people aged 10–19 on Skye in 1881 who could not be positively identified in the 1871 census, and therefore could not be linked back to their birth.

The lack of names in the 1871 I-CeM data made it difficult to gauge how many of the ‘Skye born’ aged 10–19 living elsewhere in Scotland in 1881 had left the island between 1871 and 1881. We knew that at least 373 had already left the island by 1871, but did not know which of these migrants had subsequently set sail from Scotland, moved south to England or Wales, had died, or had even returned to Skye, to be replaced by new ‘Skye born’ migrants. In just over two-thirds of cases the age and household information of individuals living elsewhere in Scotland in 1881 were sufficient to allow their birth on Skye to be identified.

Panel C in Table 13.3 combines information from Panels A and B. It shows that two-thirds of the 1860s birth cohort were still present on Skye in 1881. About 18 % had died on the island before the 1881 census and almost 10 % were resident elsewhere in Scotland. This suggests that by 3 April 1881, 267 members of the cohort were living outside Scotland or had died, unseen by our study. Overall, by that date, Skye had lost 33 % of its 1860s birth cohort; 18 % through death on the island, and a further 15 % as a result of out-migration. These figures could not have been calculated without the combination of forward and backward linkage using both the censuses and the civil registers.

Table 13.3 Linking forward and backward between those reported as 'Skye born' aged 0–9 in the 1871 census of Scotland and those aged 10–19 in the 1881 census of Scotland

PANEL A LINKING FORWARD: 'Skye born' individuals aged 0–9 years on Skye and elsewhere in Scotland in the 1871 census linked to the 1881 census

		'Skye born' individuals						
		seen on Skye in 1871, linked to a Skye birth		seen on Skye in 1871, but no birth identified		seen elsewhere in Scotland in 1871, no birth identified		Total
		N	%	*N*	%	*N*	%	*N*
A.	'Skye born' in Scotland 1871[a]	3627	100.0	182	100.0	247	100.0	4056
B.	Linked to 1881 Skye census	2568	70.8	66	36.3			2634
C.	Dead on Skye by 1881	166	4.6	2	1.1			168
D.	Accounted for on Skye by 1881	2734	75.4	68	37.4			2802
E.	Seen elsewhere in Scotland in 1881	318	8.8					474[b]
F.	'Skye born' from 1860s cohort not accounted for in 1881 (=A−(D+E))	575	15.8	114	62.6	247	100.0	936

PANEL B LINKING BACKWARD: 'Skye born' aged 10–19 years enumerated on Skye[c] and elsewhere in Scotland in 1881 linked to the 1871 census and to a Skye birth

		'Skye born' individuals			
		on Skye		elsewhere in Scotland	all in Scotland
		N	%	%	%
Skye born' enumerated on Skye 1881					
G.	All 'aged 10–19' enumerated on Skye, 1881	3214			
H.	'Aged 0–9' in 1881 but linked to 1860s cohort birth[d]	38			
I.	'Aged 20+' in 1881 but linked to 1860s cohort birth[d]	4			
J.	'Aged 10–19' in 1881, stated 'not Skye born' but linked to 1860s cohort birth[d]	17			
K.	Adjusted *N* aged 10–19 in 1881 census (=G+H+I+J)	3273	100.0		87.3
L.	Aged 10–19 linked back to 1871 census				
	L1. Linked to 1860s cohort birth[d]	2505	76.5		66.8
	L2. Not linked back to a 1860s cohort birth[d]	165	5.0		4.4
	L3. Total	2670	81.5		72.2

(continued)

Table 13.3 (continued)

PANEL B LINKING BACKWARD: 'Skye born' aged 10–19 years enumerated on Skye[c] and elsewhere in Scotland in 1881 linked to the 1871 census and to a Skye birth

	'Skye born' individuals			
	on Skye		elsewhere in Scotland	all in Scotland
	N	%	%	%
M. Aged 10–19 not linked back to 1871 census				
M1. Linked to a 1860s cohort birth[d]	80	2.4		2.1
M2. Not linked to a 1860s cohort birth[d]	523	16.0		14.0
M3. Total	603	18.4		16.1
'Skye born' enumerated elsewhere in Scotland, 1881	474		100.0	12.7
N. 'Skye born' aged 10–19 elsewhere in Scotland, 1881	318		67.1	8.5
N1. Linked to a Skye birth	156		32.9	4.2
N2. Not linked to a Skye birth				
O. Total 'Skye born' 'aged 10–19' in Scotland, 1881 (=K+N)	**3747**			**100.0**

PANEL C ACCOUNTING for the 1860s birth cohort on Skye: in the civil registers and the 1881 census

	N	%
Q. Births registered on Skye, 1860s cohort[d] (Table 13.1, line A)	4906	100.0
R. Members of cohort enumerated on Skye in 1881	3273	66.7
S. Deaths to cohort registered on Skye 1861 census–1881 census	892	18.2
T. Members of cohort enumerated elsewhere in Scotland in 1881	474	9.7
U. Number of members of cohort not accounted for by 1881	267	5.4

Note

[a] For the origin of the figures in line A, see Table 13.1, lines Q1, Q2, R and S, respectively

[b] 156 'Skye born' individuals, seen elsewhere in Scotland in 1881, could not be linked back to 1871, nevertheless they helped to account for members of the 1860s cohort, so they are included in the 'Total' column

[c] There were also 323 'Not Skye born' 10–19 year olds enumerated on Skye in 1881

[d] Births were taken from census day 1861 to the day before the 1871 census: from 7 April 1861 to 1 April 1871

13.5 Differential Rates of Successful Linkage

The national coverage of the I-CeM 1871 and 1881 Census datasets enabled the proportion of Skye's population which remained on the island but 'unaccounted for' under different methods of linkage to be estimated. Differences between groups in the rate with which they were successfully links highlighted the fact that certain groups were less likely to be linked than others.

The group for which links were most likely to be missed was 'servants'. Few young people went into service before their mid-teens, so in Table 13.4 'Skye born' 15–19 year olds on Skye in 1881 are considered. The table clearly shows that only

Table 13.4 Tracing 'Skye born' individuals, aged 15–19, resident on Skye in 1881 back to the 1871 census of Skye, by relationship to head of household

Relation to head in 1881	*N* total	*N* linked 1881–1871	*N* not linked 1881–1871	% not linked 1881–1871
Son	560	522	38	6.8
Daughter	527	496	31	5.9
Servant	181	155	166	91.7
Female servant	*132*	*5*	*127*	*96.2*
Male servant	*49*	*10*	*39*	*79.8*
Other	209	125	84	40.2
Total	**1477**	**1158**	**319**	**21.6**

6.8 % of those reported as 'sons' of their household head and 5.9 % of 'daughters' could not be linked back to 1871. However, more than 90 % of 'servants', and 46 % of 'others', could not be linked backward from 1881 to 1871. Female servants were less likely to be linked than male servants, and there were more than three female servants for every male servant. This meant that girls in their late teens were particularly likely to be 'unaccounted for' (final column of Panel A, Table 13.5).

In Table 13.5 all 9778 births comprising the Skye 1860s and 1870s birth cohorts, are differentiated into 'girls' and 'boys', and further divided into four five-year birth cohorts (Panel A), or census age groups (Panel B). As noted on the table, there were 140 cases in which the birth certificate did not give the forename of the child and therefore its sex could not be determined.[12] These cases were not included in the table. In Panel A the boys and girls born in each five-year cohort were traced forward to the Skye death registers or to the 1881 census, either on Skye or elsewhere in Scotland. Amongst the youngest cohort, born between 1876 and 1881, only 4 % of the children of either sex could not be accounted for by 1881. Amongst the cohort born between 1871 and 1876, the figure was 6–7 %. Amongst those born between 1866 and 1871, 19 % of girls and almost 17 % of boys were 'unaccounted for', while the figures for those born between 1861 and 1866, so in their late teens in 1881 were 32 % for girls and 23 % for boys. It is noticeable that boys were rather more likely to be linked to death than girls, which is not unexpected, given the higher mortality amongst males at the youngest ages. Largely for this reason, girls in the three youngest cohorts were fractionally more likely to be identified in the 1881 census, either on Skye or on the mainland.

Panel B, Table 13.5 presents all 'Skye born' boys and girls enumerated in the 1881 census of Scotland, laid out as the four age groups roughly equivalent to the birth cohorts in Panel A. The table shows the number in each group who could be linked backward to their birth certificates on Skye. Unsurprisingly, those not living on Skye in 1881 were, in general, less likely to be linked to a birth than those enumerated on the island. It is also clear once again that girls in their teens, but

[12] Almost all of these children died very young, and although they could be linked to a death certificate, no name or sex was given there either.

Table 13.5 Linking births on Skye 1861–1881 to the Skye death registers and the 1881 census of Scotland and linking individuals aged 0–19 in the 1881 census to their births in the Skye birth registers[a]

PANEL A LINKING BIRTHS on Skye 1861–1881 to the 1881 census and Skye death registers										
			Births linked to 1881 census							
Dates born	Sex	*N* born	Died on Skye before 1881		Found on Skye	Found elsewhere in Scotland	Total		Births unaccounted for	
			N	%			*N*	%	*N*	%
07/04/1861–01/04/1866	f	1226	228	*18.6*	503	101	604	*49.3*	394	32.1
	m	1255	280	*22.3*	591	95	686	*54.7*	289	23.0
02/04/1866–01/04/1871	f	1147	154	*13.4*	704	67	771	*67.2*	222	19.4
	m	1250	212	*17.0*	764	64	828	*66.2*	210	16.8
02/04/1871–02/04/1876	f	1114	143	*12.8*	847	45	892	*80.1*	79	7.1
	m	1228	208	*16.9*	897	47	944	*76.9*	76	6.2
03/04/1876–02/04/1881	f	1137	84	*7.4*	980	23	1003	*88.2*	50	4.4
	m	1281	114	*8.9*	1082	33	1115	*87.0*	52	4.1
07/04/1861–02/04/1881	f	4624	609	*13.2*	3034	236	3270	*70.7*	745	16.1
	m	5014	814	*16.2*	3334	239	3573	*71.3*	627	12.5
	Total	9638[b]	1423	*14.8*	6368	475	6843	*71.0*	1372	14.2

(continued)

Table 13.5 (continued)

PANEL B LINKING INDIVIDUALS aged 0–19 in the 1881 census to their births in the Skye birth registers

Age in 1881 census	Sex	‘Skye born’ on Skye in 1881 census				‘Skye born’ elsewhere in Scotland in 1881 census				Births unaccounted for in panel A[d]	
		Total	Linked to birth	Not linked to birth		Total[c]	Linked to birth	Not linked to birth			
		N	*N*	*N*	%	*N*	*N*	*N*	%	*N*	% of all Skye births
15–19	f	752	505	247	*32.8*	183	104	79	*43.2*	68	*5.5*
	m	722	572	150	*20.8*	136	91	45	*33.1*	94	*7.5*
10–14	f	847	708	139	*16.4*	75	63	12	*16.0*	71	*6.2*
	m	893	783	110	*12.3*	80	71	9	*11.3*	91	*7.3*
5–9	f	904	848	56	*6.2*	52	45	7	*13.5*	16	*1.4*
	m	948	894	54	*5.7*	60	50	10	*16.7*	12	*1.0*
0–4	f	983	956	27	*2.7*	32	24	8	*25.0*	19	*1.7*
	m	1075	1049	26	*2.4*	34	28	6	*17.6*	24	*1.9*
Total 0–19	f	3486	3017	469	*13.5*	342	236	106	*31.0*	174	*3.8*
	m	3638	3298	340	*9.3*	310	240	70	*22.6*	221	*4.4*
	total	7124[a]	6315	809	*11.4*	652	476	176	*27.3*	395	*4.1*

Note

[a] Births from 7 April 1861 to 2 April 1881

[b] In 140 cases out of the 9778 births in this period, the name and sex of the child were not given on the birth certificate, and they therefore could not be included in the linkage exercise

[c] The figures in this table have not been adjusted for age misreporting: the total is sum of the ‘Skye born’ aged 0–9 enumerated on Skye in 1881 (Table 13.2, line M) and the ‘Skye born’ aged 10–19 enumerated on Skye in 1881 (Table 13.3, line G)

[d] All 1881 ‘Skye born’ census entries in Scotland unlinked to a birth

particularly in their late teens, were considerably less likely to be linked back to their birth, wherever they were in Scotland.

Of the 652 males and females aged 0–19 recorded as 'Skye born' in 1881, but not resident on the island, 476 could be identified in the Skye birth registers between 1861 and 1881. Almost 8 in every 10 of those linked to a birth were related in some way to the head of the household in which they were living in 1881. Of the 176 individuals who could not be linked back to a birth on Skye, only 37 % were stated to be related to the head of their household; a further 32 % were servants and 20 % were boarders or lodgers. The births of some of these individuals may have been registered on Skye, but no definite link could be made because there were competing possibilities. As ever in reconstitution and reconstruction studies, single individuals, unattached to the rest of their household proved most difficult to link with certainty, particularly those with common names.

Panel B, Table 13.5 also shows that, while 240 boys and 236 girls could be traced from the rest of Scotland to the birth registers on Skye, 106 girls, but only 70 boys went unlinked. Considerably, more teenage girls than boys from the 1860s Skye birth cohort were thus living elsewhere in Scotland. However, when the number of 1881 census entries unlinked to a corresponding birth is subtracted from the number of births for which no census or death register can be found (final columns, Panel B, Table 13.5), it becomes apparent that teenage boys were more likely than girls to be 'missing' from the 'Skye born' population of Scotland. Boys, it may be assumed, were more likely to have left the country, or to have died away from Skye. Taken together these figures suggest that teenage girls on Skye were more likely to leave home to work in other people's houses, both on Skye and elsewhere in Scotland, but less likely to leave Scotland than 'Skye born' boys of the same age. This chimes with the fact that fishing and seafaring were common employments for Skye men, but closed to their women folk.

13.6 Conclusions

Few of the findings in this chapter are surprising, but the 'experiment' on which we embarked was to see whether family reconstruction could be extended over space now that individual-level data sources covering national populations are available. The new resources do make it possible to follow certain migrants and to gauge the extent of migration within particular cohorts, albeit with several caveats. Even when names are not available, being able to identify the number of individuals living outside their area of origin helps to 'balance' the demographic account for that area. Such an exercise has not been previously possible, certainly not for the British Isles. With both censuses and civil registers available historical demographers will be able to refine the way they calculate population change. Rather than considering population change to be the sum of natural growth (births–deaths) and net migration (immigration–out-migration) it can now be calculated as the sum of births and in-migrants, minus native deaths and in-migrant deaths, from which out-migrants are then subtracted. This will provide a much more nuanced understanding of fertility, mortality and migration flows at the local level.

The ‘experiment’ has also shown that the way different groups leave home and migrate, even within a particular community, can mean that certain groups are less likely to be successfully linked. If these differences are not properly understood any demographic rates calculated from reconstitutions or reconstructions may well be biased.

Finally, the ‘experiment’ has demonstrated that the direction of linkage is a very important consideration in reconstitution or reconstruction, as it can affect the conclusions drawn. As Ruggles argued in 1992, demographers and historians have to be very careful in shaping the questions they ask of histories constructed using record linkage methods. The analysis reported here, ‘experimental’ as it may be, has shown that a combination of ‘forward’ and ‘backward’ linkage is now possible and that this can provide a much fuller population history than linkage conducted in one direction alone.

References

Blaikie, A., Garrett, E., & Davies, R. (2005). Migration, living strategies and illegitimate childbearing: a comparison of two Scottish settings, 1871–1881. In A. Levene, T. Nutt, & S. Williams (Eds.), *Illegitimacy in Britain, 1700–1920* (pp. 141–167). Basingstoke: Palgrave MacMillan.

Church of Jesus Christ of Latter-Day Saints. (1999). *1881 British census and national index' family history resource file*. CD Rom.

Clark, P., & Souden, D. (1987). *Migration and society in early modern England*. London: Hutchinson.

Davies, R., & Garrett, E. (2005). More Irish than the Irish? Nuptiality and fertility patterns on the Isle of Skye, 1881–1891. In R. J. Morris & L. Kennedy (Eds.), *Ireland and Scotland; order and disorder, 1600–2000* (pp. 85–100). Edinburgh: John Donald.

Galley, C., Garrett, E., Davies, R., & Reid, A. (2011a). Living same-name siblings and British historical demography. *Local Population Studies, 86*, 15–36.

Galley, C., Garrett, E., Davies, R., & Reid, A. (2011b). Living same-name siblings and English historical demography; a reply to Peter Razzell. *Local Population Studies, 87*, 70–77.

Garrett, E. (2006). Urban-rural differences in infant mortality: A view from the death registers of Skye and Kilmarnock. In E. Garrett, C. Galley, N. Shelton, & R. Woods (Eds.), *Infant mortality: A continuing social problem* (pp. 119–148). Aldershot: Ashgate.

Garrett, E., & Davies, R. (2003). Birth spacing and infant mortality on the Isle of Skye, Scotland, in the 1880s: A comparison with the town of Ipswich, England. *Local Population Studies, 71*, 53–74.

Morgan, P. (1989). *Ainmean Chloinne: Scottish Gaelic names for children*. Broadford: Taigh na Teud.

Newton, G. (2011). Family reconstitution in an urban context: Some observations and methods. In *Cambridge Working Paper in Economic and Social History*, 12 (minor revisions January 2013).

Perkyn, A. (1999). Migration and mobility in six Kentish parishes, 1851–1881. *Local Population Studies, 63*, 30–70.

Reid, A., Davies, R., Garrett, E. (2002, published 2006). Nineteenth-century Scottish demography from linked censuses and civil registers: A 'sets of related individuals' approach. *History and Computing, 14*, 61–86.

Ruggles, S. (1992). Migration, marriage, and mortality: Correcting sources of bias in English family reconstitutions. *Population Studies, 46*, 507–522.

Ruggles, S. (1999). The limitations of English family reconstitution. *Continuity and Change, 14*, 105–130.

Sinclair, C. (2000). *Jock Tamson's Bairns: A history of the records of the General Register Office for Scotland*. Edinburgh: General Register Office for Scotland.

Wrigley, E. A., Davies, R. S., Oeppen, J. E., & Schofield, R. S. (1997). *English population history from family reconstitution 1580–1837*. Cambridge: Cambridge University Press.

Chapter 14
Building a Life Course Dataset from Australian Convict Records: Founders & Survivors: Australian Life Courses in Historical Context, 1803–1920

Janet McCalman, Leonard Smith, Sandra Silcot and Rebecca Kippen

Abstract Founders & Survivors is a multi-university and public collaborative project that is building a transnational and intergenerational dataset of life courses generated from the UNESCO recognised convict records of Tasmania. This chapter outlines the technical history of the project: mass digitization and archiving online of over 100,000 images, manual scholarly transcription and the building of a prosopography database. This comprises a relational genealogy database integrated with an XML (BaseX) source database. Individual life histories are compiled dynamically from diverse sources, linked by a combination of machine matching and human judgment, and managed by an independent link management module. Using Google Docs over 50 online volunteers crowdsourced the convict genealogies and coded the data. Manual linkage and scholarly verification remained essential for the collation of prosopographical data and manual coding was necessary for statistical analysis.

J. McCalman (✉) · S. Silcot · R. Kippen
Centre for Health Equity, University of Melbourne, Parkville, Australia
e-mail: janetsm@unimelb.edu.au

S. Silcot
e-mail: ssilcot@unimelb.edu.au

R. Kippen
e-mail: rkippen@unimelb.edu.au

L. Smith
Demographic and Social Research Institute, Australian National University, Acton, Australia
e-mail: leonard.smith@anu.edu.au

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_14

14.1 Introduction

The Australian colonies were settled using the forced labour of convicts. Between 1803 and 1853 almost 73,000 men, women and children were transported to the island prison of Van Diemen's Land, now the Australian state of Tasmania. The convicts' discipline, health and labour were managed by a 'paper panopticon' of manuscript records, and those convict records are recognised in UNESCO's Memory of the World[1] as part of the global cultural heritage. The Tasmanian records, in particular, are arguably the most detailed and intimate records of human beings' bodies and behaviour in the nineteenth-century world. Since 2008, the project *Founders & Survivors: Australian Life Courses in Historical Context* has been building a prosopographical life course dataset of that population transported to Van Diemen's Land, starting from the surviving usable records: an archive that comprises around 68,000 individuals (Bradley et al. 2010). Such an endeavour would have been impossible before the Internet and the explosion of digitised historical records made available by both government agencies and commercial genealogy enterprises. Information technology has enabled research on a scale that was unimaginable 20 years ago, yet the project has still needed hundreds of thousands of human hours to decipher, transcribe, interpret, code and connect data: machines could help us create the data libraries, but not in the end, do the research. This chapter outlines the design and conduct of the project and links to web documents on the information technology.

14.1.1 The Tasmanian Convict Records

In an era before photographs and fingerprints, the penal system needed precise accounts of convicts' bodies and faces to identify them: colouring, shape, the size of facial features, and of every bodily scar or deformity, tattoo or mark. On their disembarkation in the colony, a detailed manifest of the convict ships' human cargo known as an indent was created, in which the convicts provided testimony of their convictions and confessions, of their birthplaces, their families, of their religion and their past associations. The convicts' work experience and potential usefulness to the colonial economy were assessed by the convict clerks compiling the indent, records of their work assignments were maintained and they were periodically subjected to a muster or census. Finally, there was the conduct record that tracked their 'moral career' through servitude, recording infractions against convict discipline, secondary crimes and punishment. The conduct records are complex documents, created from the reports sent in weekly to the Convict Department in the

[1] http://www.unesco.org/new/en/communication-and-information/flagship-project-activities/memory-of-the-world/register/full-list-of-registered-heritage/registered-heritage-page-8/the-convict-records-of-australia/#c186408

colonial capital, Hobart, by the police magistrates who administered the system of justice.[2]

The Tasmanian records fall into two systemic categories: the *Assignment* period (to 1840 for men, 1844 for women) where the management of convict labour was outsourced to private employers, and the *Probation* period (1841–1853 for men and 1844–1853 for women) where convicts served a term of probation under penal supervision before being assigned to the labour market. The records for these periods differ in their detail and organisation. There are also records of musters (the convict censuses), of permission to marry while still under the system, and of inquests. Other records of tickets of leave (probation), pardons, absconders, secondary offences and complaints were published in the police gazettes of both Tasmania and Victoria and were also published in the newspapers. In addition to the Convict Department's records held in Tasmania, records were created before embarkation and are mostly held in the National Archives of the United Kingdom. These include records of some convicts who were withdrawn from the voyage and never embarked. In addition there are prison hulk and gaol records held in the United Kingdom and Ireland. Among the richest records are the ships' surgeons' medical logs for the voyage that record medical cases, individual character assessments as well as general remarks on health and behaviour at sea (Foxhall 2011).

The convict records in the archives are handwritten and replete with arcane terms, codes and abbreviations. They are preserved in large bound volumes on strong rag paper, with cross-hatching annotations in multi-coloured inks. The first task of the Founders & Survivors project when it received its initial funding from the Australian Research Council was the systematic imaging of the conduct records and surviving indents, linking the images to the convict index created by the Tasmanian Archives and Heritage Office (TAHO). This index lists every document and page where the convict appears in the archives. Poor quality microfilm images of some records had been created as part of an earlier LDS project, but these were unsuitable for digitization and new high-quality full colour digital images of the records were prepared for the research programme and made available online. The digital imaging and archiving were managed by technology developed by the eScholarship Research Centre at the University of Melbourne.[3]

14.1.2 Transcription

The task of data capture was complicated by the multiple sources for each convict. If a life course, or prosopography were to be assembled for an individual, it would need to incorporate information extracted from these various documents. Therefore, the next task of the project was to transcribe identifying data for each convict from

[2] http://www.linc.tas.gov.au/tasmaniasheritage/popular/convicts/convictdept

[3] http://www.esrc.unimelb.edu.au

descriptions lists and indents to define the population for study. Trained transcribers, mostly postgraduate history students from the University of Tasmania, did this work under the rigorous supervision of Dr Alison Alexander.[4] They transcribed the assignment records, recording name, place of birth, place of conviction, crime, sentence, age on arrival in the colony, height, eye colour, hair colour, scars and tattoos and religion. The assignment records were created in alphabetical volumes that included multiple voyages, hence finding a particular convict's record depended on the TAHO index references. For the probation period starting in 1841, record keeping became more elaborate and each voyage was recorded in a single volume, with a page per convict. Finding an individual was easier and the project was fortunate that Dr Deborah Oxley, who had amassed an extraordinary database in the 1990s from the microfilm of indents held in the Mitchell Library, Sydney, gave permission for it to be incorporated.[5] Although there were some incompatibilities between these two initial data capture projects that caused technical problems, the Oxley records provided valuable additional data on the convicts' families as recorded in the indent. Since then, Tasmanian project leader, Associate Professor Hamish Maxwell-Stewart, has greatly enhanced this dataset in partnership with Dr Trudy Cowley of the Female Convicts Research Centre, adding deaths found in the convict system, permissions to marry and many other records. This convict dataset resides at the University of Tasmania in a Filemaker Pro database and will eventually provide the final home of the collective dataset of Founders and Survivors.[6]

14.1.3 Linking the Convict Records

Our aim was to develop an ongoing researchable prosopography database that is shareable, maintainable, updateable and sustainable. This was achieved using open source multi-user Web based software, including a relational database with a data model in which each individual life history is made up of a chain of dynamically linked assertions (Schmitz and Pearce 2013) or factoids (Bradley and Short 2005) judged to relate to a single person. The workflow began with a registration stage of assigning a 'record type' identifier to each record and a unique ID. The content was analysed, control totals established and the registration documented. The second stage was the ingesting of the registered data into an XML (BaseX) database using the Text Encoding Initiative (TEI) protocols. The records were then matched and linked, drawing on as many points of matching as the records permitted: names, years and places of birth, ship of transportation, type of crime, etc. Once the multiple records for each individual were matched and linked, a life course could be

[4] http://alisonalexander.com.au

[5] http://www.all-souls.ox.ac.uk/people.php?personid=47

[6] http://www.utas.edu.au/arts/people/people/hamish-maxwell-stewart; http://www.researchtasmania.com.au

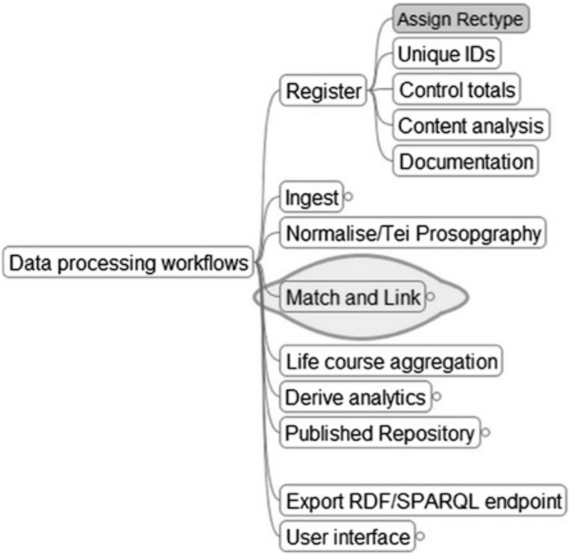


Fig. 14.1 Data processing workflow schema. See the text for a detailed explanation

aggregated and published in a repository, exported with an RDF/SPARQL endpoint to the user interface which was constructed from the open source programme, DRUPAL. Figure 14.1 demonstrates this workflow and data management practice, which conforms with the relevant best practices advanced by Mandemakers and Dillon (2000).

In contrast to linkage for statistical hypothesis testing, creating an ongoing researchable database of individual life histories requires establishing links that are definite and persistent. These links can evolve over time as more information becomes available. This is facilitated by the use of an independent link management model. This enables us to retain a simple database structure where the raw data remains close to its source format. This is a model originally devised by Dr John Bass for linking large-scale epidemiological–genealogical databases: "people", i.e. individuals are aggregations of database queries, that is they comprise the set of linked sources referring to them. He named the module "LKT" (Holman et al. 1999; Glasson et al. 2008).

To leverage what machines are good at, and to let humans do what they are good at, we use a two-step approach to linkage: computer matching is used to identify definite links, and where programmed matches are not found or are ambiguous, a customised web service is used to present possible links to a human for review and final decision in manageable batch sizes.[7] Where the linkage involves data

[7] Using a combination of XQiB and conventional HTML hyperlinks and jQuery for AJAX. A combination of batch and live searching over the web is supported. The searches are a combination of XQuery Perl and shell scripts. Data is held in BaseX, using the TEI prosopography data model. Results are entered incrementally into the Yggdrasil relational genealogy database.

containing place names, the user can map the place names interactively to assist matching where, for example different place names appear but the locations are very close on the map. The first stage of the project involves linking the records relating to each convict. There are about a dozen source record types in the archive, some for men and some for women, and an archive index, which covers all documents.[8] There are known relationships between some record types: for instance if a person appears in a conduct record they should also be in the indent and description lists, because these lists document entry into the convict system.

Linkage is assisted by the existence of Police Numbers,[9] which identified individuals within the convict system primary sources, and Archive Identification Numbers, which are secondary finding aids assigned to the convicts' records by the TAHO.[10] In the few cases where these numbers are missing or duplicated they were corrected manually.

The ship on which the convict arrived in Tasmania persisted as a basic element of their identity in public records, even after leaving the convict system, and provides a useful blocking factor in the linkage process. There were about 330 ships, each carrying between 150 and 300 convicts.

To establish the links we standardise names to incorporate the conventions used in the Police Numbers[11] and compare each record type with each other type, searching for matching individuals on the basis of forename and surname, plus any data fields the two sources have in common—all within a ship block.

If the computer finds an exact match, a link is stored in the link management system. If not, the name comparisons are made progressively fuzzier,[12] until a match is found or the process is exhausted. If only one pair matches, a link is made. If multiple possible links are identified, or if no link can be found, a human finishes the comparison. The 'individual' is a 'query' in database terminology, so an individual is totally dynamic, is recreated every time they are referenced, and will incorporate any new information added to the database tables.

The chain of links created in this way constitutes an individual's reconstructed life history, and because links to relevant records are incorporated dynamically, new information can be included simply by adding a link to the source in the link management module. Figure 14.2 maps the prosopography model.

[8] http://search.archives.tas.gov.au/default.aspx?detail=1&type=A&id=TA00060

[9] The Police Number is unique within sex, and is made up of the first letter of the surname and a numeral.

[10] Unfortunately there is no index linking the two.

[11] Surnames were normalised e.g. by removing O' from O'Connor, Mc from McAdam etc.

[12] Fuzziness in name matching is supported by NYSIIS and Soundex codes, XQuery free text searches and standard text matching.

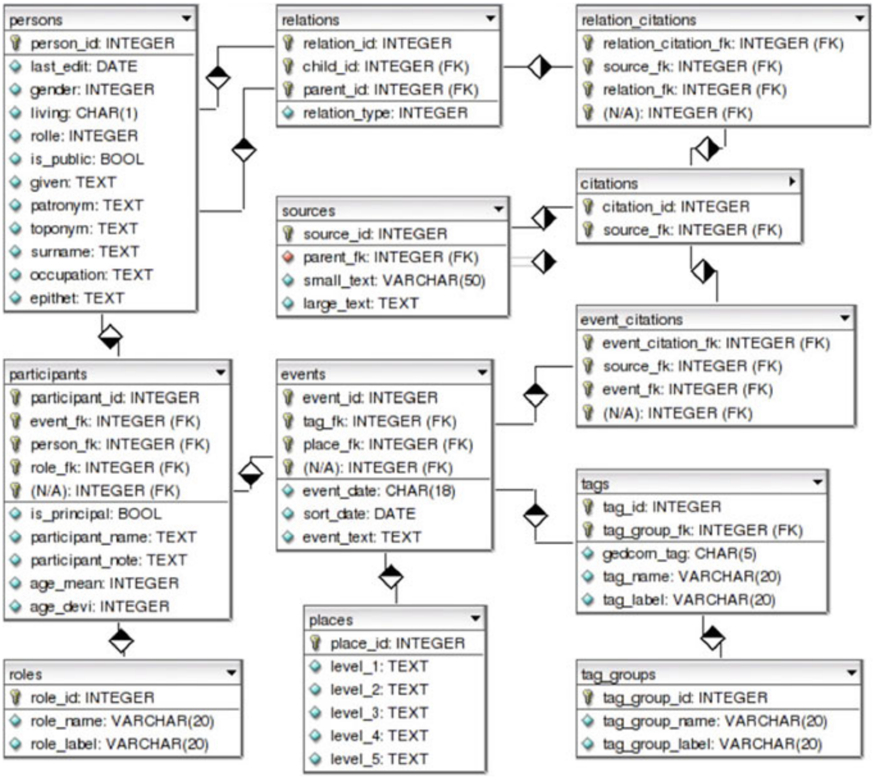


Fig. 14.2 TEIp5 prosopography model

14.1.4 Tracing Convicts Before and After Sentence

The aim of the project was to follow the life courses of the convicts not only while they were under the gaze of the paper panopticon, but also of their lives before and after sentence. Once the convict population's identification data were transcribed, that could be used to link the convict to records in the civil registration system, shipping records, newspaper reports, commercial, other genealogical databasis, human judgment and so on. This linked dataset provides the underlying database for genealogical and demographic research.

This manual linkage to historical data outside the convict system required validation against other identifiers to confirm that the record found was of the convict being investigated. The first points of identifications were ship and year of arrival, birthplace and year of birth. Arrival data was rarely included in later historical records other than police records, so birth places, occupations, names of parents and especially other relationships—siblings, spouses (marriages) and children (birth registrations)—helped to confirm identities. Often only a year and place of birth were available so that a process of manual elimination through comparing other convicts of the same name and free immigrants from arrival records might narrow the identification down to that individual.

We had access to a database of births, deaths and marriages for Tasmania from 1845 to 1899 transcribed 20 years before by Dr Peter Gunn from the original documents held in the TAHO. Dr Rebecca Kippen later expanded the linkages and used that database for her doctoral dissertation on the reporting of death in nineteenth century Tasmania (Kippen 2002). This database provides causes of death and some identifying information for individuals—sufficient to find most convicts who died in Tasmania before 1899. Since the mid 1990s digitised pioneer indexes have been available for Tasmania, Victoria and New South Wales, and since around 2010, *Ancestry.com* has published them as the Australian Indexes to Births, Deaths and Marriages with a steady updating with new data. Coverage is now good for Victoria and New South Wales, with deaths publicly searchable until 1986, and marriages and births restricted by privacy legislation to a hundred years after the event. Tasmania is still patchy after 1899.[13] Queensland and South Australia are improving and are better for deaths than for other vital events. New Zealand has recently made historical death certificates searchable online, and *Ancestry.com* has therefore provided quicker access than the use of multiple CDs. *FindMyPast* provides the best access to British and Irish data and both companies have database searches that enable searching censuses, births and baptisms, marriages and convictions. *FindMyPast* has digitised some of the Irish prison registers and its British Newspapers online has proved very useful, while *Ancestry.com* has digitised Australian electoral rolls. But the digitization project that has transformed this historical research is the National Library of Australia's historical newspapers' database that now covers 518 metropolitan and regional newspapers. The technology is OCR from microfilm that is accompanied by a tidy interactive programme that transcribes the text and invites users to correct it, while keeping a wiki-like record of corrections and registered editors.[14]

The greatest impediments to tracing convicts outside the convict system were aliases and common names, particularly among the Celts of Ireland, Scotland and Wales who had adopted patronymics. Spelling of family names was also unstable for the illiterate and semi-literate and for Gaelic names in the process of Anglicisation. Genealogical research into convicts requires a high level of both genealogical skill and historical imagination. Researchers have to second-guess people's movements around the landscape, through the economy, within family structures and associations. It is an exercise in fuzzy searching, looking for misspellings, slight alterations in spelling, phonetic spelling, deceptions, lies, mistakes and telling omissions. It is time-consuming and expensive, because outside Tasmania, vital registration certificates have to be purchased from the various state

[13] This will soon be rectified with a new transcription of the Tasmanian births, deaths and marriages organised by Rebecca Kippen.

[14] http://trove.nla.gov.au/newspaper/result?q=

registrars of births, deaths and marriages. This requires versioning the databases used for statistical analysis, while the biographical repository can be updated.

Australian state and federal governments have not retained the household schedules of past censuses, so longitudinal research using nominal census data is not possible. However, there is an abundance of other sources that are now accessible online. First, the Australian colonies were early in the Anglophone world to adopt systematic vital registration: in 1838 in Van Diemen's Land, just 12 months after it began in England and Wales; in 1853 in Victoria and similarly in New South Wales. Victoria established a system of registration that became a world gold standard when William Henry Archer was able to implement the full regime initially recommended by Dr William Farr for England and Wales (Hopper 1986). New South Wales and South Australia followed suit, although not with the rigour of the Victorian system. Victorian birth certificates, for instance, list the mother's other children, alive and deceased, with ages. All deaths had to be certified by a registered medical practitioner, and divided into immediate and up to three contributing causes of death, and their respective duration. Death certificates include full details of the deceased's place of birth, parents and their occupations, time in Australia broken down into various colonies, marriages (place, age at marriage, name of spouse) and children of those marriages, alive and dead, with ages. In the case of death certificates, the quality of the information depended (and still does) on the knowledge of the witness or the family history that has been passed down as acceptable. Therefore, death certificates often have to be read 'against the grain' to find convict pasts that were being rewritten within families and communities.

14.1.5 Crowdsourcing

It was clear from the beginning that the task of tracing large numbers of convicts was impossible without volunteers. Funding to pay research assistants on that scale is unobtainable in Australia and we estimate that the project has captured something like $4 million AUD worth of research assistance. Volunteers were also necessary to provide the links from present day families back to convicts who had concealed their past with aliases and false trails. The tracing and verification of individuals required finding triangulating data of birth dates, birthplaces, marriages, associations and occupations. Death certificates were the most reliable sources, but the researchers were able to find newspaper and other vital registration data to confirm or disprove a match. Trained data checkers undertook the final verification.

When new funding was awarded in 2011, it became feasible to extend the tracing work to build a systematic reference population by dividing the convict population into historical cohorts created by their shared experience on the convict ship, and researching every convict on that voyage: the Ships Project. This reference population would enable us to assess the significance of those found and those not traced and impute results where necessary.

Table 14.1 Percentage of convicts traced to death by selected characteristics, sample of 13,552 male and 6173 females transported to Tasmania on 101 ships, 1812–53

Characteristics		Percentage traced to death		Sample size	
		Males	Females	Males	Females
Country of birth	England	47.4	49.7	9219	1853
	Ireland	40.4	50.9	2980	3200
	Scotland	41.2	50.5	585	727
	Other British	44.6	44.4	195	90
	Other	48.4	51.0	124	51
	Not given	59.2	68.3	449	252
Place of birth	Village	47.2	52.5	5093	2880
	Town	48.4	49.1	2924	1376
	Industrial urban	44.8	46.3	2143	421
	Port cities	39.7	49.3	1160	831
	London	40.1	48.1	1637	420
	Other country	44.2	48.1	129	52
	Not given	59.0	70.5	466	193
Crime for which transported	Theft of valuables	43.6	49.6	8769	4800
	Theft of food or animals	49.5	57.4	2779	746
	Other non-violent	50.8	50.6	922	427
	Other violent	50.3	55.1	606	89
	Not given	54.0	71.2	476	111
Year of arrival in Tasmania	1810s	57.5	–	445	–
	1820s	49.8	55.5	2070	607
	1830s	46.5	49.1	3027	852
	1840–44	45.9	50.7	4307	1227
	1845–49	42.2	51.6	2504	1636
	1850–53	42.0	50.5	1199	1851
Age at arrival in Tasmania	Under 20 years	37.3	50.6	2536	1011
	20–24 years	43.8	49.7	4762	2122
	25–29 years	48.5	50.2	2494	1171
	30–34 years	47.3	50.6	1493	743
	35–39 years	53.7	50.8	843	398
	40–49 years	56.3	55.3	855	461
	50 years and over	65.7	60.5	370	177
	Not given	54.3	67.8	199	90
Total		46.0	51.1	13,552	6173

Over a research period of 20 months 2011–13, 54 volunteers from all states of Australia and even the United Kingdom completed 101 ships from 1812 to 1853 or 1:3 over the time period, yielding a research reference populations of 13,552 male and 6173 female convicts, or around 25 % of the men and 45 % of the women in the convict population. From the males 46.0 % were linked with a death record and from the females 51.1 %. Table 14.1 presents the percentages for several characteristics such as country of birth and the type of crime for which they were transported to Australia. And we will have spent more than $100,000 AUD on purchasing death certificates when the funding is exhausted.

14.2 Beyond Automated Linkage: From Prosopography to Population Analysis

The prosopography database is an XML Database (BaseX http://basex.org/), now containing millions of lines of historical data. Demographic and historical analysis required that the data be categorised and coded into multiple variables for analysis. The coding is part of the research process and is particular to the historical period and institutions under examination (Kippen and McCalman 2015). Since our interest is in the life course and the effects of different socially critical periods on survival and family formation, we have divided the convicts' life courses into (1) Life before sentence (2) Life under sentence (3) Life after sentence. Data had to be extracted from the mixture of descriptive and quantitative text collated on a convict, interpreted and coded so that one can infer life course effects.

This coding exercise was conducted by the volunteers using spreadsheets hosted by Google Docs that linked to the underlying database that was transcribed from the digitised images. The spreadsheets were generated from the convicts on each ship. Each convict's amalgamated record has a unique ID that is hot linked to the spreadsheet. When researchers worked on a convict, they filled out a new entry —"Community Contributed Content" (**CCC**)—via a Drupal interface, linked to the underlying database. The volunteers researched and linked recorded life events with their sources: births, baptisms, marriages, children's births, residences, occupations and death data. Where further sightings in other historical sources, such as inquests, newspaper reports or secondary sources were discovered, these sightings were added as free text to the prosopography in the underlying database. When descendants were found to have served in World War I, their digitised service record in the National Archives of Australia was incorporated. Many entries are extensive records of life after sentence, and standard historical sourcing has been enforced. All of this was added to the aggregated life course already created by automated linkage from the convict records, extending the prosopography to life before and after sentence. No longer were convicts to be seen purely through the view of the paper panopticon, but over their full lifespan.

The volunteers coded the spreadsheet drawing on data from the convict records in the underlying BaseX database and from the digitised images of conduct records, descriptions and appropriation lists. This required considerable paleographic skill and a knowledge of specialised language of the penal system. Rather than ask volunteers to transcribe and then code the full conduct record, it was decided that it would be more efficient to ask them to count and code clearly visible events: appearances before a police magistrate which were dated and underlined. Events relating to those like 'stripes', days in 'solitary' or the 'cells' and offences such as 'insolence', 'drunkenness', 'refusal to work' could be easily aggregated or recorded for their presence or absence. They coded for level of insults suffered and recorded offences within the system such as sexual offences and drunkenness that might point to behaviour after sentence. The two categories of 'moral' offences proved highly predictive of life expectancy for women in particular. The coders summarised family formation, geographic mobility, and end of life, details and sources from the historical linkage research they had already conducted for that convict.

Volunteers were trained in regular workshops held in Melbourne, Hobart and Launceston, and a training manual was created using animated screen capture images on Google Docs' equivalent of PowerPoint. Images were thought to be a more effective means of communication rather than textual instructions. The spreadsheet was large, with 27 columns, variously colour coded for columns that required numbers and columns that contained codes. Coding regimes were detailed on drop-down menus for each column for reference and filters applied to reduce mistakes. The spreadsheets, when cleaned, can be exported to SPSS.

14.2.1 Crowdsourcing and Building a Community Partnership

The volunteers were remarkable. They were a self-selected group of people, few of them with formal historical training, but all of them with the instincts of good historians. They were mostly retired from work, and old enough to have been schooled in copperplate handwriting (which was phased out in Australian schools in the early 1960s). The project had built a following from newspaper, television and radio publicity because the 'story' captured the imagination of many. A full-colour newsletter was produced every 4 months and distributed online to recruit volunteers, disseminate instructions and report results. It also enabled the volunteers and the research team to publish substantial articles based on the research, with full illustrations. Book reviews and news of other convict projects such as the Female Convict Research Centre and the Port Arthur Historical Site added to the feast of material reaching a rapidly growing national and international following.[15] From the beginning, it was obvious that volunteers should feel part of the research team, share in the planning and reporting, and in the intellectual life of the project. This was sustained by whole day workshops providing lunch and talks, discussion groups and coaching sessions. Each year finished with a small conference. Volunteers proved to have a wide interest in the historical issues, brought valuable insights and findings, and did some remarkable research into elusive lives. A number have emerged as born-history writers and we will be hoping to support some in producing their own books on their ship research. And it hardly needs saying that new friendships have been built for everyone. An exciting late development, through the initiative of Colette McAlpine from the Female Convict History Centre is the relationships she has built with scores of local and family

[15] http://www.foundersandsurvivors.org/content/chainletter-no-1-june-2009
http://foundersandsurvivors.org/sites/default/files/newsletters/FASNewsletter14_2013_large.pdf

history societies in the United Kingdom and everyday there are emails with information on convicts from county and local records.[16]

14.2.2 Lessons Learned

The hope of an open website with contributions from the public was naïve. The problems with spam overwhelmed our technical staff and the system itself, and email contributions that are vetted and uploaded by a volunteer research team will have to suffice in the future. The development of the interface, the data entry forms and the spreadsheet system was iterative and we did accumulate some baggage. It is very difficult, as with all multiple-user data entry, to maintain consistency, but we have now completed a large cleaning and authenticating exercise and have had sufficient funds to pay the best volunteers to do this work. Google Docs saved us from building an expensive workstation of our own, and has been cost effective and easy to use.

Drupal for the interface has struggled under the load and none of the software proved adequate to the task of managing family trees and relationships. This we hope we have resolved, with the support of the Australian National Data Service, by customising a Norwegian open source genealogical programme called Yggdrasil. Our system designer, Sandra Silcot,[17] has adapted and extended Yggdrasil so that it can manage and store individuals who are ‘created’ or ‘aggregated’ from the sightings in historical sources and where they can be studied as a population rather than in separate lineages. Underlying this entire project is the programme for link management developed by Dr John Bass for the Western Australia Health Department.[18] These software packages manipulate the data for us, but the matching and discovery, the interpretation and the coding have all needed human intervention.

Acknowledgments We extend our heartfelt thanks to our colleagues in Tasmania, Guelph, Ohio, Oxford and Flinders and to Claudine Chionh. Above all we thank our volunteers.[19] This project is

[16] http://www.femaleconvicts.org.au

[17] For further information on technical aspects contact Sandra.Silcot@unimelb.edu.au

[18] http://fasconn.blogspot.com.au

[19] **Checkers and Researchers**: Nola Beagley, Geoff Brown, Tricia Curry, Lance Dwyer, Alison Ellett, Jennifer Elliston, Leanne Goss, Dr Cheryl Griffin, Jan Kerr, Maureen Mann, Garry McLoughlin, David Noakes, Teddie Oates, Judith Price, Steve Rhodes, the late Dr Cecile Trioli, Colin Tuckerman, Jenny Wells.
FCRC Coordinator: Colette McAlpine.
FCRC Advisor and Web administrator: Dr Trudy Mae Cowley.
Ships Project Researchers: Colleen Aralappu, Maureen Austin, Vivienne Cash, Dianne Cassidy, Glenda Cox, Kathy Dadswell, Margaret Dimech, Brian Dowse, Ros Escott, Barry Files, Peter Fitzpatrick, Dr Janet Gaff, Prof Nanette Gottlieb, Stuart Hamilton, Jane Harding, Robyn Harrison, Graeme Hickey, Margaret Inglis, Bronwyn King, Dr Jenny Kisler, Darryl Massie,

funded by the Australian Research Council, the Institute for the Broadband Enabled Society at the University of Melbourne and the Australian National Data Service.

References

Bradley, J., Kippen, R., Maxwell-Stewart, H., McCalman, J., & Silcot, S. (2010). Research note: The founders and survivors project. *History of the Family, 15*(4), 467–477.

Bradley, J., & Short, H. (2005). Texts into databases: The evolving field of new-style prosopography. *Literary and Linguistic Computing, 20*, 3–24.

Foxhall, K. (2011). From convicts to colonists: The health of prisoners and the voyage to Australia, 1823–53. *Journal of Imperial & Commonwealth History, 39*(1), 1–19.

Glasson, E., de Klerk, N., Bass, A. J., Rosman, D., Palmer, L., & Holman, C. (2008). Cohort profile: The Western Australian family connections genealogical project. *International Journal of Epidemiology, 37*(1), 30–35.

Holman, C., Bass, A. J., Rouse, I., & Hobbs, M. (1999). Population-based linkage of health records in Western Australia: Development of a health services research linked database. *Australian and New Zealand Journal of Public Health, 23*(5), 453–458.

Hopper, J. (1986). The contribution of WH archer to vital statistics in the colony of victoria. *Australian Journal of Statistic, 28*, 124–137.

Kippen, R. (2002). An indispensable duty of government: Civil registration in nineteenth-century Tasmania. *Tasmanian Historical Studies, 8*(1), 42–58.

Kippen, R., & McCalman, J. (2015). A test of character: a case study of male convicts transported to Van Diemen's Land, 1826–1838. In K. Inwood & P. Baskerville (Eds.), *Lives in transition*, (pp. 19–42). McGill-Queens University Press, Montreal.

Mandemakers, K., & Dillon, L. (2000). Practices with large databases on historical populations. *Historical Methods, 33*(2), 1–6.

Schmitz, P., & Pearce, L. (2013). Berkeley prosopography services: Ancient families, modern tools. In *DH-CASE '13. Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities.* New York: ACM.

(Footnote 19 continued)

Elizabeth Nelson, Margaret Nichols, Rosemary Noble (UK), Keith Oliver, Maureen O'Toole, Margaret Parsons, Annette Sutton, Robert Tuppen, Rob Weldon, Lyn Wilkinson, Glad Wishart, Jacqueline Wisniowski, Judith Wood.

Supporters who were on board for a time: Anne Cronin, Katie Donnelly, Mary Eckhardt, Christine Hearne, John Hobbs, Lynne Hogg, Stephanie Hume, Brenda Irwin, Eileen Luscombe, Fiona McLennan, John Mugridge, Wendy Paterson, Fay Pattison, Kevin Pattison, Lorraine Polglaze, Sarah Preston, Colleen Robinson, Gary Scapin, Suzanne Smith, Claire Stevenson, Beth Stott, Edward Thomas, Sue Wyatt.

Web designers: Claudine Chionh, Robin Petterd.

Index

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2